Identifying and Measuring Internet Traffic: Techniques and Considerations

An Industry Whitepaper

Contents

sandvine®

Intelligent Broadband Networks

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.

First and foremost, CSPs must understand their use cases, as these determine tolerance for accuracy. It is likely less of a problem if reports show information that is wrong by a small margin, but it can be catastrophic if subscriber billing/charging is incorrect or management policies are applied to the wrong traffic.

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.

Faced with such variation, CSPs must understand the technologies, trade-offs (e.g., completeness and false positives), and deployment challenges (e.g., routing asymmetry; tunnels and encapsulation; encryption, obfuscation, and proxies) that exist in the context of traffic classification, and only with this detailed understanding can they ask the right questions in order to truly understand what a vendor is providing, and any limitations that would otherwise be hidden.

Version 2.20

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, or ensure correct billing and charging.

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

Traffic Identification

In general, traffic will be described as being one or more of these types:

- **Protocol:** a strict set of rules and formats that define how two or more elements share information (the information flow could be one way or bidirectional). Examples include UDP, TCP, HTTP, RTMP, SIP, FTP, and SMTP[1].
- **Application:** traffic associated with a particular software program. Examples include Skype, Netflix, PPStream, and games.
- **Website:** all the web pages that are part of a particular web domain and all content that is exchanged with a particular domain (whether or not the content corresponds to a web page)
- **Service:** a more general term that can include websites like Twitter and Facebook, cloud services like Salesforce, online storage, and many others.
- **Provider:** typically used to differentiate a brand within a type of traffic. For instance, many different video providers use RTMP, and many different voice services rely on SIP.

Even in the list above, the potential for overlap is evident. The reality is that these terms are closely related and an argument could easily be made that a particular type of traffic belongs to more than one classification: for instance, BitTorrent is a software client that uses the BitTorrent protocol as part of the BitTorrent network. As a second example, a large amount of Internet video is carried by RTMP, within HTTP, which is TCP; any of those three protocols is technically correct, but relay different degrees of information.

Consequently, these terms are frequently used interchangeably as a general label. Due to the important and technical nature of this subject, whenever there is doubt about meaning, it is a good practice to ask specific questions and define precise context.

It is also common to see sub-classifications that add further granularity to a classification. For instance, YouTube might be designated as HD or non-HD, or BitTorrent might be distinguished by being encrypted or not. Frequently, sub-classifications are used to add clarity rather than in response to convey a particular technical distinction.

There are many other terms that are important in the context of traffic classification, including:

- **Library:** the list of traffic types that are supported (i.e., identified and measured) by a solution

[1] If you would like to learn more, then these are good places to start: http://en.wikipedia.org/wiki/Communications_protocol and http://en.wikipedia.org/wiki/Lists_of_network_protocols.

- **Content Type:** typically refers to a finer level of classification of traffic as being video, text, images, audio, etc.
- **False Positive:** traffic that is incorrectly identified as being of Type B; the 'positive' identification of the traffic as being of Type B is false
- **False Negative:** traffic of Type A (that is supposed to be recognized) that is not identified as Type A; the 'negative' identification (i.e., "this is not Type A") is false
- **Unrecognized Traffic:** traffic that is not identified as belonging to any of the supported types
- **Over-the-Top (OTT):** traffic that is on a CSP's network that does not originate from a service provided by the CSP
- **Stateful:** requiring awareness of or maintaining a finite number of states
- **Data Traffic and Control Traffic** (alternatively called '**data channel**' and '**control channel**'): data traffic is the actual payload or content being exchanged, whereas control traffic governs that exchange; for instance, in a video stream the control traffic will include a feedback loop to convey user instructions (e.g., play, pause, seek) and transport quality information
- **Signature:** a pattern corresponding to a known traffic type against which observed traffic types are compared. In the most basic definition, a signature is a regular expression that is applied to packets. In the most advanced definition, a signature can be a stateful technique that monitors state changes within data and control traffic to extract information required for further identification (e.g., where the next data flow will appear) or simply requested (e.g., the provider of a video).

Traffic Categories

Typically, the protocols, applications, websites, and services will be part of a categorization hierarchy. For instance, Table 1 shows the taxonomy used in Sandvine's Global Internet Phenomena[2] reports. Different vendors will have their own taxonomies, and the nature of many traffic types makes mutually exclusive categorization impossible. CSPs can also ask if the vendor allows them to define their own hierarchies.

Table 1 - Example Traffic Categories

Traffic Category	Description	Examples
Storage	Large data transfers and online storage services	FTP, NNTP, PDBox, Rapidshare, Mega, Dropbox
Gaming	Console and PC gaming	Nintendo Wii, Xbox Live, Playstation Network, World of Warcraft
Marketplaces	Marketplaces for application and content downloads and software updates	Google Play Store, Apple iTunes, Windows Update
Administration	Protocols used to administer the network	DNS, ICMP, NTP, SNMP
File-Sharing	File-sharing applications, whether peer-to-peer or direct	BitTorrent, eDonkey, Ares, Pando, Foxy
Communications	Applications, services, and protocols that allow email, chat, voice, and video communications	Skype, ICQ, SIP, MGCP, IRC, FaceTime, WhatsApp, Gmail, SMTP
Real-Time Entertainment	Applications and protocols that allow 'on-demand' entertainment	Adaptive or progressive audio (e.g., Pandora, Rdio, Songza, Google

[2] Information about the program and the bi-annual reports is available at https://www.sandvine.com/trends/global-internet-phenomena/

		Music) and video (e.g., RTSP, RTMP, RTP), peercasting (e.g., PPStream, Octoshape), placeshifting (e.g., Slingbox), specific streaming sites and services (e.g., Netflix, Hulu, HBO Go, BBC iPlayer, SkyGo)
Tunneling	Protocols and services that allow remote access to network resources, or provide encryption or encapsulation	SSL, SSH, L2TP, Remote Desktop, VNC, PC Anywhere
Social Networking	Websites and services focused on enabling interaction and sharing	Facebook, Twitter, Habbo, Bebo
Web Browsing	Web protocols and specific websites	HTTP, WAP browsing

Data Extraction and Measurement

When a solution provides information about Internet traffic flowing on a CSP's network, some of that information has been identified, some has been extracted, and some has been measured.

While 'traffic identification' is generally used to describe the act of determining what particular Internet traffic is, the terminology can obscure what is actually happening and can mislead about what is not happening. Once more, it is important to examine in greater detail.

To illustrate, consider an adaptive video stream (i.e., a stream that can shift display quality up or down as network capacity permits) delivered via RTMP within HTTP sessions.

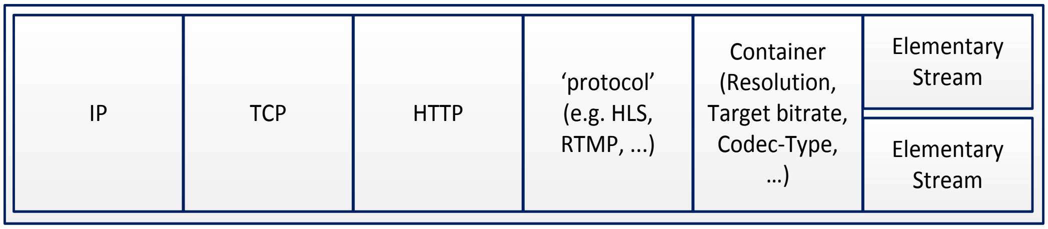

IP
TCP
HTTP
'protocol' (e.g. HLS, RTMP, …)
Container (Resolution, Target bitrate, Codec-Type, …)
Elementary Stream
Elementary Stream

Figure 1 - "Cross-section" of an adaptive video stream

Traffic identification, typically via deep packet inspection (DPI), will determine by examining the traffic that it is RTMP, within HTTP. If the recognition technology understands RTMP (i.e., is statefully aware of RTMP and can parse the protocol), then additional information can be extracted from the control traffic. For instance, information including container, codec, and resolution might be available, as shown in Figure 1.

Examination might also reveal the operating system being used, and the web browser; the device model might also be available. The solution might also provide information including the CDN from which the traffic came.

To determine the provider (e.g., brand of voice or video), a number of methods are applied. Sometimes the source domain clear, sometimes there is a 'provider' field, etc.

For the service provider, the specific mechanism via which this information becomes known is often irrelevant: for the most part, identification and extraction simply serve to tell the CSP ‘what’ the traffic is.

But the information needed by CSPs goes being the ‘what’, into the realm of measured quantities. The most basic measurements provided to CSPs are usually the volume (bytes) associated with a classification of traffic (e.g., “how many bytes of HD YouTube were used by subscribers on the iPhone 6 last week”), the number of users of a particular type of traffic, and counts of flows or connections. These are useful, certainly, but they are not the entire story.

Continuing the example of the adaptive RTMP stream, by measuring the bitrate and spotting step increases and decreases, it is possible to count the number of times a user observed an upshift or downshift in quality. Additionally, a CSP might be more interested in the number of videos or minutes of video being watched on their network than they are in the number of bytes.

Many CSPs are particularly interested in the subscriber quality of experience for video (QoE), and this calculation is only possible if many conditions are met in terms of solution capabilities.[3] For instance, first the traffic needs to be correctly identified as video, then a number of relevant fields need to be extracted from the control traffic, then bitrate changes and buffer under-runs need to be measured at a sufficient frequency, and only then can a QoE metric be calculated.

Answering questions about quantities (and questions that depend on those answers as inputs) requires advanced technology that goes far beyond ‘just’ identifying the traffic.

Techniques

Many techniques are applied, alone or in combination, to identify traffic and extract relevant fields. It is not uncommon for vendors to use the term ‘signature’ to mean any and all techniques.

Increased reliability and accuracy is typically achieved at the cost of greater processing complexity. This list introduces some popular techniques, in order of ascending reliability/accuracy:

- **Port Number:** this approach simply looks at the port number of the traffic and concludes that the traffic is of the type commonly associated with this port. Because of the certainty of false positives due to many traffic types taking random ports, this approach should not be used in any circumstances in which reliable identification is needed.
- **Regular Expression:** a byte pattern that is (assumed/expected to be) a unique identifier for a particular traffic type. The longer a regular expression, the less chance of there being a false positive due to matches against random data. Identification typically requires that one or more regular expressions be applied across multiple packets and flows.
- **Tracker:** a stateful technique that monitors state changes within data and control traffic both to extract information required for further identification (e.g., where the next data flow will appear) and to provide addition information in general.
- **Analyzer:** similar to a tracker, but with complete protocol awareness; that is, an analyzer can extract any and all meaningful pieces of information due to a complete understanding of a protocol. In the previous example of adaptive video, a tracker would be sufficient to determine from the control traffic where the data traffic would appear, but an analyzer is required to extract the resolution and codec information.

[3] More information on this topic is available in *Measuring Internet Video Quality of Experience from the Viewer’s Perspective*.

Additionally, operations (e.g., mathematical, decryption) can be used to add greater certainty to a match. For instance, analysis might reveal that byte X and byte Y multiply to give the length of the next data packet, so the identification technique could extract byte X and byte Y, multiply them together, and then compare this value to the measured length of the next data packet observed. Things can get even more complex: a particular series of bytes might be the decryption key for forthcoming content which, when decrypted, has length equal to the value of byte Z. Obviously, these operations come at a cost, but they provide practical certainty of traffic identity.

Sometimes, identification is aided with the use of heuristics, which in this context essentially mean informed guessing. When a decisive answer is needed (e.g., “do I or don’t I redirect this traffic to a video optimization service?”) but when complete information is absent, heuristics can be used. For instance, a heuristic technique might consider a collection of long-duration flows with large data payloads that are being exchanged with many other network users to be peer-to-peer file-sharing. Heuristics can be useful augmentations to other techniques, but their value should be weighed against the high potential for false positives; ultimately, heuristics are educated guesses.

Finally, recognition efforts might include IP address ranges that are known (or believed) to be associated with particular websites and services. The use of IP addresses could constitute the entire recognition signature or just a small component. Once again, it is important to ask for details.

Considerations for Traffic Classification

CSPs must understand the technologies, trade-offs, and deployment challenges that exist in the context of traffic classification. Only when CSPs have this detailed understanding can they ask the right questions in order to truly understand what a vendor is providing, and any limitations that would otherwise be hidden.

There are numerous topics that must be considered when evaluating traffic classification vendors. For ease of examination, they are examined in three subsections, below, that correspond to an objective of answering three questions:

1. What does the CSP want to do with the information?
2. What traffic classification capabilities is the vendor providing?
3. What technical challenges must the vendor overcome?

Information Requirements

In the context of network policy control, CSPs need to implement critical use cases such as accurate charging, policy-based measurements, and congestion management on a foundation of traffic classification.

The demands for accuracy vary somewhat based upon the use case. For instance:

- Charging requires the accurate counting of volume, duration, and events, to the most detailed level (e.g., application, protocol, content type, etc.) possible. Accuracy is critical: false positives and false negatives can both cause a subscriber to be overcharged and can both result in revenue leakage for the CSP (see Table 2).To be useful and actionable, business intelligence must be accurate; the solution must be able to deliver all metrics (e.g., volume, quality of experience scores, application- and subscriber-awareness, etc.) that are required to make an informed decision. Some tolerance for inaccuracy is permitted, but breadth and depth of measurements is extremely important.
- Policy enforcement must be applied accurately to avoid negative consequences; for instance, to achieve precisely targeted congestion management, only the traffic corresponding to users on a congested resource should be managed. A low false positive rate is more critical than a low false negative rate, as it is better for the CSP to fail to manage some target traffic than to manage untargeted traffic (for instance, traffic that is extremely sensitive to management).

All CSPs would do well to have a well-defined knowledge of what use cases they wish to implement, as the answer will determine traffic classification requirements and tolerance for inaccuracy or lack of measurements.[4]

Traffic Classification Capabilities

To critically examine the traffic classification capabilities of different vendors, CSPs must have a framework or a set of questions to ask.

[4] 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).****"*

This section uses the term ‘signature library’ to refer to all traffic identification and measurement capabilities.

Completeness

Many vendors promote the size of their signature library in terms of the number of signatures, but this is a useless metric for at least two reasons. First, what constitutes a ‘signature’ to one vendor is not the same as a ‘signature’ to another. Second, focusing on the number of signatures provides an incentive for vendors to artificially inflate their count. For instance[5]:

- A vendor could consider every client of a particular protocol to be a unique signature: BitComet, Azureus, Vuze, and countless other clients all use the BitTorrent protocol; are all of these counted separately as signatures?
- A vendor could consider every version of a client to be a unique signature: should Skype 6.15, Skype 6.16, etc. all be counted as separate signatures? What about Skype for Mac, Skype for Windows, Skype for Android, etc.?
- A vendor could consider every unique over-the-top provider as a unique signature: consider all SIP providers – should they count as one signature (i.e., SIP) or countless (one for every ‘provider’ that can be extracted from the control fields)?
- A vendor could easily add thousands of unique signatures by trusting that only standard services and transport protocols use the IANA port designations[6], but doing so would cause rampant false positives
- A vendor could license an enterprise signature library at relatively low cost to inflate the overall signature count: these signatures are arguably irrelevant for a consumer Internet provider

The best measure for the completeness of a signature library is the percentage of traffic that fails to be positively identified – that is, the percentage of traffic that gets a label like “unrecognized”, or “unknown”.

Best of breed solutions, including network policy control and deep packet inspection platforms for which traffic identification is a key technology, should be expected to positively identify at least 90% of traffic that they inspect. In practice, these platforms can exceed 95%. Conversely, when DPI is a technology that is just ‘switched on’ in a different platform (e.g., a GGSN or similar gateway), then recognition rates plummet; this makes sense, as DPI is not a core technology, so identification techniques are more basic due to limited processing power.

Unfortunately, assessing a solution based solely on the percentage of traffic that is positively identified provides an incentive for the vendor to accept false positives[7]. This issue is discussed in greater detail in the next section.

With the evolving nature of the Internet, it is important that a signature library keep pace. There is no perfect update frequency, though, as there are trade-offs involved: the higher the frequency of updates, the less time spent on quality assurance for a new signature and the higher the cost to the

[5] These are all real tactics used by vendors in the policy control and DPI space

[6] The full list is available here: http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml

[7] At least one vendor has stated that they can achieve 100% recognition, because in the worst case all traffic will be identified as TCP or UDP; all vendors could obviously achieve and claim this ‘recognition’, but do not do so, because it is ridiculous.

CSP of continually updating[8]; the lower the frequency of updates, the higher expectation of accuracy but the larger the risk of failing to keep pace with rapid developments. In practice, what works for one CSP might not be ideal for another; again, the recommendation is to ask questions to understand both the frequency of signature updates and how quality is ensured.

False Positives and False Negatives

False positives are extremely problematic for CSPs and introduce risk, primarily because they can cause public fall-out as a result of overbilling subscribers or managing the wrong type of traffic. Unfortunately, the pursuit of higher rates of recognition can serve as an incentive to accept false positives within a signature library.

To avoid the problems that result from false positives, CSPs must ask their vendors pointed questions.

- How do you ensure zero or minimal false positives?
- What is your priority: lower false positives or higher rates of recognition?
- Can all identified traffic be subject to congestion management policies?[9]

Ultimately, a vendor must prioritize either eliminating false positives or increasing the rate of recognition, and the CSP deserves to know the order of these priorities.

False negatives are still a problem for CSPs, but one of slightly less significance. While false positives impact business intelligence, policy management, and charging use cases, false negatives only meaningfully impact the former two.

Table 2 - Impact of False Positives and False Negatives

Use Case Family	Impact of False Positives	Impact of False Negatives
Business Intelligence	False positives cause over-counting of some traffic and undercounting of other traffic, so the significance is dependent upon the prevalence and importance of the impacted traffic.	False negatives cause undercounting of some traffic and more traffic to be 'unrecognized', so the significance is dependent upon the prevalence and importance of the impacted traffic.
Billing/Charging	False positives can cause overbilling of subscribers and revenue leakage for the CSP. For instance, a subscriber purchases a plan that includes 100 minutes of video streaming; whenever traffic is falsely identified as video streaming, it will consume some of the minutes that have been purchased. Alternatively, consider a service bundle that zero-rates all Facebook traffic; when traffic is falsely identified as Facebook, it is zero-rated rather than being charged for separately, and the CSP loses revenue.	False negatives can cause overbilling of subscribers and revenue leakage for the CSP. For instance, a subscriber purchases a service bundle that zero-rates all Facebook traffic; when Facebook traffic is not recognized, the usage will not be zero-rated and will be subjected to normal data charges. Alternatively, consider a plan that includes 100 minutes of video streaming; when a subscriber streams video and it is not recognized, the usage is not decremented and the CSP loses revenue.
Policy Management	False positives have an enormous impact on policy management, as they result in untargeted traffic being subjected to management policies. For instance, consider a congestion management policy	False negatives have only a minor impact on policy management, as they simply result in some target traffic not being subjected to management policies.

[8] For instance, many applications are carefully crafted to change their behavior in the presence of network management policies, so extensive testing under real-world conditions is critical to ensure classification remains accurate under these circumstances; if, in response to a development online, a vendor promises a new signature in hours or a few days, then serious questions should be raised about the quality of analysis and testing that went into creating and releasing that update.

[9] One embedded DPI vendor was quick to promote their ability to identify a particularly difficult application, but would tell CSPs to avoid traffic management of that application. This begs the question, why?

	that deprioritizes bulk downloads like software updates and peer-to-peer when shared resources are saturated; if deprioritization is applied to sensitive traffic like gaming or voice communications, then the result is a huge negative impact on subscriber QoE.	

Additional Data and Measurements

Beyond simply identifying traffic, what additional data can be extracted or determined and what measurements can be made?

For instance, additional data can include: service tier, IP address, MAC address, content provider, client device, media stream type, media container, video resolution, video codec, audio codec, operating system, browser, session protocol, and transport protocol, to name a few fields popular with CSPs.

If policy management is an objective, then a CSP needs to know if this data is actionable in real-time (i.e., can the data serve as a condition that triggers an action).

Measurements can add immense value, particularly for business intelligence, and can include:

- Duration of a video or audio stream
- Voice or video quality of experience
- 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)

Some vendors even include the ability for a CSP to define their own custom measurements to answer questions as they are asked.

Once again, if policy management is an objective, then a CSP needs to know if this data is actionable in real-time.

Technical Requirements

Before traffic identification signatures and techniques can even be applied, or in the course of applying such techniques, a number of technical hurdles must be overcome.

To be a truly viable solution, it is necessary that all of these challenges be addressed.

Stateful Protocols

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

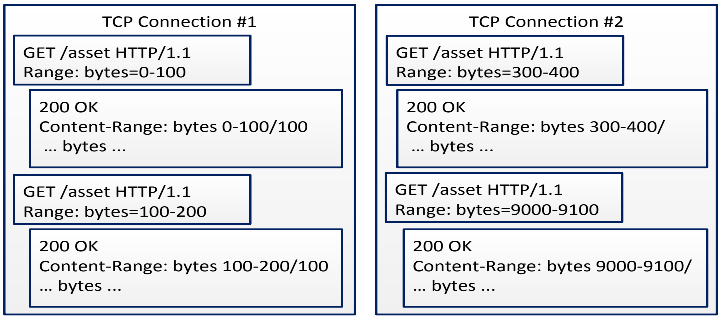
For instance, consider the example of FTP (this is also applicable to SIP, RTMP, RTSP, and many more). FTP includes a control channel that mediates the protocol data session. In this case, the identification solution must recognize the FTP control traffic and maintain a finite state machine to track 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.

Because the solution is tracking state, this type of signature is often called a 'tracker'.

Related 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.



```
GET /asset HTTP/1.1
Range: bytes=0-100

200 OK
Content-Range: bytes 0-100/100
 ... bytes ...

GET /asset HTTP/1.1
Range: bytes=100-200

200 OK
Content-Range: bytes 100-200/100
... bytes ...
```

```
GET /asset HTTP/1.1
Range: bytes=300-400

200 OK
Content-Range: bytes 300-400/
 ... bytes ...

GET /asset HTTP/1.1
Range: bytes=9000-9100

200 OK
Content-Range: bytes 9000-9100/
... bytes ...
```

Figure 2 - HTTP progressive video 'flows' across multiple TCP connections

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. CSPs must make certain that any solutions they are considering can accurately identify and measure all types of traffic in all network configurations, and CSPs should take heed that this is certainly not always the case.

This subject is explored in much greater detail in the whitepaper *Applying Network Policy Control to Asymmetric Traffic: Considerations and Solutions*.

Tunnels and Encapsulation

A significant portion of traffic that will be inspected for identification is contained within tunnels (e.g., GTP, GRE, L2TP, Q-in-Q, and IP-in-IP) or encapsulation (e.g., MPLS, EoMPLS, and VLAN). For maximum utility, the identification solution must be able to inspect (and apply policy control) within the tunnels and the encapsulation.

Devices and Tethering

In this era of the Internet of Things, there is no practical limit to the number of devices that can have an IP address. The increasing number and diversity of connected devices brings opportunity to CSPs who can identify trends and can, in turn, create services that cater to these unique demands.

With respect to traffic classification, the rich array of connected devices imposes a number of requirements.

Client and Access Devices

First, it is important to differentiate between *client device* and *access device*:

- A *client device* is the device that originates packets on the network
- An *access device* is the device that connects to the access network and owns the IP connectivity session

Consider Figure 3, below. Within the home network, there are many client devices (e.g., laptop, tablet, mobile phone), and the diagram could have included many others (e.g., game console, smart thermostat, etc.), but there is only a single access device (i.e., home router). The home router connects to the CSP's network, but the client devices actually originate packets.

In the mobile network, things can become a bit blurred. Typically, any device that connects to the mobile network is an access device and, in most cases, the mobile device is also a client device. However, in the case of tethering, a clear split is made: in this case, the mobile phone serves as an access device (as a WiFi hotspot), while the tethered laptop is the client device.

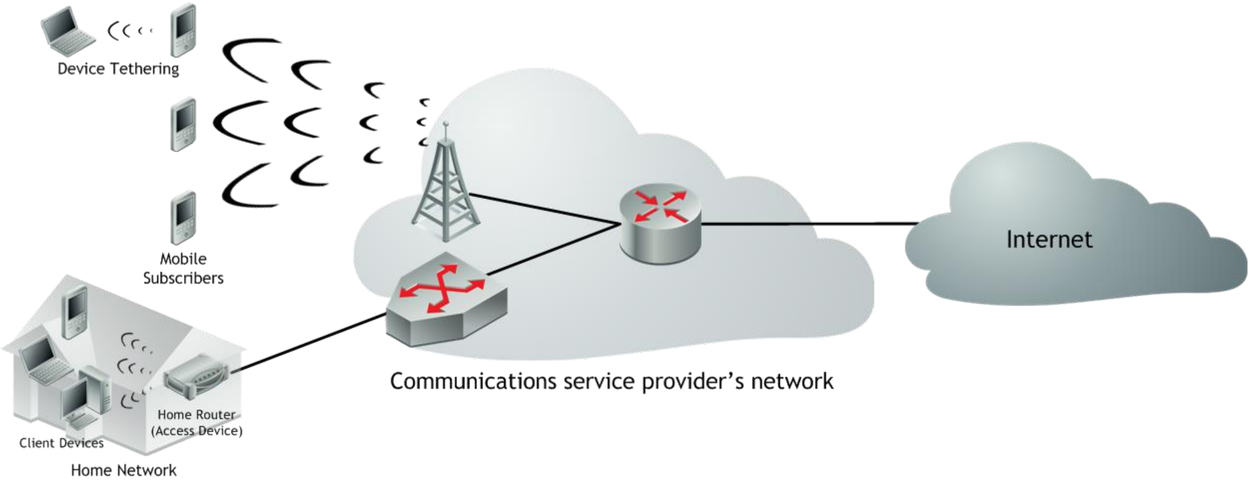

Device Tethering
Mobile Subscribers
Internet
Communications service provider's network
Home Router (Access Device)
Client Devices
Home Network

Figure 3 - Client and access devices in fixed and mobile access networks

Beyond simply differentiating between client and access devices, the next consideration is the information and insight available. For instance, are detailed measurements (e.g., application usage, video duration, quality of experience) available per-device? What can be gleaned about the device identity (e.g., manufacturer, model, operating system, browsers, etc.)? The richer this information, the richer the insight and, potentially, the richer the subscriber services that can be enabled.

Tethering Detection

Many CSPs want to offer tethering services as add-ons to existing data plans, but to do so they need to be able to detect and manage tethered devices. The most robust plans require policy control platforms that can apply separate policy to the tethered and access (i.e., hotspot) devices.

Network Address Translators

In the home network and in the case of tethering, and in other network environments (e.g., public WiFi hotspots), the client devices exist behind a network address translator (NAT). The NAT serves as a single point of access connection and, in effect, 'hides' the devices that are behind it. Detecting and identifying individual devices behind a NAT is a complex task, and one of which very few solutions are capable.

Real-Time Network Policy Control

All of this device information is useful from a business intelligence and strategy perspective, but becomes much more vital when it is available in real-time for network policy control. Naturally, passive/offline post-processing cannot be incorporate into real-time decisions and enforcement, but some solutions perform the device differentiation and tethering detection in real-time. It is important for CSPs to ask detailed questions in order to accurately understand a particular vendor's capabilities.

Encryption, Obfuscation, and Proxies

Information can be hidden or guarded in many ways, and two widespread mechanisms in the context of traffic classification are:

- **Encryption:** encoding information such that it can only be read by an authorized party
- **Obfuscation:** hiding or disguising information to prevent detection

Either or both of these general techniques might be used by any particular application, and the lines sometimes blur. For instance, consider:

- **Encryption to preserve content privacy:** Some applications encrypt user data and content as a privacy measure, but don't attempt to evade detection and management. As a significant example, Netflix traffic is carried via HTTP, but the video content itself is encrypted to prevent third-parties from inspecting video title information and revealing detailed individual viewing habits. The encryption method can be proprietary or based on a standard. Additionally, encryption is frequently employed as part of a digital rights management (DRM) strategy, in an attempt to control access to and reproduction of information[10].
- **Encryption as a means of obfuscation:** Some applications apply encryption in an attempt to evade detection and management by policy control solutions. For instance, BitTorrent clients have added increasing levels of encryption over the years[11].

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.[12]

Obfuscation measures vary widely, and are typically used to avoid detection and policy management. Early approaches randomized ports and moved information around within packets in order to overcome relatively simple pattern recognition algorithms.

Due to the prevalence of encryption and obfuscation measures, CSPs should quiz solution vendors comprehensively about specific capabilities, including how they provide traffic classification in the context of:

[10] Encryption both helps and hinders DRM, depending upon who is applying the encryption. Encrypted peer-to-peer filesharing defeats DRM strategies that inspect data for identifiers that correspond to licensed content, and laws/regulations that require CSPs to filter unlicensed content are ignorant of this technical reality. However, when the encryption is part of the DRM strategy itself it prevents unauthorized access and copying.

[11] An overview is available at http://en.wikipedia.org/wiki/BitTorrent_protocol_encryption

[12] For instance, 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.

- **SSL/TLS (Secure Sockets Layer and Transport Layer Security)[13]:** These are cryptographic protocols designed to provide secure communications, and are used extensively in applications where security is required (e.g., banking, VPNs, exchanging private data, 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. Historically, getting and maintaining an SSL certificate was cost-prohibitive for all but the larger web properties, but the Electronic Frontier Foundation's (EFF) HTTPS Everywhere initiative[14] looks to change that and will lead to wider adoption and use of SSL.
- **Virtual Private Networks (VPNs):** A VPN extends a private network across a public network, and includes security elements such as authentication and encryption (typically using SSL/TLS). VPNs are used extensively by enterprises to provide connectivity between sites and remote workers, but private VPN services are available specifically to provide encryption for Internet content.
- **SPDY[15]:** "Speedy" is an open networking protocol, developed primarily by Google, that modifies the way HTTP requests and responses are sent in the data path. The stated goals of SPDY are to reduce web page load latency and improve web security, and SPDY achieves these objectives via compression, multiplexing, and prioritization of HTTP traffic. Practically, the result of these measures is the same as if encryption was the intent, as content is obscured. Both the client (e.g., web browser) and web server need to support SPDY in order for it to be used; generally, SPDY is supported by the major web browsers and many major web services (e.g., Google, Twitter, Facebook, and WordPress). SPDY receives a great deal of attention, but in reality will only ever account for a small percentage of Internet traffic.[16]
- **QUIC (Quick UDP Internet Connections)[17]:** "Quick" is an experimental transport protocol developed by Google that has the same practical impact as encryption. QUIC is designed to provide security protection equivalent to TLS/SSL, with the added benefits of reduced latency. Like SPDY, QUIC requires both client and server support; at the time of this writing, QUIC is supported by Google Chrome and Google servers, but if the support for SPDY is any indication then QUIC will achieve widespread client and server support.
- **HTTP/2 (HTTP 2.0)[18]:** This next planned version of HTTP is based on SPDY (the draft of HTTP 2.0 published in November 2012 is based on a straight copy of SPDY); HTTP/2 is largely an effort to standardize SPDY implementations and to ensure backwards compatibility with HTTP 1.1 (the most recent standard, in use since 1999). That said, there are other differences between SPDY and HTTP/2; the main difference is that HTTP/2 allows multiplexing to happen at different hosts at the same time, to expedite downloading multiple web pages or content from multiple sources.

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. Broadly, proxies exist for a few reasons:

[13] An overview is available at: http://en.wikipedia.org/wiki/Transport_Layer_Security; the IETF RFC can be found here: http://tools.ietf.org/html/rfc5246

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

[15] General information about SPDY, including specific client and server support, is available at http://en.wikipedia.org/wiki/SPDY; protocol documentation can be found at http://www.chromium.org/spdy

[16] 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.

[17] More information can be found here: http://en.wikipedia.org/wiki/QUIC

[18] The latest IETF draft (at time of this writing) is available here: https://tools.ietf.org/html/draft-ietf-httpbis-http2-14; the Wikipedia summary can be found here: http://en.wikipedia.org/wiki/HTTP_2.0

- To decrease data usage (i.e., compression proxies)
- To increase performance by reducing latency (e.g., Google's SPDY Proxy)
- To hide from or subvert security measures

While there is sufficient overlap that mutually exclusive designations are difficult, in general CSPs should discuss the following with their solutions vendors:

- **Data Compression Proxies:** These are proxy services that provide data compression to users (with the intent of reducing bandwidth usage), and have the same practical impact to traffic classification as encryption. For instance, Google has a data compression proxy for Chrome[19], which can use a variety of protocols depending on what's available.
- **Proxy Applications:** These are applications that can be installed on (typically mobile) devices to provide users with privacy and more efficient data usage. Two of the most popular proxy applications are SPDY Proxy (see Figure 4) and Opera Max. Similarly, add-ons/plug-ins or configurations can instruct web browsers to use certain optimization protocols or techniques. For instance, Windows Phone has a Browser Optimization Service[20] that compresses data.
- **Web Proxies** (see Figure 5)**:** 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. Web proxies generally do not provide encryption[21].

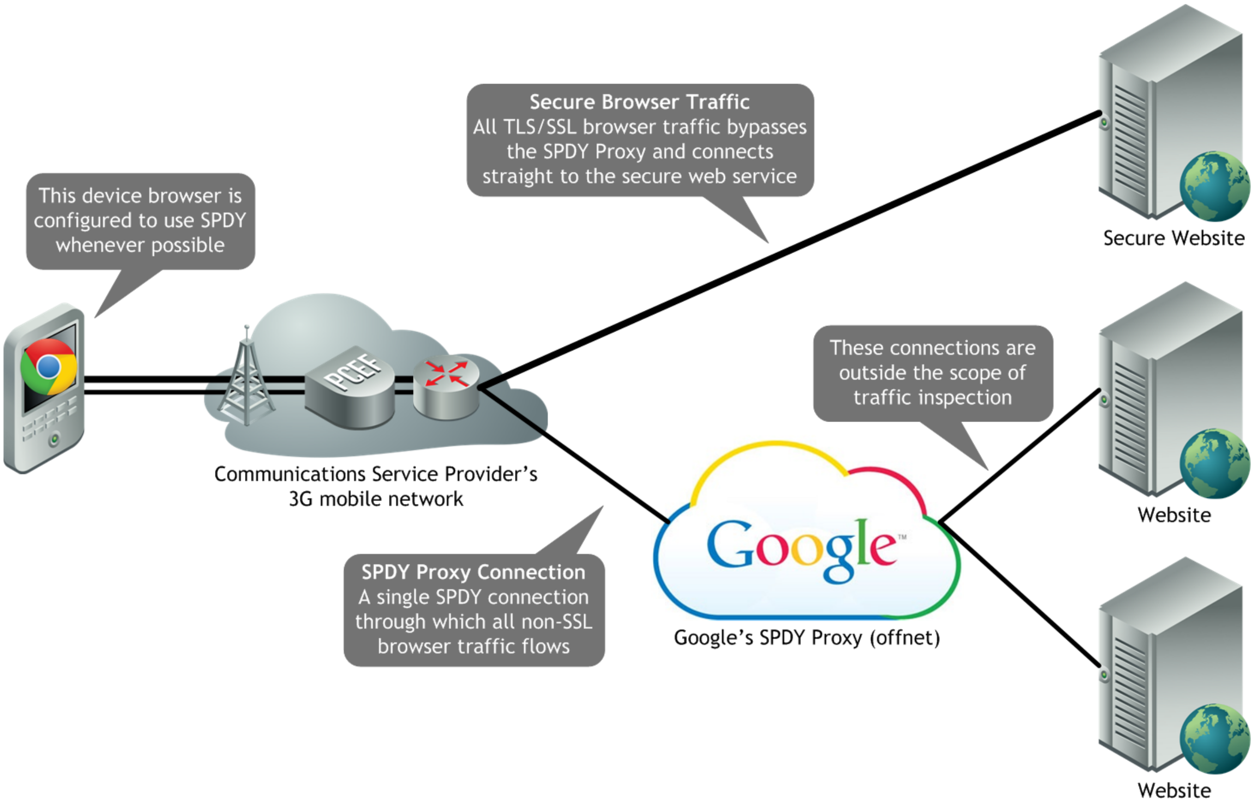

Secure Browser Traffic
All TLS/SSL browser traffic bypasses the SPDY Proxy and connects straight to the secure web service
This device browser is configured to use SPDY whenever possible
Secure Website
PCEF
Communications Service Provider's 3G mobile network
These connections are outside the scope of traffic inspection
Website
SPDY Proxy Connection
A single SPDY connection through which all non-SSL browser traffic flows
Google
Google's SPDY Proxy (offnet)
Website

Figure 4 - A device using Google's SPDY Proxy. Additionally, non-browser traffic does not use the proxy.

[19] Learn more about this service here: https://developer.chrome.com/multidevice/data-compression

[20] More information is available at https://dev.windowsphone.com/en-US/OEM/docs/Driver_Components/Browser_Optimization_Service

[21] 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

To accommodate commercial and deployment realities[22], it is possible for a network administrator (e.g., a CSP or an enterprise IT administrator) both to disable encryption and to force traffic to bypass the proxy entirely when the SPDY Proxy is in use.

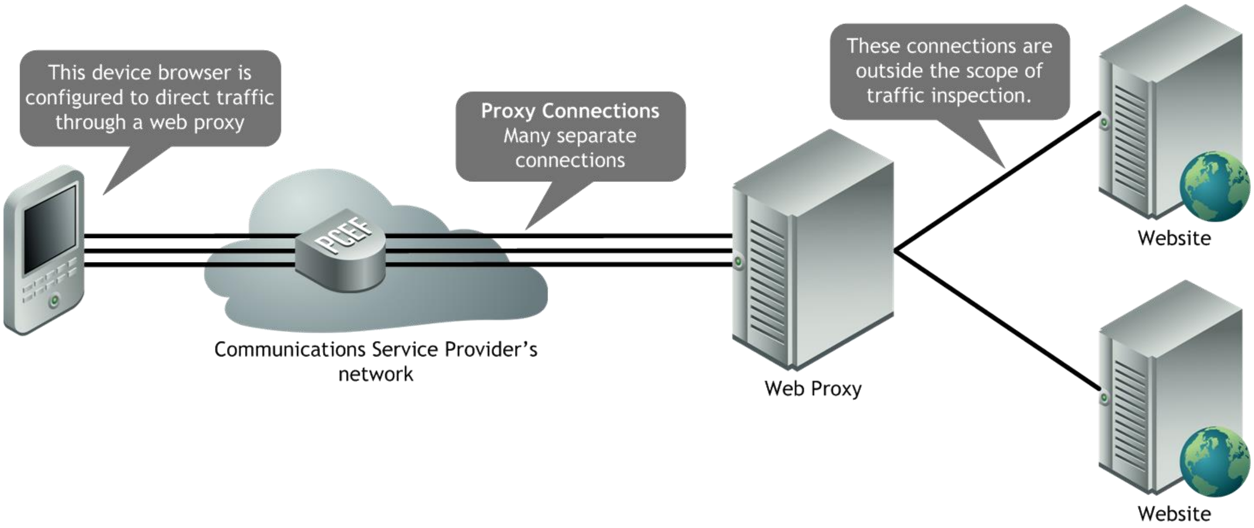

This device browser is configured to direct traffic through a web proxy
Proxy Connections
Many separate connections
These connections are outside the scope of traffic inspection.
PCEF
Communications Service Provider's network
Web Proxy
Website
Website

Figure 5 - Simplified view of a device using a web proxy service

An important difference between Figure 4 and Figure 5 is that when SPDY Proxy is in use (Figure 4), multiple web connections (on the right hand side of the diagram) are part of the same SPDY connection “SPDY Proxy Connection”); conversely, when the generic web proxy is in use, the connections remain separate.

Finally, it is worth explicitly noting that techniques to encrypt and obfuscate traffic are evolving rapidly, so it is vitally important that CSPs understand and assess their solution vendors’ capabilities to adapt to these changes. For instance, can software updates provide new capabilities in the field, or will a hardware upgrade be required? Can the traffic classification solution combine multiple techniques (e.g., measurements, analyzers, and heuristics)?[23]

[22] For instance, an IT administrator must enforce acceptable use policies and prevent data theft, and CSPs are often under regulatory requirements to filter inappropriate or illegal content. More information is available here: https://support.google.com/chrome/answer/3517349?hl=en

[23] For those especially interested in the subject, the GSMA has released a position paper called *Network Management of Encrypted Traffic*. In this paper, they (like this paper) explain various encryption technologies and then go on to make specific recommendations for “technical architects with knowledge of the operator network traffic management functions”.

Conclusions

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.

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.

Faced with such variation, CSPs must understand the technologies, trade-offs, and deployment challenges that exist in the context of traffic classification. For instance, to be useful in a modern network, solutions must be able to overcome routing asymmetry, the increasing adoption of encryption, and the prevalence of tunneling and encapsulation. To remain valuable as the Internet of Things emerges, solutions must be able to provide deep device-level insight.

Only when CSPs have this detailed understanding can they ask the right questions in order to truly understand what a vendor is providing, and any limitations that would otherwise be hidden.

Traffic Classification Requirements

After examining the many factors that must be considered when identifying and measuring Internet traffic, a number of requirements and questions emerge.

Consideration	Requirement/Question	Explanation
Traffic Classification Capabilities	Completeness: What is the percentage of traffic that fails to be positively identified (e.g., receives a label of “unrecognized”, “unknown”, “TCP”, “UDP”, etc.)?	The number of entries in a signature library is irrelevant – as there are many techniques to artificially inflate this count – so the best measure of completeness (i.e., how much traffic a CSP can expect to be recognized) is a metric about recognized and unrecognized traffic.
	Freshness: How does the vendor maintain an up-to-date signature library? With what frequency are updates made? How does the vendor ensure quality (especially when low turn-around times or frequent updates are promised)?	With the evolving nature of the Internet, it is important that a signature library keep pace. There is no perfect update frequency: the higher the frequency, the less time spent on quality assurance and the higher the cost to the CSP of continually updating; the lower the frequency, the higher expectation of accuracy but the larger the risk of failing to keep pace with rapid developments.
	False Positives: How do you ensure zero or	False positives are extremely problematic for

	minimal false positives? What is your priority: lower false positives or higher rates of recognition? Can all identified traffic be subject to congestion management policies?	CSPs and introduce risk, but the pursuit of higher rates of recognition can serve as an incentive to accept false positives within a signature library. A vendor must prioritize either eliminating false positives or increasing the rate of recognition, and the CSP should be aware of the prioritization.
	Additional Data: What data (beyond simply the identity of traffic) is made available? Is this data actionable in real-time?	Extra data (e.g., content provider, client device, media stream type, media container, video resolution, video codec, audio codec, operating system, browser, etc.) provides valuable business intelligence insight and can be useful as conditions in policy control.
	Measurements: What measured quantities are available? Are measurements actionable in real-time? Can the CSP define their own measurements?	Measurements (e.g., video duration, number of videos watched, video QoE, etc.) provide valuable business intelligence insight and can be useful as conditions in policy control.
Technical Requirements	Stateful Awareness: Does the solution have complete stateful awareness of protocols?	Many types of traffic (e.g., SIP, RTMP, FTP, etc.) can only be positively identified if the recognition technology has complete awareness of protocol state.
	Flow and Session Correlation: Does the solution correlate and apply signatures to the same asset across multiple transactions issued into the same, or different, connections?	Content is often split between and across flows. In many cases, a positive identification is only possible if the recognition solution can perform this correlation.
	Tunnels and Encapsulation: Does the solution provide identification and measurements within tunnels and encapsulation?	A significant portion of Internet traffic is carried in tunnels or is encapsulated; for maximum utility, the solution must be able to look within tunnels and encapsulation.
	Routing Asymmetry: Does traffic identification and measurement work for all types of traffic, in all routing environments, with any number of paths?	All broadband networks exhibit routing asymmetry of one form or another, but not all solutions can deliver traffic classification in the presence of this asymmetry.
	Devices: Does the solution provide device-level measurements? Can it differentiate between client and access devices? Can it see individual devices behind a NAT? Can the solution detect tethering? Can it apply separate policy to the tethered device, versus the access device?	The Internet of Things is upon us, and CSPs need deep device-level insight and policy control to accommodate and seize the opportunities provided by this enormous market shift.
	Encryption: How does the solution provide recognition when faced with encryption to preserve content privacy? Encryption or obfuscation to evade detection and management? SSL/TLS? SPDY? HTTP 2.0? Data compression proxies?	Encryption, obfuscation, and proxies come in many forms, and prevalence of each is increasing.

Additional Resources

In addition to the resources cited in the footnotes throughout this document, please consider reading these Sandvine technology showcases, all of which are available on www.sandvine.com:

- *Internet Traffic Classification* explains how Sandvine's traffic classification goes beyond traditional deep packet inspection (DPI) to deliver insightful measurements, to extract information, and to include all of these as conditions for real-time network policy control
- *Policy Control for Connected and Tethered Devices* explains how our traffic classification solution provides device-level insight and policy management

- *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

Invitation to Provide Feedback

Thank you for taking the time to read this whitepaper. We hope that you found it useful, and that it contributed to a greater understanding of the world of deep packet inspection and traffic classification.

If you have any feedback at all, then please get in touch with us at 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