APPLICATION CLASSIFICATION

CONTENTS

INTRODUCTION

Accurate classification of traffic flows is essential for network administrators to enable network tasks such as quality of service, detect threats, and restrict forbidden applications. Most tools utilize well-known signatures for network traffic classification. In recent years, more applications have begun encrypting traffic to protect subscriber content. With encrypted traffic on the rise, the amount of traffic that can be recognized by inspection of IANA-assigned port numbers and well-known application signatures has reduced drastically.

One solution to identify encrypted traffic is to decrypt the traffic before it gets into the local network. This is computationally very expensive, which brings network performance to a crawl or requires the organization to purchase expensive dedicated hardware. Since nearly half of data breaches occur within the network [Forrester, Intel], detection only at the edges is meaningless. Decrypting traffic also has legal implications as sensitive user information will be available in plaintext. Financial and healthcare regulations in many countries require PII information to be encrypted and stay hidden. Encryption does not necessarily mean the traffic cannot be identified, however it ensures that the content is private.

The 128T routers operate on the principle of applying intelligent heuristics to detect network traffic without decryption. The 128T routers can identify traffic in all routers not only at the edges. They can also share previously detected traffic information among themselves for quick detection. With a range of heuristics that can enable early detection and an accuracy of nearly 95%, the 128T routers enable networks to offer superior end-user experiences, protection, and reporting.

DETECTION TECHNIQUES

With the evolution of Internet traffic, both in terms of number and type of applications, traditional classification techniques such as those based on well-known port numbers or packet payload analysis are either no longer effective for all types of network traffic or cannot be deployed because of privacy or security concerns related to data.

There are three broad classification methods:

1. Port Numbers: This method is fast and low resource consuming. It is only useful for applications and services that use fixed port numbers. It is easy to cheat this technique by altering port numbers.
2. Deep Packet Inspection (DPI): This method inspects the actual payload of the packet. It is slow and requires huge computing resources. It can be very accurate except with applications that do not have strict classifiers. This requires up to date application signatures. It does not work when the payload is encrypted (or requires proxy decryption).
3. Statistical Classification: This method relies on statistical analysis of attributes such as byte frequencies, packet sizes and packet inter-arrival times. It can be very accurate. This technique requires implementation of machine learning algorithms.

ENCRYPTED TRAFFIC

The use of encryption for all communications on the Internet is increasing 90% year over year with roughly half of the websites today encrypting traffic by default [NSS]. It is expected that 75% of all web traffic will be encrypted by 2019. Most large websites, like Google, Twitter, and Facebook, use SSL encryption today. Google also started using HTTPS as a positive weight for websites in its search algorithm to encourage websites to use SSL or TLS.

Encrypted traffic renders DPI capabilities useless. One way to work around this is to decrypt the traffic. The client connects to a certificate they trust (SSL), the traffic is decrypted, inspected, and another connection is made to the target where the traffic is encrypted and sent (over another SSL). This computation adds a lot of performance overhead. This forces the organization to purchase expensive dedicated purpose-built hardware and still causes a huge performance impact. Most organizations usually cannot endure this.

Decrypting traffic to identify traffic also has legal implications since sensitive user data such as financial and health information is visible in plaintext. This places a huge risk on the organization on who has access to the information, where it is stored, how is it secured, and other issues.

Another possibility to prevent the performance impact of decryption for identification is to scale out the model by distributing the function to multiple devices rather than a single firewall or DPI engine. This worked well in the past when perimeter security was enough to secure the network. In the modern world where most network breaches occur from within the network, identifying applications only at the edges with a scale out model is insufficient. A scale out model also increases the amount of resources needed, which impacts budgets.

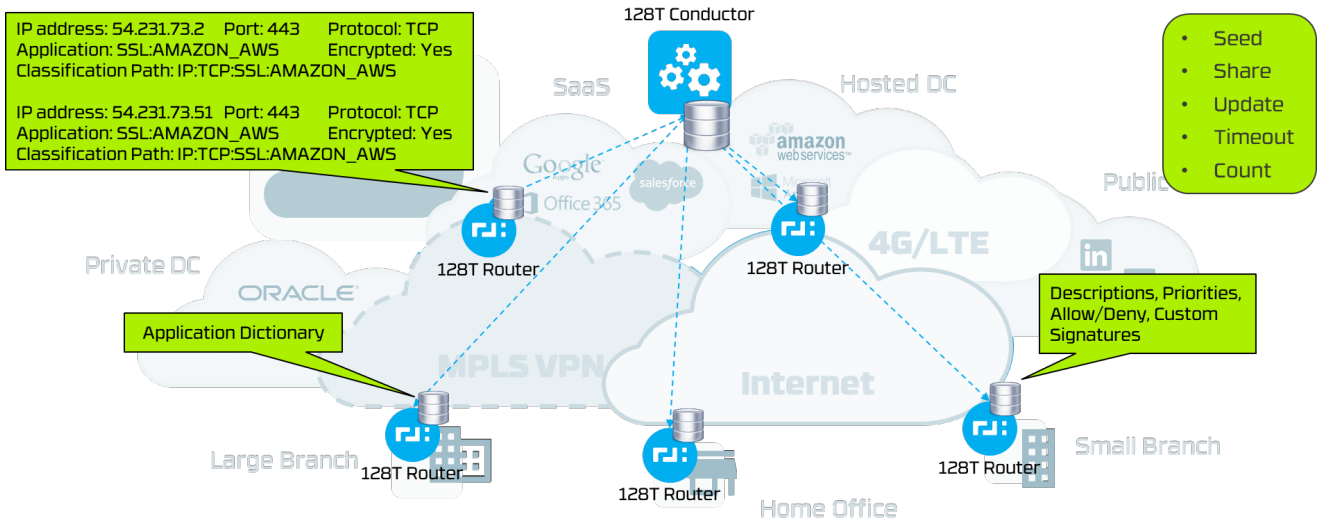
GOOGLE QUIC

One example of how traditional firewalls are unable to deal with encrypted traffic is how they work with QUIC (Quick UDP Internet Connections). QUIC is a transport layer network protocol developed by Google. Chrome browsers have the QUIC protocol enabled by default. When users try to access Google applications using the Chrome browser, a session to a Google server is established using the QUIC protocol. QUIC uses proprietary encryption methods. Traditional firewalls are unable to detect Google applications when they operate over QUIC. This results in loss of visibility and control of Google applications. Firewall vendors recommend blocking QUIC applications causing the Chrome browser to fall back to using traditional TLS/SSL. While this works today, the Chrome browser is unable to benefit from reduced connection and transport latency that QUIC brings. If Google makes it mandatory for Chrome browsers to support QUIC, the ability to cause it to fallback to TLS/SSL encryption will fail. This shows that traditional methods of decryption used by existing firewall and router vendors are insufficient and not suitable for the modern world.

128T NETWORKING PLATFORM

The 128T Networking Platform focuses on using best practices from multiple detection techniques. One way to prevent load on the router is to limit the amount of traffic that needs to be analyzed. Any flow through a

router only needs to be analyzed once (for a given period of time). If a new flow with the same characteristics is seen again, then the 128T router can intelligently decide the flow classification based on its earlier analysis. This reduces the amount of traffic that needs to be analyzed. Session based routing fits into this model very well as it operates on flows rather than on packets. The router can share this analysis with neighboring routers, enabling them to perform early detection as well.



The 128T router analyzes data and control traffic using a rich set of heuristics to identify traffic. Before any analysis, the 128T router checks to see if a session with similar properties has been previously seen in the network and has already been identified. It will do so by checking with the application dictionary which has a set of previously identified sessions. If the session has been identified then the session will be classified accordingly.

If the session has not been identified then the 128T router uses a set of heuristics to determine the session type of the traffic. The administrator has the ability to control the set of heuristics applied. The administrator can choose to speed up traffic flows with minimal or coarse detection. The administrator can also choose to enable full detection.

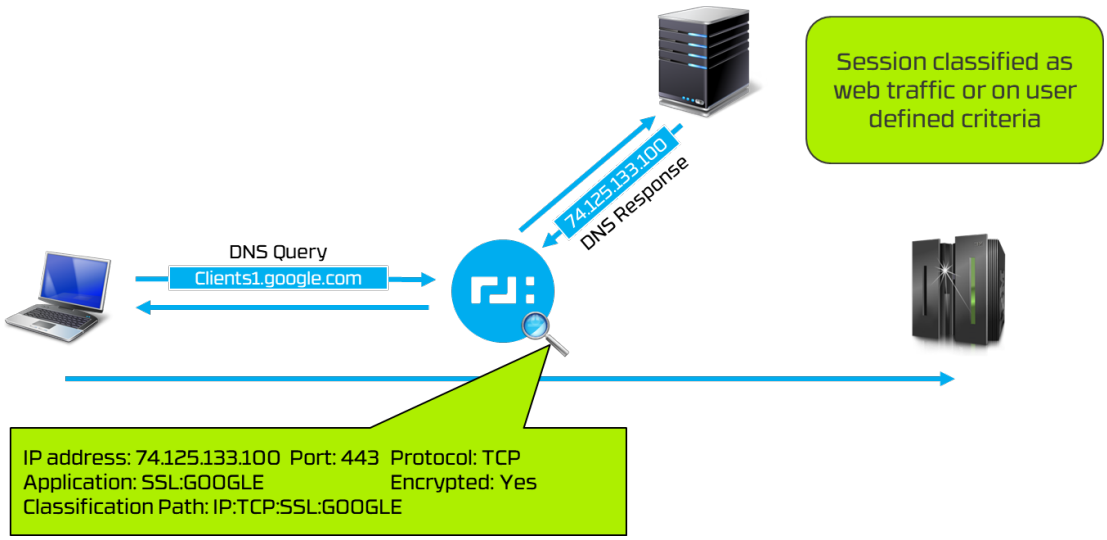
HEURISTICS

Any flow seen for the first time and not in the dictionary is classified as UNKNOWN and be treated using existing system behavior. The 128T routers then use heuristics to classify the sessions, taking appropriate action when the classification is discerned.

DNS Identification

The 128T router examines packets in DNS flows and updates the dictionary with relevant information. For many deployments where the 128T router is at the edge of a branch site, it will see many DNS queries go through it. If the 128T router receives a new flow for which no DNS query was seen then no action will be taken and the flow may continue to remain UNKNOWN if it is not identified by another heuristic. It is possible that sessions may not involve a physical query with a DNS server through the 128T router due to local

caching of DNS information. P2P and gaming applications usually do not rely on DNS queries and hence cannot be identified by this heuristic.



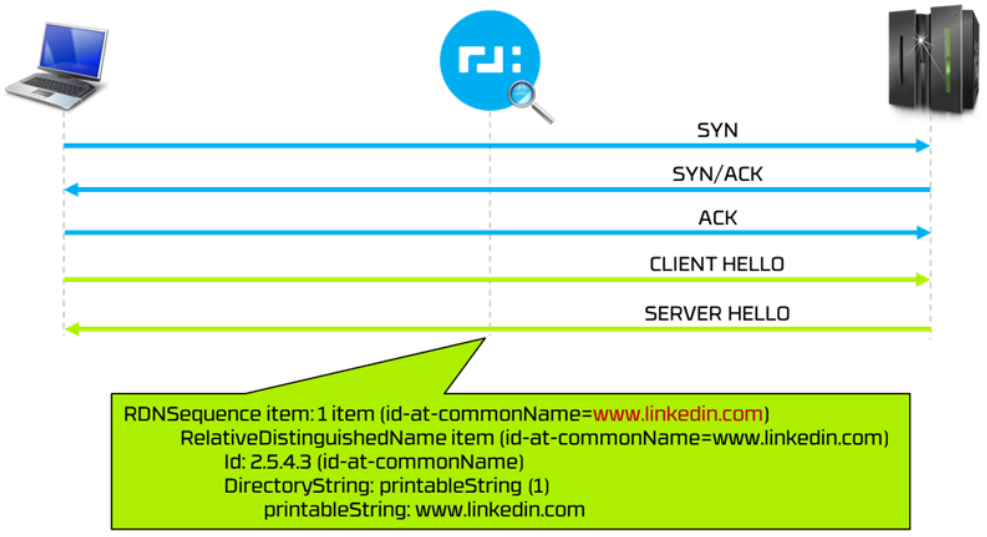
Internet connections are often anticipated by DNS query and response packets. For example, a web browser will resolve clients1.google.com to 173.194.65.104 before attempting a TCP connection to the corresponding web server. By intercepting the DNS response packet, the 128T router can uncover the original domain name entered by the user into the web browser location bar.

The packets will be dissected according to RFC1035. The 128T router will be able to identify the domain name and connection type. Combining this information with port numbers will improve the classification performance.

HTTPS Identification

The 128T router can identify applications that use HTTP over SSL/TLS or HTTPS without performing decryption. During the SSL encrypted session, the 128T router receives server "hello packets", which has the certificate details or the server can send a separate certificate packet. The 128T router will look for the X.509 digital certificate received from the server and inspects the common name field in the SSL Handshake Protocol.

For example, if a user accesses https://www.linkedin.com, the common name in the server certificate has www.linkedin.com and the 128T router will identify the application as web traffic or as defined by the user.



The common name is the server certificate must be in a complete host-domain format or equal to the name of the web address being accessed. If the common name includes a wildcard such as *.google.com then the application is identified as SSL. For example, if a user accesses http://www.youtube.com, the common name in the server certificate has *.google.com and the 128T router will identify the application as SSL. Further analysis is required to uniquely identify the application based on other heuristics.

There are other interesting elements in this exchange that are noted for analysis/reporting. The client hello message exchanged during the initial SSL/TLS handshake of the HTTPS connection contains elements that do not change with each client connection. These are noted in the dictionary and are used to make intelligent decisions associated with the session. The 128T router notes the following elements from the client hello: SSL/TLS protocol version, cipher suite list, compression, and TLS extensions.

Elements that change are not used in the dictionary for making decisions for future flows to avoid ambiguous results.

Based on HTTPS connections processed, the distribution of SSL/TLS versions is displayed for reporting. For example, the user can see data related to this for a specific time period. The administrator can configure a rule in the action to deny certain connections thereby deprecating the use of those versions. For example, the administrator could block the use of older and vulnerable SSL 3.0 protocol. The user can also see the list of cipher suites being used in the network and supported by different clients.

As more heuristics are executed, session classification continues to improve. For example, a session may go from being UNKNOWN to P2P to BITTORRENT.

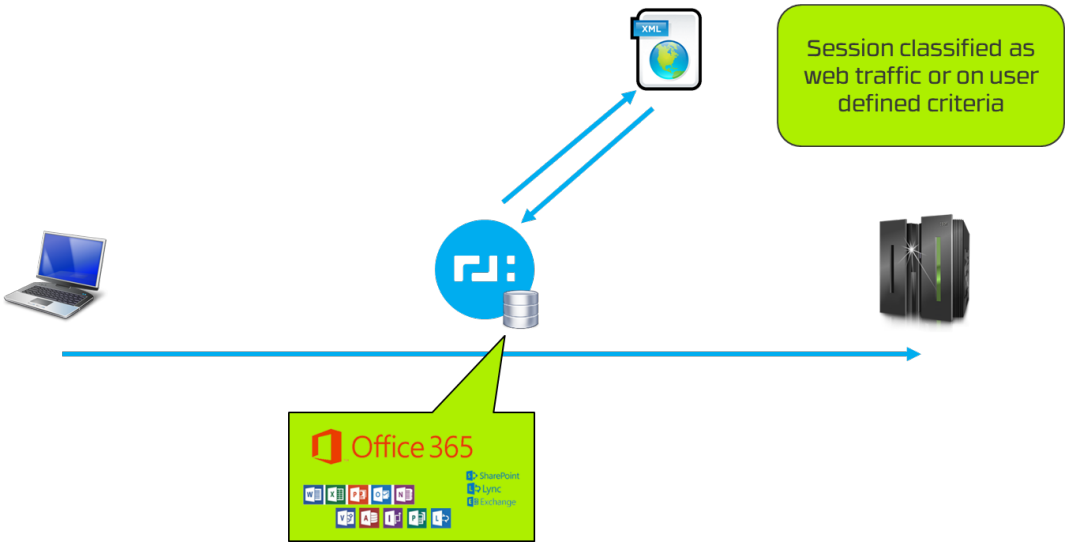
Well Known Applications Identification

Some well-known applications can be classified by using a fully up-to-date list of IP addresses, URLs, and ports. For example, to access Office 365 applications via a web browser through Microsoft's private cloud, the application uses different FQDNs which Microsoft regularly updates within a XML based file. The high

amounts of Office 365 traffic have resulted in the need for enterprises to route and classify this traffic differently than other internet bound traffic.

Office 365

As customers migrate to Office 365 they need to allow and provide special consideration to various workloads they might use in the Office 365 product sets, such as Skype for Business, OneNote, Exchange Online and so on. Microsoft publishes Office 365 over a huge range of URLs, and IP addresses. Microsoft dynamically publishes a fully up-to-date list of all IPs, URLs and ports used by each of the 17 components of Office 365 every hour.



To classify Office 365 traffic, a 128T router obtains IP addresses and FQDNs from the Microsoft site and updates the dictionary for the purpose of classifying Office 365 traffic. A network administrator can assign actions to take for such traffic depending on the needs of the organization.

APPLICATION DICTIONARY

The application dictionary maintains information needed to be able to identify flows in future based on the data in the dictionary. The dictionary is shared with 128T routers in the Authority. Local routers use part of this localized information for classification. 128T routers are able to download information from this dictionary. They are able to update the dictionary and the updates are propagated to other 128T routers as necessary. The dictionary also maintains a timeout and a count of popular flows in the network. The administrator can set a limit on the number of entries in the dictionary to limit its size.

Signatures

The application dictionary saves mappings between application type and the corresponding destination IP address, destination port, protocol type, and service. Administrators are also able to configure the period of time that mappings are maintained, and after that time, the mappings are removed from the database. To

minimize the impact on performance, application dictionary is refreshed only when TCP or UDP traffic triggers a lookup. Once an application is identified based on a heuristic, the information is stored in the dictionary. An example of this mapping is as follows:

IP address: 54.231.73.2 Port: 443 Protocol: TCP
Application: SSL:AMAZON_AWS Encrypted: Yes
Classification Path: IP:TCP:SSL:AMAZON_AWS

IP address: 54.231.73.51 Port: 443 Protocol: TCP
Application: SSL:AMAZON_AWS Encrypted: Yes
Classification Path: IP:TCP:SSL:AMAZON_AWS

The 128T routers also include other relevant information in the application dictionary such as evasiveness, vulnerabilities, descriptions, priorities, allow or deny operations, etc. based on defined application characteristics. The administrator can also enter custom application signatures into the dictionary.

These signatures are applied to classify new or subsequent flows and take appropriate action. For example, an administrator can block Facebook Farmville. In that case the 128T router would block all flows to the destination IP address identified for Facebook Farmville.

Seed

The application dictionary can be seeded with a pre-populated base library of signatures to enable early detection capabilities out of the box.

ACTION

When the 128T router sees a new flow that cannot be matched with the dictionary it is classified as UNKNOWN and assigned best-effort service.

When a flow is classified of a particular type based on the information in the dictionary or through heuristics, the following actions can be taken based on administrative preferences:

Allow or Deny

The administrator can set a policy to deny a particular session. For example, the administrator can set a policy to restrict access to www.facebook.com. By default, all sessions are set to be allowed.

Apply Traffic Shaping

By default, the 128T routers will assign best-effort service class to any UNKNOWN flows unless configured otherwise by the administrator. For identified flows the 128T router will map traffic as per the following table:

Traffic Type	Traffic Classes

Voice (G.711), Interactive Video (Telepresence), Conferencing (WebEx, Fuze)	3 (Real Time)
Network Control (BGP, OSPF), Network Management (SNMP, Syslog, SSH), Signaling (SIP, H.323), Transactional Data (ERP, CRM)	2 (Critical)
Best Effort (Default)	1 (Best Effort)
Scavenger (YouTube, iTunes, BitTorrent)	0 (Scavenger)

The administrator has the privilege to change the default mapping of flows to QoS classes.

Allow for Group

The administrator may set a particular session type to be allowed or denied only on certain groups of routers. Groups in this case refers to the groups of routers related to the 128T schema.

Allow Policy based Forwarding

The administrator can associate actions for sessions belonging to certain types to prefer one path over another for multihomed connections.

Actions can be combined in different ways to achieve desired results. For example, the administrator may choose to send Facebook traffic over only the Internet path on certain groups of routers.

SUMMARY

Traffic application classification is an essential step in the network management process to provide high availability of network services. The ability to classify applications enables the network to fine tune performance per service to provide superior end-user experiences. The 128T Networking Platform enables early detection of applications and information sharing among routers to ensure superior detection capabilities. The ability to take actions based on this classification guarantees performances per application type. The platform does not rely on decryption which results in performance bottlenecks and legal implications. The intelligent use of heuristics and the ability to continuously add new heuristics ensures the 128T Networking Platform provides continuous best of breed solution for application classification.

www.128technology.com | info@128technology.com