nDPI: Open-Source High-Speed Deep Packet Inspection

Luca Deri*†, Maurizio Martinelli*, Tomasz Bujlow‡, and Alfredo Cardigliano†
*IIT/CNR, Pisa, Italy. E-mails: {luca.deri, maurizio.martinelli}@iit.cnr.it
†ntop, Pisa, Italy. E-mails: {deri, cardigliano}@ntop.org
‡Aalborg University, Denmark. E-mail: tbu@es.aau.dk

***Abstract*—Network traffic analysis was traditionally limited to packet header, because the transport protocol and application ports were usually sufficient to identify the application protocol. With the advent of port-independent, peer-to-peer, and encrypted protocols, the task of identifying application protocols became increasingly challenging, thus creating a motivation for creating tools and libraries for network protocol classification. This paper covers the design and implementation of nDPI, an open-source library for protocol classification using both packet header and payload. nDPI was extensively validated in various monitoring projects ranging from Linux kernel protocol classification, to analysis of 10 Gbit traffic, reporting both high protocol detection accuracy and efficiency.**

Keywords*—*Passive traffic classification, Deep Packet Inspection, network traffic monitoring

I. Introduction

In the early days of the Internet, network traffic protocols were identified by a protocol and port. For instance, SMTP used TCP port 25 while telnet used TCP port 23. This well-know protocol/port association is specified in the */etc/protocols* file, which is a part of every Unix-based operating system. Over the time, with the advent of Remote Procedure Call (RPC), the use of static ports became a problem. Therefore, specific applications such as *rpcbind* and *portmap* were developed to handle dynamic mappings. Historically, application ports up to 1024 identified essential system services, such as email or remote system login and hence require super-user privileges; their port-to-protocol bindings are preserved until today. Ports above 1024 are used for user-defined services and are generally dynamic.

Protocol identification is often not reliable even when a static port is used. A case in point is TCP/80 used for HTTP. Originally, HTTP was created to carry web-related resources such as HTML pages and decorative content. However, its extensibility (in no small part due to its header flexibility and MIME type specification) along with its native integration in web browsers HTTP is now often used to carry non web-related resources. For instance, it is now the de-facto protocol for downloading/uploading files, thus replacing the File Transfer Protocol (FTP), which was designed specifically for that purpose. The pervasive use of HTTP and its native support of firewalls (i.e., a firewall recognizes and validates the protocol header), made HTTP (and its secure counterpart HTTPS) the ideal developer choice when creating a new protocol that has to traverse a firewall without restrictions. Many peer-to-peer protocols and popular applications (e.g., Skype) use HTTP as the last resort when they need to pass through a firewall in case all other ports are blocked. We created traffic reports from various networks, ranging from academic sites to commercial ISPs, and realized that HTTP is by far the most widely used protocol. This does not mean that users mostly use it for surfing the web. This protocol is extensively used by social networks, geographical maps, and video-streaming services. In other words the equation TCP/80 = web is no longer valid.

The characterization of network protocols is required not only for creating accurate network traffic reports, but increasingly, for overall network security needs. Modern firewalls combine IP/protocol/port based security with selected protocol inspection in order to validate protocols, in particular those based on UDP, e.g., Simple Network Management Protocol (SNMP) and Domain Name System (DNS). VoIP protocols, such as SIP and H.323, are inspected for specific information, e.g., the IP and port where voice and video will flow. Cisco Network-based Application Recognition (NBAR) devices [1], and Palo Alto Networks application-based firewalls [2] pioneered application-protocol based traffic management. Today, these traffic inspection facilities are available on every modern network security device, because the binding port/protocol scheme no longer holds.

The need to increase network traffic visibility created a need for Deep Traffic Inspection (DPI) libraries to replace the first generation of port-based tools [3]. Payload analysis [4], however, uses extensive computational resource [5]. This difficulty triggered the development of statistical analysis based approaches [6] often based on Machine-Learning Algorithms (MLAs) [7], [8] instead of direct payload inspection. These methods often rely on statistical protocol properties such as packet size and intra-packet arrival time distributions. Although some authors claim these algorithms provide high detection accuracy [9], [10], real-life tests [11]–[16] demonstrated that:

- Such protocols are able to classify only a few traffic categories (an order of magnitude less than DPI libraries) and thus less suitable for fine protocol granularity detection applications.
- Some tests show a significant rate of inaccuracy suggesting that such methods may be useful in passive traffic analysis, but unlikely to be used for mission critical applications, such as traffic blocking.

These needs constitute the motivation for developing an efficient open-source DPI library where efficiency is defined by the requirement to monitor 10 Gbps traffic using solely

978-1-4799-0959-9/14/$31.00 ©2014 IEEE

commodity hardware (i.e., not specialized hardware needed). The use of open-source is essential because:

- Commercial DPI libraries are very expensive both in terms of one-time license fee and yearly maintenance costs. Sometimes their price is set based on yearly customers revenues, rather than on a fixed per-license fee, thus further complicating the price scheme.
- Closed-source DPI toolkits are often not extensible by end-users. This means that developers willing to add new/custom protocols support need to request these changes to the toolkits manufacturer. In essence, users are therefore at the mercy of DPI library vendors in terms of cost and schedule.
- Open-source tools cannot incorporate commercial DPI libraries as they are subject to a Non-Disclosure Agreement (NDA) that makes them unsuitable to be mixed with open-source software and included into the operating system kernel.

Although DPI was a hot topic for a long time, beside some rare exceptions, most research works did not lead to the creation of a publicly available DPI toolkit but limited their scope to prototypes or prof-of-concept tools. The need to create an efficient open-source DPI library for network monitoring was the motivation for this work.

The rest of the paper is structured as follows. Section II describes the motivation of this work and explains the differences from its predecessor OpenDPI [17]. Section III covers the nDPI design and implementation, while section IV describes the validation process. Section V concludes the paper. The current work on nDPI and the future plans are described in Section VI.

II. Background and Motivation

DPI is defined as the analysis of a packet's data payload in real time (i.e., DPI processing must be faster than the traffic rate to be monitored as otherwise it would result in packet drops) at a given physical location. Inspection is motivated by various reasons including application protocol identification, traffic pattern analysis, and metadata (e.g., user name) extraction. Some proprietary DPI library vendors such as iPoque, QOSMOS, and Vineyard cover all aspects, whereas others, such as libprotoident [18], UPC [19], L7-filter [20], and TIE [21] limit their scope to protocol identification [22]–[24].

Protocol detection may also be implemented using pattern matching or by using specialized protocol decoders. The former approach is inefficient due to the use of regular-expressions [25] and error-prone because it does not reconstruct packets in 6-tuple flows (VLAN, Protocol, IP/-port source/destination), thus missing cross-packet matches. Searching for patterns within an un-decoded payload can also lead to out of context search data (e.g., an email including an excerpt from a HTTP connection might be confused with web-traffic) or mismatches when specific packet fields (e.g., NetBIOS host name) are encoded.

Application drive the selection of the appropriate DPI library. We chose to focus on network traffic monitoring that can range from passive packet analysis to active inline packet policy enforcement. A DPI library must include the following features:

- High-reliability protocol detection for inline, per application, protocol policy enforcement.
- Library extensibility is needed for new protocols and runtime in sub-protocols definition. This feature is required because new protocols appear from time to time or evolve (e.g., the Skype protocol changed significantly since after the Microsoft acquisition). Permanent library maintenance is, therefore, required.
- Ability to integrate under an open-source license for use by existing open-source applications and embedding into an operating system's kernel. As already discussed, full source code availability is essential to safeguard privacy.
- Extraction of basic network metrics (e.g., network and application latency) and metadata (e.g., DNS query/response) that can be used within monitoring applications thus avoiding duplicate packet decoding, once in the DPI library and also in the monitoring application.

Our focus is, therefore, reliable protocol detection using protocol decoders combined with the ability to extract selected metadata parameter for the use of applications that is this library. This enables the extraction of selected metadata parameters that can then be used by applications using the DPI library. Other open-source protocol detection libraries, such as libprotoident, are limited in scope because they do not extract metadata and only analyze the first 4 B of payload in each direction for protocol detection. Because commercial DPI libraries could not be used a starting basis, we chose OpenDPI, an open-source predecessor of the commercial iPoque Protocol and Application Classification Engine (PACE), which is no longer maintained by its developers. OpenDPI was designed to be both an application protocol detection and metadata extraction library. Because it was unmaintained for some time, the library did not include any modern protocol (e.g., Skype); the code was largely prototype quality and likely used as a proof of concept for the commercial product. A point in favor of OpenDPI was the fact that it was distributed under the GPLv3 license that allows developers to include it in software applications without being bound to an NDA or other restrictions typical of commercial DPI products. Furthermore, an open-source license allows the code to be inspected, key requirement when the packet payload is inspected and potentially private information might leak. Our choice of OpenDPI as the starting point was driven by these reasons.

A. From OpenDPI to nDPI

The OpenDPI library is written in C and it is divided in two main components: the core library (responsible for handling raw packets, decoding IP layers 3 and 4, and extracting basic information such as IP address and port) and plugin dissectors (responsible for detecting the ~100 protocols supported by OpenDPI). nDPI inherited this two-layer architecture but it addressed several issues present in the OpenDPI design:

- The OpenDPI library was designed to be extensible, but in practice the data structures used internally were

static. For instance, many data-types and bitmaps, used to keep the state for all supported protocols, were bound to specific sizes (e.g., 128 bits) and thus limiting the number of identifiable protocols.

- Whenever a protocol was detected, the library tried to find further protocol matches instead of just returning the first match. The result was a performance penalty without a real need of requiring extra detection work.
- No encrypted protocol support (e.g., HTTPS). While encryption is designed to preserve privacy and regular DPI libraries are not expected to decode the some information can be gleaned to suggest the nature of the information carried on a specific connection.
- OpenDPI was not designed to be a reentrant (i.e., thread-safe) library as it used shared global variables. This required multi-threaded applications to create several instances of the library or add semaphores in order to avoid multiple threads to modify the same data at the same time. Per thread library state was required to support reentrancy.
- Many parts of OpenDPI suggest problematic design choices. For instance, the library was performing per-flow initializations, instead of doing them at once. As the result, the applications using the library paid an unnecessary performance penalty whenever a new connection was passed to OpenDPI.
- The protocol dissection was non-hierarchical. In other words, whenever a new connection needed to be analyzed, the library was not applying the dissectors based on their matching probability. For instance, if there is a connection on TCP port 80, OpenDPI was not trying the HTTP dissector first, but it was applying dissectors in the same order as they were registered in the library.
- The library had no runtime configuration capability; the only way to define new dissectors was to code them in C. While this is usually a good policy for efficiency reasons, at times more flexibility is needed. For instance, if a given user needs to define a custom protocol Y as TCP/port-X it would be easier to have a runtime configuration directive instead of changing the library code. OpenDPI assumes that the library must have a dissector for all supported protocols, a difficult goal to achieve in reality. In particular, in closed-environments such as a LAN, specific hosts use proprietary/custom protocols that flow on specific ports/protocols. In this case, it is more convenient for the user to detect them from the packet header rather than from its payload.
- OpenDPI was not designed to extract any metadata from analyzed traffic. On one hand this preserves privacy, but on the other it requires monitoring applications to decode the application traffic again in order to extract basic information such as the URL from HTTP traffic. Reporting this information does not add any overhead to the library as it is decoded anyway when parsing the packet payload.

Summarizing, OpenDPI was a good starting point for nDPI, because we did not have to start from scratch. Many components of the original library were changed in order to address the issues we identified. This was the ground work necessary to start the creation of an efficient DPI library and extending the set of supported protocols. Not surprisingly, the number of protocols recognized has an impact on both DPI detection performance and protocol recognition. The more protocols recognized, the more time spent on detection whenever a specific traffic pattern is not identified and thus all the possible protocol decoders have to be tested for match. This means that DPI libraries supporting many protocols may be slower in specific situation than those supporting fewer. Another impact on performance is due to metadata extraction: the richer the set of extracted attributes, the slower the processing. Although specific activities such as string and pattern matching can be accelerated on specialized hardware platforms such as Cavium and RMI, or using GPUs [26], we decided not to use any of these cards, in order to let the library operate on all hardware platforms.

nDPI was designed to be used by applications that need to detect the application protocol of communication flow. Its focus is on Internet traffic, thus all the available dissectors support standard protocols (e.g., HTTP and SMTP) or selected proprietary ones (e.g., Skype or Citrix) that are popular across the Internet community. In the latter case, as protocol specifications are not publicly available, we had to create the dissectors by reverse-engineering network traffic. Although nDPI can extract specific metadata (e.g., HTTP URL) from analyzed traffic, it was not designed as a library to be used in fields such as lawful interception or data leak prevention; its primary goal is to characterize network traffic. Similarly to OpenDPI, nDPI can be used both inside the Linux kernel and in user-space applications, and it work on most operating systems including Linux, Windows, MacOS X, as well as non-Intel CPU architectures such as ARM and MIPS.

III. nDPI Design and Implementation

nDPI defines an application protocol by a unique numeric protocol Id and a symbolic protocol name (e.g., Skype). Applications using nDPI will probably use the protocol Id whereas humans the corresponding name. The protocol set includes both network protocols such as SMTP or DNS, and communications over network protocols. For instance, Facebook and Twitter are treated as two protocols, although that in fact they are communications from/to Facebook/Twitter servers. A protocol is usually detected by a traffic dissector written in C, but it can be defined also in terms of protocol/port, IP address (e.g., traffic from/to specific networks), and protocol attributes. For instance, the Dropbox traffic is identified by both the dissector for LAN-based communications, and by tagging as Dropbox the HTTP traffic on which the 'Host' header field is set to '*.dropbox.com'. As explained later in this section, nDPI supports over 170 protocols, but it can also be further extended at runtime using a configuration file.

The nDPI library inherits some of OpenDPI design, where the library code is used for implementing general functions, and protocol dissection is implemented in plugins. All the library code is now fully reentrant, meaning that applications based on nDPI do not need to use locks or other techniques to

serialize operations. All the library initialization is performed only once at startup, without a runtime penalty when a new packet needs to be dissected. nDPI expects the caller to provide the packet divided in flows (i.e., set of packets with the same VLAN, protocol, IP/port source/destination), and that the packet was decoded up to layer 3. This means that the caller has to handle all the layer-2 encapsulations, such as VLAN and MPLS, by leaving to nDPI the task of decoding the packet from the IP layer up. nDPI comes with a simple test application named pcapReader.c[1] that shows how to implement packet classification and provides utility functions for efficient flow processing. The protocol dissectors are registered with attributes, such as the default protocol and port. For instance, the HTTP dissector specifies the default TCP/80, and the DNS dissector TCP/UDP on port 53. This practice has two advantages:

- Packets belonging to a yet unclassified flow are passed to all registered dissectors, starting from the most likely one. For instance, a TCP packet on port 80 is first passed to the HTTP protocol and then (if not identified) to the remaining registered dissectors. Of course only dissectors for TCP protocols are considered, which reduces the number of dissectors that are tested, and decreases the matching time. This optimization does not prevent detecting HTTP on non-standard ports.
- When a flow is still unclassified (none of the dissectors matched), nDPI can guess the application protocol based on the transport protocol and port. Note that a flow can be unclassified not just because of protocol dissectors limitations, but also because not all flow packets where passed to nDPI. A typical example is the case when nDPI has to dissect packets belonging to a flow whose beginning was not analyzed (e.g., nDPI was activated after the flow start).

The protocol recognition lifecycle for a new flow begins from decoding layers 3 and 4 of the packet. In case there is a dissector registered for the packet protocol/port, such dissector is tried first. In case of no match, all the compatible dissectors (i.e., in case of a UDP packet, all UDP dissectors) are tried. If a dissector cannot match a packet, it has two options: either the match failed because the analyzed packet will never match (e.g., a DNS packet passed to the SNMP dissector), or it failed but it may be that future packets will match. In the former case, the dissector will not be considered for future packets belonging to the same flow, whereas in the latter the dissector will still be considered for future packets belonging to the same flow. Protocol detection ends as soon as a dissector matches.

A typical question asked by nDPI users is how many packets are needed to detect the application protocol or to decide that the given flow is unknown. From our experience, the number is protocol dependent. For most UDP-based protocols, such as DNS, NetFlow, or SNMP one packet is enough to make this decision. Unfortunately, there are other UDP-based protocols such as BitTorrent, whose signature might require up to 8 packets in order to be detected. This leads us to the rule of thumb that for nDPI at most 8 packets per direction are enough to make a decision.

[1]The application source code is available at https://svn.ntop.org/svn/ntop/trunk/nDPI/example/pcapReader.c

A. Handling Encrypted Traffic

The trend of Internet traffic is towards encrypted communications. Due to security and privacy concerns, HTTPS is slowly replacing HTTP not just for secure transactions but also for sending tweets and messages to mobile terminals, posting notes, and performing searches. Identifying this traffic as SSL is not enough, but it is necessary to better characterize it. When using encrypted communications, the only part of the data exchange that can be decoded is the initial key exchange. nDPI contains a decoder for SSL that extracts the host name from the server certificate. This information is placed in the nDPI flow metadata in a similar way as the server name from the 'Host:' HTTP header. With this approach we can identify known services and tag them according to the server name. For instance, an encrypted communication towards a server named 'api.twitter.com' is marked as Twitter, 'maps.google.com' as Google maps, and '*.whatsapp.net' as the WhatsApp messaging protocol. We can also discover self-signed SSL certificates, which is important as it might indicate that the connection is not safe, not just in terms of data leak, but also in terms of the activity behind the communication. For instance symmetric (i.e., the traffic is not predominant in one direction such as in HTTPS, where the client sends little traffic with respect to the traffic sent by the server) long standing SSL connections with self-signed certificates often hide SSL VPNs.

As described later in this section, nDPI contains internally a configuration for many known protocols that are discovered using the above technique. In addition, it is possible to add at runtime a configuration file that further extends the set of detected protocols so that new ones can be defined without changing the protocol dissector. With the advent of Content Delivery Networks (CDN) this is probably the only way of identifying the application protocol, as at any given time the same server (identified with a single IP address) can deliver two different services provided by two customers using the same CDN. As a fallback, nDPI can identify specific application protocols using the IP address. For instance, nDPI detects many Apple-provided services, such as iTunes and iMessage. In addition to that, it marks as Apple (a generic protocol) all communications that were not specifically identified, but that were exchanged with the Apple-registered IP addresses (i.e., 17.0.0.0/8).

B. Extending nDPI

As previously explained, nDPI users can define protocols not just by adding a new protocol dissector, but also providing a configuration file at runtime. New protocols are defined by name. When nDPI detects that a protocol name is already defined (e.g., in the above example SIP and HTTP are handled by the native dissector), the configuration file extends the default configuration already present in nDPI. For instance in the previous example, whenever nDPI sees TCP traffic on port 81 or 8181 it tags it as HTTP. Additionally, nDPI can also identify a protocol using strings that are matched against metadata extracted from the nDPI flow such as HTTP Host and SSL certificate server name. The defined strings are stored by

Table I. EVALUATION OF THE LATEST VERSION OF NDPI.

Protocol	Flows	% correct	% wrong	% uncl.
DNS	18 251	100.00	0.00	0.00
HTTP	42 983	97.80	0.66	1.54
ICMP	205	100.00	0.00	0.00
IMAP (Start-TLS)	35	100.00	0.00	0.00
IMAP (TLS)	103	100.00	0.00	0.00
NETBIOS (Name Service)	10 199	99.97	0.00	0.03
NETBIOS (Session Service)	11	100.00	0.00	0.00
Samba Service	42 808	100.00	0.00	0.00
NTP	42 227	100.00	0.00	0.00
POP3 (plain mode)	26	100.00	0.00	0.00
POP3 (TLS)	101	100.00	0.00	0.00
RTMP	378	70.90	15.87	13.23
SMTP (plain mode)	67	100.00	0.00	0.00
SMTP (TLS)	52	100.00	0.00	0.00
SOCKSv5	1 927	92.99	0.00	7.01
SSH	38 961	93.98	0.80	5.22
BitTorrent (encrypted mode)	96 399	54.41	0.18	45.41
BitTorrent (mixed mode)	261 527	99.41	0.02	0.57
DropBox	93	98.92	0.00	1.08
eMule (obfuscated)	12 835	11.04	2.67	86.29
eMule (mixed mode)	13 852	17.57	2.28	80.15
FTP (active mode)	126	98.41	0.00	1.59
FTP (passive mode)	122	72.95	0.00	27.05
iTunes	235	93.62	0.00	6.38
Pando Media Booster	13 453	99.26	0.63	0.11
PPLive	1 510	43.91	1.06	55.03
RDP	153 837	99.69	0.02	0.29
Skype	2 177	92.38	7.44	0.18
Sopcast	424	63.68	8.49	27.83
Steam	1 205	76.02	0.42	23.57
TOR	185	33.51	0.00	66.49
web: Amazon	602	83.89	0.00	16.11
web: Apple	477	74.63	0.00	25.37
web: Facebook	6 953	80.14	0.00	19.86
web: Google	6 541	83.03	0.02	16.95
web Wikipedia	6 092	68.96	0.23	30.81
web: Yahoo!	17 373	83.16	0.02	16.82
web: YouTube	2 534	82.16	0.00	17.84

an automata based on the Multifast[2] library that implements string matching according to the Aho-Corasick algorithm. This library is quite efficient: at startup the automata creation takes little time (i.e., almost instantaneous with tenth of strings, or some seconds with hundred thousand strings), then this library configured performs over 10 Gbps during search when configured with hundred thousand strings.

IV. NDPI VALIDATION

There are recent papers that compare the nDPI accuracy in terms of protocol detection against other DPI toolkits. Their conclusion is that "nDPI and libprotoident were successful at correctly classifying most (although admittedly not all) of the applications that we examined and only one of the evaluated applications could not be classified by both tools" [14], and "the best accuracy we obtained from nDPI (91 points), PACE (82 points), UPC MLA (79 points), and Libprotoident (78 points)" [11]. These tests showed that nDPI is pretty accurate, even more accurate than PACE, the commercial version of the old OpenDPI library on which nDPI is based. We are aware that nDPI had some false positives with Skype and BitTorrent due to the use of heuristics. In the further nDPI versions (svn revision 7249 or newer), we decided to remove the use of these heuristics, so that we basically eliminated false positives at the cost of slightly increasing the number of undetected flows when using these two protocols. We also re-implemented support for several other protocols, as FTP, SOCKSv4, SOCKSv5, eDonkey, PPLive, Steam, RTMP, and Pando. Furthermore, we added the ability to discover new web services, e.g., Wikipedia and Amazon.

```
# taskset -c 1 ./pcapReader -i ~/test.pcap
Using nDPI (r7253)
pcap file contains
IP packets: 3000543 of 3295278 packets
IP bytes:1043493248(avg pkt size 316 bytes)
Unique flows: 500
nDPI throughput: 3.42 M pps / 8.85 Gb/sec
```

Figure 1. nDPI Validation Test Outcome.

The latest version of nDPI was tested using the same methodology and the same dataset as in [27], so we could compare the results and show the impact of the latest changes. In essence, the ground-truth was established by a host-based traffic monitoring system, which marks flows by the process names obtained from the system sockets. The results in the term of percent of flows classified correctly, incorrectly, and left unclassified for the most popular protocols, applications, and web services are shown in Table I. Sometimes, the results were obtained on another level than the test presumed. For example, when we tested for the ability to detect the HTTP traffic, we also obtained results as *Google*, *Facebook*, or *DropBox*. In this evaluation, we acknowledged such results as correct, even though the protocol was not explicitly given. That could lead to a minor inaccuracy in case if the detected *Facebook* traffic was in fact not HTTP (but, for example, SSL). This is the only difference from the tests performed in [27], where only the results on the asked level were considered as correct (e.g., *HTTP*), while the others (e.g., *Facebook*) were considered as being unclassified).

As there are many extensive tests on nDPI protocol detection accuracy, this paper focuses on nDPI performance. For that purpose, we developed an application named pcapReader[3] that can both capture from a physical network device and read packets from a pcap file. To test nDPI on a physical network at 10 Gbps, we used the test application on top of PF_RING [28], which allows applications on commodity hardware to process packets in RX/TX at 10 Gbps line rate for any packet size. For our tests, we used a pcap file of over 3 million packets, captured on a heterogeneous environment, thus including both LAN protocols (e.g., NFS and NetBios) and Internet protocols (e.g., Skype and DropBox). We used a PC running Ubuntu Linux 13.10 (kernel 3.11.0-15) on a 8 core Intel i7 860. We bound the application to a single core, in order to test it in the worst case, and see how the application can scale when using multiple cores. The test outcome is depicted in Figure 1.

The outcome demonstrated that the test application processes packets at an average speed of 3.5 Mpps / 8.85 Gbps using a single core. As the test pcap file used during the test was captured on a real network, it contained some flows that began before the packet capture started. nDPI detects a flow protocol by looking at the initial flow packets, so some flows are not detected due to this reason. For undetected flows, nDPI can guess the protocol by using the flow protocol/port registered during startup or it can leave the flows undetected. When using this test application over PF_RING DNA on a 10 Gbps Intel adapter, it is possible to use the network driver with hardware flow balancing. In this way, we can start one instance of the test application per virtual queue, binding each

[2]http://multifast.sourceforge.net

[3]https://svn.ntop.org/svn/ntop/trunk/nDPI/example/pcapReader.c

instance to a different core. In sum, 10 Gbps traffic can be inspected when balanced across two cores (the above tests show DPI at 8 Gbps using a single core), using the modestly priced commodity hardware we used in our tests.

In terms of memory usage, nDPI needs some memory to load the configuration and automatas used for string-based matching. This memory used by nDPI is ~210 KB with no custom configuration loaded that increases of ~25 KB when a custom configuration is loaded. In addition to that, nDPI keeps per-flow information that is independent from the application protocols and which takes ~1 KB per flow.

V. Conclusion

This paper presents nDPI, an open-source toolkit released under GPLv3 license available at https://svn.ntop.org/svn/ntop/trunk/nDPI/. It is currently able to detect more than 170 protocols including Skype, BitTorrent, and other messaging protocols. The validation test performed by third parties demonstrated that nDPI outperforms some commercial and open-source toolkits in terms of protocol recognition accuracy. In terms of performance, using two CPU cores and commodity hardware, nDPI can handle a 10 Gbit link fully loaded with Internet traffic. This makes it suitable for scenarios where both detection accuracy and high performance are a requirement.

VI. Final Remarks

The development of nDPI is still ongoing. There are several important features, which are missing in the current version of nDPI, however, they are planned to be included in the official SVN trunk in the near future.

The most important drawback of the current implementation (as well as of the majority of the currently existing traffic classification tools) is the quality of the provided results. Habitually, the classifiers provide only result per flow, which is supposed to characterize the flow in the most detailed manner. Therefore, the output is a mix of results on various levels: IP protocols (i.e., TCP or UDP), application protocols (e.g., DNS, HTTP, SSL, BitTorrent, or SMTPS), types of the content (e.g., MPEG, or Flash), and finally, the service providers (e.g., Facebook, or YouTube). The usefulness of such result is very limited. At first, one flow can use several application protocols (as HTTP and Dropbox). At second, one application protocol (as DNS) can use several IP protocols (TCP and UDP). At third, it is impossible to judge if the most detailed level is the content (as Flash) or the service (as YouTube). Finally, this scheme does not allow to provide precise accounting of the traffic (for example, it is not possible to account the HTTP traffic, if the results are given on multiple levels, so in majority of cases HTTP is even not mentioned). Therefore, we are changing the format of the results, so all the possible classifications will be consistently provided.

References

[1] Cisco, Network Based Application Recognition (NBAR), 2008.

[2] Palo Alto Networks, Next-Generation Firewall Overview, 2011.

[3] M. Mellia, A. Pescapè, L. Salgarelli, Traffic Classification and its Applications to Modern Networks. Computer Networks, 2009.

[4] F. Risso, M. Baldi, O. Morandi, A. Baldini, and P. Monclus. Lightweight, Payload-based Traffic Classification: An Experimental Evaluation, Proceedings of IEEE ICC '08, 2008.

[5] M. Avalle, F. Risso, R. Sisto, Efficient Multistriding of Large Nondeterministic Finite State Automata for Deep Packet Inspection, Proceedings of ICC 2012, 2012.

[6] M. Crotti, M. Dusi, F. Gringoli, L. Salgarelli, Traffic Classification Through Simple Statistical Fingerprinting. ACM SIGCOMM Computer CommunicationReview, January 2007.

[7] T. Bujlow, R. Tahir, J. Myrup Pedersen, A method for classification of network traffic based on C5. 0 Machine Learning Algorithm, Proceedings of ICNC 2012, 2012.

[8] T. T. T. Nguyen and G. Armitage. A Survey of Techniques for Internet Traffic Classificationusing Machine Learning, IEEE Communications Surveys & Tutorials, 2008.

[9] S. Ubik, P. Zejdl, Evaluating application-layer classification usinga Machine Learning technique over different high speed networks. 2010 Fifth International Conference on Systems and Networks Communications, IEEE 2010, pp. 387–391.

[10] J. Li, S. Zhang, Y. Lu, Z. Zhang, Internet Traffic Classification Using Machine Learning, Proceedings of CHINACOM '07, August 2007.

[11] T. Bujlow, V. Carela-Español, P. Barlet-Ros, Comparison of Deep Packet Inspection (DPI) Tools for Traffic Classification, Technical Report, Version 3, June 2013.

[12] M. Dusi, F. Gringoli, and L. Salgarelli, Quantifying the accuracy of the ground truth associated with Internet traffic traces, International Journal of Computer and Telecommunications Networking, Vol. 55 Issue 5, 2011

[13] N. Cascarano, L. Ciminiera, and Fulvio Risso, Optimizing Deep Packet Inspection for High-Speed Traffic Analysis, Journal of Network and System Management, 2011.

[14] S. Alcock, R. Nelson, Measuring the Accuracy of Open-Source Payload-Based Traffic Classifiers Using Popular Internet Applications, IEEE Workshop on Network Measurements (WNM), 2013.

[15] R. Goss, R. Botha, Deep Packet Inspection—Fear of the unknown, Proceedings of ISSA 2010, 2010.

[16] S. Valenti, D. Rossi, A. Dainotti, A. Pescapè, A. Finamore, and M. Mellia, Reviewing Traffic Classification, Data Traffic Monitoring and Analysis, 2013.

[17] Y. Wei, Z. Yun-feng, L.Guo, Analysis of Message Identification for OpenDPI, Computer Engineering, 2011.

[18] S. Alcock, R. Nelson, Libprotoident: Traffic Classification Using Lightweight Packet Inspection ,Technical report, University of Waikato. http://www.wand.net.nz/publications/lpireport, 2013.

[19] Clear Foundation, L7-filter, http://l7-filter.clearfoundation.com/.

[20] T. Bujlow, T. Riaz, J.M. Pedersen, A Method for classification of network traffic based on C5.0 Machine Learning Algorithm, in proceedings of ICNC'12, pp. 244–248, February 2012.

[21] A. Dainotti, W. de Donato, and A. Pescapè, Tie: A Community-Oriented Traffic Classification Platform, Traffic Monitoring and Analysis, 2009.

[22] A. R. Khakpour, A. X. Liu. High-Speed Flow Nature Identification, Proceedings of ICDCS'09, 2009.

[23] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, K. Lee. Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices, Proceedings of ACM CoNEXT2008, 2008.

[24] G. Aceto, A. Dainotti, W. d. Donato, A. Pescapè, Portload: Taking the Best of Two Worlds in Traffic Classification, Proceedings of INFOCOM IEEE Conference 2010, 2010.

[25] S. Kumar, P. Crowley, Algorithms to Accelerate Multiple Regular Expressions Matching for Deep Packet Inspection, Proceedings of SIGCOMM '06, 2006.

[26] N. Cascarano, P. Rolando, F. Risso, and R. Sisto, Infant: NFAPattern Matching on GPGPUDevices, Computer Communication Review, 2010.

[27] T. Bujlow, V. Carela-Español, P. Barlet-Ros, Extended Independent Comparison of Popular Deep Packet Inspection (DPI) Tools for Traffic Classification, Technical Report, January 2014.

[28] F. Fusco, L. Deri, High Speed Network Traffic Analysis with Commodity Multi-core System, Proceedings of IMC 2010 Conference, November 2010.