Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement für Verteidigung,
Bevölkerungsschutz und Sport VBS

armasuisse
Wissenschaft und Technologie W+T

Master Thesis Task Assignment for Fabian Gasser (D-ITET)

SDN-based Source Dilution Service

Main advisor:	Dr. Vincent Lenders (armasuisse)
Advisor ETH	Dr. Bernhard Ager (ETH Zürich)
Supervisor	Prof. Dr. Adrian Perrig (ETH Zürich)
Start Date:	16. June 2014
End Date:	16. December 2014

1 Motivation

Enterprises increasingly rely on open information sources in the Internet to make intelligent and critical business decisions. To this end, crawling and information retrieval are common techniques that are used in combinations to automate and accelerate the collection and processing of data. However in practice, the crawling of open sources is not as simple as it sounds because content providers have deployed various restrictive controls which aim at limiting the download rates, showing different types of content to different classes of people, denying access to particular users that they regard as e.g. competitors, or track interests and generate user profiles as revealed by the crawling behavior. In order to circumvent these problems, this thesis shall explore the feasibility to implement a source dilution service using the software-defined networking (SDN) paradigm. The goal is to develop a scalable network enterprise service that allows crawlers to spread their requests to servers in the Internet using a diversity of source IP addresses in order to make the requests appear as if they were coming from different locations and hosts. By spreading requests from a crawler to many different source addresses, the servers of interest will have increased difficulties to limit, deny or track access to selected clients as the source IP address is often used as a key indicator for the remote identity of sessions.

2 Why Not Using Tor?

IP address anonymization has been a hot topic in research for more than a decade. Out of this research, Tor [1] has been developed and deployed as the most successful world-wide IP anonymization service. A valid question is therefore why not using Tor for circumventing rate limits, censorship, and tracking in the context of crawling?

While Tor provides IP anonymity to its users, it faces two main challenges that makes the system unsuitable for the envisioned purpose of crawling:

1. **Blacklisting:** Tor was originally designed to obfuscate the identity of a user from an eavesdropper in the network. Tor has no built-in mechanisms to hide the mere fact the user is using an anonymization service. As a result, it is fairly straightforward for a server to detect that someone is connecting through Tor. The list of Tor exit nodes is public and by blacklisting all these exit nodes, a server can easily restrict access to the entire set of Tor users. In fact, a series of popular Web services do this already. As soon as someone is trying to connect over Tor, a captcha is presented to the user or the connection is refused.

2. **Speed:** Tor was designed to provide anonymity guarantees against a strong adversary that can control a large portion of the Internet. While useful to protect user anonymity, the adversary model is different with crawling. A crawler wants to avoid getting identified and targeted by the remote server while the anonymity against an adversary sitting in the network is not a real threat. In order to provide such kinds of guarantees, Tor relies on onion routing which adds a significant overhead to the communication and unnecessarily limits the data capacity that may be achieved over the network in our case. A more lightweight approach with higher capacity is therefore desired for crawling at line speed.

3 Thesis Objectives

The main objective of this work is to provide a source address dilution service to applications which can specify how their traffic should be spread over a pool of available outgoing IP addresses. We envision a controllable service interface that every host in an enterprise can use to specify parameters such as the maximum number of requests per source IP, the data rate requirements, the desired source IP diversity, the desired geo-diversity, etc. After initial definition of the desired dilution by the crawler instance, the network then transparently routes the crawler traffic to a suitable border gateway with a flow-level outgoing IP address as determined by the policy of the crawler.

For the implementation of a network address dilution service, the student shall exploit the software-defined networking (SDN) paradigm. SDN has been proposed as a more flexible way of routing flows in packet-switched networks. Whereas traditional routers make routing decisions based on IP prefix matching rules, SDN allows a much fine-granular way of making routing decisions. In SDN, a switch/router is able to make matching rules on any header fields of the network/transport protocol headers. Additionally, SDN switches are capable of modifying the header of packets as required to perform network address translation (NAT).

The goal of the this thesis is to design, implement and evaluate such a dilution service in the context of an enterprise scenario with multiple data centers. The data centers shall each have a set of available IP prefixes that can be assumed to be managed by the border gateways or routers to the Internet. To evaluate the basic concept of a dilution service, a testbed shall be developed and tested consisting of a mix of virtual and hardware OpenFlow switches.

4 Thesis Tasks

The tasks of this thesis are

1. Set up a virtualized SDN infrastructure with at least five Open vSwitches [2] on a physical server modelling a single data center.
2. Develop a service interface for the definition of service policies that will be used by crawling applications. The interface shall allow applications to specify dilution parameters such as the maximum number of requests per IP, the data rate requirements, the desired IP diversity, the desired geo-diversity, etc.
3. Evaluate different SDN controllers (e.g., POX [3]) that would be suitable to control the SDN switches when implementing the dilution service.
4. Develop your SDN controller that switches incoming flows from crawlers to outgoing IPs from ISPs according to their service policies. In addition to switching to the correct ISP, the SDN switches shall perform NAT to translate the local source IP of the crawlers to the pool of global IPs from the ISPs. Stick to the OpenFlow specification for the control of the switches in order to facilitate deployment to hardware switches later on.
5. Implement an example crawler that makes use of your service interface to retrieve data from the Internet.
6. Test the functionality and evaluate the performance of the system when crawling data from the Internet.

7. Extend your testbed to an extended setup with a second datacenter. The second datacenter should consist of hardware SDN switches provided by armasuisse. The second datacenter shall be located at a different location with a different set of ISPs to the Internet.
8. Evaluate your system with two datacenters. In particular, evaluate the benefits and drawbacks of virtual switches vs. hardware switches as well as the impact of having a distributed architecture with geographically distributed datacenters..

5 Deliverables

- At the end of the second week, a detailed time schedule of the thesis must be given and discussed with the advisors.
- At the end of the second month, a short discussion of 15 minutes with the supervisor and the advisors will take place. The student has to talk about the major aspects of the ongoing work using slides.
- At the end of month four, another meeting with the supervisor will take place. At this point, the student should already have a preliminary version of the written report or at least a table of content to hand in to the supervisor. This preliminary version should be brought along to the short discussion.
- At the end of the thesis, a presentation of 15 minutes must be given at armasuisse and at ETH (in English). The presentations should give an overview as well as the most important aspects of the work. If possible, a demonstrator should be presented (offline after the talk).
- The final report should be written in English but may be written in German. It must contain a summary written in both English and German, the assignment and the time schedule. Its structure should include an introduction, an analysis of related work, and a complete documentation of all used hardware/software tools. Two written copies of the final report must be delivered to the main advisor along with DVD that includes developments undergone during the thesis.

References

[1] Tor Project, https://www.torproject.org/

[2] Open vSwitch, http://openvswitch.org/

[3] About POX, http://www.noxrepo.org/pox/about-pox/

armasuisse

Science and Technology
C4I Networks

Dr. Vincent Lenders
Thun, May 12th 2014