Exploration of Network Flow Processing for Geolocation and Cyber Defence

Sean Pennefather

March 11, 2013

1 Interpretation of Problem Statement

The goal of this project is to explore Flow Processing on TCP/IP networks and consider the advantages of the resulting data reduction for applications that deal with large volumes of traffic. Raw flows are already capable of being exported by many modern routers such as those developed by Cisco [4] and are used in traffic analysis for network administrators to assist in the maintenance of running a smooth network and to minimise traffic congestion. For this, applications have been developed such as NetFlow Traffic Analyser [11] to interpret collected raw flows and produce a graphic display of network usage. Though these applications and tools exist to render data for traffic analysis, the applicability of raw flow interpretation for Cyber Defence is not fully investigated. A focus of this project should be on such an investigation and an evaluation of performance gains when compared to the equivalent Cyber Defence tasks that focus on packet analysis. A prototype system should be generated that preforms the tasks using flow processing and represents data in a way that is more applicable to network defence. Another focus of this project will be on Geolocation of flows and flow comparison based on origin.

2 Objective of Research

This research has two primary objectives:

- To investigate the effectiveness of Network Flow processing in improving the performance of tasks associated with traffic analysis and Cyber Defence.
- To perform Geolocation on network flows to determine network flow origin and associate the physical region with the flow source or destination. This will help in identifying the origin of unusual network traffic and allow interpreted data to be filtered by region.

3 History and Background

The physical infrastructure and commonly implemented protocols that make network communication possible were designed to transfer information between the source and destination sinks in the shortest time while maintaining a best-effort attempt at a level of reliability in data transfer. Further high level protocols are employed by end hosts to provide reliability and authenticity of connections and data. Due to the focus on communication speed, data and connection security is not implemented at low level protocols and, as network routers operate only with the network level protocol and lower, they are unable to distinguish malicious traffic from legitimate traffic. As a result of this, network based attacks such worms, Denial of Service (DoS) and Distributed Denial of Service attacks (DDoS) are a serious threat to network dependent companies, especially those providing online services.

The goal of a DoS or DDoS attack to to deny a service provided by the victim from being accessed by legitimate clients. These DoS and DDos attacks can be further classified according to the degree of automation of bot recruitment and attack, and the vulnerability that the attack is exploiting. Further classification can happen according to rate of attack and the resulting impact to the victim [8].

In response to DoS and DDoS attacks, preventative and reactive methods are being developed to either minimise the impact and cost to a victim when a DoS attack occurs, or attempt to negate the chance of a successful DoS attack being executed. These methods can be loosely grouped into 4 categories namely; the detection of an attack, the reaction to an attack, the prevention of an attack, and the identification of the attack source [10].

Another threat to end hosts connected to a network is infection of self propagating malware code such as a worm which exploits a vulnerability in the victim to allow the malware to be executed and deliver the attached payload. The infected machine then seeds the worm by scanning other end hosts for a exploitable vulnerability and uploading the malware to those hosts [12].

Though research has been done into using flows to identify DoS intrusions, research into other fields of Cyber Defence that could benefit from the data reduction offered by flow processing is sparse. Identification and evaluation of these other areas will be a focus of this research project along with developing a prototype system to display the effectiveness of handling these tasks from a flow perspective. Worm detection using NetFlow has been researched [9] but will not be considered outside the scope of this project and may be considered when evaluating unusual traffic. The prototype system will assume input flow in either NetFlow or IPFIX format.

NetFlow is a network protocol that was developed by Cisco in the 1990s to represent network traffic as flows [2]. A network flow, as defined by Cisco, is a unidirectional stream of packets where each packet belonging to the same stream have the following characteristics in common:

- The same source and destination IP address.
- The same source and destination port number.
- Are being transmitted under the same protocol.
- Are being transmitted as the same type of service (as defined by the TOS field in the datagram).
- The packets are being transmitted from the same source interface (ifIndex).

NetFlow is already implemented in Cisco routers to log flows as packet streams are routed through them which can then be collected and processed by a stream analyser. The analysis software such as FlowScan can then interpret the raw flow data and give the user a representation of the traffic through that router [1]. This technique is more data efficient as it only records details identifying individual flows, not individual packets and so less entries need to be recorded. This data reduction also reduces the volume of raw data that must be processed for network analysis which reduces the required computational power.

Due to these advantages, The Internet Engineering Task Force (IETF) is attempting to create a standardised flow protocol named IP Flow Information Export (IPFIX) which is based on NetFlow revision 9 [5]. The IPFIX protocol defines a flow as a sequence of packets that pass through a network point and share the following common characteristics:

- Packets have components, application, and transport fields in common.
- Packets share properties such as content length.
- Packets have the same indicators of transport such as source interface.

[5]

Geolocation has become a popular research field as it allows mappings of the internet to be associated with a physical representation of the world. This has advantages to disaster prevention groups as it can show the physical location of bottlenecks and crucial points that need improved monitoring or backup network paths to minimise disruption during a disaster in that region. Advertisers can also use Geolocation to generate specific advertisements associated with regions with improves the marketing of products to target audiences. Current Geolocation techniques involve requesting the associated location for a specific IP address from a Geolocation database such as those offered by regional Internet registries [3].

Beyond the location of the subnet, other techniques are needed to attempt a greater resolution of location. Such techniques include using the propagation delays between network paths to attempt to determine distance between network nodes [7] and Time Difference of Arrival techniques [6].

4 Approach

The origin of the network flow data will not be in the scope of the experiment and tests will assume flow input from an external system. This way, collected and processed data can be reprocessed and the results can be compared.

Initially, hardware implementation will not considered and the test bed will be purely software based. Should tests prove to be successful, Hardware implantation such as an NPU, fpga or gpu may become an option to simulate the flow processing at clock speeds comparable to that of network processing units. Due to this possibility, hardware consideration will be taken into account when designing the software component of the project. The component responsible for flow processing and Geoloaction will be separated out from the remainder of the system and will be interfaced with through calls as though it were a physically separate component.

The research for this project will be broken up into 5 phases:

- Identification of possible Cyber Defence tasks that can be improved by using flow processing instead of packet analysis. Investigation of other areas that could benefit from flow processing as well as potential areas that evolve from flow processing such as identification of fake flows may also be an option to investigate.
- Investigation of NPUs and the advantages or disadvantages of separating the flow processing into a separate system. This can include performance gains and security risks.
- Development of a prototype system to test the tasks using flows.
- Comparison of the results of the system with the packet focused counterparts based on an evaluation of efficiency, speed, and accuracy of results.
- Investigation of network flow for Geolocation and associated flow density.

An important note is that the current approach could change considerably depending on what tasks are discovered and the availability of source data and comparable results.

5 Plan

The process of determining the effectiveness of flow processing for Cyber Defence and Geolocation will be broken down into 7 stages:

1. Research Cyber Defence tasks that focus on traffic analysis to determine potential candidates that would be improved with the use of network flows.

2. Gathering of source data for the selected tasks to be tested on. This data will most probably be stored packets which the results could be generated for. Those same packets could then be used to create the equivalent raw flows that would have been detected by a compatible router according to the definition of a flow put forward by Cisco [2].

3. Investigation of Network Processing Units (NPU) with focus on architecture and devlopment languages as well as how the input and output is handled.

4. Creation of a prototype system to test the identified tasks using the collected or synthesised data and generating results.

5. Comparing results of task performance based on accuracy, speed and resource use with those collected from a equivalent system using a packet based approach.

6. Further processing of received flows to generate lookup tables that would allow fast identification of flow origin and the associated Geolocation.

7. visual representation of flow origin, volume, duration, and density associated with specific physical locations.

The development of the actual system will be split into 3 components. The first component will be to either construct a synthetic network interface that can pass raw flow data to the rest of the system or acquire an external machine that will physically pass the raw flow data to the computer running the system. The second component will preform the actual processing of the raw flows to achieve the identified tasks and render the results. This component will also identify the flow origin and associate it with a physical location. The development of this component will be done with the intention of moving it onto a separate piece of hardware from which the results can be called by using an API. The third component will be create a simple front end to call the second component and give the user an interface to view the results on.

6 Project Progression Timeline

Date	Action
11 March 2013	Investigate and install open source software for packet capture and flow generation.
15 March 2013	Install and explore a flow based traffic analysis application such as FlowScan [1]
19 March 2013	Present the project proposal to staff and colleges
22 March 2013	Make headway into writing literature review
26 March 2013	Make headway into research of cyber defence tasks that can be improved with the use of network flows.
7 April 2013	Compile a list of defence tasks that the prototype system will preform or assist in preforming.
7 April 2013	Complete literature review and Plan of Action
1 May 2013	Initial working prototype of software for raw Flow analysis.
15 May 2013	First revision of prototype software to include Geolocation.
20 June 2013	Second revision of prototype software that includes complete capability of testing the defence tasks identified.
1 August 2013	Testing of tasks completed
1 August 2013	Analysis of results from testing complete
15 August 2013	Comparison and discussion between results from prototype system and current methods of preforming the evaluated tasks complete.
15 September 2013	First draft of paper handed in
25 September 2013	Final draft of paper handed in
1 October 2013	Draft of first chapters of thesis handed in
1 November 2013	Project handed in

(will be in England from June 30th to July 14th)

References

[1] *FlowScan: A Network Traffic Flow Reporting and Visualization Tool* (New Orleans, Louisiana, USA, dECEMBER 2000), no. 14 in Systems Administration Conference, USENIX.

[2] NetFlow Services Solutions Guide. White Paper, 01 2007. Available from: `http://www.cisco.com/en/US/products/sw/netmgtsw/ps1964/products_implementation_design_guide09186a00800d6a11.html#wp1030045`.

[3] The Internet Registry System. Online, Sept 2011. Available from: `http://www.ripe.net/internet-coordination/internet-governance/internet-technical-community/the-rir-system`.

[4] CisCo. Cisco Routers Product Page. Online. Available from: `http://www.cisco.com/web/EA/solutions/smb/products/routers_switches/index.html#~routers`.

[5] Ed., B. C., Inc., C. S., Ed., B. T., and Zurich, E. Specification of the ip flow information export (ipfix) protocol for the exchange of flow information. RFC, Feb 2013.

[6] Laki, S., Mátray, P., Hága, P., Csabai, I., and Vattay, G. A detailed path-latency model for router geolocation. In *Testbeds and Research Infrastructures for the Development of Networks & Communities and Workshops, 2009. TridentCom 2009. 5th International Conference on* (2009), IEEE, pp. 1–6.

[7] Laki, S., Mátray, P., Hága, P., Csabai, I., and Vattay, G. A model based approach for improving router geolocation. *Computer Networks 54*, 9 (2010), 1490–1501.

[8] Mirkovic, J., and Reiher, P. A taxonomy of ddos attack and ddos defense mechanisms. *SIGCOMM Comput. Commun. Rev. 34*, 2 (Apr. 2004), 39–53.

[9] Ms. Vidya mhaske; Dhamdhere, P. G. P. Netflow method used for internet worm detetion. *International Journal of Scientific & Engineering Research 3* (March 2012), 1–7.

[10] Peng, T., Leckie, C., and Ramamohanarao, K. Survey of network-based defense mechanisms countering the dos and ddos problems. *ACM Comput. Surv. 39*, 1 (Apr. 2007).

[11] Solarwinds. NetFlow Traffic Analyzer Product Page. Online. Available from: `http://www.solarwinds.com/netflow-traffic-analyzer.aspx`.

[12] Zou, C. C., Towsley, D., and Gong, W. On the performance of internet worm scanning strategies. *Performance Evaluation 63*, 7 (2006), 700–723.