

F (v2)

# A complete system integration of the network flow query language

Vaibhav Bajpai

Masters Thesis

reviewed by

Prof. Dr. Jürgen Schönwälder  
Dr. Heinrich Stamerjohanns

Computer Networks and Distributed Systems  
School of Engineering and Science  
Jacobs University Bremen  
Bremen, Germany

July 2012

[July 13, 2012 at 21:09]



## ABSTRACT

---

With the long dominance of Cisco’s NetFlow [1] protocol and now with the emergence of Internet Engineering Task Force (IETF)’s Internet Protocol Flow Information Export (IPFIX) [2] open standard, traffic measurement practitioners have finally settled down with using Internet Protocol (IP) flow export as the de-facto technique for sending traffic patterns. These patterns have the potential to be used for billing and mediation, bandwidth provisioning, detecting malicious attacks, network performance evaluation and overall improvement.

However, making sense of these patterns calls for sophisticated flow-analysis tools that can mine them for such a usage. Unfortunately current tools fail to deliver owing to their poor language design and naïve filtering methods. Our research group, by going clean slate has come up with the Network Flow Query Language (NFQL) [3] that can cap flow-exports to full potential. The flowquery language can process flow-records, aggregate them into groups, apply absolute (or relative) filters and invoke Allen interval algebra rules [4] on them.

Flowy [5] is the prototype implementation of NFQL, which has undergone significant changes in the last few years. The core of this Python implementation is now being rewritten in C to make it comparable to the contemporary flow processing tools. The first major release in this effort, F(v1) [6] can now read the flow-records in memory and apply absolute filters. The absolute filters, however are only one amongst the five stages of the complete NFQL processing pipeline. The rest of the stages either have a broken implementation or are completely non-existent. The flowquery used by the execution engine to filter the traces is also hardcoded in pipeline structs.

This thesis extends F(v1) to provide a complete usable implementation of NFQL with *functional* pipeline stages. The execution engine has been completely refactored with a maintainable design and profiled to eliminate any memory leaks. The stages are *robust* to work well with variety of NFQL queries. The flowqueries are no longer hardcoded in the pipeline structs, but rather the engine is *flexible* enough to read and parse them at runtime using the JSON intermediate format. F(v2) is also *portable* with automated builds using `cmake` and is *verifiable* using a full-fledged regression test-suite. An automated benchmarking suite enables the conducted performance evaluations to be repeatable.



## PUBLICATIONS

---

- [1] V. Bajpai, N. Melnikov, A. Sehgal, and J. Schönwälder, “Flow-Based Identification of Failures Caused by IPv6 Transition Mechanisms,” in *Dependable Networks and Services*, vol. 7279 of *Lecture Notes in Computer Science*, pp. 139–150, Springer Berlin Heidelberg, 2012.



*Programs must be written for people to read,  
and only incidentally for machines to execute.*

— H. Abelson and G. Sussman

## ACKNOWLEDGMENTS

---

...





## CONTENTS

---

<b>I INTRODUCTION</b>	<b>1</b>
1 TRAFFIC MEASUREMENT APPROACHES	3
1.1 Capturing Packets . . . . .	3
1.2 Capturing Flows . . . . .	4
1.3 Remote Monitoring . . . . .	5
1.4 Remote Metering . . . . .	5
2 FLOW EXPORT PROTOCOLS	7
2.1 NetFlow . . . . .	7
2.2 IPFIX . . . . .	10
2.3 sFlow . . . . .	12
<b>II STATE OF THE ART</b>	<b>15</b>
3 NFQL AND FLOWY	17
3.1 Flowy Python Framework . . . . .	17
3.2 NFQL Processing Pipeline . . . . .	18
4 FLOWY IMPROVEMENTS USING MAP/REDUCE	23
4.1 Map/Reduce Frameworks . . . . .	23
4.2 Parallelizing Flowy . . . . .	24
5 F (v1)	27
5.1 Flowy Performance Issues . . . . .	27
5.2 Preliminary Improvements . . . . .	28
5.3 Rule Interfaces . . . . .	28
5.4 Efficient Rule Processing . . . . .	29
5.5 Faster Grouping Approach . . . . .	30
5.6 Benchmarks . . . . .	30
6 NFQL: APPLICATIONS	33
6.1 Application Identification using Flow Signatures . . . . .	33
6.2 Cybermetrics: User Identification . . . . .	35
6.3 IPv6 Transition Failure Identification . . . . .	37
6.4 OpenFlow . . . . .	39
6.5 Flow Level Spam Detection . . . . .	40
<b>III IMPLEMENTATION AND EVALUATION</b>	<b>43</b>
7 DESIGN	45
7.1 Flowy Parser and F(v1) Engine Analysis . . . . .	45
7.2 Execution Workflow and Abstract Objects . . . . .	47
7.3 User Interface Design . . . . .	50
8 IMPLEMENTATION	53
8.1 Grouper Internals . . . . .	53
8.2 Robust Pipeline Execution and Runtime Complexity . . . . .	56
8.3 Merger Internals . . . . .	62
8.4 Runtime Query Evaluation . . . . .	63

8.5	Automated Builds . . . . .	65
8.6	Regression Test Suite . . . . .	67
9	PERFORMANCE EVALUATION	69
9.1	Execution Engine Profiling . . . . .	69
9.2	Benchmarking Suite . . . . .	70
9.3	Relative Comparison with SiLK . . . . .	70
10	FUTURE WORK AND CONCLUSION	75
10.1	Major Goals . . . . .	75
10.2	Minor Issues . . . . .	76
10.3	Conclusion . . . . .	77
IV	APPENDIX	79
A	NFQL INSTALLATION AND USAGE	81
B	SILK INSTALLATION AND USAGE	85
C	NFQL RELEASE NOTES	89
D	ACRONYMS	93
	BIBLIOGRAPHY	95

## LIST OF FIGURES

---

Figure 1	NetFlow: Overview [7] . . . . .	7
Figure 2	IPFIX: Overview [8] . . . . .	10
Figure 3	IPFIX: Messages [9] . . . . .	11
Figure 4	IPFIX: Templates [9] . . . . .	11
Figure 5	IPFIX: A Transport Session [9] . . . . .	11
Figure 6	sFlow: Overview [10] . . . . .	12
Figure 7	Flowy: Processing Pipeline [11] . . . . .	18
Figure 8	Parallelizing Flowy using Map/Reduce [12] . .	24
Figure 9	Slice Boundaries Aware Flowy [12] . . . . .	24
Figure 10	Flowy: Redundant Groups [12] . . . . .	25
Figure 11	Cybermetrics: Overview [7] . . . . .	35
Figure 12	Geographical Preferences [7] . . . . .	35
Figure 13	Daily Distributions for HTTP Traffic [7]. . . . .	36
Figure 14	Cross Correlation of Traces with Varying Times [7]	36
Figure 15	NAT64 Setup [1] . . . . .	37
Figure 16	OpenFlow Architecture [13] . . . . .	39
Figure 17	Spam Flow Classifier [14] . . . . .	41
Figure 18	F(v2): Base Header . . . . .	47
Figure 19	F(v2): Execution Engine Workflow . . . . .	47
Figure 20	F(v2): Verbosity Levels Workflow . . . . .	51
Figure 21	F(v2): Backtrace of Living on Exit Blocks . . . .	69
Figure 22	F(v2): DNS and mDNS over UDP . . . . .	71
Figure 23	F(v2): HTTP and HTTPs Sessions over TCP . .	72
Figure 24	F(v2): Conglomerating HTTP Traffic . . . . .	73
Figure 25	F(v2): Application Agnostic TCP Sessions . . .	73

## LIST OF TABLES

---

Table 1	NetFlow Version History . . . . .	8
Table 2	Runtime Breakup of Individual Stages [6] . . .	27
Table 3	Flowy vs F(v1) [6] . . . . .	30
Table 4	Application Flow Signatures: Results [15] . . .	34
Table 5	Features in Spam Flow [14] . . . . .	40
Table 6	F(v2): Pipeline Runtime Complexity . . . . .	61



## Part I

### INTRODUCTION

The network and user behavior traffic pattern analysis is creating a lot of traction owing to its wide applicability in accounting, resource provisioning and network monitoring purposes. This section is dedicated to perform an exhaustive study on the available techniques that can perform such traffic measurements and how they are being used today. In particular, we focus our attention to the currently favored flow-capture technique by examining the de-facto protocols that describe the semantics of this flow-export. The organization of the section is described below.

In chapter [1](#) we discuss the current state-of-the-art traffic measurement techniques, the protocols supporting them, their pros and cons and how they are being used to mine the behavioral patterns of the current network traffic.

In chapter [2](#) we discuss Cisco's proprietary and [IETF's](#) standardized protocol for [IP](#) flow export. We discuss their architecture, protocol operations, their message formats and the future they are heading towards as seen from today.



## TRAFFIC MEASUREMENT APPROACHES

Researchers, service providers and security analysts have long been interested in network and user behavioral patterns of the traffic crossing the internet backbone. They want to use this information for the purpose of billing and mediation, bandwidth provisioning, detecting malicious attacks, network performance evaluation and overall improvement. Traffic measurement techniques that have been rapidly evolving in the last decade, have matured enough today to provide such an insight. In this chapter, we discuss some of these techniques and how they are being used to shape the future of the internet.

### 1.1 CAPTURING PACKETS

In this technique, raw packets traversing a monitoring point are captured for traffic measurement. The measurements can be done either live or the packets can be saved in a trace file for offline analysis. The trace files can range from containing mere headers to entire packets depending on the level of detailed analysis required.

```
1 $ tcpdump port 80 -w $FILE
2 $ tcpdump -r $FILE
```

Listing 1: tcpdump: Example

tcpdump and wireshark are the most popular tools used for packet capture and analysis. tcpdump [16] is a premier command-line utility that uses the libpcap [17] library for packet capture. A simple example to capture and read the Hypertext Transfer Protocol (HTTP) traffic is described in listing 1. The power of tcpdump comes from the richness of its expressions, the ability to combine them using logical connectives and extract specific portions of a packet using filters. wireshark [18] is a Graphical User Interface (GUI) application, aimed at both journeymen and packet analysis ninjas. It supports a large number of protocols, has a straightforward layout, excellent documentation, and can run on all major operating systems.

*tcpdump*

*wireshark*

Several studies have made use of this approach to analyze the network traffic patterns. The authors in [19], for instance use data mining methodologies to define clusters of behavior profiles by understanding the captured traffic of end hosts. These clusters are then fed into classifiers to automatically identify anomalous behavior patterns that are of interest to network operators. Similar profiling of end-hosts traffic

*applicability*

in performed in [20], but at the transport layer. This effort focusses on making the approach tunable to strike out a balance between the amount of traffic classified and the accuracy achieved by analyzing the traffic at multiple levels of details.

*pros and cons* This approach benefits from the astounding level of detail it can provide. It allows deep packet inspection of the traces, thereby exposing even the application content being exchanged across the network. This calls for privacy concerns and can even bring in legal repercussions to make this technique unattractive for traffic analyzers today. In addition, the actual usage of this method comes at a higher price of its storage overhead and its inability to scale to larger setups.

## 1.2 CAPTURING FLOWS

In this technique, packets traversing a monitoring point are not captured raw, instead they are aggregated together based on some common characteristics. The common characteristics are learnt by inspecting the packet headers as they cross the monitoring point. Flow-records resulting from such an aggregation are then exported to a collector for further analysis.

*netflow* NetFlow and IPFIX are the two popular standards of IP flow information export. NetFlow [1] is a proprietary network protocol designed by Cisco Systems. It allow routers to generate and export flow records to a designated collector. The latest version, NetFlow v9 provides flexibility of user-tailored export templates, Multiprotocol Label Switching (MPLS) and IPv6 support and a larger set of flow keys. *ipfix* IPFIX [2] on the other hand is an open standard by IETF deemed to be the logical successor of NetFlow v9 on which it is based. The novelty of the standard lies in its ability to describe record formats at runtime using templates based on an extensible and well-defined information model. The data transfer mechanism is also simplistic and extensible by being unidirectional and transport protocol agnostic.

*applicability* The wide applicability of this approach is easily seen from the pervasive use of flow records for a vibrant set of network analysis applications. For instance, the authors in [21] use the flow characteristics in the traffic pattern to formalize a detection function that maps traffic patterns to different Denial of Service (DoS) attacks, whereas in [22], the authors use the flow-record data to exploit timing characteristics of webmail clients to classify features that could identify webmail traffic from any other traffic running over HTTPS.

*pros and cons* This is has been possible largely due to the hardware-assisted aggregation of the packets that has helped solve the storage overhead and scalability limitation of packet capturing techniques. Overcoming these limitations have eventually allowed researchers to perform network analysis over a larger dataset passing across high-speed links. However, with the ever-increasing bandwidth demands, the speed of



the network links in the internet backbone is only slated to increase further, therefore the time is not too far when this issue might again scares us of its homecoming.

### 1.3 REMOTE MONITORING

In this technique, dedicated monitoring probes are deployed on network segments to continuously collect vital statistics and perform network diagnostic operations. The probes are configured to proactively monitor the network and automatically check for error conditions to later log and notify them to the management station.

The Remote Network Monitoring (RMON) Framework [23] for Simple Network Management Protocol (SNMP) [24] defines a number of Management Information Base (MIB) objects to be used by these monitoring probes. The RMON-1 standard [25] for instance, defines a MIB module to collect statistics, capture and filter packet contents at the logical link layer. The architecture in this standard has been further extended with a feature upgrade by the RMON-2 standard [26] to support similar analysis up to the application layer.

*rmon*

The novelty of this technique lies in the ability to immediately communicate important information to the managing station using events and alarms. The constructs are extremely flexible in giving full control over what conditions will cause an alarm and subsequently what event will be generated. The event-driven nature of such a monitoring platform however still does not satisfy the requirements of traffic analysis applications since the data that is pushed out is highly aggregated and lacks enough details to be useful.

*pros and cons*

### 1.4 REMOTE METERING

In this technique, meters are deployed at the network measurement points to capture flow data according to a predefined set of rules specified by the user. The model, as defined by the Realtime Traffic Flow Measurement (RTFM) working group [27] has been designed to be protocol agnostic and restrictive in the amount of flow data that can be transmitted across the network and stored to reduce the processing time of network analysis applications.

The feature that sets this technique apart is the flexibility given to the user to specify their flow measurement requirements, thereby allowing them to filter out the traffic they do not care about. This calls for the users, to at the very outset analyze and freeze their requirements before they start off to capture the traffic. This is analogous to the flaws inherit in the waterfall model [28] of software design, whereby one need to design the design before one designs it.

*pros and cons*



## FLOW EXPORT PROTOCOLS

Flow capture today, has emerged out to be one of the favored network measurement techniques. This has largely been due to the reduction in the monitoring traffic at the flow-level and the fine-grained control which was not previously possible using [SNMP](#) interface-level queries. As a result, each networking vendor has tried to come up with a standard protocol that defines the semantics of this flow export. In this pursuit, Cisco eventually managed to make their proprietary protocol so ubiquitously available, that the next-generation universal standard is based on it. In this chapter, we discuss Cisco's de-facto proprietary and the recently standardized [IETF](#)'s open protocol for [IP](#) flow-export.

### 2.1 NETFLOW

NetFlow [\[1\]](#) by Cisco Systems is a protocol that allows network elements to export [IP](#) flow information to designated collectors from where they can be later retrieved for further analyses. The collected flow-records are flexible enough to be used for a variety of purposes such as billing and mediation, network and user monitoring, resource provisioning, security analysis and data mining research works.

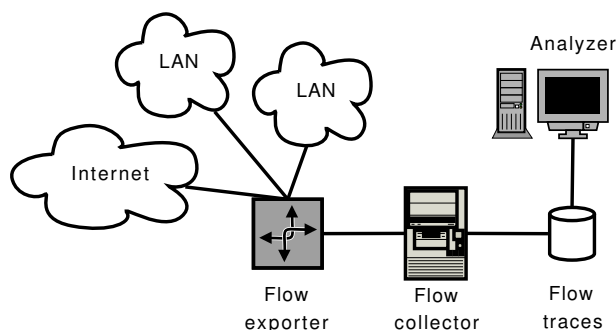


Figure 1: NetFlow: Overview [\[7\]](#)

A high-level abstracted functioning of the NetFlow protocol is shown in figure 1. The flow exporter reads the [IP](#) packets that cross its boundary to generate flow-records. The flow-records are exported based on some predefined expiration rules, such as a Transmission Control Protocol ([TCP](#)) FIN or RST, an inactivity timeout, a regular export timeout or crossing a low memory threshold. To achieve efficiency when handling large amounts of traffic, the flow-records are encapsulated in User Datagram Protocol ([UDP](#)) datagrams and are

*protocol operation*

deleted from the exporter once transmitted. On the other end, the collector on receiving these flow-records, decodes and stores them locally to be used for further processing.

```
1 (A) --> [SYN] ----->(B)
2 (A) <-- [SYN/ACK] <--(B)
3 (A) --> [ACK] ----->(B)
```

Listing 2: A Flow Example

*what is a flow?*

A flow is defined by a 7-tuple flow-key, namely: {srcIP, dstIP, srcPort, dstPort, ipProto, ifIndex, ipTOS}. IP packets with identical flow-keys become part of one flow. Two flows resulting from a three-way TCP handshake for example are shown in listing 2. In addition to the flow-key, flow-records can also contain additional accounting information such as flow start and end times, number of packets/octetes in a flow, source/destination Autonomous Systems (AS) numbers, et al.

*version history*

The NetFlow version history is summarized in table 1. NetFlow v1 was introduced in the 90s, however it was only until v5 with the introduction of Classless Inter-Domain Routing (CIDR) and AS support that the technology got mainstream. Today, NetFlow v9 is the de-facto industry standard and is the bases for IETF’s IPFIX effort to create a universal specification for IP flow-export.

version	features
v1,{2,3,4}	original format with several internal releases
v5	CIDR, AS support and flow sequence numbers
v{6,7,8}	router-based aggregation support
v9	template-based with IPv6, and MPLS support
IPFIX	universal standard, transport-protocol agnostic

Table 1: NetFlow Version History

*netflow v9*

NetFlow v9 introduces templates in its export format. With templates, the exporter only needs to send required fields to the flow collector thereby reducing the volume of flow-data exported. In addition, fields can be added/removed from the flows without changing the export format. The transmission of records encapsulated in UDP datagrams can lead to loss of flows when the link is congested and therefore the exporter and collector have usually been restricted to one-hop away dedicated links. To overcome this limitation, NetFlow v9 introduces transport support over congestion-aware Stream Control Transmission Protocol (SCTP). In addition, NetFlow v9 also provides support for MPLS and IPv6 addresses.

The ever increasing traffic volume crossing high-speed links, has been creating an enormous pressure on the routers that also engaged in NetFlow export. Sampled NetFlow was thus introduced by Cisco Systems as an extension to NetFlow v9 to tone down the gigantic computation, by allowing the routers to skip over to every  $n^{\text{th}}$  packet for flow export. The sampling rate ( $n$ ) is indicated in the export header and is either configured or randomly selected.

*sampled netflow*

Though sampled NetFlow does a good job in reducing the exported traffic volume, the sampling rate is still static which either reduces accuracy at low traffic volumes or increases memory use at high traffic volumes. An adaptive algorithm introduced in [29] helps overcome this difficulty. The introduced renormalization technique helps guarantee that the sampling rate can not only adjust to variable traffic mixes but also to network congestion. It also ensures that the flow records do not span over measurement bins to be able to guarantee statistical accuracy. The authors claim, that these updates are easily deployable to any NetFlow v9 router through a software update. In addition they say, a simple hardware add-on (flow counting extension), can also add capability to accurately count non-TCP flows, a feature long waiting to be seen in NetFlow v9.

*adaptive netflow*

Flexible NetFlow is the newest version of NetFlow v9 that incorporates Packet Sampling (PSAMP) [30] ideas to be able to select individual packets and export them in a packet record. The packet selection can be either deterministic or random depending on the chosen filters and sampling mechanism [31]. The exported packet records can even be authenticated and encrypted using either Transport Layer Security (TLS) [32] or Datagram Transport Layer Security (DTLS) [33] to prevent data manipulation across the route. Since PSAMP is based on IPFIX [34], only its limited feature set is currently supported by Flexible NetFlow. Additional features include ability to custom define flow-keys and flow-expiration rules to drastically reduce the amount of content exported by restricting it to only the needed flow-fields, and additional flows with immediate and permanent caches to suit the export timings to specific needs.

*flexible netflow*

The challenge to identify relevant records in gigantic collected datasets have fumed recent independent studies to discover flow dependencies. For instance, the authors in [35], describe a model that uses flow timing information by extending the PageRank [36] algorithm to rank and thereby extract the most relevant flows. Their model is weighted using parameters like the amount of bandwidth consumed and the likelihood of security threat a flow might result in.

*flowrank*

Today, as the industry is moving towards data center virtualization, it has become inherently critical to obtain insights into the data center network behavior for optimizations and resource provisioning. Since, Flexible NetFlow's visibility is limited to the IP protocol it currently cannot be used to monitor data-center traffic. NetFlow-lite was thus

*netflow-lite*

introduced by Cisco Systems, to flows at the layer 2/3 level to increase data center visibility. NetFlow-lite uses similar packet sampling mechanisms as introduced in Sampled NetFlow along with the combined flexibility of Flexible NetFlow v9 at the switch level. NetFlow-lite captures the layer 2 traffic, encapsulates packet samples and pushes the NetFlow cache outside the switch into a element that can convert NetFlow-lite to Flexible NetFlow records. These flow-records are then later exported to legacy collectors from where they can be used for further processing. The authors in [37] provide the first implementation of NetFlow-Lite which works as an extension to nProbe [38] to seamlessly convert NetFlow-Lite records to NetFlow/IPFIX.

## 2.2 IPFIX

[IPFIX](#) [2] by [IETF](#) is an interoperable protocol for [IP](#) flow export. It is deemed to be the logical successor of Flexible NetFlow v9. The working group defines [IPFIX](#) as, "a unidirectional, transport-independent protocol with flexible data representation and an information model covering most network management needs at layer 3 and 4". The [PSAMP](#) working group [34], that defines standards to individually sample packets in a flow export using statistical methods has adopted [IPFIX](#) as its underlying protocol for data transport.

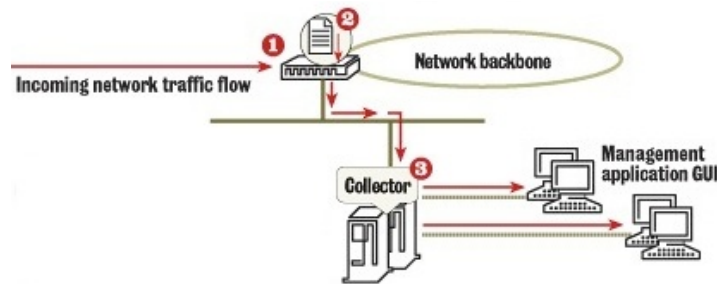


Figure 2: IPFIX: Overview [8]

*architecture*

The [IPFIX](#) architecture is described in [39] and is shown in figure 2. The architecture consists of three elements: a meter, which generates flows from [IP](#) packets, an exporter, which pushes these flows using [IPFIX](#), and a collector, that collects and saves these flows for offline storage. All these elements have a one-to-many relationship among them. The group is also working to define an intermediary element, that might work to either aggregate or anonymize the flows.

*messages and templates*

A message is a fundamental unit of data exchange in [IPFIX](#). Each such message consists of a 16-byte header along with a number of sets as shown in figure 3. A set can either be a template or a data set. Each such set in the message again contains a 16-bit header and a number of records associated with it. Each record within a template set is a template that refers to a data record. A template consists of a number

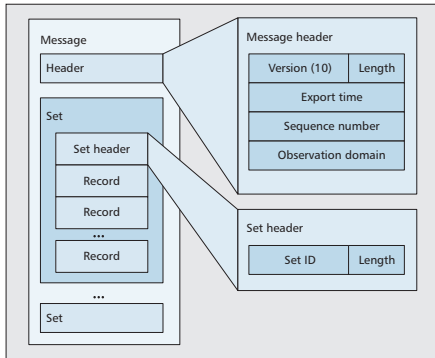


Figure 3: IPFIX: Messages [9]

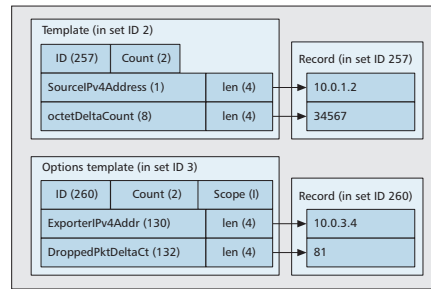


Figure 4: IPFIX: Templates [9]

of Information Elements (IE)s as shown in figure 4. These IEs are encoded using reduced-length encoding scheme. Internet Assigned Numbers Authority (IANA) keeps a registry <sup>1</sup> of all IEs with a 16-bit ID assigned to them. Templates can also contain enterprise-specific IEs that are scoped using Private Enterprise Numbers (PENs) <sup>2</sup>.

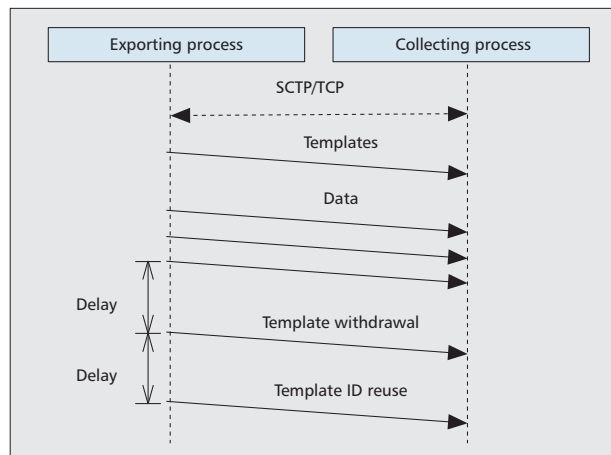


Figure 5: IPFIX: A Transport Session [9]

An IPFIX transport session is shown in figure 5. It starts off with the Exporter Process (EP) initiating a connection with the Collector Process (CP). Once the connection is established, the EP passes on the templates followed by the data that is described by them. These templates can later still be withdrawn by sending a control template of IE count zero. The transport session can use either SCTP, TCP or UDP as the underlying protocol, although SCTP is usually the preferred method given it allows selective reliability and congestion control. TCP is supported to allow secure transport over TLS, since DTLS is only supported over UDP and SCTP. The connection-less behavior of UDP calls for the template retransmission delay and template lifetime param-

*transport and  
security*

<sup>1</sup> <http://www.iana.org/assignments/ipfix/ipfix.xml>

<sup>2</sup> <http://www.iana.org/assignments/enterprise-numbers>

ters to be exchanged between EP and CP. These transport sessions can also be stored in IPFIX files and sent on top application layer protocols.

*management,  
extensions and  
future of ipfix*

A MIB to monitor IPFIX devices using SNMP is defined in [40]. A similar configuration model to be used by NETCONF and YANG is being worked upon. In addition, several extensions have been defined to expand upon the protocol's functionality. For instance, [41] defines optional templates to allow bidirectional flows in a single IPFIX export whereas [30] supports aggregating common properties of multiple flows in a single record. IPFIX is even being looked upon as the future application-layer logging protocol as well as the underlying protocol for RESTful architectures. As a result, efforts to support structured data export over IPFIX are also under way.

### 2.3 SFLOW

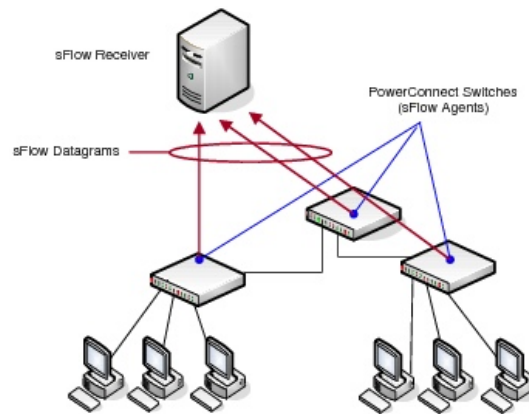


Figure 6: sFlow: Overview [10]

sFlow [42] by InMon Corporation is a competing technology used to capture traffic from switches and routers. It consists of an sFlow agent that captures traffic statistics and sends them across to a central data collector, called the sFlow analyzer as shown in figure 6. In order to be able to accurately monitor traffic at line speeds, the sFlow agent is built on a dedicated ASIC alongside the switching gear. In addition, the captured traffic is sampled before being encapsulated in sFlow datagrams and sent to the analyzer to provide scalability.

*sampling  
mechanisms*

A flow in sFlow is defined as all the packets that enter a source interface, are processed through a switching module, and eventually exit through a destination interface. Packet-based and Time-based are the two sampling methods supported by sFlow agents. Statistical packet-based sampling of switched flows uses a counter that is decremented whenever a packet crosses the switching gear. A sample is taken whenever the counter hits zero and is then reset. A sample involves copying the packet header or a packet feature extraction.



Time-based sampling of network interface statistics on the other hand involves the sFlow agent which is responsible for periodically polling each switching gear for feature extraction.

sFlow provides a standard interface to configure and monitor the sFlow agents using [SNMP](#). This subverts the need to telnet to every switch of the network infrastructure and use its Command Line Interface ([CLI](#)) to make subtle changes which can turn out to be overly complex and time-consuming. A [MIB](#) module to remotely control the sFlow agents is defined in [42]. The [MIB](#) module is Structure of Managed Information ([SMI](#)) v2 compliant and can be translated back to [SMI](#) v1 without incurring any semantic differences.

*sflow and snmp*

sFlow uses a standard format to send sampled data from the sFlow agent to the sFlow analyzer. The data format is specified using External Data Representation ([XDR](#)) [43]. [XDR](#) allows compact representation and efficient encoding (or decoding) of the sampled data. The [XDR](#) specified sampled data is sent using [UDP](#) to a well-known host and port combination specified in the sFlow [MIB](#). [UDP](#) is used as a transport mechanism owing to its less stringent buffer requirements and its robustness in delivering traffic information in a timely fashion.

*data format*

sFlow does not provide any security measures to protect the sampled data being transferred to the sFlow analyzer and is therefore at the risk of being eavesdropped. The sFlow analyzer in itself also does not verify the source addresses of the sampled data; as such the sFlow datagrams can easily be spoofed and identified as coming from one of the participating sFlow agents. In essence, now with Flexible NetFlow and [IPFIX](#) both providing [PSAMP](#) support, the packet sampling novelty of sFlow is losing significance. At one point, the capability of sFlow to monitor traffic at the layer 2 level was seen as an advantage as well, but that is also deemed to lose ground with the frequent adoption of NetFlow-lite.

*limitations and future*



## Part II

### STATE OF THE ART

The semantics and implementation of [NFQL](#) has undergone significant changes in the last few years. This section is dedicated to perform an inside-out study of the language and its prototypes, examining all their major (and minor) changes to allow us to better make a pragmatic stand towards its overall improvement. The organization of the section is described below.

In chapter [3](#) we look into the structure of [NFQL](#) by discussing each stage of the processing pipeline with their implementation details. The basic structures of the Flowy python framework are also discussed to understand its current limitations.

In chapter [4](#) we investigate the possibility of making [NFQL](#) map/reduce aware. The chapter starts off with a discussion of current map/reduce frameworks and looks into the ways to help parallelize the Flowy.

In chapter [5](#) we look into the first attempt to make [NFQL](#) implementation comparable with the state-of-the-art flow-analysis tools. After drilling down the performance hit sections of the python code, we witness how getting away with pytables and rethinking the rewrite of processing pipeline in C may help make it practically usable.

We conclude this discussion in chapter [6](#) by introducing a number of real-life application scenarios where [NFQL](#) has proved useful. We also look into a few current bleeding edge research projects where we believe [NFQL](#) could play a vital role in the near future.



Flowy [44, 5] is the first prototype implementation of the NFQL [3, 11, 45]. The query language allows to describe patterns in flow-records in a declarative and orthogonal fashion, making it easy to read and flexible enough to describe complex relationships among a given set of flows.

### 3.1 FLOWY PYTHON FRAMEWORK

Flowy is written in Python. The framework is subdivided into two main modules: the validator module and the execution module. The validator module is used for syntax checking and interconnecting of all the stages of the processing pipeline and the execution module is used to perform actions at each stage of the runtime operation.

Flowy uses PyTables [46] to store the flow-records. PyTables is built on top of the Hierarchical Data Format (HDF) library and can exploit the hierarchical nature of the flow-records to efficiently handle large amounts of flow data. The `pytables` module provides methods to read/write to PyTables files. The `FlowRecordsTable` class instance within the module exposes an iterator interface over the records stored in the HDF file. The `GroupsExpander` class instance within the same module on the other hand exposes an iterator interface over the group records and facilitates ungrouping to flow records. In addition, Flowy uses Python Lex-Yacc (PLY) for generating a Look-Ahead LR Parser (LALR) parser and providing extensive input validation, error reporting and validation on the execution modules.

*pytables and ply*

Flow-records are the principal unit of data exchange throughout Flowy's processing pipeline. The prototype implementation allows the `Record` class (defined in the `record` module) to be dynamically generated using `get_record_class(...)` allowing future implementations to easily plug in support for IPFIX or even newer versions of NetFlow [1] exports. The `FlowToolsReader` class instance (defined in `ftreader` module) provides an iterator over the records defined in `flow-tools` format. This can be plugged into the `RecordReader` class instance (defined in `record` module) to instantly get `Record` class instances.

*records*

The parser module holds definitions for the lexer and parser. The statements when parsed are implicitly converted into instances of classes defined in the `statement` module. The instances contain meta-information about the parsed statement such as the values, line numbers and sub-statements (if any).

*parsers and  
statements*

## 3.2 NFQL PROCESSING PIPELINE

The pipeline consists of a number of independent processing elements that are connected to one another using UNIX-based pipes. Each element receives the content from the previous pipe, performs an operation and pushes it to the next element in the pipeline. Figure 7 shows an overview of the processing pipeline. The flow record attributes used in this pipeline exactly correlate with the attributes defines in the [IPFIX](#) Information Model specified in RFC 5102 [47]. A complete description on the semantics of each element in the pipeline can be found in [3]

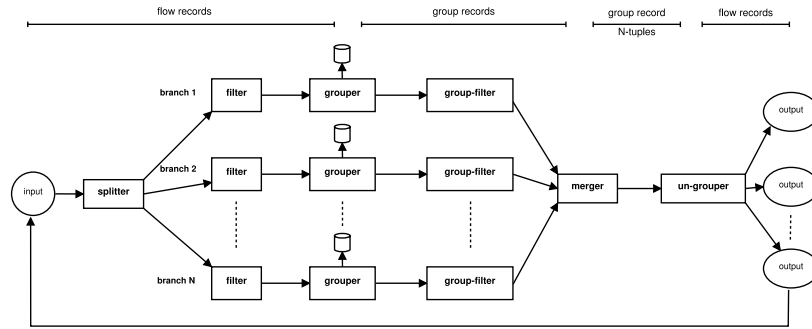


Figure 7: Flowy: Processing Pipeline [11]

The splitter takes the flow-records data as input in the flow-tools compatible format. It is responsible to duplicate the input data out to several branches without any processing whatsoever. This allows each of the branches to have an identical copy of the flow data to process it independently.

*splitter  
implementation*

The splitter module handles the duplication of the Record instances to separate branches. Instead of duplicating each flow-record to every branch (as specified in the specification), the implementation follows a pragmatic approach by filtering the records beforehand against all the defined filter rules to determine which branches a flow-record might end up in and saves this information in a record-mask tuple of boolean flags. The `go(...)` method in the `Splitter` class then iterates over all the (record, record-mask) pairs to dispatch the records to corresponding branches marked by their masks using the `split(...)` method. The class uses branch names to branch objects mapping to achieve the dispatch.

*splitter validator*

The `splitter_validator` module handles the splitter processing stage. The `SplitterValidator` class within the module uses the `Parser` and `FilterValidator` instances passed to it to create a `Splitter` instance and its child `Branch` instances.

The filter performs *absolute* filtering on the input flow-records data. The flow-records that pass the filtering criterion are forwarded to the grouper, the rest of the flow-records are dropped. The filter compares separate fields of a flow-record against either a constant value or a value on a different field of the *same* flow-record. The filter cannot *relatively* compare two different incoming flow-records

*filter*

The filter module handles the filtering stage of the pipeline. Since in the implementation the filtering stage occurs before the splitting stage, a single Filter class instance suffices for all the branches. Within the filter module, each filtering statement is converted into a Rule class instance, against which the flow-records are matched. The Rule instances are constructed using the (branch mask, logical operator, arguments) tuple. After matching the records against the rules, the record's branch mask is set and is then used by the splitter to dispatch the records to the filtered branches.

*filter implementation*

The filter\_validator module handles the filter processing stage. The FilterValidator class within the module uses the Parser instance passed to it to create a Filter instance once the check on semantical constraints have passed. The constraints involve checking whether records fields referenced in the filter definition exist, whether filters references in composite filter definitions exist and whether duplicate filter definitions are defined.

*filter validator*

The grouper performs aggregation of the input flow-records data. It consists of a number of rule modules that correspond to a specific subgroup. A flow-record in order to be a part of the group should be a part of at-least one subgroup. A flow-record can be a part of multiple subgroups within a group. A flow-record cannot be part of multiple groups. The grouping rules can be either absolute or relative. The newly formed groups which are passed on to the group filter can also contain meta-information about the flow-records contained within the group using the aggregate clause defined as part of the grouper query.

*grouper*

The grouper module handles the grouping of flow-records data. The Group class instance contains group-record's field information required for absolute filtering. It also contains the first and last records of the group required for relative filtering of the group-records. The AggrOp class instance handles the aggregation of group-records. The allowed aggregation operations are defined in aggr\_operators module. Custom-defined aggregation operations are also supported using -aggr-import command line argument.

*grouper  
implementation*

The grouper\_validator module handles the grouper processing stage. The GrouperValidator class within the module uses the Parser and SplitterValidator instances passed to it to create a Grouper instance once the check on semantical constraints such as the presence of referenced names and non-duplicate names have passed. Three aggregation operations: union(rec\_id), min(stime), max(etime) are added by default to each Grouper instance.

*grouper validator*

*group filter*

The group-filter performs *absolute* filtering on the input group-records data. The group-records that pass the filtering criterion are forwarded to the merger, the rest of the group-records are dropped. The group-filter compares separate fields (or aggregated fields) of a flow-record against either a constant value or a value on a different field of the *same* flow-record. The group-filter cannot *relatively* compare two different incoming group-records

*group filter implementation*

The groupfilter module handles the filtering of group-records. The GroupFilter class within the module iterates over the flow-records within the group and applies filtering rules across them. The filtering rules reuse the Rule class from the filter module. The flow-records are then added to the time index and stored in a pytables file for further processing. For groups that do *not* have a group-filter defined for them, run through a AcceptGroupFilter class instance. The timeindex module handles the mapping of the time intervals to the flow-records. The time index is used by the merger stage to learn about the records that satisfy the Allen relations. The add(...) method in the TimeIndex class is used to add new records to the time index. The get\_interval\_records(...) method on the other hand is used to retrieve records within a particular time interval.

*groupfilter validator*

The groupfilter\_validator module handles the group-filter processing stage. The GroupFilterValidator class within the module uses the Parser and Grouper instances passed to it to create an instance of GroupFilter. The check for the referenced fields is performed against the a gggregate clause defined in grouper statements. The class instance uses the AcceptGroupFilter instance in case a branch does *not* have a group filter defined for it.

*merger*

The merger performs relative filtering on the N-tuples of groups formed from the N stream of groups passed on from the group-filter as input. The merger rule module consists of a number of a submodules, where the output of the merger is the set difference of the output of the first submodule with the union of the output of the rest of the submodules. The relative filtering on the groups are applied to express timing and concurrency constraints using Allen interval algebra [4]

*merger implementation*

The merger module handles the merging of stream of groups passed as input. It is implemented as a nested branch loop organized in an alphabetical order where every branch is a separate for-loop over its records. During iteration, each branch loop executes the rules that matches the arguments defined in the group record tuple and subsequently passes them to the lower level for further processing. The Merger class represents the highest level branch loop and as such it must iterate over all of its records since it does not have any rules to impose restrictions on the possible records. The MergerBranch on the other hand represents an ordinary branch loop with rules.

The merger\_validator module handles the merger processing stage. The MergerValidator class within the module uses the Parser and



GroupFilterValidator instances passed to it to create a Merger instance once the check on referenced fields and branch names has passed. In addition, the validator also ensures semantic checks on Allen algebra such as whether the Allen relation arguments are correctly ordered, whether the Allen rules with the same set of arguments are connected by an OR and whether each branch loop is reachable by an Allen relation (or a chain of Allen relations) from the top level branch.

*merger validator*

The ungrouper unwraps the tuples of group-records into individual flow-records, ordered by their timestamps. The duplicate flow-records appearing from several group-records are eliminated and are sent as output only once.

*ungrouper*

The ungrouper module handles the unwrapping of the group-records. The generation of flow-records can also be suppressed using the `-no-records-ungroup` command line option. The Ungrouper class instance is initialized using a merger file and an explicit export order.

*ungrouper  
implementation*

The ungrouper\_validator module handles the ungrouper processing stage. The UngrouperValidator class within the module uses the Parser and MergerValidator instances passed to it to create a Ungrouper instance. This processing stage does *not* require any validation.

*ungrouper validator*



Flowy, although clearly setting itself apart with its additional functionality to query intricate patterns in the flows demonstrates relatively high execution times when compared to contemporary flow-processing tools. A recent study [12] revealed that a sample query run on small record set (around 250 MiB) took 19 minutes on Flowy as compared to 45 seconds on `flow-tools`. It, therefore is imperative that the application will benefit from distributed and parallel processing. To this end, recent efforts were made to investigate possibility of making Flowy Map/Reduce aware [12]

#### 4.1 MAP/REDUCE FRAMEWORKS

Map/Reduce is a programming model for processing large data sets by automatically parallelizing the computation across large-scale clusters of machines [48]. It defines an abstraction scheme where the users specify the computation in terms of a map and reduce function and the underlying systems hides away the intricate details of parallelization, fault tolerance, data distribution and load balancing behind an Application Programming Interface (API).

Apache Hadoop [49] is a Map/Reduce Framework written in Java that exposes a simple programming API to distribute large scale processing across clusters of computers. However in order to make Flowy play well with the framework, the implementation either has to use a Python wrapper around the Java API or translate the complete implementation to Java through Jython. Even more since Flowy uses HDFS files for it's I/O processing, staging the HDFS files properly in the Hadoop Distributed File System (HDFS) [50] and then later streaming them using Hadoop Streaming utility would still be an issue as suggested in [12]

*apache hadoop*

Disco [51] is a distributed computing platform using the Map/Reduce framework for large-scale data intensive applications. The core of the platform is written in Erlang and the standard library to interface with the core is written in Python. Since the map and reduce jobs can be easily written as Python functions and dispatched to the worker threads in a pre-packaged format, it is less difficult to setup Disco to utilize Flowy as a map function. In addition, the usage of HDFS files for I/O processing pose no additional modifications whatsoever since the input data files can be anywhere and supplied to the worker threads in absolute paths.

*the disco project*

#### 4.2 PARALLELIZING FLOWY

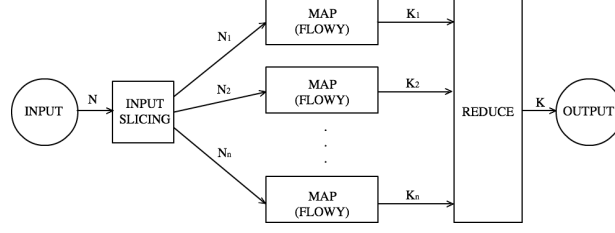


Figure 8: Parallelizing Flowy using Map/Reduce [12]

In an attempt to parallelize Flowy, it was run as a map function on a successful single node Disco installation as shown in 8. Although the setup on a multiple node cluster would be theoretically almost equivalent, Flowy has not yet been tested in such a scenario.

When running several instances of Flowy, it is imperative to effectively slice the input flow-records data in such a way so as to minimize the redundancy in distribution of input. To achieve this, the semantics of the flow-query needs to be examined from the simplest to the most complex cases. However, it is also important to realize that as of now it is not possible to *leave* out any stage in the Flowy's processing pipeline and the following examination was based on such an assumption.

*slicing inputs using  
only filters*

A flow query that involves only the filtering stage of the processing pipeline can slice its input flow data by either adding explicit export timestamps to allow each branch to skip records or separate out the input flow data into multiple input files for each branch.

*slicing inputs using  
groupers*

A flow query that also involves groupers and group-filters cannot use static slice boundaries since the grouping rules can be either absolute or relative. As a result, Flowy needs to be made aware of slice boundaries by passing the timestamps as command line parameters. In such a scenario, each branch will skip the pre-slices, whereby the actual slices and the post-slices will be processed to create relevant groups as shown in figure 9. It is advisable to slice the flow-records at low traffic spots to avoid the risk of cutting the records belonging to the same group.

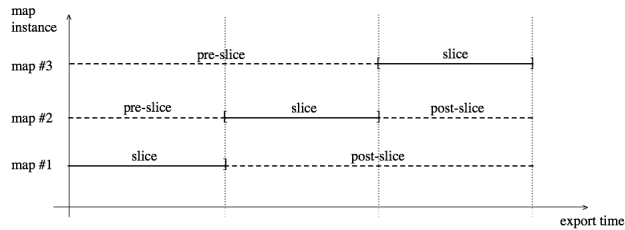


Figure 9: Slice Boundaries Aware Flowy [12]

The idea of skipping pre-slices and sweeping across post-slices can result in many fragmented redundant groups. These can be identified by the reduce function by removing the groups that are a proper subset of the previous group in the slice at the cost of additional complexity as shown in figure 10

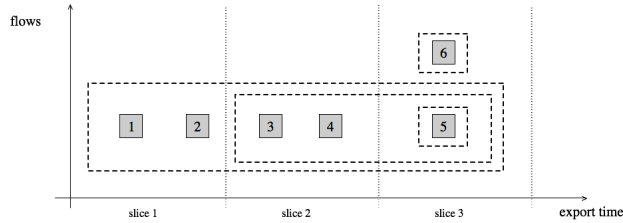


Figure 10: Flowy: Redundant Groups [12]

The relative dependency in the merger stage of the pipeline is even worse, since the comparison needs to take place between groups resulting from the output of separate map functions. This calls for inhibiting parallelism up to and including the group-filter stage. As a result each worker thread would return back its filtered groups to the master node, which then slicing inputs using mergers would apply the rules of the merger stage to all the received groups at once in a reduce function. In such a scenario, although the branch with the longest runtime complexity will become the bottleneck for the merger, the overall runtime would still be dramatically reduced when the number of branches are large as suggested in [52]

A Disco job function is created that contains the map/reduce function definitions and a location of an input file of flow-records data. A `sliceIt(...)` function within a newly defined `sliceFileCreator` module is used to create the input file. The function takes a HDF file and number of worker threads as input and writes out the slices in the input file by equally dividing HDF timespan by the number of worker threads.

*flowy as a map  
function*

In this way, the input file gets slice times for each worker thread in a separate line, which the Disco job function eventually reads to spawn a new map function with the slice times passed as arguments. The map function then starts an instance of Flowy and passes the slice times and the HDF file as command line parameters for processing.

This required modification to the `flowy_exec` module to add support for extra parameters. The filter stage of the pipeline was modified to allow for skipping of the pre-slices in the flow-records data. The grouper stage was modified as well to restrict creation of new groups that do *not* fall within the passed slice boundaries. However, the modification of the reduce function to work with the files pushed out by each Flowy instance of the map function to merge groups from each branch and eliminate duplicate records is left open.



In an attempt to develop a prototype implementation of [NFQL](#) which will be comparable with the contemporary flow-record analysis tools, the substitution of the performance hit sections of the python code was thought out. Flowy 2.0 [6] is the first attempt to try rewriting the core of the prototype implementation in C.

### 5.1 FLOWY PERFORMANCE ISSUES

no. of records	overall	filter	grouper	merger
103k	1177s	28s(2%)	240s(20%)	909s(77%)
337k	20875s	110s(1%)	2777s(13%)	17988s(86%)
656k	70035s	202s(0%)	8499s(12%)	61334s(87%)
868k	131578s	274s(0%)	15913s(12%)	115391s(87%)
1161k	234714s	1212s(1%)	25480s(11%)	208022s(88%)

Table 2: Runtime Breakup of Individual Stages [6]

The runtime breakup of individual stages of the processing pipeline as shown in table 2 reveal that the grouper and merger incur a massive performance hit. A quick investigation hints towards usage of large deep nested loops in the merger with a worst-case  $O(n^3)$  runtime complexity.

*deep nested loops*

In addition, pushing the flow-records data from one stage of the pipeline to another involved deep copying of the whole flow data whereby a mere passing across of a reference across a pipeline in a branch would have sufficed. Similar behavior is visible when the grouper when passing group records saved the individual flow-records in a temporary location tagged with the groups and/or subgroups they belonged to.

*deep copy of flow records*

The decision to use PyTables to read and write flow-records in [HDF](#) format also added to the complexity. Since, the input flow-records were most of the time in either `flow-tools` or `nfdump` file-formats, each time they had to be converted into [HDF](#) file formats prior to Flowy's execution which was unnecessary.

*pytables and hdf*

## 5.2 PRELIMINARY IMPROVEMENTS

The python parser written in [PLY](#) and the validators written for each stage of the processing pipeline that check for semantics correctness were left unmodified, since their execution time was invariant of the size of the input data and slightly varied on the query complexity.

*affinity masks, easier  
installation and  
configuration, better  
profiling and testing,  
extended command  
line switches*

Thread affinity masks were set for each new thread created to delegate the thread to a separate processor core. try/except blocks were narrowed down to only code that needed to be exception handled. A test-suite was developed with few sample queries and input traces to validate Flowy's results for regression analysis. A `setup.py` script was written to facilitate installation of Flowy and its dependencies and `options.py` was replaced with `flowy.conf` configuration file with the standard human-readable key-value pairs. The command line option handling was switched from `optparse` to `argparse` module and a switch was added for easy profiling. The profiling output was modified as well to allow standard tab delimiters which can be easily parsed by other tools. The flow query was also extended to allow file contents to be supplied using `stdin`. Variable names that are now part of Python identifiers were renamed.

*a custom c library to  
replace pytables,  
cython to connect  
the library to python*

A custom C library was written to directly read/write data in the `flow-tools` format to provide a drop-in replacement for `PyTables` and overcome the overhead of format conversions. The library sequentially reads the complete flow-records into memory to support random access required for relative filtering. Each flow-record is stored in a char array and the offsets to each field are stored in a separate struct. The array of such records are indexed allowing fast retrieval in  $O(1)$  time. The C library was connected to the Python prototype using `Cython` [\[53\]](#)[\[54\]](#). This allowed the flow-records to be easily referenced by an identifier, thereby giving away the need to every time copy all the flow-records when moving ahead in the processing pipeline. `Cython` was used since it allowed to write C extensions in a Pythonic way by strong-typing variables, calling native C libraries and allowing usage of pointers and structs, thereby providing the best of both worlds [\[55\]](#).

## 5.3 RULE INTERFACES

*filter stage struct*

A design decision was made to attempt to rewrite the entire processing pipeline in C. However, currently the core can only run absolute filters and cannot parse the flowquery file. Therefore the execution is triggered by a tedious manual filling of the struct by the contents of the query. A filter stage struct is shown in [listing 3](#). The field to be filtered is indicated using a `field_offset` and `field_length` in the char array of a records. The value to be compared against with is also supplied which can be either a static value or another field of a record. `func` is a function pointer to the operation that is to be carried out on



a record whose record identifier is passed to it. The filter runs in  $O(n)$  time as it needs to traverse through all the records of the char array.

```

1 struct filter_rule {
2     size_t field_offset;
3     uint64_t value;
4     uint64_t delta;
5     bool (*func)(
6         char *record,
7         size_t field_offset,
8         uint64_t value,
9         uint64_t delta);
10 };

```

Listing 3: Filter Rule Struct [6]

## 5.4 EFFICIENT RULE PROCESSING

The comparison operations, previously were required to make costly checks on the length of the field type passed to them, to be able to make appropriate casts. Such checks are now no longer needed. F (v1) now allows filtering of records via two methods: using specialized comparison functions or using one main fall through switch statement. The implementation defaults to using specialized comparison functions to encourage modularity in source code.

```

1 bool filter_eq_uint8_t(...);
2 bool filter_eq_uint16_t(...);
3 ...

```

Listing 4: Auto Generated Comparison Functions [52]

In the default method, there is a comparison function defined for every possible field length (33) and comparison operations (19). These functions are generated using a Python script <sup>1</sup> and are declared/defined in `auto_comps.h`, `c` as shown in listing 4. The rule definitions are now able to make calls using a function name derived from the combination of field length, delta type and operation. This subverts the need to define complex branching statements and reduces complexity. In the other method, the logic is executed by comparing the field length and the operation by falling through a huge switch statement. Such a huge switch statement is again generated using the same Python script and is defined in `auto_switch.c` as shown in listing 5.

*using function  
pointers and switch  
cases*

```

1 switch (group_modules[k].op) {
2     case RULE_EQ | RULE_S1_8 | RULE_S2_8 | RULE_ABS:
3     case RULE_EQ | RULE_S1_8 | RULE_S2_8 | RULE_REL:
4     ...

```

Listing 5: Auto Generated Switch Statement [52]

<sup>1</sup> `scripts/generate-functions.py`

## 5.5 FASTER GROUPING APPROACH

*possible grouping  
approaches*

A typical grouper module is shown listing 6. In order to be able to make comparisons on field offsets, the grouper initially creates a copy of the pointers in the filtered recordset. A naïve approach is to linearly walk through all the pointers against each pointer in the copy leading to a complexity of  $O(n^2)$ . A smarter approach is to put the copy in a hash table and then try to map each pointer while walking down the recordset once, leading to a complexity of  $O(n)$ . The hash table approach, although will work on this specific example, will fail badly on other relative comparisons.

```

1 grouper g1 {
2   srcIP = srcIP
3   dstIP = dstIP
4 }
```

Listing 6: A Grouper Module

*using quick sort and  
binary search*

It is clear that a middle ground compromise was needed. As a result, using a binary search after a quick sort on the filtered recordset was thought out. To achieve this, the array of pointers to the copy are sorted according to the offset on the right side of the comparison of the first grouping rule. Such a sorted array of pointers is then traversed linearly to find unique values. This helps the grouper perform binary searches to find records that will eventually group together. The preprocessing step takes  $O(n * \lg(n)) + O(n)$  in the worst case, with a  $O(n * \lg(k))$  for binary search on the entire recordset. The implementation of this idea is currently broken and segfaults at multiple stages. In essence, this approach is not currently used in practise and quite some effort is needed to bring it to life.

## 5.6 BENCHMARKS

Number of records	Flowy	F (v1)
103k	1177s	0.3s
337k	20875s	3.4s
656k	70035s	13s
868k	131578s	23s
1161k	234714s	86s

Table 3: Flowy vs F(v1) [6]

The benchmarks in table 3 show the relative comparison of Flowy [5] with F (v1) [6]. As the author appropriately points out, an accurate

comparison of Flowy with F v1 cannot be done accurately, since, the the C execution engine lacks many feature-set and pipeline stages. However, running a common flow query that can be run on both the implementations was used and it was not surprising that F (v1) was in orders of magnitude faster in comparison.

*vs flowy*

```
1 src port 80
2 src port 80 or dst port 25
3 src port 443 or (src port 80 and dst port 25)
```

Listing 7: F(v1) vs flow-tools, nfdump [52]

In order to evaluate how well F now performs with these added improvements, the authors decided to compare it with the state-of-the-art flow-processing tools: flow-tools [56] and nfdump [57]. A set of 3 queries involving only absolute filters was defined as shown in listing 7 and evaluated on a set of 500K – 10M flow-records. It turned out that F (v1) performed as well if not better than the other flow tools.

*vs flow-tools and  
nfdump*



## NFQL: APPLICATIONS

NFQL helped to underpin a number of recent research efforts to solve real-world application problems that were deemed difficult before. This was possible due to the power and flexibility of the flow-query language to suit itself from generic to specific needs thereby opening doors of innovation. This section documents such efforts that use the in-house flow query language as well as a few others that exploit the flow level characteristics of the traffic patterns in general.

### 6.1 APPLICATION IDENTIFICATION USING FLOW SIGNATURES

The idea behind this study was to identify applications using flow traces on a network by analyzing potential left-behind signatures that describe them [58, 15]. This was based on the hypothesis that each application type generates unique flow signatures that might work as a fingerprint feature. To achieve this, a collection of network traces were recorded from several users and subsequently analyzed. The identified signatures were formalized by writing flow queries that were executed on Flowy [44]. Several separate instances of the network traces were queried to evaluate the approach and come to a conclusion.

```

1  splitter S {}
2
3  ...
4
5  merger M {
6    module m1 {
7      branches A, B
8      A.srcip = B.srcip
9      A o B OR B o A
10   }
11   export m1
12 }
13
14 ungroup U {}
15
16 "input" -> S
17 S branch A -> F_SSDP -> G_SSDP -> GF_SSDP -> M
18 S branch B -> F_NAT_PMP -> G_NAT_PMP -> GF_NAT_PMP -> M
19 M -> U -> "output"

```

Listing 8: Skype Application Signature [15]

A formalized Flowy query to identify Skype from the flow traces for an instance is described in listing 8. The filter, grouper and group-filter sections of each branch are shown separately in listings 10 and 9. Additional queries identifying variety of web browsers, mail clients, IM clients and media players can be found in [15].

```

1  filter F_SSDP {
2      dstport = 1900
3      port = protocol("UDP")
4      dstip = 239.255.255.250
5  }
6
7  grouper G_SSDP {
8      module g1 {
9          srcip = scrip
10         dstip = dstip
11         srcport = srcport
12     }
13     aggregate srcip, sum(bytes) as B
14 }
15
16 groupfilter GF_SSDP {
17     B = 321
18 }

```

Listing 9: Branch A [15]

```

1  filter F_NAT_PMP {
2      dstport = 5351
3      port = protocol("UDP")
4  }
5
6  grouper G_NAT_PMP {
7      module g1 {
8          srcip = scrip
9          dstip = dstip
10     }
11     aggregate srcip, sum(bytes) as B
12 }
13
14 groupfilter GF_NAT_PMP {
15     B = 160
16 }

```

Listing 10: Branch B [15]

skype application  
signature

The filter F\_SSDP is used to identify the four identical [UDP](#) multicast messages the client sends out using Simple Service Discovery Protocol ([SSDP](#)) [59]. Similarly F\_NAT\_PMP filter is used to identify four Network Address Translation Port Mapping Protocol ([NAT-PMP](#)) [60] messages sent over UDP. The groupers G\_SSDP and G\_NAT\_PMP group together flow records with the same source and destination IP address and the aggregate clauses describe the meta information with unique source IP addresses for each group records along with the total bytes carried within each group. The meta information is used to further filter the group-records in GF\_SSDP and GF\_NAT\_PMP modules.

UserID	Skype	Opera	Amarok	Chrome	Live
u0	✓	○	✗	○	○
u1	✓	○	○	○	○
u2	○	○	○	○	○
u3	✓	○	✗	○	○
u4	○	○	○	○	○
u5	✓	○	✓	✓	○
u6	○	○	○	○	○
u7	○	✓	✓	○	○
u8	○	○	○	○	○
u9	✓	✓	✓	✓	○

Table 4: Application Flow Signatures: Results [15]

The identification results obtained from the analysis of flow-traces from ten unique users are compiled together in table 4. The results demonstrate a success rate of 96% for the five applications tested. This study reveals that it is possible to identify applications from their network flow fingerprints and is a first step towards automating the complete process whereby machine learning techniques would be used to automatically generate flow-queries and identify new applications and even more so newer versions of the same application.

success rate

## 6.2 CYBERMETRICS: USER IDENTIFICATION

The idea of identification of users based on biometric patterns such as keystroke dynamics [61], mouse interactions [62] or activity cycles in online games [63] has been long known. This study takes the idea even further by using flow-record patterns as a characteristic (cybermetrics) to identify a user on a network [7, 64]. Such a cybermetric user identification can be used for the purpose of providing secure access, system administration and network management. The feature extraction module of the analyzer as shown in figure 11 uses three distinct feature sets that could possibly be used to identify a user from a flow-record trace.

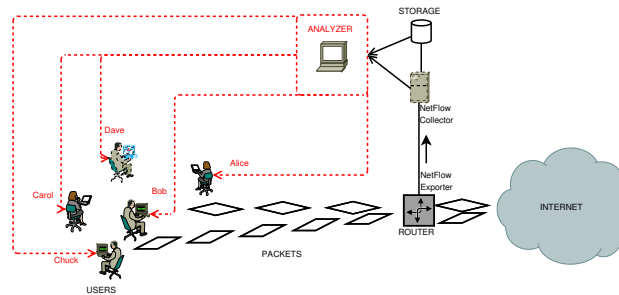


Figure 11: Cybermetrics: Overview [7]

Initial research efforts started with identifying application signatures in flow-records in [58, 15] and became relevant because different people have different preferences in the applications they use and as such a set of applications in flow-records is a characteristic feature of a user. Flowy queries were formalized for four different set of applications and tested against a known set of users. The evaluation results of the derived queries as shown in table 4 demonstrated a strong evidence of presence (or absence) of applications and thereby provided an eventual marker for user identification.

*application  
signatures*

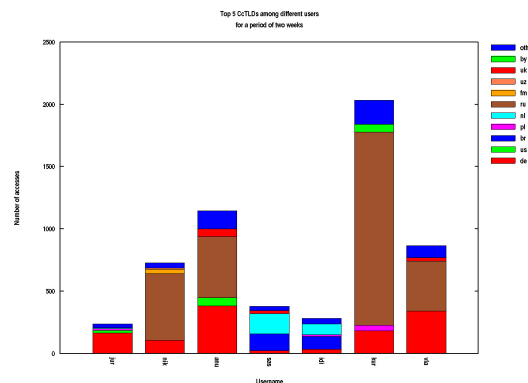


Figure 12: Geographical Preferences [7]

geographical  
preferences

The authors also looked into the geographical affiliations of different users by analyzing the Country Code Top-Level Domain (ccTLD) of the browsed websites. They proposed a hypothesis that a user's origins strongly influences their browsing activity. The analysis of the results established that the top five visited ccTLDs constituted more than 85% of the overall number of a user's visits as shown in figure 12.



Figure 13: Daily Distributions for HTTP Traffic [7]

flow-record statistics

In the end, the authors introduced a proof-of-concept method of user differentiation based on statistical features. These features considered daily distributions of parameters that were based on different port numbers. For instance, figure 13 shows the daily distribution of different users based on their HTTP traffic usage. It was also witnessed that the time duration also played a key role in the process of feature formation, whereby the number of longer flows increased with the duration and consequently resulted in higher cross-correlations as shown in figure 14



Figure 14: Cross Correlation of Traces with Varying Times [7]

This research is a first attempt to identify users based on their network flow fingerprints and the on-going effort is focussing on sophisticated machine learning techniques to learn behavioral patterns of known users to identify them in the future from their current network-flow traces.



## 6.3 IPV6 TRANSITION FAILURE IDENTIFICATION

The IPv4 address space depletion is upon us and has become more imminent in the last few years. While IPv6 can readily expand the extent of the Internet, deploying it alone is clearly not a solution today and hence there are a continuum of transitioning solutions that would help in this migration. In this study [65] we evaluated the compatibility of popular applications with such transitioning solutions: NAT64 [66] and Dual-Stack Lite [67]. The goal was to find potential failures by identifying application failure signatures left behind in the flow-record traces using Flowy. These failure signatures could later be used by service providers to automate the detection and eventually shorten the deployment verification cycle.

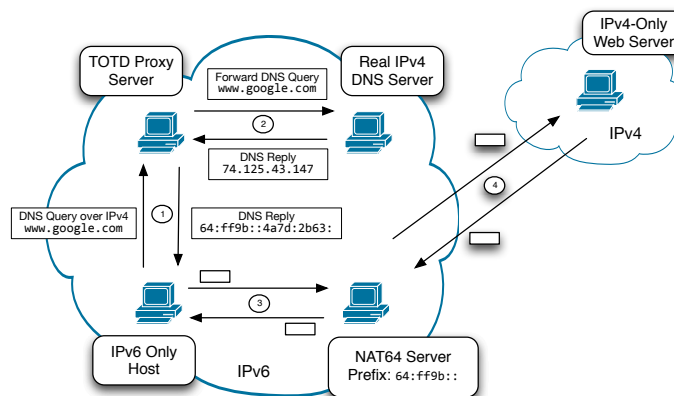


Figure 15: NAT64 Setup [1]

In the NAT64 deployment testbed as shown in figure 15, the authors witnessed failure in 3 applications: Skype, OpenVPN and Transmission. Flowy queries were defined to establish failure signatures for each application. A formalized Flowy query to identify Skype failure signature for an instance is described in listing 13. The filter sections of each branch are shown separately in listings 11 and 12.

*application operation  
under NAT64*

```
1 filter f-mDNS {
2   dstport = 5353
3   srcport = 5353
4   dstip = 224.0.0.251
5   duration > 1 sec
6   duration < 5 sec
7 }
```

Listing 11: Branch A [65]

```
1 filter f-login1 {
2   dstport = 443
3   duration > 55 sec
4   duration < 59 sec
5 }
```

Listing 12: Branch B-C-D [65]

Filter `f-mDNS` is used to filter multicast messages used by Skype to discover clients in the link-local network sent to the destination IP address-port combination (224.0.0.251 : 5353). Filter `f-login1` is used

*skype failure signature*

to filter 3 unsuccessful attempts to contact the login server each in a separate branch. The source port and the duration increases with decreasing number of packets for each subsequent flow.

```

1  splitter S {}
2
3  ...
4
5  grouper g-login1 {
6    module g1 {
7      srcport = srcport
8      dstip = dstip
9      dstport = dstport
10   }
11   aggregate srcip, dstip, srcport, td,
12   sum(packets) as pkt-sum, count(rec_id) as n
13 }
14
15 merger M {
16   branches mDNS, LOGIN1, LOGIN2, LOGIN3
17
18   LOGIN1.srcip = LOGIN2.srcip
19   LOGIN2.srcip = LOGIN3.srcip
20   LOGIN1.dstip = LOGIN2.dstip
21   LOGIN2.dstip = LOGIN3.dstip
22
23   LOGIN1.srcport = LOGIN2.srcport rdelta 1
24   LOGIN2.srcport = LOGIN3.srcport rdelta 1
25
26   LOGIN1.pkt-sum > LOGIN2.pkt-sum
27   LOGIN2.pkt-sum > LOGIN3.pkt-sum
28
29   mDNS.td < LOGIN1.td
30   mDNS.td < LOGIN2.td
31   mDNS.td < LOGIN3.td
32
33   mDNS < LOGIN1
34   mDNS < LOGIN2
35   mDNS < LOGIN3
36 }
37
38 "input"-> S
39 S br mDNS -> f-mDNS -> g-mDNS -> gf-mDNS -> M
40 S br LOGIN1 -> f-login1 -> g-login1 -> gf-login1 -> M
41 S br LOGIN2 -> f-login2 -> g-login2 -> gf-login2 -> M
42 S br LOGIN3 -> f-login3 -> g-login3 -> gf-login3 -> M
43 M -> U -> "output"

```

Listing 13: Skype Failure Signature [65]

The groupers count the number of packets in each flow-records using `pkt-sum` which is later utilized by the merger stage to distinguish the branches. The group-filter stage finally is used to filter out groups with more than one record.

The NAT64 translation works when the applications running on the IPv6-only host explicitly make DNS requests to allow DNS64 to capture and masquerade them as fake IPv6 addresses that are eventually sent to the NAT64 box. If the applications use IPv4 literals to contact the servers, DNS64 is skipped and therefore NAT64 cannot perform the translation. This was reason behind the failure of the other two applications (OpenVPN and Transmission).

This study sets across a baseline to automate the failure detection by formalizing queries against flow-records. While a more exhaustive study encompassing wider set of applications still needs to be carried out, it is imperative that this unique approach is not just limited to IPv6 transition technologies, but can be utilized to identify failures in more generic cases.

*failure when using  
IPv4 literals*

## 6.4 OPENFLOW

OpenFlow [68] is an open standards protocol that runs between an Ethernet switch and an OpenFlow controller (a software designed to run on a x86 server) to securely manage the forwarding plane of the switch over the network as shown in figure 16. This enables the controller to push out policies that dictate how to process flow-records crossing the networking infrastructure to eventually improve bandwidth, reduce latency and save power.

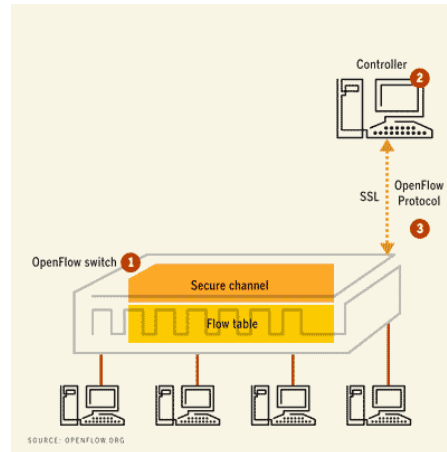


Figure 16: OpenFlow Architecture [13]

OpenFlow initially started as a way to allow researchers to experiment with new ideas in sufficiently realistic settings by allowing the live production networking gear to open a narrow programmable external interface to it whereby at the same time keeping the inner workings of the gear hidden and proprietary. The idea took off outside the academic setting in recent years with the need of data centers requiring to run large-scale map/reduce jobs with full cross-sectional bandwidth. Such a requirement called for flexible forwarding and programmable networks to meet the application-specific needs. Today, the commercial underpinning of OpenFlow are driven by the Infrastructure as a Service (IaaS) providers trying to virtualize their network architecture to solve the issue of multi-tenancy to implement Network as a Service (NaaS) architectures [69].

*motivation*

An OpenFlow switch manages a flow table to keep record of the flows crossing it. A flow table contains a packet header, an action and some statistical information about the flow. OpenFlow defines a common set of methods to program such flow tables irrespective of the way different vendors internally defined them. This allows a network administrator to partition the incoming traffic into numerous Virtual Local Area Networks (vLANs) thereby isolating the production and several experimental networks at the layer 2 level. Now, with the a complete suite of OpenFlow software stack defined on top of the

*programming flows*

controller, such a power is also available at the hands of the developers that gives them the ability to control the flow tables themselves and even decide the routes for their flow.

*software stack*

The OpenFlow protocol in itself is like an x86 instruction-set by itself. However, there is a lot of innovation possible at the software stack layer that can be built on top the controller that exposes the API and pushes this low-level instruction-set to the networking gear. For instance, the stack can deploy network-wide policies and administer Access Control Lists (ACLs) for each incoming flow or allow seamless handover of mobile hosts by rerouting requests making the networking gear location-aware in itself. As such, it is conspicuous that the possibilities are endless and is the beginning of a kick-start of a new internet evolution.

*flowy and openflow*

It is not difficult to anticipate that Flowy could be of much use for OpenFlow. It could be envisaged that the controller would define Flowy queries to get to a specific flow-entry in the flow table before sending action level instructions to the networking gear. In addition, Flowy could be extended to allow flow manipulation constructs to define the action instructions themselves which can be sent out by the controller. In a future outlook, Flowy can even be envisioned to allow procedural constructs (variables, functions, loops, conditions) around the declarative query to add power to what can be retrieved or sent back to the switches.

## 6.5 FLOW LEVEL SPAM DETECTION

Feature	Description
Pkts	Packets
Rxmits	Retransmissions
RSTs	Packets with RST bit set
FINs	Packets with FIN bit set
Cwndo	Times o-window advertised
CwndMin	Minimum window advertised
MaxIdle	Maximum idle time between packets
RTT	Initial round trip time estimate
JitterVar	Variance of inter-packet delay

Table 5: Features in Spam Flow [14]

Classical methods to mitigate spam such as content filtering and reputation analysis utilize the the weakness of spam messages and the places from where they originate from. Though currently effective, it's only a matter of time when spammers find a way to subvert around these vantage points. In this study [14, 70], the authors analyze the transport level characteristics of the email flows to differentiate between spam and legitimate email. These characteristics exploit the fundamental weakness of each spam: the requirements to send large amounts of the same email on resource constrained links owned by

compromised botnets which is unlikely to change in the near future. They reason that a spammer's traffic is more likely to experience TCP timeouts, retransmissions, resets and variable Round Trip Time (RTT) estimates. Based on this hypothesis they extract 13 learning features as shown in table 5 to formalize a machine learning problem.

*spamflow features*

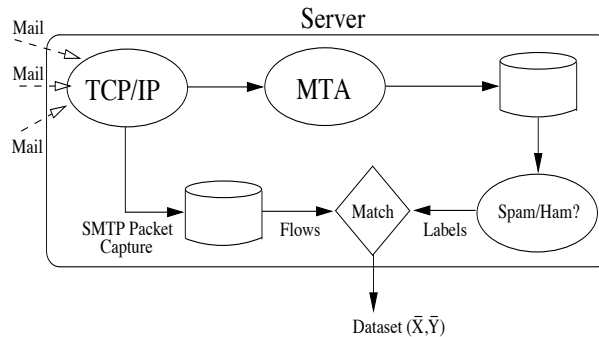


Figure 17: Spam Flow Classifier [14]

The data collection methodology is depicted in figure 17 where TCP packets corresponding to email messages are extracted and examined on a per-email flow basis. The packets in an email flow are coalesced together by using TCP port numbers in the email headers. Using machine learning feature selection, a spam classifier is built that matches each flow to a binary spam/ham ground-truth label. Support Vector Machines (SVMs) [71] are used for classification and Greedy Forward Fitting (FF) [72] is used for feature selection to find a set of features that provide the least training error. It turns out the classifier achieves 90% accuracy with 78% detectability of false-negatives from a particular content filter.

*a spam classifier*

One possible limitation of this approach is the inability to distinguish between botnets sending large quantities of spam and innocent busy hosts that happen to be on a congested network. This is most probably because of the naïve Simple Mail Transfer Protocol (SMTP) flow aggregation and filtering. We believe, that Flowy can help overcome this shortcoming by automated flow-queries generated by another trained classifier that filters out these innocent hosts before passing them to the spam classifier thereby reducing the number of false negatives.

*flowy and spamflow*

This study presented a content and IP reputation agnostic scheme based on SMTP flow-level analysis of traffic stream. It is imperative, augmented with Flowy capabilities, this approach can be extended to identify any botnet generated traffic. Such a novel approach could then be used to also identify phishing attacks, scam infrastructure hosting, Distributed Denial of Service (DDoS), dictionary attacks and Completely Automated Public Turing Test to Tell Computers and Humans Apart (CAPTCHA) solvers.

*extending spamflow*



## Part III

### IMPLEMENTATION AND EVALUATION

This section dives deep into the implementation of the next iteration F(v2) of the [NFQL](#) execution engine. This iteration provides a functional and robust implementation of the complete processing pipeline. It is flexible to allow runtime evaluation of [NFQL](#) flowqueries and is backed up with automated builds, regression and benchmarking suites. The organization of the section is described below.

In chapter [7](#) we begin by analyzing the current implementation snapshot. The analysis involves reverse-engineering the parser and the execution engine. This is followed by a discussion on an early visualization of a complete engine refactor using abstract objects and reasonings on an envisaged high-level execution workflow.

In chapter [8](#) we introduce the inner workings of the code. It begins by an explanation of how each stage is brought to life to result in a robust pipeline execution. The grouper and merger internals, being non-trivial are explained in more details. This is followed by an explanation on how the engine is engineered to make runtime query evaluation a possibility. The chapter concludes with a discussion on automated builds using `cmake` and a full-fledged regression-test suite using python scripts.

In chapter [9](#) we begin by a discussion of the execution engine profiling results. This is followed by how python scripts are used to completely automate the process of benchmarking the engine against SiLK. A number of queries are run a collection of varied-sized flow traces. The chapter concludes with a discussion of the evaluation graphs. We conclude the discussion in chapter [10](#) by documenting the future work by segregating it into major goals and minor issues that need to be resolved to carry the implementation work forward.





## DESIGN

With a software that has underwent such significant iterative lifecycles over the past few years, it is imperative to understand and analyze the inner working of the application before diving in to add more functionality. Reverse engineering the current snapshot not only helped identify glitches to give a head start with preliminary improvements, but also enabled understanding the design of the eventual execution workflow. It also helped in early visualization of a complete engine refactor to introduce abstract objects that made it possible to evaluate the flowquery at runtime. This chapter starts off with the analysis to set a platform for reasoning out the design patterns and the user functionality envisaged from the finished product.

### 7.1 FLOWY PARSER AND F(v1) ENGINE ANALYSIS

Since both the parser and the engine were developed in an isolated sandboxed environments, an extensive validation of how their functionality (or errors) would regress was always needed. In this pursuit, the first challenge was to get F(v1) engine to compile smoothly. Since, the engine was using linux-specific integer types to read the flow record offsets, its compilation was an issue on other Unix flavors. As such moving to C99 [73] fixed-width integer types increased portability. In addition a number of extraneous files that were not part of the build plagued the source directory and were removed after thorough inspection. Boolean enums were replaced by C99 bool types and include guards were added in the headers to remove circular dependencies. These changes led to succesful compilation of the engine and an initial run iteration is shown in listing 14. It appears that the execution engine can read the flow records in memory and successfully filter records in each thread. However, it segfaults at the grouper stage, thereby ending the execution.

*compilation and  
runtime issues*

```

1 $ ./flowy2 < trace.ft
2 number of filtered records: 556
3 number of filtered records: 166
4 segmentation fault ./flowy2 < trace.ft
5
6 (gdb) backtrace
7 ...
8 #1 0x00000001000134fa in build_record_trees
9 #2 0x00000001000138a0 in grouper
10 #3 0x0000000100011eb9 in branch_start
11 ...

```

Listing 14: F(v1): Segmentation Fault

*missing pipeline  
stages, hardcoded  
rules, assumptions*

In addition, the implementations for group filter, merger and un-grouper are missing. A major issue is that the complete flow query is hardcoded in pipeline structs as shown in listing 15. Similar rules are hardcoded for each branch. In addition the functions that evaluate the filter and the grouper rule also assume offsets of a specific integer type and result in undefined behavior once the parameters in the flow query are altered.

```

1 struct filter_rule filter_rules_branch1[1] = {
2   { data->offsets.dstport, 80, filter_eq_uint16_t },
3 };
4
5 struct grouper_rule group_module_branch1[2] = {
6   { data->offsets.srcaddr, data->offsets.srcaddr, grouper_eq_uint32_t_uint32_t_rel },
7   { data->offsets.dstaddr, data->offsets.dstaddr, grouper_eq_uint32_t_uint32_t_rel },
8 };
9
10 struct grouper_aggr group_aggr_branch1[2] = {
11   { data->offsets.srcaddr, aggr_static_uint32_t },
12   { data->offsets.dstaddr, aggr_static_uint32_t },
13 };
14
15 binfos[0].branch_id = 0;
16 binfos[0].filter_rules = filter_rules_branch1;
17 binfos[0].num_filter_rules = 0;
18 binfos[0].group_modules = group_module_branch1;
19 binfos[0].num_group_modules = 2;
20 binfos[0].aggr = group_aggr_branch1;
21 binfos[0].num_aggr = 2;

```

Listing 15: F(v1): Flow Query Hardcoded in Pipeline Structs

*reverse-engineering*

To analyze the call flow and data structure collaboration and dependency, the execution engine was reverse engineered to generate Unified Modeling Language (UML) using doxygen. A similar technique was used to generate UML for the parser using pylint and pyreverse. Makefile targets were added to ease documentation generation for future developers as shown in listing 16

```

1 [engine] $ make doc
2 [parser] $ make doc

```

Listing 16: F(v2): High Level Documentation

*arguments parsing  
in parser*

Flowy parser tools assumed correct number and format of command line arguments and poorly exited out of execution with IndexError exceptions. The python argparse module is now used to exit gracefully with usage instructions on bad input as shown in listings 17.

```

1 [parser] $ python src/flowy.py
2 usage: flowy.py [options] query.flw
3
4 [parser] $ python src/ft2hdf.py
5 usage: ft2hdf.py [options] input_path1 [input_path2 [\cdots]] output_file.h5
6
7 [parser] $ python src/printhdf.py
8 usage: printhdf.py trace.h5
9
10 [parser] $ python print_hdf_in_step.py
11 usage: print_hdf_in_step [options] input_files

```

Listing 17: Flowy Interfaces

It was clear from the generated [UMLs](#) that the current snapshot required multiple stages of refactoring before it can be deemed maintainable. As such forward declarations were removed and thus arising circular dependencies were resolved by reorganizing the code in multiple files. For instance, a base header was added to include common library headers as shown in figure 18. `error_functions` module was added to avoid plaguing error handlers everywhere. Each stage of the pipeline was moved into its separate module, while utility functions were moved to `utils` module. All the pipeline structs were also moved to a specific pipeline header to increase readability.

*minor refactor*

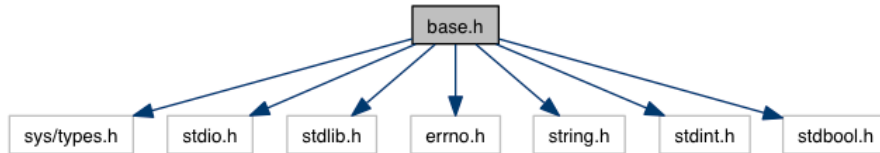


Figure 18: F(v2): Base Header

## 7.2 EXECUTION WORKFLOW AND ABSTRACT OBJECTS

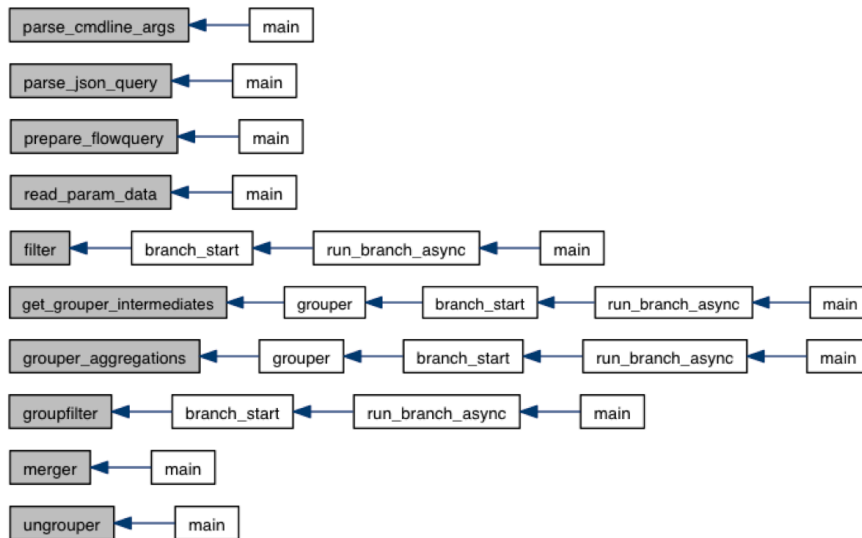


Figure 19: F(v2): Execution Engine Workflow

In order to keep the codebase maintainable, it was essential to design the execution engine workflow in such a way so as to naturally map it to the original pipeline model specification [11] as shown in figure 19. Each stage of the pipeline is a separate independent module blackboxed into one public interface function. Each stage is

also wrapped around conditional compilation macros to allow them to be easily enabled/disabled during development if desired so.

```

1 struct flowquery {
2     size_t          num_branches;
3     struct branch** branchset;
4
5     size_t          num_merger_rules;
6     struct merger_rule** mruleset;
7
8     struct merger_result* merger_result;
9     struct ungrouper_result* ungrouper_result;
10 };

```

Listing 18: F(v2): Flow Query Struct

*flowquery and  
branch struct*

The abstract objects that store the JSON query and the results that incubate from each stage are designed to be self-descriptive and hierarchically chainable. The complete JSON query information for instance, is held within the `flowquery` struct as shown in listing 18. Each individual branch of the `flowquery` itself is described in a `branch` struct. A collection of these `branch` structs are referenced in the parent `flowquery` struct. All the query rules are clubbed into `X_ruleset`, where `X` can be any stage as shown in listing 19.

```

1 struct branch {
2     size_t          num_grouper_rules;
3     size_t          num_aggr_rules;
4     size_t          num_gfilter_rules;
5
6     struct filter_rule** filter_ruleset;
7     struct grouper_rule** grouper_ruleset;
8     struct aggr_rule** aggr_ruleset;
9     struct gfilter_rule** gfilter_ruleset;
10
11     struct filter_result* filter_result;
12     struct grouper_result* grouper_result;
13     struct groupfilter_result* gfilter_result;
14 };

```

Listing 19: F(v2): Branch Struct

*public interfaces*

A call to the public interface function of each stage returns a `X_result` struct object as shown in listing 20. The `X_result` objects encapsulate all elements of the stage into one single entity as shown in listing 20 to easily allow them to be passed around and for easy maintainability of in-memory object stores.

```

1 struct filter_result*
2 filter(...) {...}
3
4 struct grouper_result*
5 grouper(...) {...}
6
7 struct groupfilter_result*
8 groupfilter(...) {...}
9
10 struct merger_result*
11 merger(...) {...}
12
13 struct ungrouper_result*
14 ungrouper(...) {...}

```

Listing 20: F(v2): Public Interfaces

Each result struct holds information about the number of flow records that passed the stage and pointers to each such flow records. Since the group filter and merger stages do not work on the individual flows but on a collection; they take the group struct that encapsulates a collection of similar flows as input arguments. It is important to realize that the flow records themselves are never carried forward from each stage to its subsequents, but only offset pointers to the original flow trace are.

*result structs*

```

1 struct filter_result {
2     size_t          num_filtered_records;
3     char**          filtered_recordset;
4 };
5
6 struct grouper_result {
7     size_t          num_unique_records;
8     char**          sorted_recordset;
9     char**          unique_recordset;
10    size_t          num_groups;
11    struct group**   groupset;
12 };
13
14 struct groupfilter_result {
15     size_t          num_filtered_groups;
16     struct group**   filtered_groupset;
17 };
18
19 struct merger_result {
20     size_t          num_group_tuples;
21     size_t          total_num_group_tuples;
22     struct group***  group_tuples;
23 };
24
25 struct ungrouper_result {
26     size_t          num_streams;
27     struct stream** streamset;
28 };

```

Listing 21: F(v2): Result Structs

The query fragment structs (*X\_ruleset*) used to get the result is greedily deallocated soon after the stage returns to keep the in-memory usage to the minimum. The *filter\_ruleset* although are kept until the end of the grouper stage since it helps the grouper aggregation stage make decisions on whether a linear pass through the flow trace is required to aggregate a column that may have been already a criterion for the filter stage.

*greedy ruleset  
deallocation*

```

1 branch->grouper_result = grouper(...);
2 if (branch->grouper_result == NULL) ...
3 else {
4     /* free filter rules */
5     /* free grouper rules */
6     /* free grouper aggregation rules */
7 }
8
9 branch->gfilter_result = groupfilter(...);
10 if (branch->gfilter_result == NULL) ...
11 else {
12     /* free group filter rules */
13 }
14
15 fquery->merger_result = merger(...);
16 if (fquery->merger_result == NULL) ...
17 else {
18     /* free merger rules */
19 }

```

Listing 22: F(v2): Greedy Deallocation

## 7.3 USER INTERFACE DESIGN

*pretty usage help,  
tracking invalid  
options*

It is essential to allow the interface to be intuitive to any new user who is interested in using the tool for network analysis. In essence, this is achieved using the standard `getopt_long` call to allow both short and long option arguments. The execution engine appropriately displays the usage help when insufficient number of arguments are provided as shown in listing 23. The engine is also interactive to help one choose the right switches with required options.

```

1  $ bin/engine
2  usage: bin/engine [OPTIONS] queryfile tracefile      query the specified trace
3                or: bin/engine [OPTIONS] queryfile -    read the trace from stdin
4
5  OPTIONS:
6  -d, --debug                enable debugging mode
7  -v, --verbose              increase the verbosity level
8  -h, --help                 display this help and exit
9  -V, --version              output version information and enable exit
10
11 $ bin/engine queryfile tracefile --foo
12 bin/engine: invalid option --foo
13
14 $ bin/engine queryfile tracefile --verbose
15 bin/engine: option --verbose requires an argument
16
17 $ bin/engine queryfile tracefile --verbose=5
18 ERROR: valid verbosity levels: (1-3)

```

Listing 23: F(v2): User Interface

*consistency checks*

Since the execution engine largely depends on the sanity of the query and trace files passed to it as arguments, it is essential to let the input files pass through a level of consistency check before going forward with the processing pipeline to avoid any undefined behavior as shown in listing 24.

```

1  $ bin/engine README.md tracefile
2  ERROR: json_tokenizer_parse_ex(...)
3
4  $ bin/engine queryfile README.md
5  ERROR: ftio_init(...)

```

Listing 24: F(v2): Consistency Checks

*backtrace on graceful  
exits*

With a software undergoing such a rapid pace of development, it's helpful to be able to see the inner workings of each stage of pipeline during a debugging lifecycle. As such, the engine echoes a backtrace whenever it fails gracefully as shown in listing 25.

```

1  $ bin/engine foo bar
2  ERROR: open(...)
3  BACKTRACE:
4  0  engine                0x000000010bc891c2 print_trace + 34
5  1  engine                0x000000010bc893cb errExit + 395
6  2  engine                0x000000010bc898de read_param_data + 174
7  3  engine                0x000000010bc8c600 main + 80
8  4  engine                0x000000010bc6e054 start + 52

```

Listing 25: F(v2): Backtraces

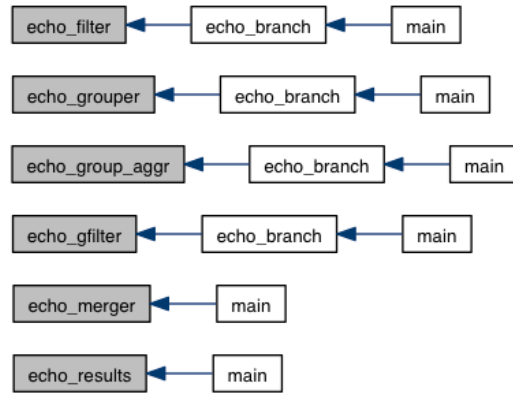


Figure 20: F(v2): Verbosity Levels Workflow

The engine also allows to increase the amount of echo using a number of verbosity levels. A specific function is designed to handle the echo of each stage of the pipeline as shown in figure 20. In its default state, the engine only echoes the resultant streams of flow records. A debug (or `-verbose=3`) level engine execution is shown in listing 26. In addition to echoing the flow (or group) records resulting from each stage, it also echoes the results of each intermediate stage alongwith the original trace that was passed to it. With `-verbose=2`, the echo of the original trace is pruned, while intermediate results get pruned with `-verbose=1`.

*debug and verbosity levels*

```

1  $ bin/engine tracefile queryfile --debug
2
3  # mode:                normal
4  # capture hostname:    ihp.jacobs.jacobs-university.de
5  ...
6
7  No. of Filtered Records: ...
8  No. of Sorted Records: ...
9  No. of Unique Records: ...
10 No. of Groups: (Verbose Output): ...
11
12 ... 0      216.137.61.203  80  0      192.168.0.135 ...
13 ... 0      216.137.61.203  80  0      192.168.0.135 ...
14
15 ... 0      8.12.214.126    80  0      192.168.0.135 ...
16 ... 0      8.12.214.126    80  0      192.168.0.135 ...
17
18 No. of Groups: 32 (Aggregations): ...
19
20 ... 0      216.137.61.203  80  0      192.168.0.135 ...
21 ... 0      8.12.214.126    80  0      192.168.0.135 ...
22
23 No. of Filtered Groups: (Aggregations): ...
24 No. of (to be) Matched Groups: ...
25
26 ... 0      192.168.0.135  0  0      204.160.123.126 80  ...
27 ... 0      87.238.86.121  80  0      192.168.0.135  0  ...
28
29 No. of Merged Groups: 3 (Tuples): ...
30
31 ... 0      192.168.0.135  0  0      216.46.94.66  80  ...
32 ... 0      216.46.94.66  80  0      192.168.0.135  0  ...
33
34 No. of Streams: ...

```

Listing 26: F(v2): Debugging





## IMPLEMENTATION

The effort to provide a clean usable implementation of the language was from the initial outset backed up by three goals. The first goal was to allow the implementation to flawlessly walk through all stages of the pipeline without incurring major performance overhead. The second goal was to abstract out the engine functionality in such a way so as to allow runtime evaluation of the flow query. The third goal was to provide a clean layout of the working code with a seamless build process to allow future developers to quickly get started on top of the current snapshot. This was supplemented by a thorough regression and benchmarking suite to make the code verifiable. This chapter introduces the inner workings of the code to explain how these goals were set into practise and brought to life.

## 8.1 GROUPEL INTERNALS

In order to be able to make comparisons on field offsets, the grouper initially creates a copy of the pointers in the filtered recordset. A naïve approach is to linearly walk through all the pointers against each pointer in the copy leading to a complexity of  $O(n^2)$ , where  $n$  is the number of flow records. A smarter approach is to put the copy in a hash table and then try to map each pointer while walking down the recordset once, leading to a complexity of  $O(n)$ . The hash table approach, although will work on this specific example, will fail badly on other relative comparisons. The F(v1) execution engine [6] formulates an approach to grouping records using a binary search after a quick sort on the field offset of the first grouping rule of each record in the filtered recordset. This helps the grouper perform faster search lookups to find records that must group together in  $O(n * \lg(k))$  time with a preprocessing step taking  $O(n * \lg(n)) + O(n)$  in the worst case, where  $n$  is the number of flow records and  $k$  is the number of unique flow records that passed the filter. The implementation of this idea was broken and used to segfault at multiple steps within the grouping stage. F(v2) introduces a completely functional implementation of this approach that does not incorporate any assumptions on the type of field offsets and is robust enough to work with different types of grouping queries.

The approach of grouping records using `qsort` and `bsearch` definitely is an improvement. However, it appears having an actual binary search tree would benefit more, whereby one of the recordset will be traversed sequentially in  $O(n)$  time and for each record, the compar-

*functional grouping  
using qsort and  
bsearch*

using search trees  
and hash tables

ison will be performed by tree lookups in  $O(\log(n))$  time bringing down the complexity to  $O(n\log(n))$  and is a future action item. In addition, it would be best if the engine can figure out the type of the relative comparison to trigger a hash table lookup for equality comparisons to bring down the complexity to  $O(n)$  for this specific case.

```

1 struct grouper_type {
2 #if defined (__APPLE__) || defined (__FreeBSD__)
3     int (*qsort_comp)(
4         void*          thunk,
5         const void*     e1,
6         const void*     e2
7     );
8 #elif defined (__linux)
9     int (*qsort_comp)(
10        const void*     e1,
11        const void*     e2,
12        void*           thunk
13    );
14 #endif
15 ...
16
17 #if defined (__APPLE__) || defined (__FreeBSD__)
18     qsort_r(
19         sorted_recordset_ref,
20         num_filtered_records,
21         get_grouper_intermediates(
22             (void*)&grouper_ruleset[0]->field_offset2,
23             gtype->qsort_comp
24         );
25 #elif defined (__linux)
26     qsort_r(
27         sorted_recordset_ref,
28         num_filtered_records,
29         sizeof(char **),
30         gtype->qsort_comp,
31         (void*)&grouper_ruleset[0]->field_offset2
32     );
33 #endif

```

Listing 27: F(v2): qsort\_r Invocation

cross platform  
qsort\_r and  
bsearch\_r

The reentrant `qsort_r` was used, since it can pass an additional argument `thunk` to the comparator, which in our case is the field offset used for comparing two flow records. Since the order of arguments of `qsort_r` are different for `glibc` and `BSD`, the function invocation had to be wrapped around platform specific macros as shown in listing 27. More Surprisingly, there is currently no equivalent `bsearch_r` to complement `qsort_r`. As such, the contemporary `bsearch` function from the `glibc` library was adapted to accommodate the void and is defined in `utils` module.

```

1 struct group {
2     size_t          num_members;
3     char**          members;
4     struct aggr_result* aggr_result;
5 };
6
7 struct aggr_result {
8     char*           aggr_record;
9     struct aggr**   aggrset;
10 };

```

Listing 28: F(v2): Group Struct

Group records are a conglomeration of several flow records with some common characteristics defined by the flow query. Some of the non-common characteristics can also be aggregated into a single value using group aggregations as shown in listing 29. Since, the execution engine supports multiple verbosity levels, it is useful if a single group record can be again mapped into a NetFlow v5 record template, so that it can be echoed as the representative of all its members. This was achieved using a struct group as shown in listing 28.

*groups as cooked  
netflow v5 records*

```

1 grouper g_www_res {
2   module g1 {
3     srcip = srcip
4     dstip = dstip
5   }
6   aggregate srcip, dstip, sum(bytes) as bytes, bitOR(tcp_flags) as flags,
7 }
8
9 $ bin/engine queryfile tracefile --verbose=1
10 ...
11 No. of Groups: ...
12
13 ... SrcIPAddress ... DstIPAddress OR(FL) Sum(Octets)
14 ... 4.23.48.126 ... 192.168.0.135 3 81034
15 ... 8.12.214.126 ... 192.168.1.138 2 5065

```

Listing 29: F(v2): Aggregations Example

There can be a situation where the query designer might incorrectly ask for aggregation on a field already specified in a grouper (or filter) module. If the relative operator is an equality comparison, the aggregation on such a field becomes less useful, since the members of the grouped record will always have the same value for that field. The engine is now smart to realize this redundant request and ignores such aggregations as shown in listing 29.

*ignoring redundant  
aggregation requests*

```

1 grouper g_www_res {
2   module g1 {}
3   aggregate sum(bytes) as bytes
4 }
5
6 $ bin/engine queryfile tracefile --verbose=1
7 ...
8 No. of Groups: 1 (Aggregations)
9
10 ... Sum(Octets)
11 ... 2356654

```

Listing 30: F(v2): Clubbing Records with No Grouper Rules

Records are clubbed together into one group if no group modules are defined. Previously such a query used to form groups for each individual filtered record. That was less useful since then it was not possible meaningful aggregations on all the records that passed the filter stage. Now, when the group modules are empty, all the filtered records are clubbed into one group to allow aggregations as shown in listing 30.

*clubbing records  
with no grouper  
rules*

## 8.2 ROBUST PIPELINE EXECUTION AND RUNTIME COMPLEXITY

```

1 grouper g {
2   module g1 {
3     srcip = srcip
4     dstip = dstip
5   }
6   aggregate srcip, dstip, sum(packets) as pkts
7 }
8
9 groupfilter gf {
10  pkts > 200
11 }

```

Listing 31: Group Filter Query Example

*group filter  
implementation*

A simple query to filter the groups with `sum(pkts) > 200` is shown in listing 31. The struct rule holds information about the flow record offset, the value being compared to and the operator which maps to a unique enum value. This enum value is later used to map the operation to a specific group-filter function using a switch case case as shown in listing 32. The group-filter functions are auto-generated using a python script `scripts/generate-functions.py`.

```

1 struct gfilter_rule gfilter_branch2[1] = {
2   {trace_data->offsets.dPkts, 200, 0, RULE_GT, NULL}
3 };
4 ...
5
6 /* for loop for the group-filter */
7 for (int j = 0; j < binfos[i].num_gfilter_rules; j++) {
8   switch (binfos[i].gfilter_rules[j].op) {
9     ...
10    case RULE_GT:
11      binfos[i].gfilter_rules[j].func = gfilter_gt;
12      break;
13    ...
14   }
15 }

```

Listing 32: F(v2): Group Filter Implementation

*merger  
implementation*

The merger is used to relate the groups from different branches according to a criterean. A query to merge groups of flow records creating a session between two endpoints is shown in listing 33. Similar to the group filter, a unique enum value is used to map the operator to a specific specific merger function using a switch case as shown in listing 34. The merger functions are again auto-generated using `scripts/generate-functions.py`.

```

1 merger m {
2   module m1 {
3     branches A, B
4     A.srcip = B.dstip
5     A.dstip = B.srcip
6   }
7 }

```

Listing 33: Merger Query Example

```

1 struct merger_rule mfilter[2] = {
2     {&binfos[0], trace_data->offsets.srcaddr,
3      &binfos[1], trace_data->offsets.dstaddr,
4      RULE_EQ | RULE_S1_32 | RULE_S2_32, 0, NULL},
5
6     {&binfos[0], trace_data->offsets.dstaddr,
7      &binfos[1], trace_data->offsets.srcaddr,
8      RULE_EQ | RULE_S1_32 | RULE_S2_32, 0, NULL},
9 };
10 ...
11
12 /* for loop for the merger */
13 for (int j = 0; j < fquery->num_merger_rules; j++) {
14     switch (fquery->mrules[j].op) {
15         ...
16         case RULE_EQ | RULE_S1_32 | RULE_S2_32:
17             fquery->mrules[j].func = merger_eq_uint32_t_uint32_t;
18             break;
19         ...
20     }
21 }

```

Listing 34: F(v2): Merger Implementation

The ungrouper implementation is straightforward. Given the group tuples, the ungrouper returns a set of stream of flow records. The matched group tuples are generated by the merger. The ungrouper returns as many streams as there are number of matched group tuples. An example output is shown in listing 35.

*ungrouper  
implementation*

```

1 $ bin/engine tracefile queryfile
2
3 No. of Streams: 2
4 -----
5
6 No. of Records in Stream (1): 24
7
8 ... Sif      SrcIPaddress      SrcP  DIf      DstIPaddress      DstP  ...
9
10 ... 0        192.168.0.135      56225 0        216.46.94.66      80    ...
11 ... 0        216.46.94.66      80     0        192.168.0.135     56228 ...

```

Listing 35: F(v2): Ungrouper Result Echo

The public interface function grouper(...) call was plagued with hardcoded uint32\_t type assumptions on the field offset. These field offsets are used to make grouper rule comparisons. The function now internally calls get\_gtype(...) to fall through a switch case to determine the type of the field offset at runtime as shown in listing 36.

*flexible grouper with  
no type assumptions*

```

1 struct grouper_type* get_gtype(uint64_t op) {
2     ...
3     switch (op) {
4         case RULE_S2_8:
5             gtype->qsort_comp = comp_uint8_t;
6             gtype->bsearch = bsearch_uint8_t;
7             gtype->alloc_uniqresult = alloc_uniqresult_uint8_t;
8             gtype->get_uniq_record = get_uniq_record_uint8_t;
9             gtype->dealloc_uniqresult = dealloc_uniqresult_uint8_t;
10            break;
11            case RULE_S2_16: ...
12            case RULE_S2_32: ...
13            case RULE_S2_64: ...
14        }
15        return gtype;
16    }

```

Listing 36: F(v2): Flexible Grouper

*flexible group  
aggregations*

The group aggregation functions were hardcoded in `group_aggr` struct. The functions are now replaced with rules that map to a specific aggregation function. The mapping of the rule to the function is done using a switch case as shown in listing 37. The aggregation functions are auto-generated using `scripts/generate-functions.py`.

```

1 struct grouper_aggr group_aggr_branch1[4] = {
2
3     - { 0, trace_data->offsets.srcaddr, aggr_static_uint32_t },
4     - { 0, trace_data->offsets.dPkts, aggr_sum_uint32_t },
5     + { 0, trace_data->offsets.srcaddr, RULE_STATIC | RULE_S1_32, NULL },
6     + { 0, trace_data->offsets.dPkts, RULE_SUM | RULE_S1_32, NULL },
7 };
8 ...
9
10 /* for loop for the group-aggregation */
11 for (int j = 0; j < binfos[i].num_aggr; j++) {
12     switch (binfos[i].aggr[j].op) {
13         ...
14         case RULE_SUM | RULE_S1_32:
15             binfos[i].aggr[j].func = aggr_sum_uint32_t;
16             break;
17         ...
18     }
19 }

```

Listing 37: F(v2): Flexible Group Aggregations

*flexible group  
aggregation  
redundancy checks*

The aggregation function also needs to know the type of the offsets used previously in the filter and grouper rules to be able to fill in common fields in its cooked netflow v5 group aggregation record. As a result a `get_aggr_fptr(...)` function was defined that accepts those previous rules to fall through a switch to return a function pointer to an aggregation function of the correct type as shown in listing 38. This aggregation function is then later used to fill in the common fields. A similar call is made for the grouper rules as well.

```

1 struct aggr*
2 (*get_aggr_fptr(
3     bool ifgrouper,
4     uint64_t op
5 )) (
6     char** records,
7     char* group_aggregation,
8     size_t num_records,
9     size_t field_offset,
10    bool if_aggr_common
11 ) {
12     ...
13     switch (op) {
14 +     case RULE_EQ | RULE_S1_8:
15 +     case RULE_NE | RULE_S1_8:
16 +     case RULE_GT | RULE_S1_8:
17         ...
18 +     aggr_function = aggr_static_uint8_t;
19 +     break;
20     ...
21     }
22     ...
23     aggr_function = get_aggr_fptr(binfo->filter_rules[i].op);
24     (*aggr_function)(
25         group->members,
26         group_aggregation,
27         group->num_members,
28         field_offset,
29         TRUE
30     );

```

Listing 38: F(v2): Flexible Group Filters

The group filter struct `gfilter_rule` now accepts a `uint_X` enum when mapping functions. The additional enum is used to map to a function that knows the type of the offset at runtime. The additional switch cases and comparison functions are automatically generated using `scripts/generate-functions.py`.

*flexible group filters*

```

1  struct gfilter_rule gfilter_branch1[1] = {
2  - {trace_data->offsets.dPkts, 200, 0, RULE_GT, NULL}
3  + {trace_data->offsets.dPkts, 200, 0, RULE_GT | RULE_S1_32, NULL}
4  };
5
6  switch (binfos[i].gfilter_rules[j].op) {
7  - case RULE_EQ:
8    - binfos[i].gfilter_rules[j].func = gfilter_eq;
9  + case RULE_EQ | RULE_S1_32:
10   + binfos[i].gfilter_rules[j].func = gfilter_eq_uint32_t;
11   + break;
12   ...
13  }

```

Listing 39: F(v2): Flexible Group Filters

Each branch of the pipeline is executed by a separate thread. Since the branches do *not* have a copy of the trace but point to the original records, they *cannot* free the records that failed the filter rule, since they can pass in some other branch. As a consequence, the records that failed in all branches can only be free'd once all threads join `main(...)` ie. before calling the `merger(...)`. The naïve approach to linearly search non-filtered records by falling through filtered recordset of each branch is costly and runs in worst case  $O(nkm)$  time where  $n$  is the number of records in the trace,  $k$  is the number of branches, and  $m$  is the number of filtered records in each branch. Instead, it is better to trade space for lower runtime complexity. Listing 40 shows how the trace structure is extended to allow a flag that stores meta-information about the record. The non-filtered records can now be free'd in worst case  $O(n)$  time.

*greedily deallocating  
non-filtered records*

```

1  struct ft_data {
2  - char**                      records;
3  + struct record**            recordset;
4  + int                        num_records;
5  };
6
7  struct record {
8  + char*                      record;
9  + bool                       if_filtered;
10 };
11 ...
12 + for (int i = 0; i < param_data->trace->num_records; i++) {
13 +   struct record* recordinfo = param_data->trace->recordset[i];
14 +   if (recordinfo->if_filtered == false)
15 +     free(recordinfo->record); recordinfo->record = NULL;
16 + }

```

Listing 40: F(v2): Greedy Deallocation of Non-Filtered Records tabsize

There are dedicated comparator functions for each `uintX_t` type of the field offset. Up until now, the choice for the function was made using a single function, `assign_fptr(...)`, which was called before

the start of the pipeline to ensure all function pointers point to the right functions for each stage as shown in listing 41.

```

1 assign_fp_ptr(struct flowquery *fquery) {
2     for (int i = 0; i < fquery->num_branches; i++) {
3         /* for loop for the filter */
4         for (int j = 0; j < branch->num_filter_rules; j++) {...}
5         /* for loop for the grouper */
6         for (int j = 0; j < branch->num_grouper_rules; j++) {...}
7         /* for loop for the group-aggregation */
8         for (int j = 0; j < branch->num_aggr_rules; j++) {...}
9         /* for loop for the group-filter */
10        for (int j = 0; j < branch->num_gfilter_rules; j++) {...}
11    }
12 }

```

Listing 41: F(v1): Early Comparator Assignments

*lazy comparator  
assignments*

This function is computationally expensive, since it falls through a *huge* switch statement to determine the function of right type. It is not guaranteed that given the type of the query and trace, the program will eventually go through each stage of the pipeline. It is also possible that the program exits before, because there is nothing more for the next stage to compute. The function pointers should therefore be set as late as possible as shown in listing 42. Each of these functions are called from their respective stages just before the comparison. As a result, we save the computation time wasted in setting the function pointer for stage X if X is never executed.

```

1 assign_filter_func(struct filter_rule* const frule) {...}
2 assign_grouper_func(struct grouper_rule* const grule) {...}
3 assign_aggr_func(struct aggr_rule* const arule) {...}
4 assign_gfilter_func(struct gfilter_rule* const gfrule) {...}
5 assign_merger_func(struct merger_rule* const mrule) {...}

```

Listing 42: F(v2): Lazy Comparator Assignments

*early thread exits*

Each branch runs in its own thread. If any of the stages of the branch return a NULL when returning from their public interface function, there is no reason to continue the thread. The subsequent stages of the branch cannot do much with a NULL result. Therefore, the branch thread returns with a EXIT\_FAILURE if either stage returns NULL, and with EXIT\_SUCCESS on normal exit as shown in listing 43.

```

1 void *
2 branch_start(void *arg) {
3     ...
4     branch->filter_result = filter(...);
5     if (branch->filter_result == NULL)
6         pthread_exit((void*)EXIT_FAILURE);
7     branch->grouper_result = grouper(...);
8     if (branch->grouper_result == NULL)
9         pthread_exit((void*)EXIT_FAILURE);
10    branch->gfilter_result = groupfilter(...);
11    if (branch->gfilter_result == NULL)
12        pthread_exit((void*)EXIT_FAILURE);
13    ...
14    pthread_exit((void*)EXIT_SUCCESS);
15 }

```

Listing 43: F(v2): Early Thread Exits



Each stage of the processing pipeline is dependent on the result of the previous one. As a result, the stages should only proceed and process, when the previous returned results. Implementing such a response was straightforward for the grouper and group filter as shown in listing 44, the merger although was a little trickier. The merger stage proceeds only when every branch has non-zero filtered groups. The iterator initializer `iter_init(...)` deallocates and returns NULL if any one branch has 0 filtered groups. Consequently a check is performed in the merger to make sure `iter` is *not* NULL.

*context-aware  
pipeline stages*

```

1  /* grouper */
2  struct grouper_result*
3  grouper(...) {
4
5      /* go ahead if there is something to group */
6      if (fresult->num_filtered_records > 0) {...}
7  }
8
9  /* group filter */
10 struct groupfilter_result*
11 groupfilter(...) {
12
13     /* go ahead if there is something to group filter */
14     for (int i = 0, j = 0; i < gresult->num_groups; i++) {...}
15 }
16
17 /* merger */
18 struct merger_result*
19 merger(...) {
20
21     /* initialize the iterator */
22     struct permut_iter* iter = iter_init(num_branches, branchset);
23     if (iter == NULL)
24         return mresult;
25     ...
26 }

```

Listing 44: F(v2): Context-Aware Pipeline Stages

A rundown of the runtime complexity of each stage of the processing pipeline is shown in table 6. In the table,  $n$  is the total number of flow records in the trace, while  $k$  is the number of unique flow records. The number of branches (or threads) spawned by the execution engine is  $m$ . It is clear that the merger is currently the bottleneck of the pipeline and needs further optimizations.

*runtime complexity*

Pipeline Stage	Runtime Complexity
Filter (worst case)	$O(n)$
Grouper (average case)	$O(n * \lg(n)) + O(n) + O(n * \lg(k))$
Group Aggregation (worst case)	$O(n)$
Merger (worst case)	$O(n^m)$
Ungrouper (worst case)	$O(n)$

Table 6: F(v2): Pipeline Runtime Complexity

## 8.3 MERGER INTERNALS

```

1  get_module_output_stream(module m) {
2      (branch_1, branch_2, ..., branch_n) = get_input_branches(m);
3      for each g_1 in group_records(branch_1)
4          for each g_2 in group_records(branch_2)
5              ...
6              ...
7                  for each g_n in group_records(branch_n)
8                      if match(g_1, g_2, ..., g_n, rules(m))
9                          output.add(g_1, g_2, ..., g_n);
10     return output;
11 }

```

Listing 45: Merger Pseudocode [11]

The merger pseudocode as defined in the [NFQL](#) specification [11] is shown in listing 45. Implementing this pseudocode in C is not straightforward. The level of nesting depends on the number of branches, and is therefore not known at compile time. The information on the number of branches comes from the query which is passed to the execution engine at runtime.

```

1  /* initialize the iterator */
2  struct permut_iter *iter = iter_init(bininfo_set, num_branches);
3
4  /* iterate over all permutations */
5  unsigned int index = 0;
6  while(iter_next(iter)) {
7      index++
8      for (int j = 0; j < num_branches; j++) {
9          /* first item */
10         if(j == 0)
11             printf("\n%d: (%zu ", index, iter->filtered_group_tuple[j]);
12         /* last item */
13         else if(j == num_branches - 1)
14             printf("%zu)", iter->filtered_group_tuple[j]);
15         else
16             printf("%zu ", iter->filtered_group_tuple[j]);
17     }
18 }
19
20 /* free the iterator */
21 iter_destroy(iter);

```

Listing 46: F(v2): Merger Iterator Utility

As a result, an iterator that can provide all possible permutations of  $N$ -tuple (where  $N$  is the number of branches) group record IDs was needed. The result of the iterator can then be used to make a match. The merger stage, therefore begins by initializing this iterator passing it the number of branches, and information about each branch. Then, it loops over to get a new  $N$ -tuple group record IDs on each iteration until the iterator returns false. A sample to print all possible group ID permutation is shown in listing 46, with the output in listing 47

*merger iterator  
utility*

```

1  1: (1 1 1)
2  2: (1 1 2)
3  ...
4  12: (3 2 2)

```

Listing 47: F(v2): Merger Iterator Utility Output

## 8.4 RUNTIME QUERY EVALUATION

The complete query is now read in at *runtime*. The query is supplied as a JSON file. The branchsets and each ruleset of the pipeline is a JSON array. A sample JSON query is shown in listing 48.

```

1 {
2   branchset: [
3     {
4       filter: { ruleset: [...] },
5       grouper: { ruleset: [...] },
6       aggregation: { ruleset: [...] },
7       groupfilter: { ruleset: [...] },
8     },
9     {
10      ...
11    }
12  ],
13  merger: { ruleset: [...] },
14 }

```

Listing 48: F(v2): Flow Query in JSON

json-c is used to parse such a query file read into memory by calling `parse_json_query(...)`. The `json_query` is then used to prepare the struct `flowquery` used by the pipeline stages as shown in listing 52. The `json_query` struct is just an intermediate and shouldn't be needed. Ideally, `parse_json_query(...)` can directly fill in and create the `flowquery` struct and is a future refactor item.

*parsing using json-c*

```

1 struct json {
2   size_t          num_branches;
3   size_t          num_mrules;
4
5   struct json_branch_rules** branchset;
6   struct json_merger_rule**  mruleset;
7 };
8
9 struct json_branch_rules {
10  size_t          num_frules;
11  size_t          num_grules;
12  size_t          num_arules;
13  size_t          num_gfrules;
14
15  struct json_filter_rule** fruleset;
16  struct json_grouper_rule** gruleset;
17  struct json_aggr_rule**  aruleset;
18  struct json_gfilter_rule** gfruleset;
19 };
20
21 struct json*
22 json_query = parse_json_query(param_data->query_mmap);
23
24 struct flowquery*
25 fquery = prepare_flowquery(param_data->trace, json_query);

```

Listing 49: F(v2): Parsing JSON query using json-c

The JSON query is verbose and cumbersome to write manually. The python parser will eventually emit this intermediate format, so the next logical step is to generate the query from python. A python module (`scripts/queries/pipeline.py`) that encapsulates each pipeline stage as a separate class is shown in listing 50. Scripts that generate JSON queries can import this module to reduce code redundancy.

*generating json  
queries using python*

```

1 def protocol(name):
2     return socket.getprotobyname(name)
3
4 class FilterRule: ...
5 class GrouperRule: ...
6 class AggregationRule: ...
7 class GroupFilterRule: ...
8 class MergerRule: ...

```

Listing 50: F(v2): Python Pipeline Module

*sample scripts*

A sample script to generate such a query is shown in listing 51. Each ruleset is a list of python objects of a specific class of the pipeline module. At this point, the python parser just needs to create each stage rule objects and the script will take care to emit the JSON. Example scripts to generate different queries are provided in `scripts/queries/`.

```

1 import json
2 from pipeline import FilterRule, GrouperRule, AggregationRule
3 from pipeline import GroupFilterRule, MergerRule
4 from pipeline import protocol
5
6 if __name__ == '__main__':
7
8     fruleset = []
9     fruleset.append(vars(FilterRule(...))) ...
10    filter = {'ruleset': fruleset}
11
12    gruleset = []
13    gruleset.append(vars(GrouperRule(...))) ...
14    grouper = {'ruleset': gruleset}
15
16    aruleset = []
17    aruleset.append(vars(AggregationRule(...))) ...
18    a = {'ruleset': aruleset}
19
20    gfruleset = []
21    gfruleset.append(vars(GroupFilterRule(...))) ...
22    gfilter = {'ruleset': gfruleset}
23
24    branchset = []
25    branchset.append({'filter': filter,
26                      'grouper': grouper,
27                      'aggregation': a,
28                      'gfilter': gfilter,
29                      })
30
31    mruleset = []
32    mruleset.append(vars(MergerRule(...))) ...
33    merger = {'ruleset': mruleset}
34    query = {'branchset': branchset, 'merger': merger}
35    fjson = json.dumps(query, indent=2)

```

Listing 51: F(v2): Python Scripts to Generate JSON queries

*runtime query  
internals*

The mapping of the JSON query to the structs defined in the execution engine is trickier than it looks. When reading the JSON query at runtime, the field offsets of the NetFlow v5 record struct are read in as char pointers. A utility function `get_offset(...)` was thus introduced that maps the read names to struct offsets. In addition, the type of each offset and the operations are also read in as char pointers. This information is saved and thus used by the engine using an enum defined in `pipeline.h`. Therefore, another utility function `get_enum(...)` was defined to map this information to the unique enum members.

```

1 size_t
2 get_offset(
3     const char * const name,
4     const struct fts3rec_offsets* const offsets
5 ) {
6
7     #define CASEOFF(memb) \
8     if (strcmp(name, #memb) == 0) \
9         return offsets->memb
10
11     CASEOFF(unix_secs);
12     CASEOFF(unix_nsecs);
13     ...
14
15     return -1;
16 }
17
18 uint64_t
19 get_enum(const char * const name) {
20
21     #define CASEENUM(memb) \
22     if (strcmp(name, #memb) == 0) \
23         return memb
24
25     CASEENUM(RULE_S1_8);
26     CASEENUM(RULE_S1_16);
27     ...
28     CASEENUM(RULE_S2_8);
29     CASEENUM(RULE_S2_16);
30     ...
31     CASEENUM(RULE_ABS);
32     CASEENUM(RULE_REL);
33     CASEENUM(RULE_NO);
34     ...
35     CASEENUM(RULE_EQ);
36     CASEENUM(RULE_NE);
37     ...
38     CASEENUM(RULE_STATIC);
39     CASEENUM(RULE_COUNT);
40     ...
41     CASEENUM(RULE_ALLEN_BF);
42     CASEENUM(RULE_ALLEN_AF);
43     ...
44     return -1;
45 }

```

Listing 52: F(v2): JSON Parsing Utilities

## 8.5 AUTOMATED BUILDS

A considerate amount of attention was paid to make sure the execution engine utilize only standard ANSI libraries to allow it to seamlessly build across Unix platforms. Since the engine depends on the flow-tools library that uses BSD extensions, it proved useful to include the GNU\_SOURCE feature test macro. GNU\_SOURCE allows to request and let the compiler enable a larger class of features.

*feature test macros*

CMake was used to ensure a compiler and platform independent build process <sup>1</sup>. Since the execution engine requires some headers/sources that are auto-generated by a python script, a custom command was added to run the script on each compilation to add the generated files in .build/ as shown in listing 53. These files are automatically included during the compilation and linked to the final binary. CMake also runs the build query scripts defined in scripts/queries/ to generate some example JSON queries and moves them to the examples/ folder ready to be used by the binary as shown in listing 53.

*cmake custom  
commands*

<sup>1</sup> detailed engine installation instructions are available in the appendix.

```

1  # custom command to prepare auto-generated sources
2  add_custom_command (
3      OUTPUT ${CMAKE_CURRENT_BINARY_DIR}/auto-assign.h
4              ${CMAKE_CURRENT_BINARY_DIR}/auto-assign.c
5              ${CMAKE_CURRENT_BINARY_DIR}/auto-comps.h
6              ${CMAKE_CURRENT_BINARY_DIR}/auto-comps.c
7      COMMAND python ${CMAKE_SOURCE_DIR}/scripts/generate-functions.py
8      COMMENT "Generating: auto-comps{h,c} and auto-assign.{h,c}"
9  )
10
11 # custom command to generate examples
12 file(GLOB pyFILES ${CMAKE_SOURCE_DIR}/scripts/queries/*.py)
13 foreach(pyFILE ${pyFILES})
14     set(query "${pyFILE}_query")
15     add_custom_command (
16         OUTPUT ${query}
17         WORKING_DIRECTORY ${CMAKE_SOURCE_DIR}/examples/
18         COMMAND python ${pyFILE}
19         COMMENT "Generating: JSON example query using ${pyFILE}"
20     )
21     list(APPEND queryFILES ${query})
22 endforeach(pyFILE)

```

Listing 53: F(v2): CMake Custom Commands

CMake build process requires one to invoke quite a number of bash commands as shown in listing 54. In essence, a user does not need to know the CMake *way* to working around the build to use the execution engine. As such a Makefile is included that can make CMake calls to automate this operation. Additional targets to clean and generate doxygen documentation. The generated documentation goes in doc/ and is subsequently deleted by a cleanup.

*makefile to automate  
cmake*

```

1  [engine] $ mkdir .build
2  [engine] $ cd .build
3  [.build] $ cmake ..
4  [.build] $ make
5  [.build] $ cd ..
6  [engine] $ rm -r .build
7
8  [engine] $ cat Makefile
9
10 make: build/Makefile
11      (cd .build; make)
12 build/Makefile: build
13      (cd .build; cmake -D CMAKE_PREFIX_PATH=$(CMAKE_PREFIX_PATH) ..)
14 build:
15      mkdir -p .build
16 doc: Doxyfile
17      (mkdir -p doc; doxygen Doxyfile)
18 clean:
19      rm -f -r .build/ bin/ doc/
20      rm -f -r examples/*.json

```

Listing 54: F(v2): Automating CMake Invocations

The Makefile can also take CMAKE\_PREFIX\_PATH as an argument and pass it on to CMake. CMAKE\_PREFIX\_PATH is used to supply arbitrary location of external libraries and include PATH. This can be useful since the flow-tools installation from source dispatches the library and headers in /usr/local/flow-tools.

*cmake prefix path*

```

1  [engine] $ make CMAKE_PREFIX_PATH=/usr/local/flow-tools

```

Listing 55: F(v2): CMake Prefix Paths

There has never been a clean seamless way to install python flowy. Since the parser in the flowy implementation is eventually going to converge with the new execution engine, it is essential to provide an easy way to install and manage the parser. The software tool used in the python ecosystem to manage packages is pip. It uses a *flat* requirements.txt file to install all the package dependencies in one go. However, it requires that all the (to be) installed dependencies do not import external packages in their egg files. This turned to be the case for numexpr which is required by the parser, thereby resulting in failed installation. To circumvent the issue, a custom Makefile<sup>2</sup> was created that virtually adds a preprocessing pass to install numexpr dependencies before going forward with installation from requirements.txt as shown in listing 56.

*packaging the parser*

```

1 make: numexpr
2     (pip install -r requirements.txt)
3 numexpr: numpy
4     (pip install numexpr==2.0.1)
5 numpy: cython
6     (pip install numpy==1.6.1)
7 cython:
8     (pip install Cython==0.15.1)
9 clean:
10     rm -f -r build/
11     rm -f -r src/*.pyc
12     rm -f -r flowy-run/
13     rm -f -r parsetab.py parser.out
14     rm -f -r examples/output.h5

```

Listing 56: F(v2): Automating Parser Installation

## 8.6 REGRESSION TEST SUITE

A regression test-suite has been added in tests/. The suite asserts the numbers of results in each stage for a query-trace combination. It also looks for any segmentation faults if they may have occurred. Tests can be run either individually or as a complete suite as shown in listing 57. The suite can also run in a verbose mode to see the expected and achieved result combination for run each test case.

```

1 [engine] $ make
2 [engine] $ tests/regression.py [-v]
3 .....
4 -----
5 Ran 60 tests in 32.533s
6
7 OK
8
9 [engine] $ tests/test-query-http-tcp-session.py [-v]
10 .....
11 -----
12 Ran 10 tests in 8.672s
13
14 OK

```

Listing 57: F(v2): Regression Test Suite

<sup>2</sup> detailed parser installation instructions are available in the appendix





## PERFORMANCE EVALUATION

### 9.1 EXECUTION ENGINE PROFILING

The F(v1) execution engine had chunks of memory leaks. The blocks of heap memory were still reachable when the engine exited. As such, it was essential to profile the engine to properly deallocate all blocks before exit. Listing 58 shows the valgrind profile output of both versions. The 20kB of created and still living blocks in the current snapshot are due two libraries. The dyld library makes 81 malloc invocations that are not free'd by the library as shown in figure 21. On GNU/Linux, dyld is replaced by dlopen which does not have this issue. The other set of libraries, libsystem\_c, libsystem\_notify, libdispatch make 10 malloc invocations that are again not free'd as shown in figure 21. These malloc calls invoke localtime(...) which uses tzset(...) to initialize and return struct tm\*. This structure is never free'd apparently due to a bug in these libraries.



Figure 21: F(v2): Backtrace of Living on Exit Blocks

```

1 $ git checkout v0.1; make
2 $ valgrind bin/engine queryfile tracefile
3
4 ==19000== LEAK SUMMARY:
5 ==19000==    definitely lost: 6,912 bytes in 472 blocks
6 ==19000==    indirectly lost: 0 bytes in 0 blocks
7 ==19000==    possibly lost: 0 bytes in 0 blocks
8 ==19000==    still reachable: 124,607 bytes in 710 blocks
9 ==19000==    suppressed: 0 bytes in 0 blocks
10
11 $ git checkout master; make
12 $ valgrind bin/engine queryfile tracefile
13
14 ==19164== LEAK SUMMARY:
15 ==19164==    definitely lost: 0 bytes in 0 blocks
16 ==19164==    indirectly lost: 0 bytes in 0 blocks
17 ==19164==    possibly lost: 0 bytes in 0 blocks
18 ==19164==    still reachable: 20,228 bytes in 37 blocks
19 ==19164==    suppressed: 0 bytes in 0 blocks

```

Listing 58: F(v2): Valgrind-based Engine Profiling

## 9.2 BENCHMARKING SUITE

```

1 [engine] $ make; sudo benchmarks/nfql.py bin/engine trace[s]/ querie[s]/
2 benchmarking nfql ...
3 executing: [engine tcp-session trace-2012]:  1 2 3 4 5 6 7 8 9 10 (3.315148 secs)
4 executing: [engine tcp-session trace-2009]:  1 2 3 4 5 6 7 8 9 10 (0.034624 secs)
5 ...
6
7 [engine] $ sudo benchmarks/silk.py trace[s]/ querie[s]/
8 benchmarking silk ...
9 executing: [silk http-tcp-session trace-2009]:  1 2 3 4 5 6 7 8 9 10 (0.102465 secs)
10 executing: [silk http-tcp-session trace-2012]:  1 2 3 4 5 6 7 8 9 10 (0.279106 secs)
11 ...

```

Listing 59: F(v2): Automated Benchmarking

To be able to run and reproduce the benchmarking results as and when required it was essential to automate the whole process. The target design was to be able to use one script to run all sets of query-trace combination in one go for each network analysis application as shown in listing 59. The directories containing the traces and the queries required by the script can be supplied as command line arguments. Few examples are provided in `examples/`. The benchmarking suite only runs on python 2.7 and above. Each query-trace combination is run 10 times and the timings are averaged and echoed on the screen. Attention is given to clear page caches, dentries and inodes before each iteration invocation as shown in listing 60. The script, therefore needs to run with sudo privileges. The results are saved in `benchmarks/results/`. SiLK query files are simply bash commands separated by a delimiter and are further discussed in the next section.

```

1 ...
2 # clear pagecache, dentries and inodes
3 os.system('sync')
4 try:
5     with open('/proc/sys/vm/drop_caches', 'w') as stream:
6         stream.write('3\n')
7 except IOError as e:
8     print 'run the script with sudo privileges'
9     exit(1)
10 ...

```

Listing 60: F(v2) Benchmarking: Clearing Kernel Caches

## 9.3 RELATIVE COMPARISON WITH SILK

The benchmarking suite was used to run a number of queries over a varied size of flow traces. The smallest trace was 500K flow records, going upto 10M flow records with a step size of 500K. This accounts for a week worth of flow data in a university environment. The suite was run on a high-end machine with 24 cores of 2.5 GHz clock speed and 18GiB of memory. The results for NFQL and SiLK <sup>1</sup> [74] are available on the benchmarks branch of the git repository.

<sup>1</sup> detailed silk installation and usage instructions are available in the appendix.

The first query attempts to find all the DNS (over UDP) traffic of the supplied trace. One of the branch, filters and groups the flow records that contain DNS requests, while the other holds the DNS responses. The request-response pairs are then merged according to the endpoints and unfolded to display the flow-record streams in each such pair. Such a query in SiLK<sup>2</sup> can be represented by piping multiple tools together as shown in listing 61. The subsequent NFQL query is available at [examples/](#) and is omitted here for brevity reasons. The query is similar for mDNS whereby the port 53 is replaced by 5353. The results are shown in figure 22.

*dns and mdns  
request-response  
over udp*

```

1 $ cat dns-udp.txt
2 rm -f /tmp/A.raw /tmp/B.raw /tmp/result.raw; \
3 rfilter --sport=53 --proto=17 --pass=stdout %s | \
4 rwsort --fields=sIP,dIP | \
5 rwgroup --id-fields=sIP,dIP --summarize | \
6 rfilter --input-pipe=stdin --pass=/tmp/A.raw --packets=10-; \
7 rfilter --dport=53 --proto=6 --pass=stdout %s | \
8 rwsort --fields=sIP,dIP | \
9 rwgroup --id-fields=sIP,dIP --summarize | \
10 rfilter --input-pipe=stdin --pass=/tmp/B.raw --packets=10-; \
11 rwmatch --relate=1,2 --relate=2,1 /tmp/A.raw /tmp/B.raw /tmp/result.raw;
12
13 $ cat mdns-udp.txt
14 rm -f /tmp/result.raw; \
15 rfilter --sport=5353 --dport=5353 --proto=17 --pass=stdout %s | \
16 rwsort --fields=sIP,dIP | \
17 rwgroup --id-fields=sIP,dIP --summarize | \
18 rfilter --input-pipe=stdin --pass=/tmp/result.raw --packets=500-; \

```

Listing 61: SiLK Query: DNS and mDNS over UDP

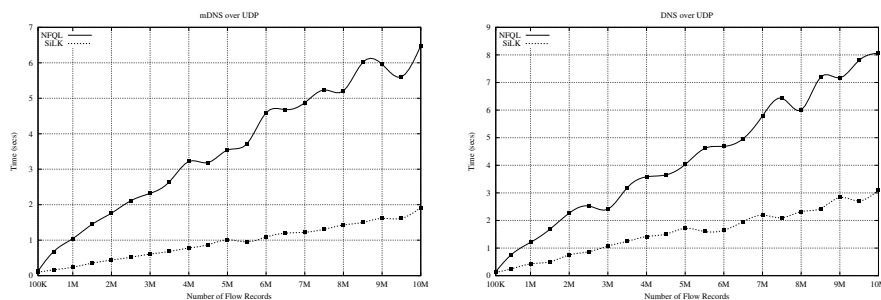


Figure 22: F(v2): DNS and mDNS over UDP

The second query searches for all the HTTP and HTTPs sessions over TCP on a supplied trace. There are separate branches for requests and responses. Filters in each such branch search for all HTTP (or HTTPs) traffic over port 80 (or 443). The groupers group all the flow-records with respect to the endpoints, ie. source-destination IP pairs. The group filters filter the groups to keep only those that generate at-least 200 packets, while the merger relates the request-response pairs to create a session. Each such session is then unfolded to display the stream of flow-records in each session. SiLK queries that tend to emulate the behavior of finding HTTP (and HTTPs) sessions is shown in listing 62. The comparison results are shown in figure 23.

*http and https  
sessions over tcp*

<sup>2</sup> detailed instructions to convert flow-tools traces to silk are available in the appendix

```

1 $ cat http-tcp-session.txt
2 rm -f /tmp/A.raw /tmp/B.raw /tmp/result.raw; \
3 rwsort --sport=80 --proto=6 --pass=stdout %s | \
4 rwsort --fields=sIP,dIP | \
5 rwgroup --id-fields=sIP,dIP --summarize | \
6 rwfilter --input-pipe=stdin --pass=/tmp/A.raw --packets=200-; \
7 rwfilter --dport=80 --proto=6 --pass=stdout %s | \
8 rwsort --fields=sIP,dIP | \
9 rwgroup --id-fields=sIP,dIP --summarize | \
10 rwfilter --input-pipe=stdin --pass=/tmp/B.raw --packets=200-; \
11 rwmatch --relate=1,2 --relate=2,1 /tmp/A.raw /tmp/B.raw /tmp/result.raw;
12
13 $ cat https-tcp-session.txt
14 rm -f /tmp/A.raw /tmp/B.raw /tmp/result.raw; \
15 rwsort --sport=443 --proto=6 --pass=stdout %s | \
16 rwsort --fields=sIP,dIP | \
17 rwgroup --id-fields=sIP,dIP --summarize | \
18 rwfilter --input-pipe=stdin --pass=/tmp/A.raw --packets=200-; \
19 rwfilter --dport=443 --proto=6 --pass=stdout %s | \
20 rwsort --fields=sIP,dIP | \
21 rwgroup --id-fields=sIP,dIP --summarize | \
22 rwfilter --input-pipe=stdin --pass=/tmp/B.raw --packets=200-; \
23 rwmatch --relate=1,2 --relate=2,1 /tmp/A.raw /tmp/B.raw /tmp/result.raw;

```

Listing 62: SiLK Query: HTTP and HTTPs Sessions over TCP

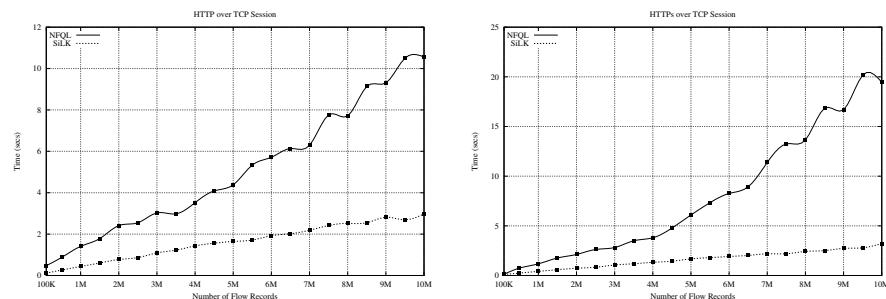


Figure 23: F(v2): HTTP and HTTPs Sessions over TCP

The third query attempts to collect an overall statistics of the amount of HTTP traffic in the supplied flow trace. The two branches represent HTTP requests and responses. The filters in each such branch, filters in all HTTP traffic over port 80. The result from both the branches are concatenated together and a sum of all the packets and octets is generated as a summary. In NFQL such a query is defined by omitting the grouper module and aggregating all the traffic generated by the filters. The merger and ungrouper modules are also omitted. The SiLK query is shown in listing 63 and the results are shown in figure 24.

*conglomerating http  
traffic*

```

1 $ cat http-octets.txt
2 rm -f /tmp/A.raw /tmp/B.raw /tmp/result.raw /tmp/result.txt; \
3 rwfilter --sport=80 --pass=stdout %s | \
4 rwfilter --input-pipe=stdin --pass=/tmp/A.raw --packets=0-; \
5 rwfilter --dport=80 --pass=stdout %s | \
6 rwfilter --input-pipe=stdin --pass=/tmp/B.raw --packets=0-; \
7 rwcats --output=/tmp/result.raw /tmp/A.raw /tmp/B.raw; \
8 rwmstats --overall-stats --output-path=/tmp/result.txt /tmp/result.raw

```

Listing 63: SiLK Query: Conglomerating HTTP Traffic

The fourth query searches for all application-protocol agnostic TCP sessions in the supplied trace. The two branches represent TCP requests and responses. The filters in each branch filter in all the TCP

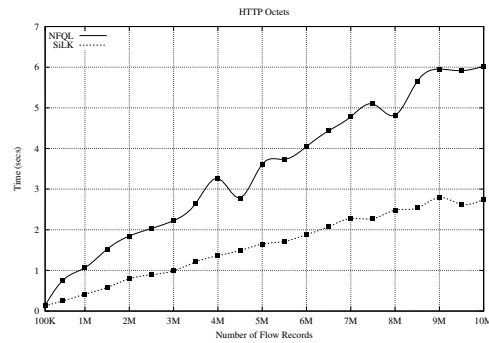


Figure 24: F(v2): Conglomerating HTTP Traffic

traffic, while the grouper clubs the flow-records coming to (or from) a specific end-point into one group record. The group filter, filters in all groups that have at-least 20K packets to reduce the merger load. The merger eventually relates the branches together to form request-response pairs. The ungrouper unfolds these pairs to print the stream of flow-records members in each pair. The SiLK query is shown in listing 64 and the NFQL comparison results are shown in figure 25.

*application agnostic  
tcp sessions*

```

1 $ cat tcp-session.txt
2 rm -f /tmp/A.raw /tmp/B.raw /tmp/result.raw; \
3 rfilter --proto=6 --pass=stdout %s | \
4 rwsort --fields=sIP,dIP | \
5 rwgroup --id-fields=sIP,dIP --summarize | \
6 rfilter --input-pipe=stdin --pass=/tmp/A.raw --packets=20000-; \
7 rfilter --proto=6 --pass=stdout %s | \
8 rwsort --fields=sIP,dIP | \
9 rwgroup --id-fields=sIP,dIP --summarize | \
10 rfilter --input-pipe=stdin --pass=/tmp/B.raw --packets=20000-; \
11 rmatch --relate=1,2 --relate=2,1 /tmp/A.raw /tmp/B.raw /tmp/result.raw;

```

Listing 64: SiLK Query: Application Agnostic TCP Session

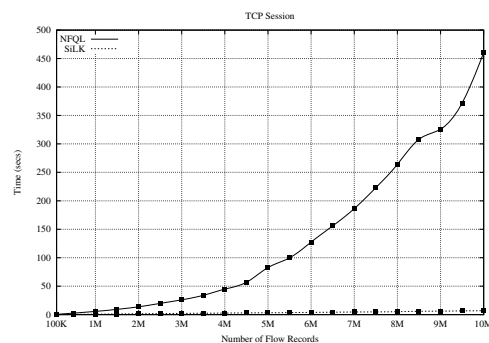


Figure 25: F(v2): Application Agnostic TCP Sessions

It is clear that the merger is the bottleneck of the processing pipeline. The near exponential growth is due the  $O(n^m)$  complexity of the merger. It appears that using a hash table for an equality relative operation and search trees for generic operators will go a long way in improving the merger speed and is a future work item.



## FUTURE WORK AND CONCLUSION

---

The C execution engine is a major leap forward to a fast and robust implementation of [NFQL](#). It is the first time when a [NFQL](#) implementation can actually be used on practically sized flow traces. However it is far from being deemed complete. The fast-paced development is clear from the fact that engine's issue tracker <sup>1</sup> has a 1 : 1 ratio of closed to open issues. As a result, the future outlook of the engine is divided into major goals and minor issues that still need to be addressed.

### 10.1 MAJOR GOALS

The execution engine uses the flow-tools [API](#) to read and parse the flow-records. The flow-tools [API](#) can only parse NetFlow v5 records, thereby sandboxing the engine's functionality. It restricts the engine's understanding of a flow to a fixed NetFlow v5 format and inhibits the capability to parse IPv6 flows. Since NetFlow v9 and now with [IPFIX](#), the flow format is dynamically parsed using its accompanied template. It relaxes the definition of a flow and gives more power to the sender on how the data should be conglomerated together. There are three well-known implementations of [IPFIX](#): CERT' [libfixbuf](#) <sup>2</sup>, Fraunhofer FOKUS' [libipfix](#) <sup>3</sup> and WAND' [Maji](#) <sup>4</sup>. The former two [APIs](#) appear to be under heavy development and one of them is foreseen to be used by the engine to provide [IPFIX](#) capability in the future.

*ipfix support*

The python parser is currently unused. The idea at this stage is to allow the parser to parse and validate the flow query and generate an equivalent intermediate JSON format. This is supposed to be a preprocessing step performed offline. The JSON flowquery file can then later be supplied to the execution engine at runtime. It can also be foreseen to push parts of the JSON flowquery using a RESTful interface to multiple map/reduce jobs running the execution engine to completely distribute the workflow for faster processing. The first step to achieve such a convergence will be by pruning the processing pipeline out of the python implementation. The reverse engineered package and class diagrams generated using [pyreverse](#) are available in [parser/docs/](#) and will help one get started with this task. The installation and usage instructions available in the appendix will make the convergence head start a breeze.

*parser and engine  
convergence*

---

<sup>1</sup> <https://github.com/vbajpai/mthesis-src/issues>

<sup>2</sup> <http://tools.netsa.cert.org/fixbuf/>

<sup>3</sup> <http://libipfix.git.sourceforge.net/>

<sup>4</sup> <http://research.wand.net.nz/software/maji.php>

*search tree and hash  
table lookups*

The runtime complexity is deemed to increase exponentially as the number of records, and the modules in the grouper and merger increase. The current grouper implementation runs in  $O(n * \lg(n)) + O(n) + O(n * \lg(k))$ , while the merger runs in  $O(n^m)$  time, where  $n$  is the number of flow records,  $m$  is the number of branches, and  $k$  is the number of unique flow records that passed the filter stage. Therefore, a search tree lookup would help bring the runtime costs down, whereby one of the fields will be traversed sequentially in  $O(n)$  time and for each field, a comparison will be performed by search tree lookups in  $O(\log(n))$  time bringing down the overall complexity to  $O(n \log(n))$ . In addition, letting the execution engine override the search tree lookups by hash table lookups for equality operators will further bring down the runtime to  $O(n)$  for this specific case.

*efficient  
multithreading*

The execution engine currently has limited multithreading. Each branch in the pipeline runs on a separate thread. However, this implies that the merger and ungrouper stages still remain single-threaded. The merger stage in the processing pipeline is currently the most computationally intensive operation with an exponential runtime complexity of  $O(n^m)$  for  $n$  flow records and  $m$  branches. It is possible to handle the merger's outermost branch loop using multiple threads in a non-blocking fashion to improve performance. The situation can further be improved by writing a pthreads wrapper that auto detects the number of available cores, creates a appropriate size thread pool and equally divides the tasks among the threads. This would also lead to an increased complexity of managing mutual-exclusion of shared memory, but the performance gains will go a long way.

## 10.2 MINOR ISSUES

```

1  merger M {
2    module m1 {
3      branches a, b
4      a.srcIP = b.dstIP
5      a.dstIP = b.srcIP
6    }
7    module m2 {
8      branches a, b
9      a.srcIP = b.srcIP
10     a.dstIP = b.dstIP
11   }
12 }
```

Listing 65: [NFQL](#): Multiple Modules

*multiple modules  
and or expressions*

Each module of the pipeline in an [NFQL](#) query is a Conjunctive Normal Form ([CNF](#)) expression. All the rules of a submodule are AND'd together while all the submodules themselves are OR'd. The engine currently reads all the rules of a stage and applies an AND operation. It cannot handle submodules in a module that can be OR'd together as shown in listing [65](#). An atomic rule in a query, for instance in a filter module also cannot have an OR expression as shown in listing [66](#)



```

1 filter http {
2     srcPort = 80 OR dstPort = 80
3 }

```

Listing 66: NFQL: OR Expressions in Atomic Rules

The START and END timestamps of the cooked NetFlow v5 group records are currently unset. These timestamps should be superset of their member flow-records. This will allow the group filter to skip checking the timestamps of each member of the group and only focus on the superset interval of the group record itself when performing allen interval operations.

*superset intervals in  
group records*

```

1 struct json*
2 json_query = parse_json_query(param_data->query_mmap);
3
4 struct flowquery*
5 fquery = prepare_flowquery(param_data->trace, json_query);

```

Listing 67: F(v2): Redundant Structs

The execution engine currently uses an additional data structure (struct json) to hold the parsed JSON query. This data structure is then read by prepare\_flowquery(...) to generate struct flowquery which is eventually used by the pipeline stages as shown in listing 67. In essence, the intermediate struct is not needed, and is a redundant datastructure. There is no reason why parse\_json\_query(...) cannot directly read the query elements into struct flowquery and is a future refactor item.

*eliminating  
redundant structs*

The ungrouper reads the tuples of merged group records and unfolds each one of them to create a stream. Each such stream is a collection of flow-records that went pass the whole pipeline. However, it is possible that a flow record is part of multiple groups in a single group tuple, and is therefore outputted multiple times. The engine currently does not eliminate such flow records repetitions. It is also does not order the flow records according to their timestamps as defined in the NFQL specification.

*eliminating  
redundant flows  
from a stream*

### 10.3 CONCLUSION

The NFQL execution engine has come a long way in a short time. It now consists of a robust implementation of the processing pipeline that adapts itself to the kind of query provided at runtime to dynamically decide the type of data and the type of operation to be performed. It is flexible to be able to read and parse an entire flowquery at runtime. It is fast to be able to process millions of flow traces in matter of seconds. It is portable and can seamlessly build on multiple Unix flavors and is verifiable using a regression test-suite that will allow future developers to work further to improve the engine with confidence.



## Part IV

### APPENDIX

The appendix is used to supplement the work done in the thesis with clear instructions on how to use the end-product. It also includes instructions on how to build and use other products that were used as reference points when conducting performance evaluation to allow the process to be easily repeatable by others. The organization of the appendix is described below.

In section [A](#) we discuss step-by-step instructions on how to install [NFQL](#) parser and the execution engine on Debian-based systems and OS X. This walkthrough is going to help not only the users, but also future developers to quickly get started with [NFQL](#) implementations.

In section [B](#) we discuss SiLK installation on Debian-based systems. We then investigate which SiLK tool can be used to achieve the desired effect of each stage of the [NFQL](#) processing pipeline. Additional SiLK analysis and capture tools are also discussed. We end the section, by enumerating the process of converting `flow-tools` compatible traces to SiLK proprietary format.

In section [C](#) we enlist the major iterations of the development lifecycle of the execution engine starting from F(v2.0) to F(v2.5). An explanation of the feature set in each iteration is followed by instructions on how `git` tags can be used to go back to a previous version. The appendix is concluded by a list of acronyms used in this work.



## NFQL INSTALLATION AND USAGE

The [NFQL](#) Engine uses CMake to prepare the build. As a result, the installation process is quite seamless. The external dependencies that need to be installed are shown in listing [68](#). The build and usage instructions are also listed. The build environment was tested on 64-bit machines running Debian Squeeze, Ubuntu 10.04 and 12.04.

*nfql engine on  
debian/ubuntu*

```
1 $ sudo apt-get install cmake
2 $ sudo apt-get install flow-tools-dev
3 $ sudo apt-get install zlib1g-dev
4 $ sudo apt-get install libjson0-dev
5 $ sudo apt-get install doxygen
6 $ sudo apt-get install graphviz
7
8 [engine] $ make
9 [engine] $ bin/engine querfile tracefile
10 [engine] $ make doc
11 [engine] $ make clean
```

Listing 68: NFQL Engine on Debian/Ubuntu

The [NFQL](#) parser uses the pip packaging and installation environment to set itself up. Besides some of the external dependencies required by the parser, it is highly recommended to use `virtualenv` and `virtualenvwrapper` to create a virtual environment where all the python libraries will be installed. This helps isolate the user's system-level python libraries and avoids any conflicts that may occur otherwise. The external dependencies and pip environment installation is shown in listing [69](#).

*nfql parser on  
debian/ubuntu*

```
1 $ sudo apt-get install libhdf5-serial-dev
2 $ sudo apt-get install liblz2-dev
3
4 $ sudo apt-get install python-pip
5 $ sudo pip install pip --upgrade
6 $ sudo pip install virtualenv
7 $ sudo pip install virtualenvwrapper
```

Listing 69: NFQL Parser Dependencies on Debian/Ubuntu

A python virtual environment is setup from the parent directory of the parser by running `mkvirtualenv`. Once within the virtual environment, `make` takes care of installing all the python libraries required by the parser in one go. It installs `cython`, `numpy`, `numexpr` in a preprocessing step, and then lets `pip` handle the rest of the installation using a requirements file. The list of python libraries installed can be checked by running `pip freeze`. The parser usage instructions are provided in the next section. The virtual environment can be deactivated using `deactivate` and brought back again using `workon $NAME`. It can be destroyed using `rmvirtualenv` as shown in listing [70](#)

```

1 [parser] $ mkvirtualenv parser
2 (parser)
3 [parser] $ make
4 (parser)
5 [parser] $ pip freeze
6
7 (parser)
8 [parser] $ python ft2hdf.py traces/ output.h5
9 (parser)
10 [parser] $ python printhdf.py output.h5
11 (parser)
12 [parser] $ python print_hdf_in_step.py output.h5
13 (parser)
14 [parser] $ python flowy.py queryfile
15
16 (parser)
17 [parser] $ make clean
18 (parser)
19 [parser] $ deactivate
20
21 [parser] $ rmvirtualenv parser

```

Listing 70: Platform-Agnostic NFQL Parser Build

*nfql engine on osx*

The [NFQL](#) engine installation on OS X is quite similar. The external dependencies can either be installed from source or using a package manager. It is recommended to use Homebrew <sup>1</sup> to install the external dependencies since it does not require and plague the install with sudo privileges and installs the packages at /usr/local/. The external dependencies are shown in listing 71. The build and usage instructions are also listed. The build environment was tested on OS X Lion 10.7.

```

1 $ brew install cmake
2 $ brew install json-c
3 $ brew install doxygen
4 $ brew install graphviz
5
6 $ wget http://dl.dropbox.com/u/500389/flow-tools-0.68.4.tar.bz2
7 $ tar -xvf flow-tools-0.68.4.tar.bz2
8
9 [flow-tools-0.68.4] $ ./configure
10 [flow-tools-0.68.4] $ make
11 [flow-tools-0.68.4] $ make install
12
13 [engine] $ make CMAKE_PREFIX_PATH=/usr/local/flow-tools/
14 [engine] $ make doc
15
16 [engine] $ bin/engine queryfile tracefile
17 [engine] $ make clean

```

Listing 71: NFQL Engine on OS X

*flow-tools issues on  
osx*

There is a ruby gem for flow-tools in the Homebrew package repository, however currently it fails to build. The v0.68 stable build at [splintered.net](http://splintered.net) <sup>2</sup> has a bug <sup>3</sup> that inhibits it from correctly parsing the trace timestamps on 64-bit machines. A forked flow-tools branch <sup>4</sup> resolves the 64-bit issues but then the latest stable release v0.68.5.1 does not build successfully. The previous release of the forked branch v0.68.4 works well and is used by the execution engine on OS X.

- 1 <http://mxcl.github.com/homebrew/>
- 2 <http://www.splintered.net/sw/flow-tools/>
- 3 <http://ensight.eos.nasa.gov/FlowViewer/faq.html#14>
- 4 <https://code.google.com/p/flow-tools/>

```
1 $ brew install hdf5
2 $ brew install lzo
3
4 $ brew install python --framework
5 $ export PATH=/usr/local/share/python:$PATH
6
7 $ easy_install pip
8 $ pip install pip --upgrade
9 $ pip install virtualenv
10 $ pip install virtualenvwrapper
11 $ source /usr/local/bin/virtualenvwrapper.sh
```

Listing 72: NFQL Parser Dependencies on OS X

The [NFQL](#) parser installation on OS X only differs in the *way* how external libraries are installed. Homebrew is used to install the external packages. The default python framework on OS X is fairly old, it is in the best interest to also upgrade it using Homebrew. The older python packaging environment `easy_install` is used to install `pip`, and then `pip` is used to upgrade itself to the latest revision. The packaging from there is on handled by the `pip` environment. The external dependencies and `pip` environment installation is shown in [listing 72](#), while the virtual environment setup and parser build process is exactly the same as shown in [listing 70](#).

*nfql parser on osx*





## SILK INSTALLATION AND USAGE

SiLK <sup>1</sup> is a network traffic collection and analysis tool developed and maintained by the CERT Network Situational Awareness Team (CERT NetSA) at Carnegie Mellon University. Since SiLK is the only tool that comes even remotely closer to the functionality offered by the [NFQL](#), it was used as a reference point to compare the performance of the [NFQL](#) execution engine. This section illustrates the instructions on how to install and use SiLK. The installation was tried on Debian Squeeze and is pretty straight forward as shown in listing 73.

```

1 $ wget http://tools.netsa.cert.org/releases/silk-2.4.7.tar.gz
2 $ sha1sum silk-2.4.7.tar.gz | grep 2ff0cd1d00de70f667728830aa3e920292e99aec
3
4 $ ./configure
5 $ make
6 $ sudo make install
7 $ sudo ldconfig

```

Listing 73: SiLK Installation on Debian

SiLK design and implementation differs a lot from [NFQL](#). SiLK believes in the philosophy of a tool performing a single task well. For instance, in SiLK there are separate tools to perform the task of each stage of the [NFQL](#) processing pipeline. The stage functionality is not full-fledged though. The grouping and merging operations can only be performed using an equality operator. This is assumed in the tool, thereby allowing it to perform optimization such as using hash tables to perform lookups. The usage instruction of tools that can perform (if not) an equivalent [NFQL](#) stage operation is given in listing 74.

*nfql equivalent silk  
tools*

```

1 # absolute filtering (dst port 80 or (src port 80 and dst port 25)
2 $ rfilter --dport=80 --pass=out1.rwf.gz in.rwf.gz
3 $ rfilter --sport=80 --dport=25 --pass=out2.rwf.gz in.rwf.gz
4
5 # concatenating flow records
6 $ rwcatt --output=out.rwf.gz out1.rwf.gz out2.rwf.gz
7 $ rwcatt out1.rwf.gz out2.rwf.gz >> out.rwf.gz
8 $ rwappend [--create] out.rwf.gz out1.rwf.gz out2.rwf.gz
9
10 # grouping flow records and setting thresholds
11 $ rwuniq --field=1 out.raw --bytes --packets=1000 --flows=200
12 $ rwgroup --id-field=1,2,3,4 --delta-field=9 --delta-value=3600 in.rwf.gz > out.rwf.gz
13
14 # merging flow records
15 $ rwmatt --relate=1,2 --relate=2,1 \
16         --relate=3,4 --relate=4,3 \
17         query.rwf.gz response.rwf.gz stdout

```

Listing 74: NFQL Equivalent SiLK Tools

<sup>1</sup> <http://tools.netsa.cert.org/silk/>

*additional silk  
analysis and capture  
tools*

There are also stringent requirements to how a flow-data needs to be organized before it can be piped into a tool. The grouping tool, for instance assumes that the to-be supplied input flow data is already sorted on the field column. These requirements made it a little cumbersome to design a full-fledged [NFQL](#) query. The final query had over a dozen of SiLK tools piped together and saved as a bash script. These bash scripts were then called by the benchmarking suite for performance evaluation. Few additional tools available in the SiLK's repertoire are shown in listing 75.

```

1 # reading flow records
2 $ rwcut out.rwf.gz
3
4 # generating statistical summary
5 $ rwstats --overall-stats out.rwf.gz
6
7 # creating time series (10 minute interval)
8 $ rwcoun --bin-size=600 out.rwf.gz
9
10 # sorting flow records (on srcIP)
11 $ rwsort --fields=1 --output=out-sort.rwf.gz out.rwf.gz
12
13 # remove duplicate flow fecords
14 $ rwdedupe --stime-delta=100 out1.rwf.gz out2.rwf.gz > out.rwf.gz
15
16 # splitting flow records
17 $ rwsplit out.rwf.gz --basename=splits --flow-limit=1000
18
19 # show silk file characteristics
20 $ rwfileinfo out.rwf.gz
21
22 # generate flows from text files
23 $ rwtuc --fields=1-9 out.txt > out.rwf.gz
24
25 # generate flows from tcpdump traces
26 $ rwptoflow out.pcap > out.rwf.gz

```

Listing 75: Additional SiLK Analysis and Capture Tools

*flow-tools to nfdump*

SiLK uses its own proprietary format for reading flow traces. In order to be able to perform the evaluation on the same flow traces, it was essential to convert the flow-tools <sup>2</sup> format trace files to SiLK proprietary format. The best way is to replay the original trace files on a host-port combination, and let SiLK flow capturing daemon pick it up to save it in its proprietary format. Unfortunately, flow-tools does not have any such replay tool. Although, one is provided by the nfdump <sup>3</sup> package. Listing 76 shows how flow-tools traces were converted to nfdump format to allow them to be replayed later.

```

1 # install nfdump and ft2nfdump
2 $ sudo apt-get install nfdump
3 $ sudo apt-get install nfdump-flow-tools
4
5 # convert flow-tools traces to nfdump
6 $ flow-cat $INPUT | ft2nfdump | nfdump -w $OUTPUT
7 $ nfdump -r $OUTPUT

```

Listing 76: flow-tools to nfdump

<sup>2</sup> <http://www.splintered.net/sw/flow-tools/>

<sup>3</sup> <http://nfdump.sourceforge.net/>

The converted `nfdump` traces were then replayed using `nfreplay`. By default, the tools replays the traces to 127.0.0.1 at port 9995. A sensor configuration file was created on the other end for SiLK to collect the replayed data. The configuration file defines the NetFlow protocol used in the replay, host-port combination and definition of internal and external IP blocks to separate the flow-traces accordingly. SiLK's `rwflowpack` was then used to regenerate the traces in the proprietary format. The tool segregates the flows into multiple hierarchically arranged directories. Since, only the runtime analysis of SiLK on the original query was of interest, a combination of `find` and `xargs` was used to flatten the directory and combine the flows into one file as shown in listing 77.

*nfdump to silk*

```

1  # replay the nfdump trace
2  $ nfreplay -r $TRACE
3
4  # configure a sensor to collect replayed data
5  $ cat sensors.conf
6  probe S1 netflow-v5
7      listen-on-port 9995
8      protocol udp
9      accept-from-host 127.0.0.1
10 end probe
11 sensor S1
12     netflow-v5-probes S1
13     internal-ipblocks 10.0.0.0/8
14     external-ipblocks remainder
15 end sensor
16
17 # collect flow data and save in binary silk files
18 $ rwflowpack \
19     --site-config-file=/usr/local/share/silk/generic-silk.conf \
20     --sensor-configuration=sensors.conf \
21     --root-directory=/var/log/silk/ \
22     --log-destination=both
23
24 rwflowpack[14830]: Forked child 14831. Parent exiting
25 rwflowpack[14831]: Using packing logic from ...
26 rwflowpack[14831]: Creating stream cache
27 rwflowpack[14831]: Starting flow processor #1 for PDU Reader
28 rwflowpack[14831]: Creating PDU Reader Source Pool
29 rwflowpack[14831]: Creating PDU Reader for probe 'S1' on 0.0.0.0:9995
30 rwflowpack[14831]: Starting flush timer
31 rwflowpack[14831]: Started manager thread for PDU Reader
32
33 # flatten the silk root directory
34 $ find /var/log/silk -type f -exec cp {} /var/log/silkflat/ \;
35
36 # combine all silk files into a single archive
37 $ ls | rwcat --xargs --output-path=/var/log/silk.gz

```

Listing 77: `nfdump` to SiLK





## NFQL RELEASE NOTES

---

The execution engine is now verifiable using a full-fledged regression suite, alongwith an automated benchmarking suite. The python parser implementation has been properly packaged to allow seamless single-step installation using make and pip. The installation and usage instructions for both Debian/Ubuntu and OS X are included. The workset of the release is shown in listing 78.

*it's verifiable*

```
1 >> git show v0.5
2
3 tag v0.5
4 Tagger: Vaibhav Bajpai <contact@vaibhavbajpai.com>
5 Date:   Wed Jul 11 10:33:58 2012 +0200
6 Commit: 8d2f9b374a1e104e97398de47542cd5c0479a0dc
7
8 * better engine usage on run.
9 * evaluation of query ruleset lengths at RUNTIME.
10 * python pipeline module to encapsulate pipeline stage classes.
11 * painless parser installation using make.
12 * parser installation instructions on debian/ubuntu and osx.
13 * regression test-suite for the execution engine.
14 * silk installation and usage instructions.
15 * instructions to convert flow-tools traces to silk.
16 * automated benchmarking suite.
17 * resolved issues:
18   * no segfault on srcIP = dstIP in a grouper rule.
19   * no segfault when no grouper rules are defined.
```

Listing 78: Release Notes: v0.5

The execution engine is now portable to be able to seamlessly build on multiple Unix flavors, tagged as v0.4. CMake is used to orchestrate the buid process. It uses custom commands to invoke scripts that can prepare the auto-generated sources and sample JSON queries. A Makefile is used to automate the CMake invocations. Feature-test and platform-specific macros allow the code to become portable. The snapshot contains installation instructions for both Debian/Ubuntu and OS X. A workset of the release is shown in listing 79.

*it's portable*

```
1 $ git show v0.4
2
3 tag v0.4
4 Tagger: Vaibhav Bajpai <contact@vaibhavbajpai.com>
5 Date:   Fri May 18 15:07:42 2012 +0200
6 Commit: 00c17385e37dd944c9139205a5eb3660c707858a
7
8 * _GNU_SOURCE feature test MACRO and -std=c99
9 * (__FreeBSD, __APPLE__) and __linux MACROS around qsort_r(...)
10 * reverted to a flat source structure for the CMake build process.
11 * CMake custom command to call a script to create auto-generated sources and headers.
12 * CMake custom command to call a scripts in queries/ to save sample JSON queries in
   examples/
13 * Makefile to automate invocation of CMake commands.
14 * installation instruction for Ubuntu.
15 * installation instruction for Mac OS X.
```

Listing 79: Release Notes: v0.4

*it's flexible*

The execution engine is now flexible to read and parse the entire flowquery at *runtime*, tagged as v0.3. The flowquery is written in an intermediate JSON format and read using json-c. The JSON queries are themselves generated using python scripts. The engine is fail-safe using consistency checks that allow it to fail gracefully when either the trace or the flowquery cannot be read. Each branch thread can now exit early on failure and each stage of the pipeline only proceeds when the previous stage returned results. A workset of the release is shown in listing 80.

```

1 $ git show v0.3
2
3 tag v0.3
4 Tagger: Vaibhav Bajpai <contact@vaibhavbajpai.com>
5 Date:   Wed May 16 18:25:22 2012 +0200
6 Commit: 1c323fa66b9aaaad56ad7c4127b8d187eaf4ec0c
7
8 * complete query is read at RUNTIME using JSON-C
9 * JSON queries are generated using python scripts
10 * glibc backtrace(...) to print the back trace on errExit(...)
11 * gracefully exiting when trace cannot be read
12 * gracefully exiting when JSON query cannot be parsed
13 * branch thread returns EXIT_FAILURE if either stage returns NULL
14 * branch thread returns EXIT_SUCCESS on normal exit
15 * each stage proceeds only when previous returned results
16 * flow-cat ... | flowy-engine $QUERY -

```

Listing 80: Release Notes: v0.3

*it's robust*

The execution engine is robust to work well with variety of queries, tagged as v0.2. The complete engine has been refactored with a maintainable design to allow better execution workflow and abstract objects. Each stage of the pipeline has one public interface function that takes a ruleset as input and returns a result abstract object. The engine has been profiled to eliminate any memory leaks and allow early deallocation of objects as soon as they are not required. The engine is now smart enough to realize and ignore redundant aggregation requests. All the hardcoded rules of the flowquery are clubbed together in one header file for easy maintainability. The dedicated rule function pointer assignments are lazy. A workset of the release state is shown in listing 81.

```

1 $ git show v0.2
2
3 tag v0.2
4 Tagger: Vaibhav Bajpai <contact@vaibhavbajpai.com>
5 Date:   Wed Apr 18 13:24:16 2012 +0200
6 Commit: 2c571f80cd076172cbd00ef7f9976b88cb44b425
7
8 * complete engine refactor.
9 * complete engine profiling (no memory leaks).
10 * issues closed:
11   - greedily deallocating non-filtered records in O(n) before merger(...).
12   - resolved a grouper segfault when NO records got filtered.
13   - all records are grouped into 1 group when no grouping rule specified.
14   - aggregation on common fields touched by filter/grouper rules is ignored.
15   - no uintX_t assumptions for field offsets.
16   - rules are clubbed together and assigned using a loop.
17   - function parameters are as minimum as required.
18   - function parameters are safe using [const] ptr and ptr to [const].
19   - lazy rule->func(...) assignment when the stage is entered.

```

Listing 81: Release Notes: v0.2

This tag starts off the F(v2) branch. The complete pipeline now works for the first time, tagged as v0.1. The grouper segfaults have been resolved. The resulting group records are cooked as NetFlow v5 records with their field offsets representing group aggregations. This snapshot contains the first-ever group filter, merger and ungrouper implementation in C. The stages do not assume type of the field offsets that are not known until *runtime*. The engine now works with multiple verbosity levels increasing the amount of echo on each level. A workset of the release is shown in listing 82.

*it's functional*

```

1 $ git show v0.1
2
3 tag v0.1
4 Tagger: Vaibhav Bajpai <contact@vaibhavbajpai.com>
5 Date:   Fri Apr 6 19:07:49 2012 +0200
6 Commit: a8a67a13aa07f671d21d062537a2ef17e58dcc07
7 ...
8
9 * reverse engineered parser to generate UML.
10 * froze requirements to allow single step installation of the python parser.
11 * doxygen documentation of the engine.
12 * prelim JSON parsing framework for the parser and engine to spit and parse the JSON
    queries.
13 * replaced GNU99 extensions dependent code with c99.
14 * resolved numerous segfaults in the grouper.
15 * generated group aggregations as a separate (cooked) NetFlow v5 record.
16 * flexible group aggregations with no uintX_t assumptions on field offsets.
17 * first-ever group filter implementation.
18 * reorganized the src/ directory structure
19 * enabled multiple verbosity levels in the engine.
20 * first-ever merger implementation.
21 * flexible filters and group filters with no uintX_t assumptions on field offsets.
22 * first-ever ungrouper implementation.
```

Listing 82: Release Notes: v0.1

This tag represents an evolution of the core of the former Python implementation (Flowy) in C. It is also referred as F(v1). It can efficiently read flow records into memory and filter them. Each branch of the pipeline runs in a separate thread. Each rule type has its own dedicated function pointer generated using a python script. The idea of using qsort/bsearch for efficient grouper processing is in place, but the implementation is broken with numerous segfaults. The engine has concerns as shown in listing 83.

*it's a start*

```

1 $ git show v0.0
2
3 tag v0.0
4 Tagger: Vaibhav Bajpai <contact@vaibhavbajpai.com>
5 Date:   Thu May 17 10:48:02 2012 +0200
6 Commit: 8cb309c8a956c99e6b1494eddb601c8f6a520696
7
8 * read flow-records into memory
9 * rewrite of the execution pipeline in C (non functional)
10 * efficient rule processing with dedicated function pointers
11 * reduced grouper complexity using qsort(...) and bsearch(...)
12 * concerns
13   - flow query is currently hardcoded in pipeline structs
14   - functions assume specific uintX_t offsets
15   - numerous grouper segfaults
16   - no group filter, no ungrouper
17   - commented out merger (segfaults when uncommented)
18   - code dependent on GNU99 extensions
19   - some headers are missing include guards
20   - unused extraneous source files and headers
```

Listing 83: Release Notes: v0.0





ACRONYMS

---

IPFIX	Internet Protocol Flow Information Export
HDF	Hierarchical Data Format
LALR	Look-Ahead LR Parser
PLY	Python Lex-Yacc
HDFS	Hadoop Distributed File System
API	Application Programming Interface
CNF	Conjunctive Normal Form
SSDP	Simple Service Discovery Protocol
IP	Internet Protocol
UDP	User Datagram Protocol
TCP	Transmission Control Protocol
NAT-PMP	Network Address Translation Port Mapping Protocol
ccTLD	Country Code Top-Level Domain
HTTP	Hypertext Transfer Protocol
IaaS	Infrastructure as a Service
NaaS	Network as a Service
vLANs	Virtual Local Area Networks
ACLs	Access Control Lists
MPLS	Multiprotocol Label Switching
RTT	Round Trip Time
SVMs	Support Vector Machines
FF	Greedy Forward Fitting
SMTP	Simple Mail Transfer Protocol
DDoS	Distributed Denial of Service
CAPTCHA	Completely Automated Public Turing Test to Tell Computers and Humans Apart

RMON	Remote Network Monitoring
MIB	Management Information Base
SNMP	Simple Network Management Protocol
RTFM	Realtime Traffic Flow Measurement
GUI	Graphical User Interface
IETF	Internet Engineering Task Force
DoS	Denial of Service
AS	Autonomous Systems
CIDR	Classless Inter-Domain Routing
SCTP	Stream Control Transmission Protocol
PSAMP	Packet Sampling
TLS	Transport Layer Security
DTLS	Datagram Transport Layer Security
IE	Information Elements
IANA	Internet Assigned Numbers Authority
PENs	Private Enterprise Numbers
EP	Exporter Process
CP	Collector Process
SMI	Structure of Managed Information
CLI	Command Line Interface
XDR	External Data Representation
UML	Unified Modeling Language
NFQL	Network Flow Query Language

## BIBLIOGRAPHY

---

- [1] B. Claise, "Cisco Systems NetFlow Services Export Version 9." RFC 3954 (Informational), Oct. 2004.
- [2] B. Claise, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information." RFC 5101 (Proposed Standard), Jan. 2008.
- [3] V. Marinov, "Design of an IP Flow Record Query Language," Master's thesis, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany, August 2009.
- [4] J. F. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, pp. 832–843, November 1983.
- [5] K. Kanev, "Flowy - Network Flow Analysis Application," Master's thesis, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany, August 2009.
- [6] J. Schauer, "Flowy 2.0: Fast Execution of Stream based IP Flow Queries," bachelor's thesis, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany, May 2011.
- [7] N. Melnikov, "Cybermetrics: Identification of Users through Network Flow Analysis," Master's thesis, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany, August 2010.
- [8] P. Kohler and B. Claise, "IPFIX Fine-Tunes Traffic Analysis," *Network World*, Aug. 2003.
- [9] B. Trammell and E. Boschi, "An Introduction to IP Flow Information Export (IPFIX)," *Communications Magazine, IEEE*, vol. 49, pp. 89–95, april 2011.
- [10] Dell, Texas, *Dell PowerConnect M6220, M6348, M8024, and M8024âk Switch User's Configuration Guide*.
- [11] V. Marinov and J. Schönwälder, "Design of a Stream-Based IP Flow Record Query Language," in *Proceedings of the 20th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management: Integrated Management of Systems, Services, Processes and People in IT, DSOM '09*, (Berlin, Heidelberg), pp. 15–28, Springer-Verlag, 2009.
- [12] P. Nemeth, "Flowy Improvements using Map/Reduce," bachelor's thesis, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany, May 2010.

- [13] B. Daviss, "Building a Crash-Proof Internet," *New Scientist*, vol. 26, pp. 38–41, June 2009.
- [14] R. Beverly and K. Sollins, "Exploiting Transport-Level Characteristics of Spam," in *Proceedings of the Fifth Conference on Email and Anti-Spam (CEAS)*, Aug. 2008.
- [15] V. Perelman, "Flow signatures of Popular Applications," bachelor's thesis, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany, May 2010.
- [16] V. Jacobson, C. Leres, and S. McCanne, *tcpdump - dump traffic on a network*. Lawrence Berkeley National Laboratory, University of California, Berkeley, CA.
- [17] V. Jacobson, C. Leres, and S. McCanne, *pcap - Packet Capture library*. Lawrence Berkeley National Laboratory, University of California, Berkeley, CA.
- [18] G. Combs, *wireshark - Interactively dump and analyze network traffic*. University of Missouri, Kansas City.
- [19] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, "Profiling Internet Backbone Traffic: Behavior Models and Applications," in *Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, SIGCOMM '05*, (New York, NY, USA), pp. 169–180, ACM, 2005.
- [20] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark," in *Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, SIGCOMM '05*, (New York, NY, USA), pp. 229–240, ACM, 2005.
- [21] M.-S. Kim, H.-J. Kong, S.-C. Hong, S.-H. Chung, and J. Hong, "A Flow-based Method for Abnormal Network Traffic Detection," in *Network Operations and Management Symposium, 2004. NOMS 2004. IEEE/IFIP*, vol. 1, pp. 599–612 Vol.1, april 2004.
- [22] D. Schatzmann, W. Mühlbauer, T. Spyropoulos, and X. Dimitropoulos, "Digging into HTTPS: Flow-based Classification of Webmail Traffic," in *Proceedings of the 10th annual conference on Internet measurement, IMC '10*, (New York, NY, USA), pp. 322–327, ACM, 2010.
- [23] S. Waldbusser, R. Cole, C. Kalbfleisch, and D. Romascanu, "Introduction to the Remote Monitoring (RMON) Family of MIB Modules." RFC 3577 (Informational), Aug. 2003.
- [24] J. Case, M. Fedor, M. Schoffstall, and J. Davin, "Simple Network Management Protocol (SNMP)." RFC 1157 (Historic), May 1990.

- [25] S. Waldbusser, "Remote Network Monitoring Management Information Base." RFC 2819 (Standard), May 2000.
- [26] S. Waldbusser, "Remote Network Monitoring Management Information Base Version 2." RFC 4502 (Draft Standard), May 2006.
- [27] N. Brownlee, C. Mills, and G. Ruth, "Traffic Flow Measurement: Architecture." RFC 2722 (Informational), Oct. 1999.
- [28] W. W. Royce, "Managing the Development of Large Software Systems: Concepts and Techniques," in *Proceedings of the 9th International Conference on Software Engineering, ICSE '87*, (Los Alamitos, CA, USA), pp. 328–338, IEEE Computer Society Press, 1987.
- [29] C. Estan, K. Keys, D. Moore, and G. Varghese, "Building a Better NetFlow," in *Proceedings of the 2004 conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, SIGCOMM 2004*, (New York, NY, USA), pp. 245–256, ACM, 2004.
- [30] N. Duffield, D. Chiou, B. Claise, A. Greenberg, M. Grossglauser, and J. Rexford, "A Framework for Packet Selection and Reporting." RFC 5474 (Informational), Mar. 2009.
- [31] T. Zseby, M. Molina, N. Duffield, S. Niccolini, and F. Raspall, "Sampling and Filtering Techniques for IP Packet Selection." RFC 5475 (Proposed Standard), Mar. 2009.
- [32] T. Dierks and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2." RFC 5246 (Proposed Standard), Aug. 2008. Updated by RFCs 5746, 5878, 6176.
- [33] E. Rescorla and N. Modadugu, "Datagram Transport Layer Security." RFC 4347 (Proposed Standard), Apr. 2006. Updated by RFC 5746.
- [34] B. Claise, A. Johnson, and J. Quittek, "Packet Sampling (PSAMP) Protocol Specifications." RFC 5476 (Proposed Standard), Mar. 2009.
- [35] S. Wang, R. State, M. Ourdane, and T. Engel, "FlowRank: Ranking NetFlow Records," in *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference, IWCMC '10*, (New York, NY, USA), pp. 484–488, ACM, 2010.
- [36] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.
- [37] L. Deri, E. Chou, Z. Cherian, K. Karmarkar, and M. Patterson, "Increasing Data Center Network Visibility with Cisco NetFlow-Lite," in *Network and Service Management (CNSM), 2011 7th International Conference on*, pp. 1–6, oct. 2011.

- [38] L. Deri, “nprobe: an open source netflow probe for gigabit networks,” in *In Proceedings of Terena TNC 2003*, 2003.
- [39] G. Sadasivan, N. Brownlee, B. Claise, and J. Quittek, “Architecture for IP Flow Information Export.” RFC 5470 (Informational), Mar. 2009. Updated by RFC 6183.
- [40] T. Dietz, A. Kobayashi, B. Claise, and G. Muenz, “Definitions of Managed Objects for IP Flow Information Export.” RFC 5815 (Proposed Standard), Apr. 2010.
- [41] B. Trammell and E. Boschi, “Bidirectional Flow Export Using IP Flow Information Export (IPFIX).” RFC 5103 (Proposed Standard), Jan. 2008.
- [42] P. Phaal, S. Panchen, and N. McKee, “InMon Corporation’s sFlow: A Method for Monitoring Traffic in Switched and Routed Networks.” RFC 3176 (Informational), Sept. 2001.
- [43] S. Microsystems, “XDR: External Data Representation standard.” RFC 1014, June 1987.
- [44] K. Kanev, N. Melnikov, and J. Schönwälder, “Implementation of a stream-based IP flow record query language,” in *Proceedings of the Mechanisms for autonomous management of networks and services, and 4th international conference on Autonomous infrastructure, management and security, AIMS’10*, (Berlin, Heidelberg), pp. 147–158, Springer-Verlag, 2010.
- [45] V. Marinov and J. Schönwälder, “Design of an IP Flow Record Query Language,” in *Proceedings of the 2nd international conference on Autonomous Infrastructure, Management and Security: Resilient Networks and Services, AIMS ’08*, (Berlin, Heidelberg), pp. 205–210, Springer-Verlag, 2008.
- [46] F. Alted and M. Fernández-Alonso, “PyTables: Processing And Analyzing Extremely Large Amounts Of Data In Python,” 2003.
- [47] J. Quittek, S. Bryant, B. Claise, P. Aitken, and J. Meyer, “Information Model for IP Flow Information Export.” RFC 5102 (Proposed Standard), Jan. 2008. Updated by RFC 6313.
- [48] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” in *Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation - Volume 6*, (Berkeley, CA, USA), pp. 10–10, USENIX Association, 2004.
- [49] T. White, *Hadoop: The Definitive Guide*. Definitive Guide Series, O’Reilly, 2010.

- [50] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, pp. 1–10, May 2010.
- [51] P. Mundkur, V. Tuulos, and J. Flatow, "Disco: A Computing Platform for Large-Scale Data Analytics," in *Proceedings of the 10th ACM SIGPLAN workshop on Erlang, Erlang '11*, (New York, NY, USA), pp. 84–89, ACM, 2011.
- [52] J. Schauer, N. Melnikov, and J. Schönwälder, "F." 2012.
- [53] D. S. Seljebotn, "Fast numerical computations with Cython," in *Proceedings of the 8th Python in Science Conference* (G. Varoquaux, S. van der Walt, and J. Millman, eds.), (Pasadena, CA USA), pp. 15–22, 2009.
- [54] I. Wilbers, H. P. Langtangen, and Å. Ødegård, "Using Cython to Speed up Numerical Python Programs," in *Proceedings of MekIT'09* (B. Skallerud and H. I. Andersson, eds.), pp. 495–512, NTNU, Tapir, 2009.
- [55] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. Seljebotn, and K. Smith, "Cython: The Best of Both Worlds," *Computing in Science Engineering*, vol. 13, pp. 31–39, march-april 2011.
- [56] S. Romig, "The OSU Flow-tools Package and CISCO NetFlow Logs," in *Proceedings of the 14th USENIX conference on System administration*, (Berkeley, CA, USA), pp. 291–304, USENIX Association, 2000.
- [57] P. Haag, "Netflow Tools NfSen and NFDUMP," in *Proceedings of the 18th Annual FIRST conference*, 2006.
- [58] V. Perelman, N. Melnikov, and J. Schönwälder, "Flow signatures of Popular Applications," in *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*, pp. 9–16, May 2011.
- [59] M. Bodlaender, "UPnP 1.1 - Designing for Performance Compatibility," *Consumer Electronics, IEEE Transactions on*, vol. 51, pp. 69–75, feb. 2005.
- [60] S. Cheshire, M. Krochmal, and K. Sekar, "NAT port mapping protocol (NAT-PMP)," Internet-Draft draft-cheshire-nat-pmp-03.txt, IETF Secretariat, Fremont, CA, USA, Apr. 2008.
- [61] F. Bergadano, D. Gunetti, and C. Picardi, "User Authentication through Keystroke Dynamics," *ACM Trans. Inf. Syst. Secur.*, vol. 5, pp. 367–397, November 2002.
- [62] A. Ahmed and I. Traore, "A New Biometric Technology Based on Mouse Dynamics," *IEEE Transactions on Dependable and Secure Computing*, vol. 4, pp. 165–179, July-Sept 2007.

- [63] K.-T. Chen and L.-W. Hong, "User identification based on Game-Play Activity Patterns," in *Proceedings of the 6th ACM SIGCOMM workshop on Network and System Support for Games*, NetGames '07, (New York, NY, USA), pp. 7–12, ACM, 2007.
- [64] N. Melnikov and J. Schönwälder, "Cybermetrics: User Identification through Network Flow Analysis," in *Proceedings of the Mechanisms for autonomous management of networks and services, and 4th international conference on Autonomous infrastructure, management and security*, AIMS'10, (Berlin, Heidelberg), pp. 167–170, Springer-Verlag, 2010.
- [65] V. Bajpai, N. Melnikov, and J. Schönwälder, "Automated Failure Identification under IPv6 Transition Mechanisms." 2012.
- [66] M. Bagnulo, P. Matthews, and I. van Beijnum, "Stateful NAT64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers." RFC 6146 (Proposed Standard), Apr. 2011.
- [67] A. Durand, R. Droms, J. Woodyatt, and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion." RFC 6333 (Proposed Standard), Aug. 2011.
- [68] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: Enabling Innovation in Campus Networks," *SIGCOMM Computer Communications Review*, vol. 38, pp. 69–74, March 2008.
- [69] T. Benson, A. Akella, A. Shaikh, and S. Sahu, "CloudNaaS: A Cloud Networking Platform for Enterprise Applications," in *Proceedings of the 2nd ACM Symposium on Cloud Computing*, SOCC '11, (New York, NY, USA), pp. 8:1–8:13, ACM, 2011.
- [70] G. Kakavelakis, R. Beverly, and J. Young, "Auto-learning of SMTP TCP Transport-Layer Features for Spam and Abusive Message Detection," in *LISA 2011, 25th Large Installation System Administration Conference* (T. A. Limoncelli and D. Hughes, eds.), (Berkeley, CA, USA), USENIX, LOPSA, USENIX Association, Dec. 2011.
- [71] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995. 10.1007/BF00994018.
- [72] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," 1997.
- [73] ISO, "The ANSI c standard (c99)," Tech. Rep. WG14 N1124, ISO/IEC, 1999.
- [74] CERT/NetSA at Carnegie Mellon University, "SiLK (System for Internet-Level Knowledge)." [Online]. Available: <http://tools.netsa.cert.org/silk> [Accessed: July 10, 2012].



## DECLARATION

---

I hereby declare that this thesis is the result of my independent work, has not been previously accepted in substance for any degree and is not concurrently submitted for any degree. This thesis is being submitted in fulfillment of the requirements for the degree of Master of Science in Computer Science.

*Bremen, Germany, July 2012*

---

Vaibhav Bajpai

## COLOPHON

This thesis was typeset with  $\text{\LaTeX} 2_\epsilon$  using Hermann Zapf's *Palatino* and *Euler* type faces. The listings are typeset in *Bera Mono*, originally developed by Bitstream, Inc. The typographic style was inspired by *The Elements of Typographic Style*.