Internet Draft
Category: Informational
Document: <draft-ietf-psamp-framework-08.txt>
Expires: March 2005

Nick Duffield (Editor)
AT&T Labs û Research

September 2004

A Framework for Packet Selection and Reporting

Status of this Memo

Abstract

This document specifies a framework for the PSAMP (Packet SAMPling) protocol. The functions of this protocol are to select packets from a stream according to a set of standardized selectors, to form a stream of reports on the selected packets, and to export the reports to a collector. This framework details the components of this architecture, then describes some generic requirements, motivated the dual aims of ubiquitous deployment and utility of the reports for applications. Detailed requirements for selection, reporting and exporting processes are described, along with configuration requirements of the PSAMP functions.

Comments on this document should be addressed to the PSAMP Working Group mailing list: psamp@ops.ietf.org

To subscribe: psamp-request@ops.ietf.org, in body: subscribe

Archive: https://ops.ietf.org/lists/psamp/

Table of Contents

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

1. Introduction

This document describes the PSAMP framework for network elements to select subsets of packets by statistical and other methods, and to export a stream of reports on the selected packets to a collector.

The motivation for the PSAMP standard comes from the need for measurement-based support for network management and control across multivendor domains. This requires domain wide consistency in the types of selection schemes available, the manner in which the resulting measurements are presented, and consequently, consistency of the interpretation that can be put on them.

The motivation for specific packet selection operations comes from the applications that they enable. Development of the PSAMP standard is open to influence by the requirements of standards in related IETF Working Groups, for example, IP Performance Metrics (IPPM) [RFC-2330] and Internet Traffic Engineering (TEWG).

The name PSAMP is a contraction of the phrase Packet Sampling. The word ôsamplingö captures the idea that only a subset of all packets passing a network element will be selected for reporting. But PSAMP selection operations include random selection, deterministic selection (filtering), and deterministic approximations to random selection (hash-based selection).

2. PSAMP Documents Overview

PSAMP-FRAMEWORK: ôA Framework for Packet Selection and Reportingö: this document. This document describes the PSAMP framework for network elements to select subsets of packets by statistical and other methods, and to export a stream of reports on the selected packets to a collector. Definitions of terminology and the use of the terms ômustö, ôshouldö and ômayö in this document are informational only.

[PSAMP-TECH]: ôSampling and Filtering Techniques for IP Packet Selectionö, describes the set of packet selection techniques supported by PSAMP.

[PSAMP-MIB]: ôDefinitions of Managed Objects for Packet Samplingö describes the PSAMP Management Information Base

[PSAMP-PROTO]: ôPacket Sampling (PSAMP) Protocol Specificationsö specifies the export of packet information from a PSAMP Exporting Process to a PSAMP Colleting Process

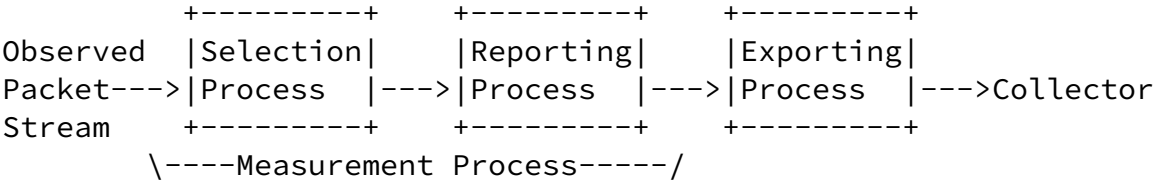
[PSAMP-INFO]: ôInformation Model for Packet Sampling Exportsö defines an information and data model for PSAMP.

3. Elements, Terminology and High-level Architecture

3.1 High-level description of the PSAMP Architecture

Here is an informal high level description of the PSAMP protocol operating in a PSAMP device (all terms will be defined presently). A stream of packets is observed at an observation point. A selection process inspects each packet to determine whether it should be selected. A reporting process constructs a report on each selected packet, using the packet content, and possibly other information such as the packet treatment or the arrival timestamp. An exporting process sends the reports to a collector, together with any subsidiary information needed for their interpretation.

The following figure indicates the sequence of the three processes (selection, reporting, and exporting) within the PSAMP device. The composition of the selection process followed by the reporting process is known as the measurement process.



The following sections give the detailed definitions of each of all the objects just named.

3.2 Observation Points, Packet Streams and Packet Content

This section contains the definition of terms relevant to obtaining the packet input to the selection process.

* Observation Point

  An observation point is a location in the network where packets can be observed. Examples include:

  (i) a line to which a probe is attached;
  (ii) a shared medium, such as an Ethernet-based LAN;
  (iii) a single port of a router, or set of interfaces (physical or logical) of a router;
  (iv) an embedded measurement subsystem within an interface.

  Note that one observation point may be a superset of several other observation points. For example one observation point can be an entire line card. This would be the superset of the individual observation points at the line card's interfaces.

* Observed Packet Stream

  The observed packet stream is the set of all packets observed at the observation point.

* Packet Stream

  A packet stream denotes a subset of the observed packet stream.

* Packet Content

  The packet content denotes the union of the packet header (which includes link layer, network layer and other encapsulation headers) and the packet payload.

Note that packets selected from a stream, e.g. by sampling, do not necessarily possess a property by which they can be distinguished from packets that have not been selected. For this reason the term ôstreamö is favored over ôflowö, which is defined as set of packets with common properties [IPFIX-REQUIRE].

3.3 Selection Process

This section defines the selection process and related objects.

* Selection Process

  A selection process takes a packet stream as its input and selects a subset of that stream as its output.

* Selection State:

  A selection process may maintain state information for use by the selection process and/or the reporting process. At a given time, the selection state may depend on packets observed at and before that time, and other variables. Examples include:

  (i) sequence numbers of packets at the input of selectors;

  (ii) a timestamp of observation of the packet at the observation point;

  (iii) iterators for pseudorandom number generators;

  (iv) hash values calculated during selection;

  (v) indicators of whether the packet was selected by a given selector;

  Selection processes may change portions of the selection state as a result of processing a packet. Selection state for a packet is to reflect the state after processing the packet.

* Selector:

  A selector defines the action of a selection process on a single packet of its input. A selected packet becomes an element of the output packet stream of the selection process.

  The selector can make use of the following information in determining whether a packet is selected:

(i) the packetÆs content;

(ii) information derived from the packet's treatment at the observation point;

(iii) any selection state that may be maintained by the selection process.

* Composite Selection Process:

A composite selection process is an ordered composition of selection processes, in which the output stream issuing from one component forms the input stream for the succeeding component.

* Composite Selector:

A selector is composite if it defines a composite selection process.

* Primitive Selection Process:

A selection process is primitive if it is not a composite a selection process.

* Primitive Selector:

A selector is primitive if it defines a primitive selection process.

3.4 Reporting Process

* Reporting Process:

A reporting process creates a report stream on packets selected by a selection process, in preparation for export. The input to the reporting process comprises that information available to the selection process per selected packet, specifically:

(i) the selected packetÆs content;

(ii) information derived from the selected packet's treatment at the observation point;

(iii) any selection state maintained by the inputting selection process, reflecting any modifications to the

selection state made during selection of the packet.

* Packet Reports:

Packet reports comprise a configurable subset of a packetÆs input to the reporting process, including the packetÆs content, information relating to its treatment (for example, the output interface), and its associated selection state (for example, a hash of the packetÆs content)

* Report Interpretation:

Report interpretation comprises subsidiary information, relating to one or more packets, that is used for interpretation of their packet reports. Examples include configuration parameters of the selection process and of the reporting process.

* Report Stream:

The report stream is the output of a reporting process, comprising two distinguished types of information: packet reports, and report interpretation.

3.5 Measurement Process

* A Measurement Process is the composition of a selection process that takes the observed packet stream as its input, followed by a reporting process.

3.6 Exporting Process

* Exporting Process:

An exporting process sends, in the form of export packet, the output of one or more measurement processes to one or more collectors.

* Export Packets:

a combination of report interpretation and/or one or more

packet reports are bundled by the exporting process into a export packet for exporting to a collector.

3.7 PSAMP Device

A PSAMP Device is a device hosting at least an observation point, a measurement process and an exporting process. Typically, corresponding observation point(s), measurement process(es) and exporting process(es) are co-located at this device, for example at a router.

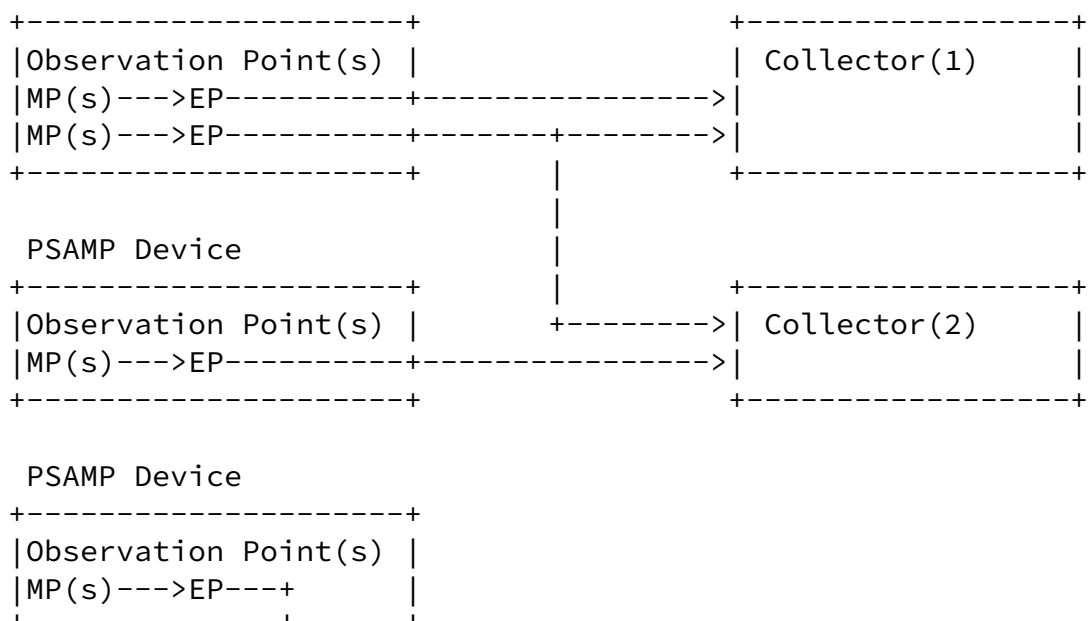
3.8 Collector

A collector receives a report stream exported by one or more exporting processes. In some cases, the host of the measurement and/or exporting processes may also serve as the collector.

3.9 Possible Configurations

Various possibilities for the high level architecture of these elements are as follows.

MP = Measurement Process, EP = Exporting process



```
|Collector(3)<-+      |
+--------------------+
```

3.10 PSAMP and IPFIX Interaction

The PSAMP measurement process can be viewed as analogous to the IPFIX metering process. The PSAMP measurement process takes an observed packet stream as its input, and produces packet reports as its output. The IPFIX metering process produces flow records as its output. The distinct name ômeasurement processö has been retained in order to avoid potential confusion in settings where IPFIX and PSAMP coexist, and in order to avoid the implicit requirement that the PSAMP version satisfy the requirements of an IPFIX metering process (at least while these are under development). The relationship between PSAMP and IPFIX is described more in [PSAMP-INFO].

4. Generic Requirements for PSAMP

This section describes the generic requirements for the PSAMP protocol. A number of these are realized as specific requirements in later sections.

4.1 Generic Selection Process Requirements.

* Ubiquity: The selectors must be simple enough to be implemented ubiquitously at maximal line rate.

* Applicability: the set of selectors must be rich enough to support a range of existing and emerging measurement based applications and protocols. This requires a workable trade-off between the range of traffic engineering applications and operational tasks it enables, and the complexity of the set of capabilities.

* Extensibility: the protocol must be able to accommodate additional packet selectors not currently defined.

* Flexibility: the protocol must support selection of packets using various network protocols or encapsulation layers, including Internet Protocol Version 4 (IPv4) [IPv4], Internet Protocol Version 6 (IPv6) [RFC-2460], and Multiprotocol Label Switching (MPLS) [RFC-3031].

* Robust Selection: packet selection must be robust against attempts to craft an observed packet stream from which packets are selected disproportionately (e.g. to evade selection, or overload measurement systems).

* Parallel Measurement Processes: the protocol must support simultaneous operation of multiple independent measurement processes at the same host.

* Non-Contingency: the selection decision for each packet must not depend on future packets.

* Encrypted Packets: selection operations based on interpretation of packet fields must be configurable to ignore (i.e. not select) encrypted packets, when they are detected.

Selectors are outlined in Section 5, and described in more detail in the companion document [PSAMP-TECH].

4.2 Generic Reporting Process Requirements

* Self-defining: the report stream must be complete in the sense that no additional information need be retrieved from the observation point in order to interpret and analyze the reports.

* Indication of Information Loss: the reports stream must include sufficient information to indicate or allow the detection of loss occurring within the selection, reporting or exporting processes, or in transport. This may be achieved by the use of sequence numbers.

* Accuracy: the report stream must include information that enables the accuracy of measurements to be determined.

* Faithfulness: all reported quantities that relate to the packet treatment must reflect the router state and configuration encountered by the packet at the time it is received by the measurement process.

* Privacy: selection of the content of packet reports will be cognizant of privacy and anonymity issues while being

responsive to the needs of measurement applications, and in accordance with [RFC-2804]. Full packet capture of arbitrary packet streams is explicitly out of scope.

A specific reporting process meeting these requirements, and the requirement for ubiquity, is described in Section 6.

4.3 Generic Exporting process Requirements

* Timeliness: configuration must allow for limiting of buffering delays for the formation and transmission for export reports. See Section Error! Reference source not found. for further details.

* Congestion Avoidance: export of a report stream across a network must be congestion avoiding in compliance with [RFC-2914].

* Secure Export:

  (i) confidentiality: the option to encrypt exported data must be provided.

  (ii) integrity: alterations in transit to exported data must be detectable at the collector

  (iii) authenticity: authenticity of exported data must be verifiable by the collector in order to detect forged data.

The motivation here is the same as for security in IPFIX export; see Sections 6.3 and 10 of [IPFIX-REQUIRE].

4.4 Generic Configuration Requirements

* Ease of Configuration: of sampling and export parameters, e.g. for automated remote reconfiguration in response to collected reports.

* Secure Configuration: the option to configure via protocols that prevent unauthorized reconfiguration or eavesdropping on configuration communications must be available. Eavesdropping

on configuration might allow an attacker to gain knowledge that would be helpful in crafting a packet stream to evade

subversion, or overload the measurement infrastructure.

Configuration is discussed in Section 9. Feasibility and complexity of PSAMP operations is discussed in Section 10.

5. Packet Selection Operations

5.1 Two Types of Selection Operation

PSAMP categorizes selection operations into two types:

* Filtering: a filter is a selection operation that selects a packet deterministically based on the packet content, its treatment, and functions of these occurring in the selection state. Two examples are:

    (i) Mask/match filtering.

    (ii) Hash-based selection: a hash function is applied to the packet content, and the packet is selected if the result falls in a specified range.

* Sampling: a selection operation that is not a filter is called a sampling operation. This reflects the intuitive notion that if the selection of a packet cannot be determined from its content alone, there must be some type of sampling taking place.

  Sampling operations can be divided into two subtypes:

    (i) Content-independent Sampling, which does not use packet content in reaching sampling decisions. Examples include periodic sampling, and uniform pseudorandom sampling driven by a pseudorandom number whose generation is independent of packet content. Note that in content-independent sampling it is not necessary to access the packet content in order to make the selection decision.

    (ii) Content-dependent Sampling, in which the packet content is used in reaching selection decisions. Examples include pseudorandom selection according to a probability that depends on the contents of a packet field; note that this is not a filter.

5.2 PSAMP Packet Selection Operations

A spectrum of packet selection operations is described in detail in [PSAMP-TECH]. Here we only briefly summarize the meanings for completeness.

A PSAMP selection process must support at least one of the following selectors.

* Systematic Time Based Sampling: packet selection is triggered at periodic instants separated by a time called the spacing. All packets that arrive within a certain time of the trigger (called the interval length) are selected.

* Systematic Count Based Sampling: similar to systematic time based expect that selection is reckoned with respect to packet count rather than time. Packet selection is triggered periodically by packet count, a number of successive packets being selected subsequent to each trigger.

* Uniform Probabilistic Sampling: packets are selected independently with fixed sampling probability p.

* Non-uniform Probabilistic Sampling: packets are selected independently with probability p that depends on packet content.

* Probabilistic n-out-of-N Sampling: form each count-based successive block of N packets, n are selected at random.

* Mask/match Filtering: this entails taking the masking portions of the packet (i.e. taking the logical ôandö with a binary mask) and selecting the packet if the result falls in a range specified in the selection parameters of the filter. This specification does not preclude the future definition of a high level syntax for defining filtering in a concise way (e.g. TCP port taking a particular value) providing that syntax can be compiled into the bitwise expression.

  Mask/match operations should be available for different protocol portions of the packet header:

  (i) the IP header (excluding options in IPv4, stacked headers in IPv6)

  (ii) transport header

  (iii) encapsulation headers (e.g. the MPLS label stack) if present)

When the PSAMP device offers mask/match filtering, and, in its usual capacity other than in performing PSAMP functions, identifies or processes information from one or more of the above protocols, then the information should be made available for filtering. For example, when a PSAMP device routes based on destination IP address, that field should be made available for filtering. Conversely, a PSAMP device that does not route is not expected to be able to locate an IP address within a

packet, or make it available for filtering, although it may do so.

Since packet encryption alters the meaning of encrypted fields, Mask/Match filtering must be configurable to ignore encrypted packets, when detected.

Hash-based Selection: Hash-based selection will employ one or more hash functions to be standardized. A hash function is applied to a subset of packet content, and the packet is selected of the resulting hash falls in a specified range. With a suitable hash function, hash based selection approximates uniform random sampling. Applications of hash-based sampling are described in Section 11.

* Router State Filtering: the selection process may support filtering based on the following conditions, which may be combined with the logical "and", "or" or "not" operators:

  (i) Ingress interface at which packet arrives equals a specified value
  (ii) Egress interface to which packet is routed to equals a specified value
  (iii) Packet violated Access Control List (ACL) on the router
  (iv) Failed Reverse Path Forwarding (RPF)
  (v) Failed Resource Reservation (RSVP)
  (vi) No route found for the packet
  (vii) Origin Border Gateway Protocol (BGP) Autonomous System (AS) equals a specified value or lies within a given range
  (viii) Destination BGP AS equals a specified value or lies within a given range

Router architectural considerations may preclude some

information concerning the packet treatment, e.g. routing state, being available at line rate for selection of packets. However, if selection not based on routing state has reduced down from line rate, subselection based on routing state may be feasible.

This section detailed specific requirements for the selection process, motivated by the generic requirement of Section 3.3.

5.3 Selection Rate Terminology

The proportion of packets that are selected by a selection operation is figured in two ways:

* Attained Selection Frequency: the actual frequency with which packets are selected by a selection process. When packets are selected from a set of packets in a stream, the attained sampling frequency is calculated as ratio of the number of packets selected to the number of packets in the set.

* Target Selection Frequency: the average frequency with which packets are expected to be selected, based on selector parameter settings.

  For sampling operations, due to the inherent statistical variability of sampling decisions, the target and attained selection frequencies will not in general be equal, although they may be close in some circumstances, e.g., when the population size is large.

5.4 Input Sequence Numbers for Primitive Selection Processes

Each instance of a primitive selection process must maintain a count of packets presented at its input. The counter value is to be included as a sequence number for selected packets. The sequence numbers are considered as part of the packet's selection state.

Use of input sequence numbers enables applications to determine the attained selection frequency, and hence correctly normalize network usage estimates regardless of loss of information, regardless of whether this loss occurs because of discard of packet reports in the measurement or reporting process (e.g. due to resource contention in the host of these processes), or loss

of export packets in transmission or collection. See [RFC-3176] for further details.

As an example, consider a set of n consecutive packet reports r1, r2,... , rn, selected by a sampling operation and received at a collector. Let s1, s2,..., sn be the input sequence numbers reported by the packets. The attained selection frequency, taking into account both packet sampling at the observation point and selection arising from loss in transmission, is R = (n-1)/(sn-s1). (Note R would be 1 if all packets were selected and there were no transmission loss).

The attained selection frequency can be used to estimate the number bytes present in a portion of the observed packet stream. Let b1, b2,..., bn be the bytes reported in each of the packets that reached the collector, and set B = b1+b2+...+bn. Then the total bytes present in packets in the observed packet stream whose input sequence numbers lie between s1 and sn is estimated by B/R, i.e, scaling up the measured bytes through division by the attained selection frequency.

With composite selectors, and input sequence number must be reported for each selector in the composition.

5.5 Composite Selectors

The ability to compose selectors in a selection process should be provided. The following combinations appear to be most useful for applications:

* filtering followed by sampling

* sampling followed by filtering

Composite selectors are useful for drill down applications. The first component of a composite selector can be used to reduce the load on the second component. In this setting, the advantage to be gained from a given ordering can depend on the composition of the packet stream.

5.6 Constraints on the Sampling Frequency

Sampling at full line rate, i.e. with probability 1, is not excluded in principle, although resource constraints may not support it in practice.

6. Reporting Process

This section detailed specific requirements for the reporting process, motivated by the generic requirement of Section 3.4

6.1 Mandatory Contents of Packet Reports

The reporting process must include the following in each packet report:

(i) the input sequence number(s) of any sampling operation that acted on the packet in the instance of a measurement process of which the reporting process is a component.

The reporting process must support inclusion of the following in each packet report, as a configurable option:

(ii) a basic report on the packet, i.e., some number of contiguous bytes from the start of the packet, including the packet header (which includes link layer, network layer and other encapsulation headers) and some subsequent bytes of the packet payload.

Some devices hosting reporting processes may not have the resource capacity or functionality to provide more detailed packet reports that those in (i) and (ii) above. Using this minimum required reporting functionality, the reporting process places the burden of interpretation on the collector, or on applications that it supplies. Some devices may have the capability to provide extended packet reports, described in the next section.

6.2 Extended Packet Reports

The reporting process may support inclusion in packet reports of the following information, inclusion any or all being configurable as an option.

(iii) fields relating to the following protocols used in the

packet: IPv4, IPV6, transport protocols, MPLS.

(iv) packet treatment, including:

- identifiers for any input and output interfaces of the observation point that were traversed by the packet

- source and destination BGP AS

(v) selection state associated with the packet, including:

- the timestamp of observation of the packet at the observation point. The timestamp should be reported to microsecond resolution.

- hashes, where calculated.

It is envisaged that selection of fields for extended packet reporting may be used to reduce reporting bandwidth, in which case the option to report information in (ii) may not be exercised.

6.3 Extended Packet Reports in the Presence of IPFIX

If an IPFIX metering process is supported at the observation point, then in order to be PSAMP compliant, extended packet reports must be able to include all fields required in the IPFIX information model [IPFIX-INFO], with modifications appropriate to reporting on single packets rather than flows.

6.4 Report Interpretation

Information for use in report interpretation must include

(i) configuration parameters of the selectors of the packets reported on.

(ii) format of the packet report;

(iii) indication of the inherent accuracy of the reported quantities, e.g., of the packet timestamp.

(iv) identifiers for observation point, measurement process, and exporting process.

The accuracy measure in (iii) is of fundamental importance for estimating the likely error attached to estimates formed from the packet reports by applications.

Identifiers in (iv) are necessary, e.g., in order to match packet reports to the selection process that selected them. For example, when packet reports due to a sampling operation suffer loss (either during export, or in transit) it may be desirable to reconfigure downwards the sampling rate on the selection process that selected them.

The requirements for robustness and transparency are motivations for including report interpretation in the report stream. Inclusion makes the report stream self-defining. The PSAMP framework excludes reliance on an alternative model in which interpretation is recovered out of band. This latter approach is not robust with respect to undocumented changes in selector configuration, and may give rise to future architectural problems for network management systems to coherently manage both configuration and data collection.

It is not envisaged that all report interpretation be included in every packet report. Many of the quantities listed above are expected to be relatively static; they could be communicated periodically, and upon change.

6.5 Export Packet Compression

To conserve network bandwidth and resources at the collector, the export packets may be compressed before export. Compression is expected to be quite effective since the sampled packets may share many fields in common, e.g. if a filter focuses on packets with certain values in particular header fields. Using compression, however, could impact the timeliness of packet reports. Any consequent delay must not violate the timeliness requirement for availability of packet reports at the collector.

7. Parallel Measurement Processes

Because of the increasing number of distinct measurement applications, with varying requirements, it is desirable to set up parallel measurement processes on given observed packet stream. A device capable of hosting a measurement process should be able to support more than one independently configurable measurement process simultaneously. Each such measurement process should have the option of being equipped with its own exporting process; otherwise the parallel measurement processes may share the same exporting process.

Each of the parallel measurement processes should be independent. However, resource constraints may prevent complete reporting on a packet selected by multiple selection processes. In this case,

reporting for the packet must be complete for at least one measurement process; other measurement processes need only record that they selected the packet, e.g., by incrementing a counter. The priority amongst measurement processes under resource contention should be configurable.

It is not proposed to standardize the number of parallel measurement processes.

8. Exporting Process

This section detailed specific requirements for the exporting process, motivated by the generic requirements of Section 3.6

8.1 Use of IPFIX

PSAMP will use the IP Flow Information eXport (IPFIX) protocol for export of the report stream. The IPFIX protocol is well suited for this purpose, because the IPFIX architecture matches the PSAMP architecture very well and the means provided by the IPFIX protocol are sufficient.

8.2 Congestion-aware Unreliable Transport

The export of the report stream does not require reliable export. Section 5.4 shows that the use of input sequence number in packet selectors means that the ability to estimate traffic rates is not impaired by export loss. Export packet loss becomes another form of sampling, albeit a less desirable, and less controlled, form of sampling.

On the contrary, retransmission of lost export packets consumes additional network resources. The requirement to store unacknowledged data is an impediment to having ubiquitous support for PSAMP.

In order to jointly satisfy the timeliness and congestion avoidance requirements of Section 4.3, a congestion aware unreliable transport protocol must be used. IPFIX is compatible with this requirement, since it mandates support of the Stream

Control Transmission Protocol (SCTP) [SCTP] and the SCTP Partial Reliability Extension [RFC-3758]. IPFIX also allows the use of User Datagram Protocol (UDP) [UDP] although it is not a congestion aware protocol. However, in this case, the Export Packets must remain wholly within the administrative domains of the operators [IPFIX-PROTO].

8.3 Limiting Delay for Export Packets

Low measurement latency allows the traffic monitoring system to be more responsive to real-time network events, for example, in quickly identifying sources of congestion. Timeliness is

generally a good thing for devices performing the sampling since it minimizes the amount of memory needed to buffer samples.

Keeping the packet dispatching delay small has other benefits besides limiting buffer requirements. For many applications a resolution of 1 second is sufficient. Applications in this category would include: identifying sources associated with congestion; tracing denial of service attacks through the network and constructing traffic matrices. Furthermore, keeping dispatch delay within the resolution required by applications eliminates the need for timestamping by synchronized clocks at observation points, or for the observation points and collector to maintain bi-directional communication in order to track clock offsets. The collector can simply process packet reports in the order that they are received, using its own clock as a "global" time base. This avoids the complexity of buffering and reordering samples. See [DuGeGr02] for an example.

The delay between observation of a packet and transmission of a export packet containing a report on that packet has several components. It is difficult to standardize a given numerical delay requirement, since in practice the delay may be sensitive to processor load at the observation point. Therefore, PSAMP aims to control that portion of the delay within the observation point that is due to buffering in the formation and transmission of export packets.

In order to limit delay in the formation of export packets, the exporting process must provide the ability to close out and enqueue for transmission any export packet in formation as soon as it includes one packet report. This could be achieved, for

example, by the following means:

- the number of packet reports per export packet is not to exceed a maximum value, which can be configured to take the value 1.

- the ability to exclude report interpretation from any export packet that contains a packet report;

In order to limit the delay in the transmission of export packets, a configurable upper bound to the delay of an export packet prior to transmission must be provided. If the bound is exceeded the export packet is dropped. This functionality can be provided by the timed reliability service of the SCTP Partial Reliability Extension [RFC-3758].

The exporting process may queue the report stream in order to export multiple packet reports in a single export packet. Any consequent delay must still allow for timely availability of packet reports as just described. The timed reliability service of the SCTP Partial Reliability Extension [RFC-3758] allows from

the dropping of packets from the export buffer once their age in the buffer exceeds a configurable bound.

8.4 Configurable Export Rate Limit

The exporting process must have an export rate limit, configurable per exporting process. This is useful for two reasons:

(i) Even without network congestion, the rate of packet selection may exceed the capacity of the collector to process reports, particularly when many exporting processes feed a common collector. Use of an export rate limit allows control of the global input rate to the collector.

(ii) IPFIX provides export using UDP as the transport protocol in some circumstances. An export rate limit allows the capping of the export rate to match both path link speeds and the capacity of the collector.

8.5 Collector Destination

When exporting to a remote collector, the collector is identified by IP address, transport protocol, and transport port number.

8.6 Local Export

The report stream may be directly exported to on-board measurement based applications, for example those that form composite statistics from more than one packet. Local export may be presented through an interface direct to the higher level applications, i.e., through an API, rather than employing the transport used for off-board export. Specification of such an API is outside the scope of the PSAMP framework.

A possible example of local export could be that packets selected by the PSAMP measurement process serve as the input for the IPFIX protocol, which then forms flow records out of the stream of selected packets.

9. Configuration and Management

A key requirement for PSAMP is the easy reconfiguration of the parameters of the measurement process: those for selection, packet reports and export. Examples are

(i) support of measurement-based applications that want to drill-down on traffic detail in real-time;

(ii) collector-based rate reconfiguration.

To facilitate reconfiguration and retrieval of parameters, they are to reside in a Management Information Base (MIB). Mandatory configuration, capabilities and monitoring objects will cover all mandatory PSAMP functionality.

Secondary objects will cover the recommended and optional PSAMP functionality, and must be provided when such functionality is offered by a PSAMP device. Such PSAMP functionality includes configuration of offered selectors, composite selectors, multiple measurement processes, and report format including the choice of fields to be reported. For further details concerning the PSAMP MIB, see [PSAMP-MIB].

PSAMP requires a uniform mechanism with which to access and configure the MIB. SNMP access must be provided by the host of the MIB.

10. Feasibility and Complexity

In order for PSAMP to be supported across the entire spectrum of networking equipment, it must be simple and inexpensive to implement. One can envision easy-to-implement instances of the mechanisms described within this draft. Thus, for that subset of instances, it should be straightforward for virtually all system vendors to include them within their products. Indeed, sampling and filtering operations are already realized in available equipment.

Here we give some specific arguments to demonstrate feasibility and comment on the complexity of hardware implementations. We stress here that the point of these arguments is not to favor or recommend any particular implementation, or to suggest a path for standardization, but rather to demonstrate that the set of possible implementations is not empty.

10.1 Feasibility

10.1.1 Filtering

Filtering consists of a small number of mask (bit-wise logical), comparison and range (greater than) operations. Implementation of at least a small number of such operations is straightforward. For example, filters for security access control lists (ACLs) are widely implemented. This could be as simple as an exact match on certain fields, or involve more complex comparisons and ranges.

10.1.2 Sampling

Sampling based on either counters (counter set, decrement, test for equal to zero) or range matching on the hash of a packet (greater than) is possible given a small number of selectors,

although there may be some differences in ease of implementation for hardware vs. software platforms.

10.1.3 Hashing

Hashing functions vary greatly in complexity. Execution of a small number of sufficient simple hash functions is implementable at line rate. Concerning the input to the hash function, hop-invariant IP header fields (IP address, IP identification) and TCP/UDP header fields (port numbers, TCP sequence number) drawn from the first 40 bytes of the packet have been found to possess a considerable variability; see [DuGr01].

10.1.4 Reporting

The simplest packet report would duplicate the first n bytes of the packet. However, such an uncompressed format may tax the bandwidth available to the reporting process for high sampling rates; reporting selected fields would save on this bandwidth. Thus there is a trade-off between simplicity and bandwidth limitations.

10.1.5 Export

Ease of exporting export packets depends on the system architecture. Most systems should be able to support export by insertion of export packets, even through the software path.

10.2 Potential Hardware Complexity

We now comment on the complexity of possible hardware implementations. Achieving low constants for performance while minimizing hardware resources is, of course, a challenge, especially at very high clock frequencies. Most of these operations, however, are very basic and their implementations very well understood; in fact, the average ASIC designer simply uses canned library instances of these operations rather than design them from scratch. In addition, networking equipment generally does not need to run at the fastest clock rates, further reducing the effort required to get reasonably efficient implementations.

Simple bit-wise logical operations are easy to implement in hardware. Such operations (NAND/NOR/XNOR/NOT) directly translate to four-transistor gates. Each bit of a multiple-bit logical operation is completely independent and thus can be performed in parallel incurring no additional performance cost above a single bit operation.

Comparisons (EQ/NEQ) take O(lg(M)) stages of logic, where M is the number of bits involved in the comparison. The lg(M) is required to accumulate the result into a single bit.

Greater than operations, as used to determine whether a hash falls in a selection range, are a determination of the most significant not-equivalent bit in the two operands. The operand with that most-significant-not-equal bit set to be one is greater than the other. Thus, a greater than operation is also an O(lg(M)) stages of logic operation. Optimized implementations of arithmetic operations are also O(lg(M)) due to propagation of the carry bit.

Setting a counter is simply loading a register with a state. Such an operation is simple and fast O(1). Incrementing or decrementing a counter is a read, followed by an arithmetic operation followed by a store. Making the register dual-ported does take additional space, but it is a well-understood technique. Thus, the increment/decrement is also an O(lg(M)) operation.

Hashing functions come in a variety of forms. The computation involved in a standard Cyclic Redundancy Code (CRC) for example are essentially a set of XOR operations, where the intermediate result is stored and XORed with the next chunk of data. There are only O(1) operations and no log complexity operations. Thus, a simple hash function, such as CRC or generalizations thereof, can be implemented in hardware very efficiently.

At the other end of the range of complexity, the MD5 function uses a large number of bit-wise conditional operations and arithmetic operations. The former are O(1) operations and the latter are O(lg(M)). MD5 specifies 256 32b ADD operations per 16B of input processed. Consider processing 10Gb/sec at 100MHz (this processing rate appears to be currently available). This requires processing 12.5B/cycle, and hence at least 200 adders, a sizeable number. Because of data dependencies within the MD5 algorithm, the adders cannot be simply run in parallel, thus requiring either faster clock rates and/or more advanced architectures. Thus, selection hashing functions as complex as MD5 may be precluded for ubiquitous use at full line rate. This motivates exploring the use of selection hash functions with complexity somewhere between that of MD5 and CRC. However, identification hashing with MD5 on only selected packets is feasible at a sufficiently low sampling frequency.

11. Applications

We first describe several representative operational applications that require traffic measurements at various levels of temporal

and spatial granularity. Some of the goals here appear similar to those of IPFIX, at least in the broad classes of applications supported. The major benefit of PSAMP is the support of new network management applications, specifically, those enabled by the packet selectors that it supports.

11.1 Baseline Measurement and Drill Down

Packet sampling is ideally suited to determine the composition of the traffic across a network. The approach is to enable measurement on a cut-set of the network links such that each packet entering the network is seen at least once, for example, on all ingress links. Unfiltered sampling with a relatively low frequency establishes baseline measurements of the network traffic. Packet reports include packet attributes of common interest: source and destination address and port numbers, prefix, protocol number, type of service, etc. Traffic matrices are indicated by reporting source and destination AS matrices. Absolute traffic volumes are estimated by renormalizing the sampled traffic volumes through division by either the target sampling frequency, or by the attained sampling frequency (as derived by interface packet counters included in the report stream)

Suppose an operator or a measurement-based application detects an interesting subset of a packet stream, as identified by a particular packet attribute. Real-time drill-down to that subset is achieved by instantiating a new measurement process on the same packet stream from which the subset was reported. The selection process of the new measurement process filters according to the attribute of interest, and composes with sampling if necessary to manage the frequency of packet selection.

11.2 Trajectory Sampling

Trajectory sampling is the selection of a subset of packets at either all of a set of observation points or none of them. Trajectory sampling is realized by hash-based sampling if all observation points in the set apply a common hash function to a portion of the packet content that is invariant along the packet path. (Thus, fields such at TTL and CRC are excluded).

The trajectory followed by a packet is reconstructed from PSAMP reports on it that reach the collector. Reports on a given packet are associated either by matching a label comprising the invariant reported packet content, or possibly some digest of it. The reconstruction of trajectories, and methods for dealing with possible ambiguities due to label collisions (identical labels reported by different packets) and potential loss of reports in transmission are dealt with in [DuGr01], [DuGeGr02] and [DuGr04].

11.3 Passive Performance Measurement

Trajectory sampling enables the tracking of the performance experience by customer traffic, customers identified by a list of source or destination prefixes, or by ingress or egress

interfaces. Operational uses include the verification of Service Level Agreements (SLAs), and troubleshooting following a customer complaint.

In this application, trajectory sampling is enabled at all network ingress and egress interfaces. Rates of loss in transit between ingress and egress are estimated from the proportion of trajectories for which no egress report is received. Note that loss of customer packets is distinguishable from loss of packet reports through use of report sequence numbers. Assuming synchronization of clocks between different entities, delay of customer traffic across the network may also be measured; see [Zs02].

Extending hash-selection to all interfaces in the network would enable attribution of poor performance to individual network links.

11.4 Troubleshooting

PSAMP reports can also be used to diagnose problems whose occurrence is evident from aggregate statistics, per interface utilization and packet loss statistics. These statistics are typically moving averages over relatively long time windows, e.g., 5 minutes, and serve as a coarse-grain indication of operational health of the network. The most common method of obtaining such measurements are through the appropriate SNMP MIBs (MIB-II [RFC-1213] and vendor-specific MIBs.)

Suppose an operator detects a link that is persistently overloaded and experiences significant packet drop rates. There is a wide range of potential causes: routing parameters (e.g., OSPF link weights) that are poorly adapted to the traffic matrix, e.g., because of a shift in that matrix; a denial of service attack or a flash crowd; a routing problem (link flapping). In most cases, aggregate link statistics are not sufficient to distinguish between such causes, and to decide on an appropriate corrective action. For example, if routing over two links is unstable, and the links flap between being overloaded and inactive, this might be averaged out in a 5 minute window, indicating moderate loads on both links.

Baseline PSAMP measurement of the congested link, as described in Section 11.1, enables measurements that are fine grained in both space and time. The operator has to be able to determine how many bytes/packets are generated for each source/destination address, port number, and prefix, or other attributes, such as protocol number, MPLS forwarding equivalence class (FEC), type of service, etc. This allows the precise determination of the nature of the offending traffic. For example, in the case of a Distributed Denial of Service(DDoS) attack, the operator would see a

significant fraction of traffic with an identical destination address.

In certain circumstances, precise information about the spatial flow of traffic through the network domain is required to detect and diagnose problems and verify correct network behavior. In the case of the overloaded link, it would be very helpful to know the precise set of paths that packets traversing this link follow. This would readily reveal a routing problem such as a loop, or a link with a misconfigured weight. More generally, complex diagnosis scenarios can benefit from measurement of traffic intensities (and other attributes) over a set of paths that is constrained in some way. For example, if a multihomed customer complains about performance problems on one of the access links from a particular source address prefix, the operator should be able to examine in detail the traffic from that source prefix which also traverses the specified access link towards the customer.

While it is in principle possible to obtain the spatial flow of

traffic through auxiliary network state information, e.g., by downloading routing and forwarding tables from routers, this information is often unreliable, outdated, voluminous, and contingent on a network model. For operational purposes, a direct observation of traffic flow provided by trajectory sampling is more reliable, as it does not depend on any such auxiliary information. For example, if there was a bug in a router's software, direct observation would allow the diagnosis the effect of this bug, while an indirect method would not.

12. Security Considerations

Security considerations are addressed in:

- Section 4.1: item Robust Selection
- Section 4.3: item Secure Export
- Section 4.4: item Secure Configuration

13. Normative References

[PSAMP-TECH] T. Zseby, M. Molina, F. Raspall, N. G. Duffield, Sampling and Filtering Techniques for IP Packet Selection, RFC XXXX. [Currently Internet Draft, draft-ietf-psamp-sample-tech-04.txt, work in progress, February 2004.

[PSAMP-MIB] T. Dietz, B. Claise, Definitions of Managed Objects for Packet Sampling, RFC XXXX. [Currently Internet Draft, draft-ietf-psamp-mib-03.txt, work in progress, July 2004.]

[PSAMP-PROTO] B. Claise (Ed.) Packet Sampling (PSAMP) Protocol Specifications, RFC XXXX. [Currently Internet Draft draft-ietf-psamp-protocol-01.txt, work in progress, February 2004.]

[PSAMP-INFO] T. Dietz, F. Dressler, G. Carle, B. Claise, Information Model for Packet Sampling Exports, RFC XXXX. [Currently Internet Draft, draft-ietf-psamp-info-02, July 2004

14. Informative References

[B88] R.T. Braden, A pseudo-machine for packet monitoring and statistics, in Proc ACM SIGCOMM 1988

[IPFIX-INFO] Calato, P, Meyer, J, Quittek, J, "Information Model for IP Flow Information Export" draft-ietf-ipfix-info-04, November 2003

[ClPB93] K.C. Claffy, G.C. Polyzos, H.-W. Braun, Application of Sampling Methodologies to Network Traffic Characterization, Proceedings of ACM SIGCOMM'93, San Francisco, CA, USA, September 13-17, 1993

[IPFIX-PROTO] B. Claise, B. Stewart, G. Sadasivan, M. Fullmer,P. Calato , R. Penno, IPFIX Protocol Specifications , Internet Draft, draft-ietf-ipfix-protocol-05.txt, August 2004.

[RFC-2460] S. Deering, R. Hinden, Internet Protocol, Version 6 (IPv6) Specification, RFC 2460, December 1998.

[DuGr01] N. G. Duffield and M. Grossglauser, Trajectory Sampling for Direct Traffic Observation, IEEE/ACM Trans. on Networking, 9(3), 280-292, June 2001.

[DuGeGr02] N.G. Duffield, A. Gerber, M. Grossglauser, Trajectory Engine: A Backend for Trajectory Sampling, IEEE Network Operations and Management Symposium 2002, Florence, Italy, April 15-19, 2002.

[DuGr04] N. G. Duffield and M. Grossglauser, Trajectory Sampling with Unreliable Reporting, Proc IEEE Infocom 2004, Hong Kong, March 2004,

[RFC-2914] S. Floyd, Congestion Control Principles, RFC 2914, September 2000.

[RFC-2804] IAB and IESG, Network Working Group, IETF Policy on Wiretapping, RFC 2804, May 2000

[RFC-1213] - K. McCloghrie, M. Rose, Management Information Base for Network Management of TCP/IP-based internets:MIB-II, RFC 1213, March 1991.

[RFC-3176] P. Phaal, S. Panchen, N. McKee, InMon Corporation's sFlow: A Method for Monitoring Traffic in Switched and Routed Networks, RFC 3176, September 2001

[RFC-2330] V. Paxson, G. Almes, J. Mahdavi, M. Mathis, Framework for IP Performance Metrics, RFC 2330, May 1998

[RFC-791] J. Postel, "Internet Protocol", STD 5, RFC 791, September 1981.

[UDP] Postel, J., "User Datagram Protocol" RFC 768, August 1980

[IPFIX-REQUIRE] J. Quittek, T. Zseby, B. Claise, S. Zander, Requirements for IP Flow Information Export, Internet Draft draft-ietf-ipfix-reqs-16.txt, work in progress, June 2004.

[RFC1771] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[RFC-3031] Rosen, E., Viswanathan, A. and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[SPSJTKS01] A. C. Snoeren, C. Partridge, L. A. Sanchez, C. E. Jones, F. Tchakountio, S. T. Kent, W. T. Strayer, Hash-Based IP Traceback, Proc. ACM SIGCOMM 2001, San Diego, CA, September 2001.

[RFC-2960] R. Stewart, (ed.) "Stream Control Transmission Protocol", RFC 2960, October 2000.

[RFC-3758] R. Stewart, M. Ramalho, Q. Xie, M. Tuexen, P. Conrad, "SCTP Partial Reliability Extension", RFC 3758, May 2004.

[Zs02] T. Zseby, ``Deployment of Sampling Methods for SLA Validation with Non-Intrusive Measurements'', Proceedings of Passive and Active Measurement Workshop (PAM 2002), Fort Collins, CO, USA, March 25-26, 2002

15. Authors' Addresses

Derek Chiou
Avici Systems
101 Billerica Ave
North Billerica, MA 01862
Phone: +1 978-964-2017
Email: dchiou@avici.com

Benoit Claise
Cisco Systems
De Kleetlaan 6a b1
1831 Diegem
Belgium
Phone: +32 2 704 5622
Email: bclaise@cisco.com

Nick Duffield
AT&T Labs - Research
Room B-139
180 Park Ave
Florham Park NJ 07932, USA
Phone: +1 973-360-8726
Email: duffield@research.att.com

Albert Greenberg
AT&T Labs - Research
Room A-161
180 Park Ave
Florham Park NJ 07932, USA
Phone: +1 973-360-8730
Email: albert@research.att.com

Matthias Grossglauser
School of Computer and Communication Sciences
EPFL
1015 Lausanne
Switzerland
Email: matthias.grossglauser@epfl.ch

Peram Marimuthu
Cisco Systems
170, W. Tasman Drive
San Jose, CA 95134
Phone: (408) 527-6314

Email: peram@cisco.com

Jennifer Rexford
AT&T Labs - Research
Room A-169
180 Park Ave
Florham Park NJ 07932, USA
Phone: +1 973-360-8728
Email: jrex@research.att.com

Ganesh Sadasivan
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
Phone: (408) 527-0251
Email: gsadasiv@cisco.com

16. Intellectual Property Statements

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of RFC 3668.

The IETF has been notified by AT&T Corp. of intellectual property rights claimed in regard to some or all of the specification contained in this document. For more information, see http://www.ietf.org/ietf/IPR/att-ipr-draft-ietf-psamp-framework.txt

The IETF has been notified by Cisco Corp. of intellectual property rights claimed in regard to some or all of the specification contained in this document. For more information, see http://www.ietf.org/ietf/IPR/cisco-ipr-draft-ietf-psamp-protocol.txt

17. Full Copyright Statement

rights.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns.
