Network Working Group
Internet-Draft
Intended status: Informational
Expires: 5 December 2026

L. Zhu
X. Chen
Zhejiang University

3 June 2026

A Framework for Co-Designing Sketch and In-Band Network Telemetry for Accurate Network Measurement
draft-zhu-sketch-int-codesign-00

Abstract

Network measurement is a fundamental building block for network management applications. Existing measurement techniques face a trade-off between accuracy and resource efficiency: sketch-based techniques achieve high accuracy for large flows but degrade for small flows, while In-band Network Telemetry (INT) measures every flow accurately but at the cost of significant bandwidth and control plane resources.

This document describes a framework for co-designing sketches and INT to measure large and small flows respectively, achieving both high accuracy and resource efficiency. It addresses two key challenges: (1) where to deploy measurement functions when routing information is incomplete, and (2) how to collect measurement data without causing network congestion.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 5 December 2026.

Copyright Notice

Table of Contents

1. Introduction

Network measurement collects traffic statistics such as per-flow packet counts from switches and periodically reports them to the control plane. The control plane provides these data to network management applications that identify events of interest and make corresponding decisions, including heavy hitter detection, DDoS detection, congestion control, and flow size distribution estimation.

Accurate measurement of both large flows and small flows is critical. Large flows (i.e., flows comprising many packets) are important for volumetric applications such as heavy hitter and superspreader detection. Small flows are essential for applications such as flow size distribution estimation and congestion control, which require visibility into the long tail of the flow size distribution.

However, existing measurement techniques face a fundamental trade-off between accuracy and resource efficiency:

* Sketch-based techniques [CM-Sketch] [Count-Sketch] [Elastic-Sketch] use compact probabilistic data structures to achieve accurate measurement of large flows with low resource consumption. However, due to hash collisions in memory-constrained switch environments, sketches exhibit significant errors when measuring small flows.

* In-band Network Telemetry (INT) [INT-Spec] records per-flow statistics within packet headers and extracts them at network egress. INT preserves full accuracy for every flow but generates high volumes of telemetry data that consume significant bandwidth and control plane resources.

Recent hybrid approaches [SketchINT] [LightGuardian] attempt to combine sketches and INT but still inherit limitations from one or both techniques.

This document describes a measurement framework that co-designs sketches and INT by assigning each technique to the flow type it handles best: sketches for large flows and INT for small flows. The framework addresses two optimization challenges: measurement point selection under incomplete routing knowledge, and congestion-free collection of measurement data.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Sketch: A probabilistic data structure that maintains approximate traffic statistics using compact memory. Examples include Count-Min Sketch and Count Sketch.

In-band Network Telemetry (INT): A technique where each switch along

a packet's path appends measurement metadata to the packet header. At the egress switch, the accumulated metadata is extracted and reported to the control plane.

Large Flow: A flow whose packet count exceeds a predefined threshold, indicating it contributes a significant portion of total traffic volume.

Small Flow: A flow whose packet count is below the large flow threshold.

Measurement Point: A programmable switch that executes sketch and/or INT functions to collect traffic statistics.

Control Plane Node: A server or controller that receives measurement data from switches and runs network management applications.

OD Pair: An Origin-Destination pair representing the ingress and egress switches of a specific flow.

Flow Coverage: The fraction of flows in the network that are measured by at least one measurement point.

3. Problem Statement

3.1. Limitations of Sketch-Only Measurement

Sketches achieve high accuracy for large flows because large flow counters dominate over noise from hash collisions. However, when measuring small flows, the few packets in each small flow are easily overwhelmed by collisions with large flow data.

In resource-constrained switch environments, sketch memory is typically limited to a few megabytes per switch. Under such constraints, experimental evaluation shows that fewer than 50% of flows achieve measurement errors below 10% when using state-of-the-art sketches with 10 MB memory per switch.

Recent techniques such as compressive sensing-based sketches and learning-based sketches attempt to mitigate this issue but require specific sketch designs (limiting generality), involve complex data structures that hinder hardware implementation, or require recovery times on the order of tens of seconds (unsuitable for latency-sensitive applications).

3.2. Limitations of INT-Only Measurement

INT achieves full accuracy for every flow by piggybacking metadata on each packet. However, this per-packet monitoring generates significant overhead.

According to the INT protocol specification [INT-Spec], each switch adds a 12-byte INT header to each packet. In modern networks transferring Tbps-level traffic, the total number of packets per second exceeds 10^9, producing a corresponding volume of INT headers. This accumulation creates non-trivial pressure on both network bandwidth (for transferring INT headers) and control plane resources (for processing them).

Existing optimizations such as sampling-based INT reduce overhead but degrade accuracy, particularly for small flows with few packets. Path-based INT optimizations reduce redundancy but still suffer from per-packet overhead for large flows.

3.3. Limitations of Existing Hybrid Approaches

Two categories of hybrid approaches have been proposed:

* Control-plane sketch aggregation: These approaches activate INT at data plane switches and build sketches at the control plane to aggregate collected INT data. However, they inherit the bandwidth overhead of full INT and additionally lose small flow accuracy through sketch aggregation.

* INT-embedded sketch data: These approaches encode sketch data into INT headers for efficient collection. While this reduces bandwidth overhead, it retains the fundamental accuracy limitations of sketches for small flows.

Neither approach effectively leverages the complementary strengths of sketches and INT.

4. Framework Overview

4.1. Design Goals

The framework aims to achieve two goals:

G1: High Accuracy. Measure both large flows and small flows with high accuracy.

G2: Resource Efficiency. Avoid excessive bandwidth and control plane resource consumption during measurement data collection.

4.2. Architecture

The core observation is that sketches and INT are complementary:

* Sketches offer high accuracy and resource efficiency for large flows, but fall short for small flows.

* INT provides full accuracy for all flows, but at high resource cost that scales with total packet volume.

In modern networks, traffic is typically skewed: most packets come from a small number of large flows [Traffic-Skew]. This skewness enables the following assignment:

* Large flows are measured by sketches, which provide high accuracy and resource efficiency for these flows.

* Small flows are measured by INT, which preserves full accuracy. Because small flows collectively contribute few packets, INT's resource consumption remains bounded.

This co-design achieves both G1 and G2 simultaneously.

The framework operates in a general architecture comprising two planes:

Data Plane: Programmable switches execute both sketch and INT functions. Each incoming flow is initially measured by both sketches and INT. Once a flow is identified as a large flow by the sketch, subsequent packets of that flow are recorded only by the sketch, and INT processing for that flow is deactivated.

Control Plane: A cluster of servers receives measurement data from switches, runs network management applications, and provides query interfaces for traffic statistics.

4.3. Workflow

The framework operates in four steps:

Step 1: Configuration. The administrator specifies which sketch and INT techniques to deploy. The framework supports arbitrary combinations of sketch types (e.g., Count-Min, Count Sketch, Elastic Sketch) and INT variants.

Step 2: Measurement Point Selection. Given the network topology and

a set of OD pairs, the framework selects which programmable switches to deploy sketch and INT functions on. The selection maximizes flow coverage while minimizing the distance between measurement points and control plane nodes. This step handles the case where precise routing information is unknown (Section 5).

Step 3: Data Plane Execution. On each selected switch, every incoming packet is processed as follows:

a. The packet is inserted into the sketch data structure.

b. The sketch checks whether the flow is a large flow (i.e., exceeds the threshold).

c. If the flow is NOT a large flow, INT metadata is appended to the packet header.

d. If the flow IS a large flow, the packet is forwarded without INT headers. The sketch maintains approximate statistics for this flow.

e. At the egress switch, INT headers (if present) are extracted and queued for transmission to the control plane.

Step 4: Data Collection. The framework selects network paths to transfer measurement data (both periodic sketch dumps and INT reports) from switches to control plane nodes. Path selection ensures that measurement data traffic does not congest normal data plane traffic (Section 6).

5. Measurement Point Selection

5.1. Problem Formulation

The measurement point selection problem determines which programmable switches in the network should deploy sketch and INT functions.

Input:

* Network topology G = (V, E), where V is the set of switches and E is the set of links.

* The set P of programmable switches (P is a subset of V) capable of running sketches and INT.

* The set C of control plane nodes (C is a subset of V).

* A set F of flows, where each flow f is characterized by an OD pair (o_f, d_f), representing the ingress and egress switches.

* For each flow f, the set P_f of programmable switches on any shortest path between o_f and d_f.

* A distance metric delta(p, c) between switch p and control plane node c (e.g., hop count).

Objectives:

* Maximize flow coverage: measure as many flows as possible.

* Minimize collection distance: reduce the distance between measurement points and control plane nodes to enable timely data collection.

The two objectives are balanced by a user-configurable parameter alpha in [0, 1].

Constraints:

* A flow is covered if at least one measurement point exists on any shortest path connecting its OD pair.

* Each selected measurement point MUST be assigned to exactly one control plane node for data reporting.

* Decision variables are binary: each switch is either selected or not.

This problem is NP-hard, as it reduces to a combination of the set cover problem (for coverage maximization) and the uncapacitated facility location problem (for distance minimization).

5.2. Optimization via Lagrangian Relaxation

Given the NP-hardness, the framework employs Lagrangian relaxation to obtain near-optimal solutions in polynomial time.

The coverage constraints are relaxed using Lagrange multipliers. The relaxed problem decomposes into independent per-switch decisions:

* For each switch p, a "switch penalty" is computed as the sum of Lagrange multipliers over all flows that could be measured at p.

* A "collection cost" is computed as the minimum distance from p to any control plane node.

* Switch p is selected if and only if the switch penalty exceeds the collection cost.

The Lagrange multipliers are iteratively updated using subgradient optimization. At each iteration:

1. The Lagrangian subproblem is solved to obtain primal variables and the dual bound.

2. Subgradients are computed based on constraint violations.

3. Multipliers are updated with a diminishing step size.

4. The best solution across iterations is recorded.

This procedure yields solutions with bounded optimality gaps, as guaranteed by weak duality: the optimal dual value provides a lower bound on the optimal primal value.

6. Congestion-Free Data Collection

After selecting measurement points, the framework determines how to transfer measurement data from switches to control plane nodes without causing network congestion.

6.1. Worst-Case Rate Estimation

The framework estimates the maximum possible sending rate of measurement data at each switch.

For sketch data, the worst-case sending rate is determined by the sketch memory size divided by the collection interval: gamma_sketch = S / T, where S is the sketch size in bytes and T is the collection window in seconds. For example, a 10 MB sketch with a 1 ms window produces a worst-case rate of 8 Gbps.

For INT data, the worst-case rate at switch p is: gamma_INT = (C_p * phi / mu) * B_INT, where C_p is the link bandwidth capacity, phi is the maximum fraction of bandwidth consumed by small flows (obtainable from historical traffic analysis), mu is the average packet size of small flows, and B_INT is the INT header size per packet.

The total worst-case rate at switch p is the sum of sketch and INT rates across all deployed measurement functions.

6.2. Dynamic Path Selection

Given the worst-case rate estimates, the framework selects network paths for measurement data transfer with the goal of avoiding congestion.

The path selection problem is formulated as: minimize the total queue depth across all links, subject to the constraint that measurement data traffic on each link, combined with normal data traffic, MUST NOT exceed a safety threshold (e.g., 80% of link capacity).

The framework computes candidate paths (e.g., K-shortest paths) between each measurement point and each control plane node. It then determines splitting ratios that distribute measurement data across these paths.

At each time step, the following inputs are collected:

* Current data plane traffic utilization on each link.

* Current queue depth at each switch.

* Worst-case measurement data rates.

Based on these inputs, the path selection algorithm outputs:

* Selected paths from each measurement point to the assigned control plane node.

* Splitting ratios for distributing measurement data.

If the measurement data rate on any link would exceed the safety threshold, the splitting ratios are scaled down proportionally and renormalized to ensure compliance.

The path selection process operates at sub-second timescales to adapt to changing traffic conditions.

7. Applicability

The framework is applicable to the following network management scenarios:

Volumetric Applications: Heavy hitter detection, superspreader detection, DDoS flow detection, and per-flow counting benefit from accurate large flow measurement (provided by sketches) combined with accurate small flow visibility (provided by INT).

Aggregated Applications: Entropy estimation and flow size distribution estimation require accurate statistics across all flow sizes. The framework provides near-ideal flow size distributions by preserving small flow accuracy.

Troubleshooting Applications: Microburst detection and congestion control require per-flow, per-hop metadata. INT provides this metadata for small flows while sketches efficiently summarize large flow behavior.

The framework is designed to be general-purpose:

* It supports arbitrary sketch types as pluggable components.

* It supports standard INT as well as variants such as probabilistic INT and delta-based INT.

* It operates on programmable switches (e.g., those based on the Protocol-Independent Switch Architecture) without requiring modifications to the forwarding pipeline.

Implementation experience on 12.8 Tbps programmable switches demonstrates that the framework is feasible on production-grade hardware.

8. Security Considerations

The framework inherits the security properties and risks of the underlying sketch and INT mechanisms.

Measurement data transmitted from switches to control plane nodes SHOULD be integrity-protected to prevent tampering. In environments where measurement data traverses untrusted network segments, encryption SHOULD be applied.

The large flow identification mechanism at each switch could be targeted by adversaries who craft traffic to evade classification (e.g., splitting a large flow into many small flows to force excessive INT processing). Implementations SHOULD incorporate rate limiting on INT data generation to mitigate such attacks.

The measurement point selection algorithm takes OD pairs as input. In deployments where OD pair information is sensitive, access to this information SHOULD be restricted to authorized control plane components.

9. IANA Considerations

This document has no IANA actions.

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

10.2. Informative References

[INT-Spec] The P4.org Applications Working Group, "In-band Network Telemetry (INT) Dataplane Specification, Version 2.1", 2020, <https://p4.org/p4-spec/docs/INT_v2_1.pdf>.

[CM-Sketch] Cormode, G. and S. Muthukrishnan, "An Improved Data Stream Summary: the Count-Min Sketch and its Applications", Journal of Algorithms Vol. 55, No. 1, pp. 58-75, 2005.

[Count-Sketch] Charikar, M., Chen, K., and M. Farach-Colton, "Finding Frequent Items in Data Streams", Theoretical Computer Science Vol. 312, No. 1, pp. 3-15, 2004.

[Elastic-Sketch] Yang, T., "Elastic Sketch: Adaptive and Fast Network-wide Measurements", Proceedings of ACM SIGCOMM pp. 561-575, 2018.

[SketchINT] Yang, K., "SketchINT: Empowering INT with TowerSketch for Per-flow Per-switch Measurement", IEEE TPDS Vol. 34, No. 11, 2023.

[LightGuardian] Zhao, Y., "LightGuardian: A Full-Visibility, Lightweight, In-Band Telemetry System Using Sketchlets", Proceedings of USENIX NSDI pp. 991-1010, 2021.

[Traffic-Skew]
Roy, A., "Inside the Social Network’s (Datacenter) Network", Proceedings of ACM SIGCOMM pp. 123-137, 2015.

Authors’ Addresses

Longlong Zhu
Zhejiang University
College of Computer Science and Technology
Hangzhou
Zhejiang,
China
Email: pocofo@foxmail.com

Xiang Chen
Zhejiang University
College of Computer Science and Technology
Hangzhou
Zhejiang,
China
Email: wasdnsxchen@gmail.com