A Data-Plane-Only Approach to Accurate Persistent Flow Detection on Programmable Switches in High-Speed Networks

Weihe Li[1], Beyza Bütün[2], Tianyue Chu[2], Marco Fiore[2], and Paul Patras[1]

[1]University of Edinburgh, Edinburgh, United Kingdom
[2]IMDEA Networks Institute, Madrid, Spain

Abstract

In high-speed data center networks, persistent flows are repeatedly observed over extended periods, potentially signaling threats such as stealthy DDoS or botnet attacks. Monitoring every flow in production-grade hardware switches that feature limited memory, however, is challenging under typical high flow rates and data volumes. To tackle this, approximate data structures—i.e., sketches—are often employed in cutting-edge programmable switches. Yet many existing methods rely on per-time-window flag resets, which require frequent control-plane interventions that make them unsuitable for high-speed traffic. This paper introduces Pallas, a fully data-plane-implementable sketch for detecting persistent flows in high-speed networks with high accuracy.

Introduction

Detecting persistent flows—those that remain active over extended periods—is essential for network management and security applications, including the detection and mitigation of stealthy DDoS attacks. For instance, some adversaries transmit malicious traffic at a controlled, low rate over a long duration, thereby evading traditional volume-based anomaly detectors. However, in high-speed environments, the sheer flow rates and massive data volumes make it impractical to track every flow within the constrained memory of switching infrastructure, even in cutting-edge programmable switches.

To overcome this challenge, a variety of sketch-based methods have been proposed. A sketch is a hashing-based data structure that stores flow information in limited memory while retaining acceptable accuracy. Meanwhile, programmable switches are increasingly adopted in high-speed networks due to their rapid processing capabilities and high flexibility, enabling adaptable forwarding strategies through P4-based prototyping. However, programmable switches still face stringent constraints—such as a limited number of pipeline stages and a narrow range of arithmetic operations—which complicate the implementation of existing sketch-based methods for persistent flow detection.

To surmount these constraints, we present Pallas, a new sketch specifically designed for accurate persistent flow detection on programmable switches. To the best of our knowledge, this is the first method that maintains reliable performance at high data rates in a production-grade switch. Instead of relying on per-time-window flag resets—which necessitate communication between the control plane and the data plane—we eliminate the use of flags altogether. We track persistence by using a global counter to record the current time window and storing each flow's latest arrival time window in the sketch buckets. When a flow arrives, its tracked time window is compared against the global one to determine whether it is the flow's first arrival within that window. Also, programmable switches have limited support for arithmetic operations, requiring an update strategy for sketches that is feasible to implement. To do so, we adopt a probability-based update mechanism that assigns flows with higher persistence a lower probability of being decayed and evicted, ensuring reliable performance under highly skewed traffic in practical scenarios.

We rigorously model and analyze Pallas by providing theoretical guarantees. We then implement a prototype in P4 on a production-grade Intel Tofino switch. Evaluations show that Pallas supports traffic data rates that are over 60× higher than those of the state-of-the-art method Pontus [1], while achieving 5.28% higher F1 score in low-speed networks. Furthermore, Pallas utilizes only 8.5% of the switch's total resources on average while maintaining low per-packet processing latency, underscoring its practical efficiency and effectiveness.

[1] Li, W., Li, Z., Bütün, B., Diallo, A. F., Fiore, M., & Patras, P. Pontus: A Memory-Efficient and High-Accuracy Approach for Persistence-Based Item Lookup in High-Velocity Data Streams. In THE WEB CONFERENCE 2025.