SCONE
Internet-Draft
Intended status: Informational
Expires: 20 September 2025

K. Zarifis
S. Jaiswal
I. Purushothaman
Meta
J. Varsanik
Google
A. Tiwari
M. Joras
Meta

19 March 2025

SCONEPRO Taxonomy of throttling policies used worldwide
draft-zarifis-scone-taxonomy-01

Abstract

This document provides a description of traffic throttling and a taxonomy of throttling policies used by CSPs worldwide.

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 20 September 2025.

Copyright Notice

Table of Contents

1. Conventions and Definitions 2
2. Throttling Background and Introduction 2
2.1. Background 3
2.2. Throttler Implementations 3
2.2.1. Token Bucket Filters 3
2.2.2. Policers 3
2.2.3. Shapers 4
2.3. Impact of throttling 4
2.3.1. Impact of throttling on protocols 4
2.3.2. Impact of throttling on involved parties 4
3. Presence of throttling globally 5
4. Taxonomy of throttling policies 6
5. Design considerations for throughput advice signaling 7
6. Security Considerations 8
7. IANA Considerations 8
8. References 8
8.1. Normative References 8
8.2. Informative References 8
Acknowledgments 9
Authors' Addresses 9

1. Conventions and Definitions

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.

2. Throttling Background and Introduction

2.1. Background

Video traffic constitutes 70-80% of all traffic on the internet. Communications Service Providers (CSPs) throttle video traffic for a variety of reasons. Reasons cited by CSPs include service differentiation for their subscribers, enforcement of data limited plans, accommodation of mobility of users from well provisioned cells to purely provisioned or congested cells, efficient operation of radio access networks etc.

[draft-tomar-sconepro-ecn-01] provides additional background on the motivation and challenges around CSP traffic throttling.

2.2. Throttler Implementations

2.2.1. Token Bucket Filters

Achieving a controlled flow rate, referred to as Committed Information Rate (CIR), is commonly done using a Token Bucket Filter (TBF). In the simplest abstraction, network devices maintain a token bucket per flow, and tokens are added to a bucket at specified intervals. If the bucket is full with tokens, no more tokens are added. Token buckets control the burst size and mean rate of a flow by only allowing through as many packets as the number of tokens available in the flow's bucket at any given time. The burst size is thus equal to the bucket size, and arriving packets are forwarded immediately as long as there are enough tokens to serve them.

There are two main types of traffic throttlers: policers and shapers. Both typically implement TBFs, and they mostly differ in how they handle non-conforming packets, i.e. packets that arrive in a token bucket with insufficient tokens to forward them.

2.2.2. Policers

Policers drop (or deprioritize) non conforming packets. A burst is propagated as long as it fits within the TBF's capacity, otherwise excess packets are discarded. While this converges to a consistent mean transmission rate in the long term, it generates high packet loss and triggers retransmissions by the content delivery network (CDN) endpoint, leading to transmission rates with fluctuations.

2.2.3. Shapers

Shapers, on the other hand, employ an additional outbound queue and add non-conforming packets to it instead of dropping them. The enqueued packets are then scheduled for transmission at the configured rate of the shaper. This results in a more stable output rate close to the CIR, but also adds to the round trip time (RTT) measured by the clients by introducing queueing delays. The additional delay is bounded by the size of the queue used by the shaper.

2.3. Impact of throttling

2.3.1. Impact of throttling on protocols

In the transport layer, throttling complicates estimation mechanisms. For example, BBR can base its initial bandwidth estimation for a new connection on the initial burst that a full token bucket will allow. Once the tokens are depleted and throttling kicks in, the bandwidth estimates will drop considerably, triggering a reaction in the congestion control.

As a result, in the application layer, ABR algorithms can be adversely affected due to fluctuating bandwidth estimation. This can lead to video players flapping between video bitrates, causing degraded QoE and even video stalls.

2.3.2. Impact of throttling on involved parties

By influencing transport and application layer mechanisms, throttling has a direct impact on content providers, service providers, and end users.

Throttling affects Content Application Providers (CAPs):

* The high packet loss rate introduced by policers leads to aggressive and unnecessary retransmissions, generating CPU overhead on CDN servers.

* Both shapers and policers can lead to degraded QoE caused respectively by RTT inflation and packet loss

Throttling also affects Communications Service Providers (CSPs):

* Both policers and shapers require the ability to identify candidate flows for throttling (Deep Packet Inspection (DPI), Server Name Identification (SNI) parsing), which can be CPU and memory intensive, especially for high rate flows where the packet inspection has to be done on-path at the rate of the flow.

* Although packets are dropped on arrival by policers (TBFs are on inbound interfaces), carriers can still carry bytes to the edge of their network just to drop them there, which can waste bandwidth, especially due to unnecessarily retransmitted packets. With retransmission rates up to 20-30%, network overhead can add up.

* Shapers require slightly more complex middleboxes that maintain outbound queues on egress interfaces. The queues create artificial and unnecessary congestion. On the Radio Access Network side, while smoothing bursts can maintain consistent throughput, removing bursts can actually decrease radio access network (RAN) scheduling efficiency.

Lastly, throttling affects clients:

* The most obvious impact is QoE degradation, since video playback mechanisms are limited to artificially imposed bandwidth bounds.

* Additionally, the temporal characteristics of shaped traffic does not allow mobile device modems to schedule sleep cycles (Discontinuous Reception (DRX)) leading to higher device battery consumption.

An overall impact of throttling, as well as the benefits obviating the need to throttle is summarized in [YouTube]

3. Presence of throttling globally

Considering the breadth of implementations and policies used by CSPs to meet their specific needs, detecting throttling in the wild is not trivial. Some studies have quantified the prevalence of throttling using various methodologies which come with limitations or vantage point biases This section summarizes some key findings.

An analysis of millions of video streams served by Google's CDN over a period of 7 days in 2015 ([flach]) revealed that ~2% of video segments served in North America were impacted by policing, while the number increased to ~7% for Africa and APAC. 30% of the connections were throttled between 0.5 and 2Mbps, and for all the connections that were throttled, a 15% increase in rebuffering time was observed compared to non-throttled counterparts.

In 2019 [li] identified 30 (out of 144 analyzed) CSPs across 7 countries that throttled at least one of several video streaming services, by conducting crowd-sourced experiments from mobile phones. Specifically in the US, where the examined content providers were popular, it was found that most CSPs throttle several of the streaming services, with video traffic typically being throttled at 1.5Mbps. However, notably, CSPs used different throttling rates for different content providers (mainly due to the inaccuracy in identifying candidate flows), and throttled constantly both during peak as well as low traffic hours when congestion was not an issue.

Analyzing goodput data across connections served by Meta to countries outside of the US in 2022 identified ~200 ASNs that use throttling. Among them, 60% throttled at rates up to 1Mbps, 30% between 1-2Mbps, and ~10% between 2-5Mbps. A more recent analysis in 2024 identified ~274 globally that throttle content served by Meta. The detected throttling rates were mostly consistent, with a slight shift towards lower rates: 68% of the ASNs now throttled at rates up to 1Mbps, 30% between 1-2Mbps, and 7% between 2-5Mbps. In most cases, especially for ASNs in Latin America and Asia/Pacific, throttling was attributed to carrier-side Fair Usage Policies, with carriers implementing application-specific data caps for the various Meta apps. Within the US, most major CSPs throttled at specific rates (predominantly 2Mbps or 4Mbps), without imposing data caps for most of the users.

Across the different studies, a few key take-aways are highlighted:

* Over the last ~10 years there has been an increasing trend in use of throttling by CSPs, particularly for video traffic.

* Most US carriers throttle video traffic constantly, regardless of time of day or network conditions or user usage patterns.

* User Data Caps are prevalent, especially in APAC/LATAM

* CSPs around the world employ different mechanisms and diverse throttling policies, depending on the market and business needs.

4. Taxonomy of throttling policies

This section provides a taxonomy of the various throttling policies implemented by Communication Service Providers (CSPs) in the wild. These policies are designed to manage network traffic effectively and are influenced by a range of factors, including business objectives, network conditions, and regulatory requirements.

* *Constant, Application-Based Throttling*: CSPs may implement constant throttling policies that specifically target video flows from Content Application Providers (CAPs). These policies enforce a fixed rate, often utilizing Server Name Indication (SNI) to identify and manage specific applications. This approach ensures consistent bandwidth allocation for certain types of traffic, regardless of network conditions.

* *Time-of-Day (ToD) Based Throttling*: To manage peak traffic periods, CSPs may employ ToD-based throttling. For example, network traffic may be throttled during peak hours, such as between 6 PM and 10 PM, to alleviate congestion and maintain service quality for all users. This strategy allows CSPs to optimize network performance during times of high demand.

* *Data Capped User Throttling*: Throttling policies may also be applied based on data usage caps. Users may experience throttling once they exceed predefined data limits, which can be set on a daily or monthly basis. Some CSPs offer manual top-ups to allow users to purchase additional data. Additionally, app-specific data caps may be enforced, such as throttling Facebook or Instagram or YouTube video traffic after a user consumes 25GB or 30GB of app-specific data per month.

* *User-Specific Policies*: CSPs may implement user-specific throttling policies based on individual user profiles or subscription tiers. This allows for personalized bandwidth management, where users on different plans may experience varying levels of throttling. Such policies enable CSPs to offer differentiated services and pricing models.

* *Network topology specific policies*: CSP may implement different throttling policies based on the network topology. For example in the case of a cellular network as users move from a 5G network to a 4G network the CSP may change the throttling policies. Another example is users accessing the internet through a wifi access point using a cellular backbone (fixed wireless access) vs a satellite backbone, vs a fiber/cable backhaul are subject to very different throttling policies.

5. Design considerations for throughput advice signaling

Understanding the throttling ecosystem on the Internet today is crucial for designing a protocol to communicate throughput advice. In particular, to sufficiently communicate the kinds of policies in use today any throughput advice signalling must be able to achieve certain degrees of frequency and granularity.

* *Frequency of signal*: As discussed above, policies vary from being constantly active to being applied only during certain hours of the day. They can also be reactive to user-specific changes such as exceeding a subscription's data allotment. This variability implies a requirement on the advice signaling protocol that is able to alter the advice in step with these changes in policy.

* *Granularity of signal*: Similarly to frequency, the actual rates used in these policies vary as wildly as the policies themselves. Some, such as in the time-of-day policies, are based on network usage and demand. Others, such as the application-based constant policies, are aimed at an application serving a certain video quality. These rates also change over time with no standardization. A solution which seeks to communicate these rates as throughput advice needs to have the ability to encode the wide range of policies used.

* *Target unit*: Different data plans require the ability to signal desired rates at a per-user level. While signalling could be applied at a network or prefix level, signalling at the user level is required to satisfy subscriber-specific policies.

6. Security Considerations

General SCONE security considerations are discussed in the other documents covering the specific network-to-host signaling methods. This document only addresses questions regarding use of ECN for SCONE.

7. IANA Considerations

This document has no IANA actions.

8. References

8.1. 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/rfc/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/rfc/rfc8174>.

8.2. Informative References

[draft-tomar-sconepro-ecn-01]
"SCONEPRO Need for Defining A New On-Path Signaling Mechanism", n.d., <https://www.ietf.org/archive/id/draft-tomar-sconepro-ecn-01.html>.

[flach] Flach, T., "An Internet-Wide Analysis of Traffic Policing", n.d., <https://dl.acm.org/doi/pdf/10.1145/2934872.2934873>.

[li] Li, F., "A Large-Scale Analysis of Deployed Traffic Differentiation Practices", n.d., <https://dl.acm.org/doi/pdf/10.1145/3341302.3342092>.

[YouTube] YouTube, "YouTube Plan Aware Streaming", 21 March 2024, <https://datatracker.ietf.org/meeting/119/materials/slides-119-sconepro-youtube-plan-aware-streaming-01>.

Acknowledgments

This document represents collaboration, comments, and inputs from others, including:

* Tom Saffell (Youtube)

* Wesley Eddy (Meta)

Authors' Addresses

Kyriakos Zarifis
Meta
Email: kzarifis@meta.com

Sharad Jaiswal
Meta
Email: sj77@meta.com

Ilango Purushothaman
Meta
Email: ipurush@meta.com

Jon Varsanik
Google
Email: jvarsanik@google.com

Abhishek Tiwari
Meta
Email: atiwari@meta.com

Matt Joras
Meta
Email: mjoras@meta.com