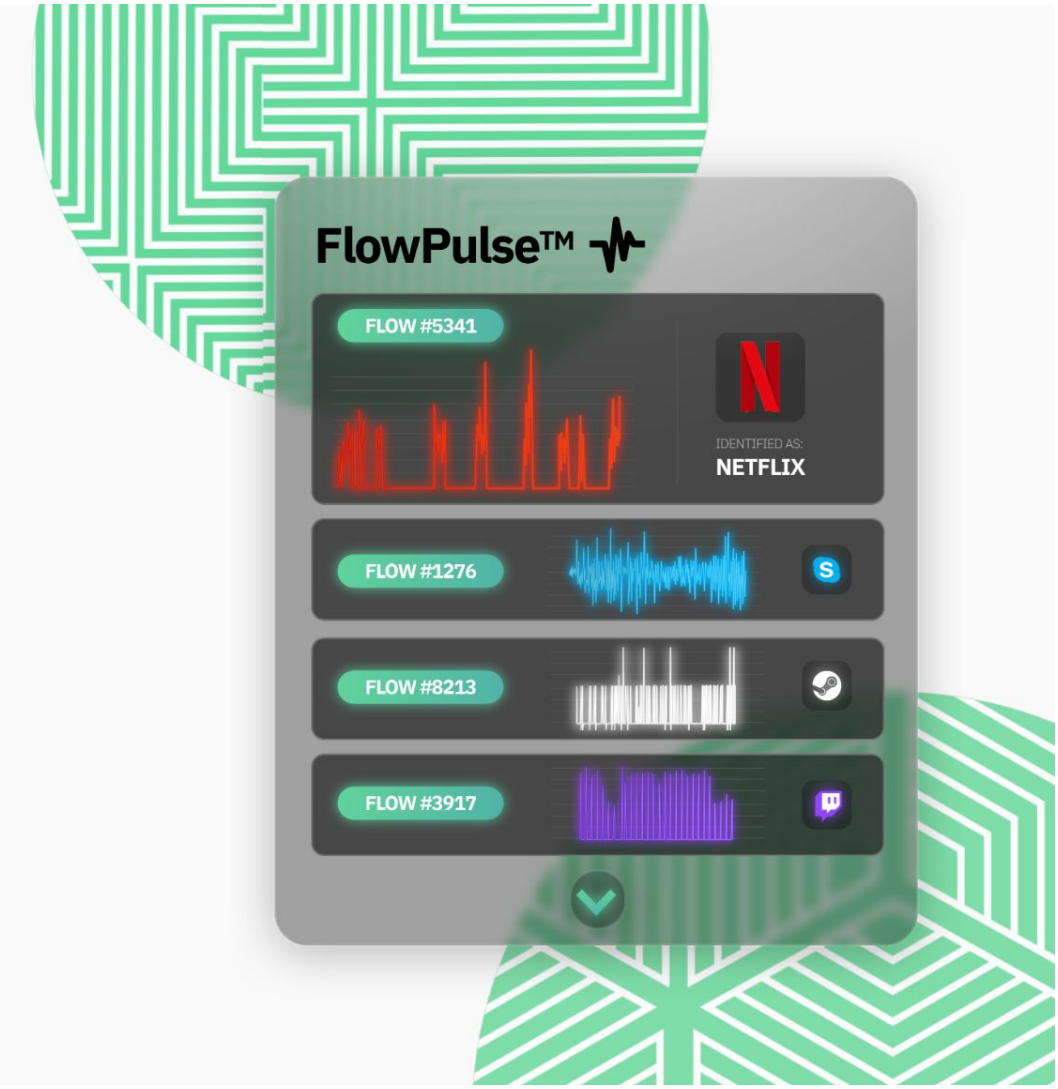
White Paper | Jun 2022

DPI RIP: Why Telcos need a new approach to visibility.

Deep packet inspection (DPI) fails with encrypted Terabit-speed traffic; Programmable Switches coupled with AI models are the way forward.



CANOPUS
NETWORKS

Contents.

Executive Summary.

Telcos operating multi-billion dollar wireline and wireless network infrastructure are under pressure to increase profitability, requiring them to optimise infrastructure efficiency, reduce customer support costs, and create upsell products. In the absence of continuous visibility into application usage patterns (e.g. streaming video, gaming, conferencing, downloads, etc.) and user experience (e.g. video freeze, game lag, conference stutter, download speed, etc.), Telcos typically operate in the blind, ceding value of their infrastructure investments to over-the-top beneficiaries. Unfortunately, visibility relying on traditional methods of examining packet contents (aka deep packet inspection or DPI) is starting to fail as payloads get increasingly encrypted and data rates grow to Terabits-per-second and beyond. A new approach is needed.

In this whitepaper we describe how the challenges of encryption and scale can be overcome using a combination of two emerging technologies: programmable switching and artificial intelligence. Rather than inspecting individual packets, programmable switches track the behaviour of each communication "flow" at millisecond-level time-scales to export a "pulse" akin to a heartbeat. The "FlowPulse" of various applications exhibit distinct characteristics, allowing trained AI engines to distinguish thousands of applications such as Netflix videos, Twitch live streams, Call-Of-Duty games, and Zoom teleconferences; and to further infer user experience explicitly in terms of video freeze, game lag spikes, and conference stutters. The result: a Terabit-speed platform that gives accurate sub-second real-time visibility into encrypted traffic streams using low-cost commodity hardware.

The implications are profound - Telcos can get deep application insights across their entire network footprint in a future-proof and affordable manner to increase their profits immediately and lower their costs. They can tune their networks to maximise user experience on high-value applications like gaming; increase customer satisfaction, loyalty, and retention; and create new revenues from premium services for emerging immersive applications like cloud gaming, virtual reality, and the Metaverse.

At about 300 km/hr speeds, engineers realised diminishing returns in attempting to design faster cars. In transportation, the breakthrough came with air travel.

At Terabit-speed encrypted communications, Deep Packet Inspection is showing diminishing returns. In networking, the breakthrough is here with AI and programmable switching.

DPI's problem with encryption.

Content Encryption: With widespread adoption of HTTPS in recent years, analysts estimate that roughly 80-90% of Internet packets carry encrypted payloads. This has dramatically reduced the efficacy of DPI techniques that look for patterns in the payload, such as URLs, to identify content type.

Fig. 1(a) Payload encryption using SSL.

DNS Encryption: The DNS system that maps domain names to IP addresses has traditionally been operated in clear text, allowing DPI tools to deduce the websites visited. However, recent years have seen DNS over TLS (DoT), DNS over HTTPS (DoH), and DNS over QUIC (DoQ) increase in uptake, making it impossible to deduce the domains visited by inspecting DNS packets. Even the IP address of the content server is becoming less meaningful as many services are hosted in shared public cloud infrastructure.

Fig. 1(b) DNS encryption over HTTPS

SNI Encryption: During secure connection setup, the client specifies the name of the server for which it seeks to exchange security certificates. The server name indication (SNI) happens in clear text today, and is widely used by DPI tools to infer the content provider and type. The new version (TLS1.3) encrypts the client hello message, and when deployed at scale, will render most existing DPI useless.

Fig. 1(c) SNI encryption with TLS1.3.

DPI's problem with scale.

Household broadband traffic is increasing 26% year-on-year, and mobile traffic 46%, per Statista [1]. With Telco networks carrying tens to hundreds of Terabits-per-second (Tbps) of traffic, DPI becomes prohibitively expensive, no matter which of the following approaches is used:

Hardware Appliances built with custom ASICs/FPGAs for inspecting packet contents (including byte patterns, URLs, SNI, etc.) can easily cost millions of dollars per Tbps, as illustrated by Roscomnadzor's nationwide rollout of DPI in Russia in 2019 which was estimated at US$300m, as reported by the BBC [2].

Integrated DPI from vendors of network switching and routing hardware attempts to analyse packet contents and provide application visibility, but these are often under-resourced functions (secondary to packet routing) that give low granularity sampled views of application traffic type and performance.

Software Functions for DPI are available both as open source and from vendors, but these run on general purpose CPUs that are simply not optimised for high speed packet processing. As shown in Fig. 2, CPU is about 100x slower than a switching ASIC for processing packets [3]; consequently, software DPI will require unjustifiably high volumes of server compute to process Tbps traffic, while also being very power hungry.

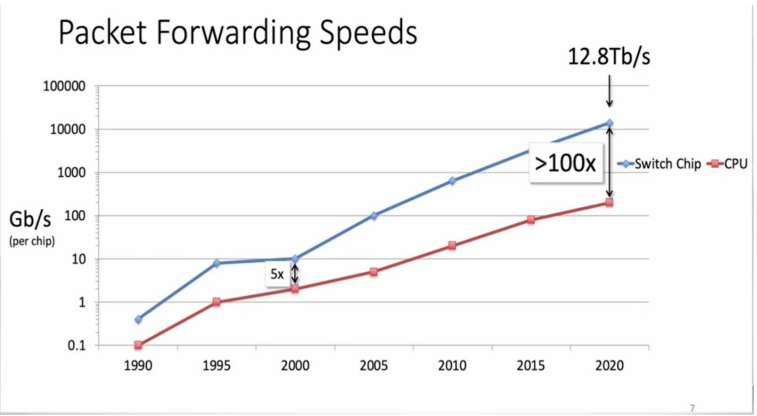


Fig. 2 CPU is 100x slower than switching ASICs at processing packets [3].

Tackling encryption with AI models.

A new approach is emerging that is agnostic to encryption and can be implemented at Terabit scale, leveraging the fact that applications have unique behaviour patterns on the network in the form of a “heartbeat” or “pulse”. As an illustration, consider the typical network behavior of a Netflix video-on-demand stream in terms of bit-rate every 100ms (Fig. 3a): once the client playback buffer is filled, the stream periodically fetches a “chunk” of video, manifesting in the form of a network activity spike every 4-8 seconds, and remains idle otherwise. In contrast, a Twitch live-stream (Fig. 3b) also fetches chunks, but typically at 2-second intervals, without any idle periods. At Canopus Networks, we have trained AI engines [4] on these stochastic behavioural patterns to not only distinguish various traffic types (on-demand video, live video, gaming, conferencing, etc.), but also the various providers (e.g. Netflix from Disney+, Zoom from Teams, etc.) with high confidence. Further, by examining the pulse, such as the height, width, and spacing of the spikes, the AI engines can also deduce the quality of the on-demand [5] or live [6] video stream in terms of resolution and client buffer health. The methods have been further extended to measure latencies and lag spikes for online games [7], and stutters for conferencing session, even when all traffic is encrypted.

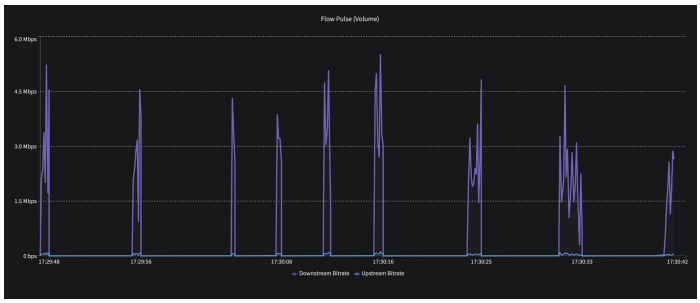


Fig. 3(a) Typical pulse of a Netflix on-demand video stream.

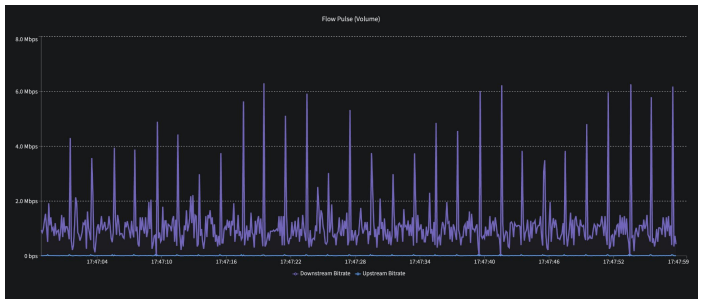


Fig. 3(b) Typical pulse of a Twitch live video stream.

Tackling scale with programmable switches.

The ability for AI engines to make inferences about the traffic type and user experience as described above depend on the ability to extract the pulse for millions of flows every 100 ms at Terabit speeds. Traditional fixed-function switches do not have this capability, and custom-building it would be prohibitively expensive. Fortunately, a new breed of programmable switching silicon is emerging, exemplified by the Tofino (Fig. 4a) series of chips from Intel [8], capable of 12.8 Tbps and available in 32 x 400G port switches (Fig. 4b) today. Such switches, low in cost and commoditised by the hypescalers, can be programmed in a language called P4 to extract fine-grained per-flow telemetry at scale.

Fig. 4(a) Intel Tofino 12.8 Tbps P4 chip.

Fig. 4(b) Edgecore 32 x 400G P4 switch.

Key Takeaway

Telcos with billions of dollars invested into infrastructure cannot afford to operate blind to the application mix and user experience on their network. With DPI becoming unviable for encrypted traffic at Terabit speeds, Canopus FlowPulse™ technology combines commodity programmable switching with AI engines trained on stochastic behavioural models to provide Telcos with accurate, scalable, and cost-effective visibility well into the future.

References

[1] Statista, 2022, https://www.statista.com/statistics/271405/global-mobile-data-traffic-forecast/ and https://www.statista.com/statistics/267194/forecast-of-internet-traffic-by-subsegment/
[2] "Roskomnadzor to deploy new blocking technology (in Russian)". BBC News Русская Служба. 18 December 2018, https://www.bbc.com/russian/features-46596673
[3] Nick McKeown Keynote at NetDev 2020, https://www.youtube.com/watch?v=fiBuao6YZl0
[4] R. Babaria, S. Madanapalli, H. Kumar, V. Sivaraman, “FlowFormers: Transformer-based Models for Real-time Network Flow Classification”, Intl. Conf. Mobility, Sensing, and Networking (MSN), UK, Dec 2021, https://ieeexplore.ieee.org/document/9751578
[5] S. Madanapalli, H. Habibi Gharakheili and V. Sivaraman, “Inferring netflix User Experience from Broadband Network Measurement", IFIP Traffic Measurement and Analysis (TMA) Conference, Paris, France, Jun 2019.
[6] S. Madanapalli, A. Mathai, H. Habibi Gharakheili, V. Sivaraman, "ReCLive: Real-Time Classification and QoE Inference of Live Video Streaming Services", IEEE IWQoS, Tokyo, Japan, Jun 2021.
[7] S. Madanapalli, H. Habibi Gharakheili, V. Sivaraman, “Know Thy Lag: In-Network Game Detection and Latency Measurement”, Passive and Active Measurement (PAM) Conference, Mar 2022, https://dl.acm.org/doi/abs/10.1007/978-3-030-98785-5_17
[8] Intel Intelligent Fabric Processors, https://www.intel.com/content/www/us/en/products/network-io/programmable-ethernet-switch.html