Hot Topic Abstract: *Marina*: Realizing ML-driven Real-time Network Traffic Monitoring at Terabit Scale

Michael Seufert [*] [†], Katharina Dietz[†], Nikolas Wehner[†], Stefan Geißler[†], Joshua Schüler[†], Manuel Wolz[†], Andreas Hotho[†], Pedro Casas[‡], Tobias Hoßfeld[†], and Anja Feldmann[§]

[*] University of Augsburg, Augsburg, Germany
Email: michael.seufert@uni-a.de

[†]University of Würzburg, Würzburg, Germany

[‡]AIT Austrian Institute of Technology, Vienna, Austria

[§]Max Planck Institute for Informatics, Saarbrücken, Germany

Abstract

Network operators need real-time traffic monitoring to ensure high performance and security across their infrastructure. While artificial intelligence and machine learning (ML) have proven effective in enhancing network visibility, especially for encrypted traffic, existing solutions struggle to handle the scale and speed of modern high-throughput networks. To address this, we present *Marina*, a system designed for ML-driven traffic monitoring at terabit scale. *Marina* distributes the monitoring workload between a high-performance data plane, capable of extracting traffic statistics at line rate, and a powerful ML server that performs inference using complex ML models. By applying temporal microaggregation over sub-second intervals and computing moment-based statistics, *Marina* enables timely and flexible monitoring decisions during the next time slot. We demonstrate the scalability of our approach through a prototype implemented on a Barefoot Wedge 100BF-65X P4 switch, capable of monitoring over 520,000 concurrent flows at full switching capacity of 6.4 Tbps. Finally, we validate *Marina* 's analytics capabilities across four real-time ML-based monitoring tasks using standard ML models, achieving results that are comparable or better than state-of-the-art methods.

Motivation and Contributions of the Paper

Network operators increasingly require real-time insights regarding performance and security in order to effectively manage their networks. The growing complexity and heterogeneity of applications as well as the adoption of network-level encryption, however, make obtaining this kind of information an especially challenging task. This has motivated a surge in the research and conception of Artificial Intelligence (AI) and Machine Learning (ML) driven approaches to in-network traffic monitoring.

A typical ML-driven network traffic monitoring workflow involves three main steps, as depicted in Figure 1: analyzing the *(encrypted) network traffic*, performing *feature extraction* to create a vector representation, and feeding it into an *ML data analytics model*. Traditional approaches often rely on *offline or out-of-band processing* (top part of Figure 1), where traffic is mirrored to an external server for analysis. This method provides flexibility and can reach excellent monitoring performance when sufficient compute resources are available but lacks real-time capabilities, making it unsuitable for deployment in operational networks with high traffic rates and volumes.

Modern network telemetry increasingly leverages programmable switches and software-based packet processors that can analyze massive traffic volumes at line rate without degrading forwarding performance. These systems often perform *in-network feature extraction* to reduce data volume before sending it to out-of-band stream processors (middle part of Figure 1). While this setup offers better scalability and real-time capabilities, the scope of monitoring tasks remains constrained by the capabilities of the hardware and the extracted features.

In-network ML (bottom part of Figure 1) pushes integration further by embedding the ML model directly into the network's data plane. This can involve offloading feature extraction or embedding both the feature extraction and ML model within the device itself. Although this method offers high-speed, real-time analysis, it demands significant simplification of the ML models due to hardware limitations, which compromises flexibility and analytical depth.

DOI: 10.22032/dbt.67116 *DOI (proceedings): 10.22032/dbt.66316*

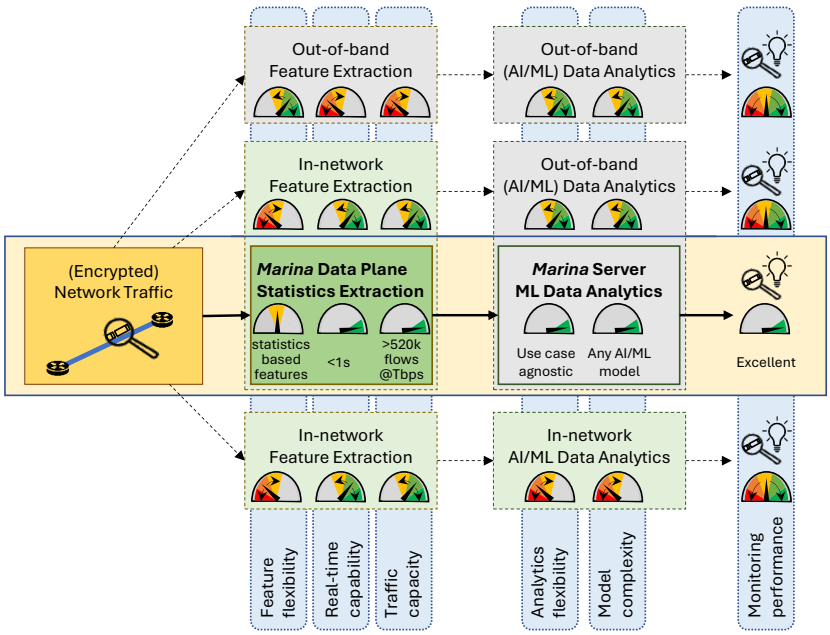

Fig. 1: Design options for ML-based network telemetry systems.

To support large-scale, real-time traffic monitoring with rich analytics, our system ***Marina*** **(MAchine-learning-based Real-time Network traffic Analytics)** focuses on efficient in-network extraction of monitoring information and out-of-band processing on a powerful server, as shown in the middle part of Figure 1.

Marina is designed explicitly for ML-based real-time network traffic monitoring at terabit scale, dedicating limited data plane resources to extract meaningful moment-based traffic statistics at sub-second granularity, even from encrypted traffic. These statistics are then transformed into high-dimensional feature sets on the server enabling accurate and flexible real-time traffic monitoring tasks, such as traffic/device classification, application health (e.g., Quality of Experience), and fault/anomaly detection (e.g., intrusion detection), which we demonstrate in this work. In combination with a powerful ML server, *Marina* ensures fine-grained flow-level visibility and delivers actionable insights with sub-second delay.

The contributions of this paper are as follows:

1) *Marina* system design: We introduce a novel approach that combines data plane extraction of moment-based traffic statistics at line rate with ML-based analytics on a powerful server. Using temporal microaggregation of packets into sub-second intervals, the design enables flexible, scalable, and accurate real-time monitoring for various tasks, even for encrypted traffic.
2) *Marina* data plane implementation and evaluation: We implement the *Marina* data plane on a Barefoot Wedge 100BF-65X P4 switch, fully utilizing its data plane to monitor up to $6.4\,\mathrm{Tbps}$ across 524,288 concurrent flows on 65 QSFP $100\,\mathrm{Gbps}$ ports. The system generates under $385\,\mathrm{Mbps}$ monitoring traffic and achieves a monitoring granularity and delay until obtaining monitoring results for all flows as low as $500\,\mathrm{ms}$. To support reproducibility, we publicly release *Marina*'s source code at: https://github.com/lsinfo3/Marina
3) *Marina* ML-based real-time traffic monitoring: We demonstrate *Marina*'s effectiveness across four use cases: encrypted traffic classification, video streaming application health/Quality of Experience, intrusion detection, and IoT device classification. In all cases, *Marina* delivers results on par with or better than existing approaches using standard ML models.

This work was published in [1]. All details and results can be found in the original publication.

References

[1] M. Seufert, K. Dietz, N. Wehner, S. Geißler, J. Schüler, M. Wolz, A. Hotho, P. Casas, T. Hoßfeld, and A. Feldmann, "Marina: Realizing ML-Driven Real-Time Network Traffic Monitoring at Terabit Scale," *IEEE Transactions on Network and Service Management*, vol. 21, no. 3, pp. 2773–2790, 2024. [Online]. Available: https://doi.org/10.1109/TNSM.2024.3382393