Systems@ETH Zürich

# Real-time verification of datacenter security policies via online log analysis

Master Thesis Proposal

IT giants like Google and Facebook store and process vast amounts of data containing confidential information about real persons and organizations. As a result, access to these data must be granted only to authorized system users. Such security requirements are prevalent in today's data infrastructures, and emerge also in medical and financial institutions, which often have to comply with acts like HIPAA and SOX.

Formal system specification is an active area of research and several solutions have been proposed; a recent example is the Metric First-order Temporal Logic (MFOTL) [1]. MFOTL enables a rigorous description of the expected system behavior over time and can be used along with the system-generated logs to monitor and verify the enforcement of security policies on user actions. The problem becomes challenging in modern enterprise datacenters due to their big size and complexity. Modern datacenters are multi-layered distributed systems that serve hundreds of thousands of concurrent users with different profiles and non-trivial access patterns. They are also heavily instrumented (for troubleshooting purposes) to log almost every interaction among the different system components, which often have many explicit and implicit dependencies. The result is to generate TBs of traces within minutes, and checking whether these traces comply with large collections of MFOTL rules becomes hard.

For this reason, most of the existing approaches focus on offline log processing [2] and they cannot be used when fast reactions are needed, e.g. in case of an intrusion. Our novel system, Strymon [4], aims to fill this gap via scalable and intelligent processing of datacenter logs in real time. The key idea is to take advantage of the log streams already generated in datacenters to continuously verify access policies and provide an additional security layer at a negligible cost.

Strymon is based on Timely Dataflow that was first introduced in the Naiad data processing system [3]. Timely is a programming framework for writing and executing data-parallel computations in the form of dataflow graphs; nodes in these graphs represent data operators whereas edges denote the flow of data between operators. Timely's model is general and supports arbitrary dataflows that may contain User-Defined Functions (UDFs) as well as cycles (iterations). In addition, Timely adopts a pure event-driven execution model that forms an ideal basis for Strymon's use cases, including the online verification scenario we address here.

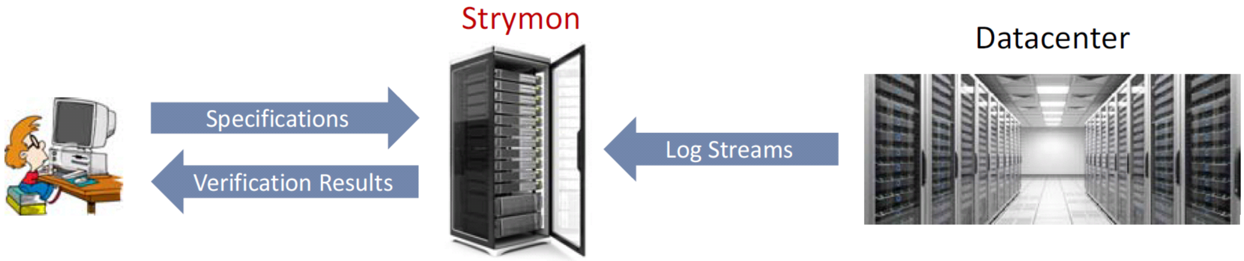
The goal of this thesis is the design and implementation of a Strymon module for the online verification of security policies in distributed systems. The module will ingest logs of events coming from the instrumented systems in a streaming fashion, and will apply the user-defined policies – modeled with MFOTL – on the fly. Real logs will be provided by our industry partners but for the purposes of the thesis we can also use synthetic data generators. The core part of the work lies in the development of the MFOTL monitor; essentially, a state machine that we plan to implement as a Timely dataflow.

[1] David Basin et al. “Monitoring Metric First-Order Temporal Properties”. In: *J. ACM* 62.2 (May 2015), 15:1–15:45.

[2] David Basin et al. “Scalable Offline Monitoring of Temporal Specifications”. In: *Form. Methods Syst. Des.* 49.1-2 (Oct. 2016), pp. 75–108.

[3] Derek G Murray et al. “Naiad: a timely dataflow system”. In: *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*. ACM. 2013, pp. 439–455.

[4] *Strymon: Queryable Online Simulations for Modern Datacenters*. URL: `http://strymon.systems.ethz.ch`.



**Interested? Please contact John Liagouris (liagos@inf.ethz.ch) and Dmitriy Traytel (traytel@inf.ethz.ch). The proposed thesis will be supervised by Dr. John Liagouris and Prof. Timothy Roscoe.**