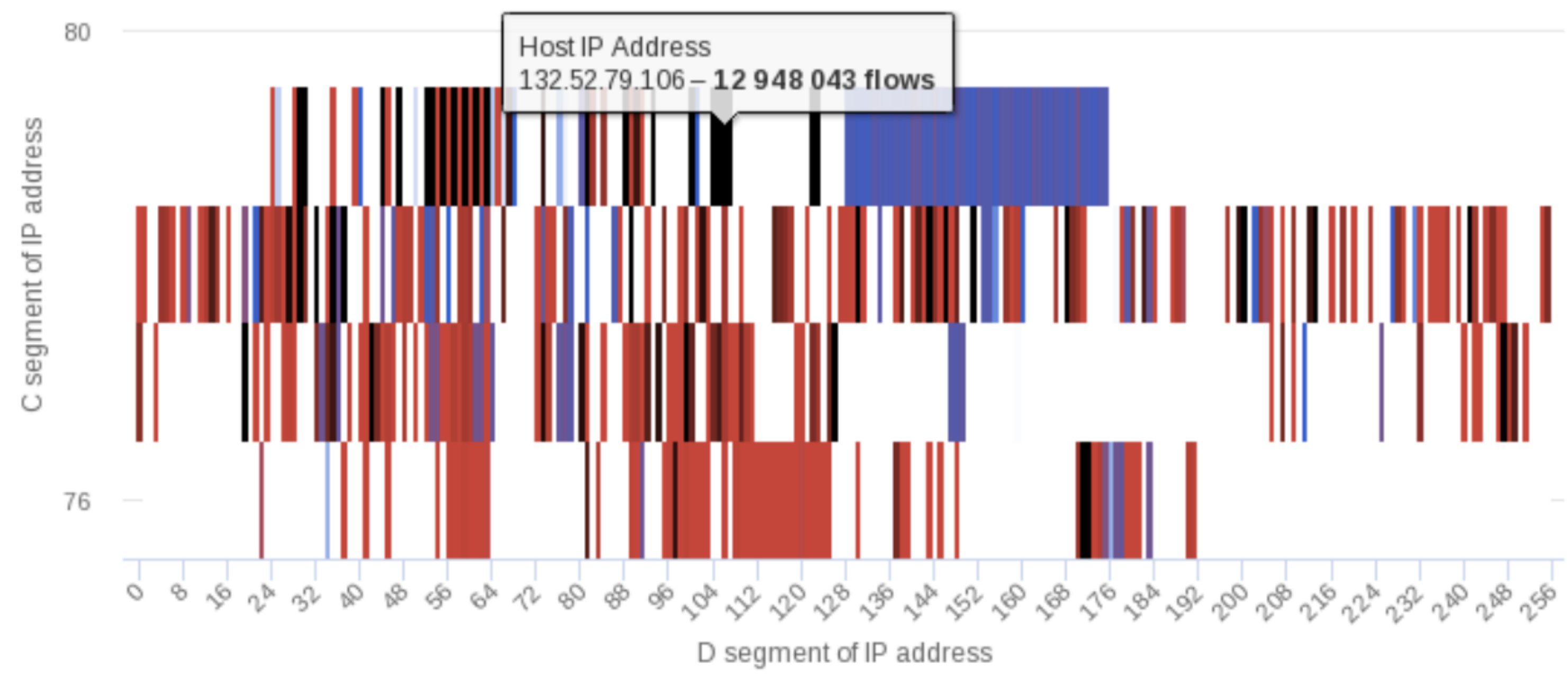
Stream4Flow: Real-time IP Flow Host Monitoring using Apache Spark

Tomas Jirsik

Institute of Computer Science & Faculty of Informatics,
Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic
e-mail: jirsik@ics.muni.cz

Abstract – *we present Stream4Flow, a framework for cyber situational awareness based on Apache Spark Streaming. We demonstrate utilization of Stream4Flow for real-time IP flow host monitoring in a large campus network. Contemporary IP flow analysis systems are not designed for the continuous host monitoring. Gaining the detailed overview of each host is not straightforward with these systems due to connection-based paradigm and performance challenges. We show that distributed stream processing is a natural solution for detailed IP flow host monitoring. Moreover, we describe a real-time host monitoring workflow in data streams in detail and present advantages of flow-based host monitoring in Apache Spark including real-time host profiling, dynamic level of detail and granularity.*

Zoomable Host Heatmap



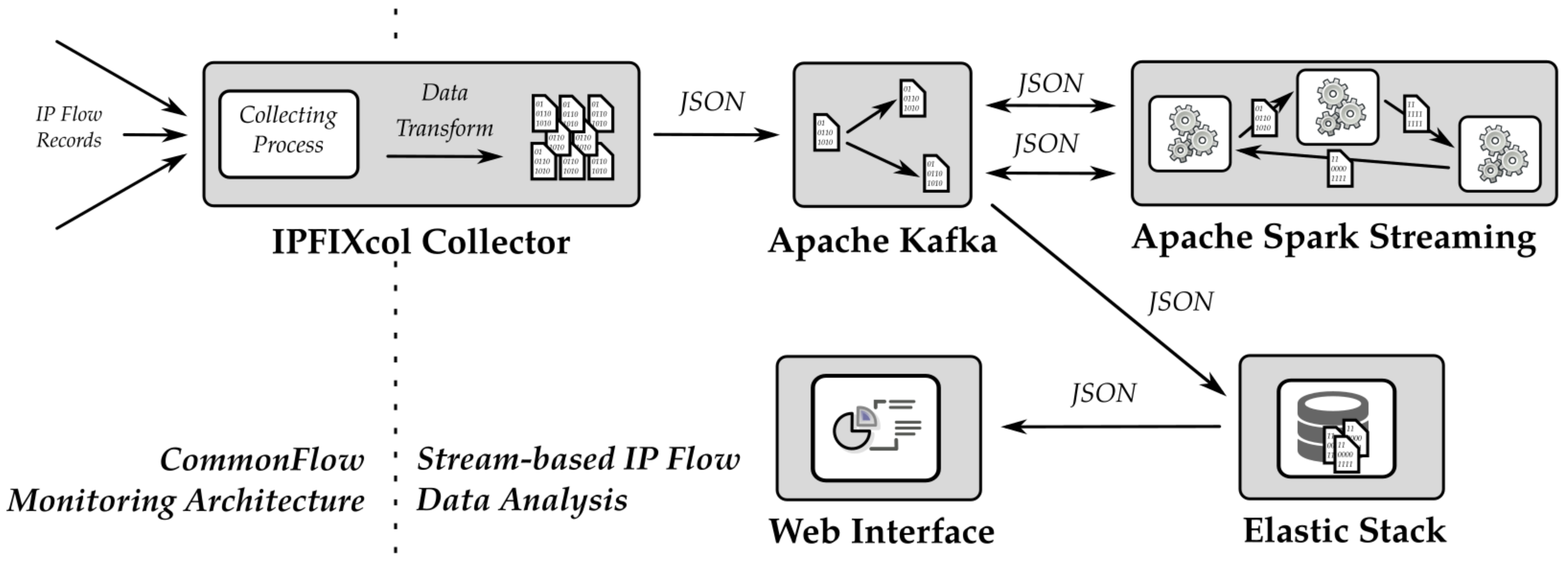
Major Issues of Micro View IP Flow Monitoring

Detection oriented analyses - the goal of the majority of the analyses in IP flow monitoring is to identify a traffic of interest, e.g., malicious one. Considering this goal, we comprehend the traffic of interest, but we have only limited information about the other network traffic, which prevents us from a complex understanding of the network.

Analysis delay - the majority of current IP flow monitoring tools analyses data in five-minute intervals. The information of hosts is then available with a five-minute aggregation. The aggregation level needs to be reduced as important behavior characteristics are lost due to the aggregation

Connection-based data paradigm - IP flow data are provided per connection and stored in five-minute bins. A transformation from connection-based view to host-based view is time-consuming and includes aggregation and filtration over each bin of data. Therefore, complex information about a host is not available for an operator for everyday use.

Stream4Flow Architecture



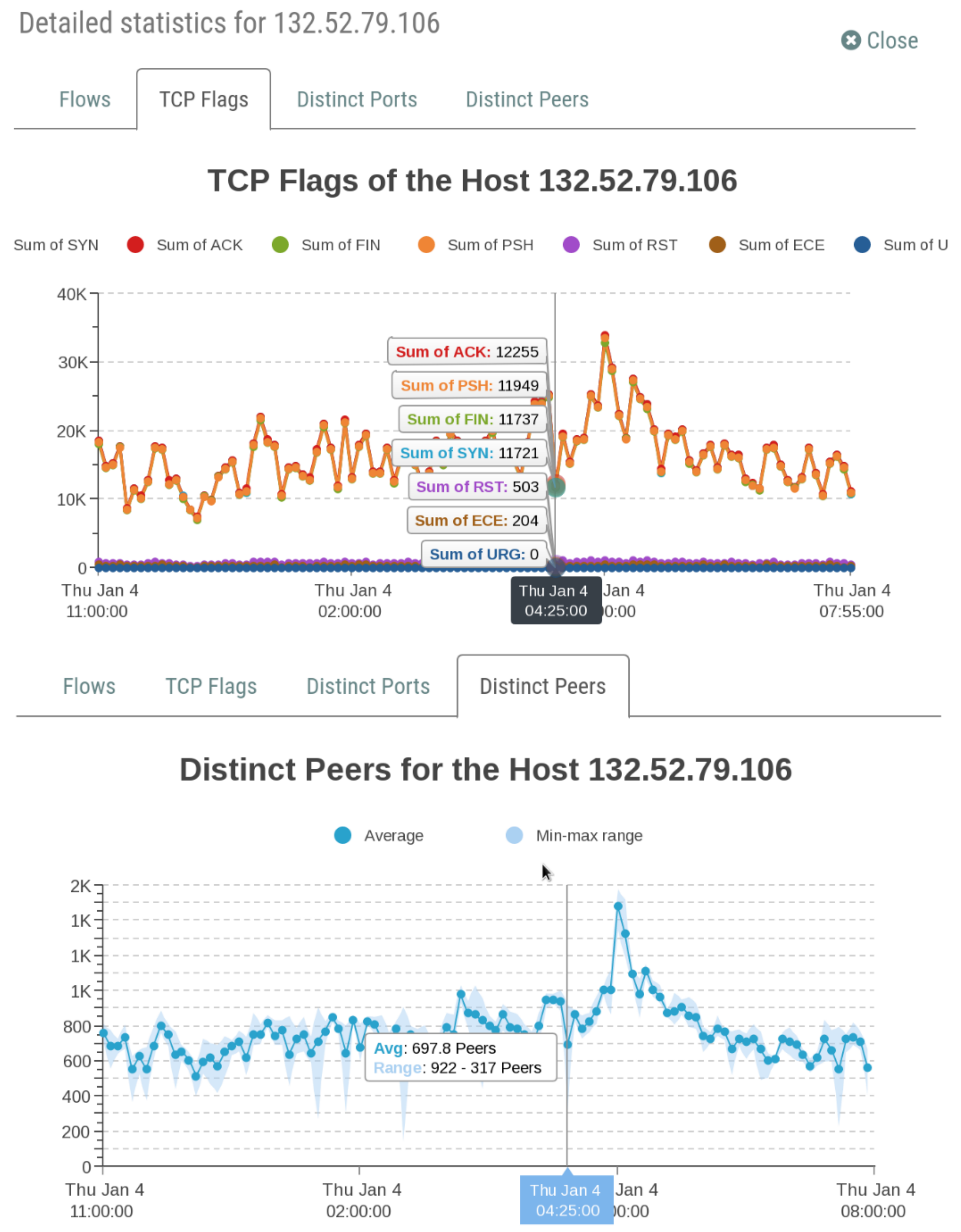
The Transformation of the Connection- to Host-based View

Transformation is done in each data stream by using **map-reduce** principle:

- A map is created with Source IP addres as a Key
- The map is then reduced by the key and various statistics (e.g., number of flows) are computed for each IP address.
- The multiplied data streams are united into a single data stream containing host-based records
- A reduce operation is applied to the data stream to obtain a collection of statistics for a host.

The data stream now contains one record of all computed statistics per host for a given analysis window. The data stream is passed back to Kafka. From there host-based records are

Real-Time Host Statistics


Detailed statistics for 132.52.79.106

https://stream4flow.ics.muni.cz/ @csirtmu https:/www.kypo.cz/

This research was supported by the Security Research Programme of the Czech Republic 2015-2020 (BV III/1--VS) granted by the Ministry of the Interior of the Czech Republic under No. VI20162019014.

CSIRT-MU