Streaming Data Analysis on the Wire

Dimitrios Katramatos, Shinjae Yoo, Meng Yue, Kerstin Kleese van Dam
Brookhaven National Laboratory

Motivation

In the era of Big Data, in an ever increasing amount of cases in science and industry, more data can be found at any given moment in transit than in storage media ranging from memory to tape. This raises the question whether it would be feasible to extract value out of this data while it is still in transit. There is a strong incentive for on-the-wire processing since such processing can provide near real time information to speed up decision making processes, optimize and prioritize information routing and correlation, and offer additional processing cycles that can free up data center resources. A related concept comes from the world of cybersecurity, where packet streams passing through firewalls are inspected in an attempt to detect patterns of intrusion and avert cyber attacks. Suitably designed algorithms can exploit available computing power and perform specific forms of computation on streaming data while they are “on the wire”, i.e., being transported in the network. This requires taking the concept of processing on the wire to levels beyond the common cybersecurity applications and investigate equipment and methods to enable generic, statically or even dynamically programmable data analysis and/or transformation of data streams while going through network devices.

Concept

The basic concept for analyzing data on the wire is that one or more network devices can be programmed to recognize specific data flows and transparently apply a certain type of computation on the data of a flow before forwarding it to its destination (see figure 1). Depending on the case, the recipient may receive data transformed in some expected way or original data, while analysis information may be gathered from the data and sent to a different recipient. The performed computation, in the general case, is of low overhead so that it does not adversely impact a flow (e.g., cause timeouts). Therefore, it is important to use algorithms designed to be fast and match the capabilities of the equipment they are meant to run on. There are many potential streaming algorithms that can be adapted for processing on the wire; for example, classical online algorithms for streaming outlier detection, approximated summary statistics, such as billing, or for lightweight dimensionality reduction using problem characteristics; batch supervised and unsupervised learning algorithms; and adaptive supervised and unsupervised learning algorithms.

From the perspective of networking, there are several potential solutions. Software Defined Networking (SDN) environments already support mechanisms for Network Function Virtualization (NFV) and Service Function Chaining (SFC) [1]. Such mechanisms could be utilized, after potential modifications, to support streaming data analysis. Several vendors offer equipment mainly targeted at enabling sophisticated cybersecurity functions, i.e., Deep Packet Inspection (DPI), Deep Packet Processing (DPP) [2] which, could be potentially adapted to support analysis algorithms. Standard networking devices could be used in conjunction with external computing systems to perform similar analysis tasks. Conceivably, future designs of

standard devices may be influenced so as to include native support for some level of user-programmable transparent data stream processing.

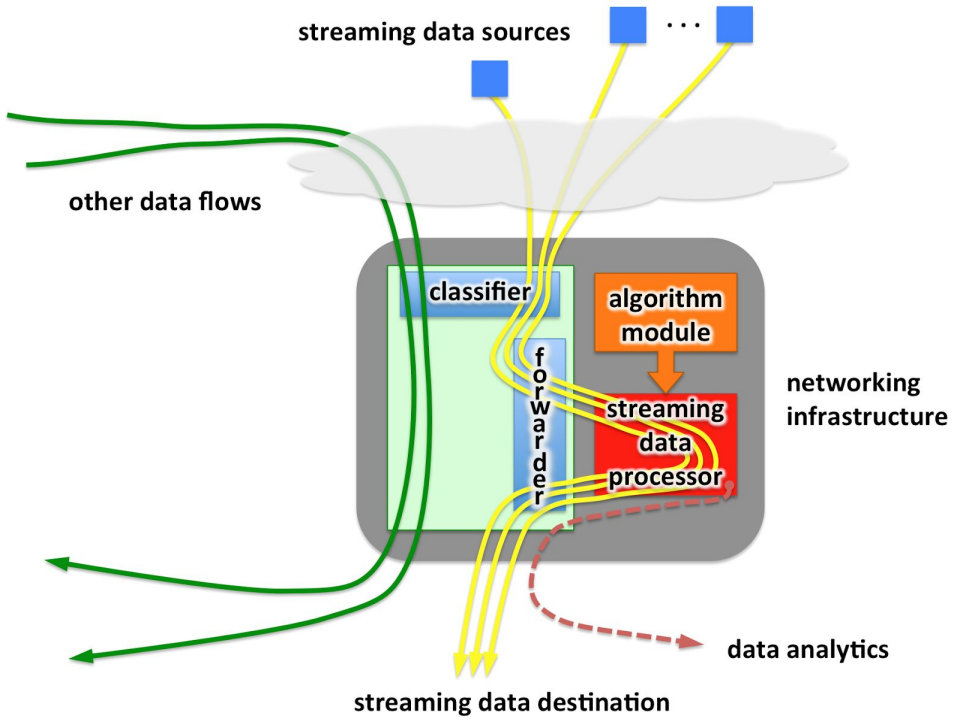

Figure 1: Analysis on the wire concept

Use Cases

Processing on the wire offers additional processing cycles that can be contributed to a computation and free up data center resources. The cases that such kind of processing makes sense instead of standard data transport and processing in a data center are several. In the commercial world, a company may require real time processing on data before that data arrives at their data center for decision making purposes; or a network provider may want to offer added value services to its subscribers by performing a certain processing on data that source and/or destination subscribers don't have sufficient resources to perform. In the scientific world, processing on the wire may perform, e.g., simple data transformations and free up data center resources or be part of a programmable data acquisition system. Of particular interest are cases involving sensor network analysis, such as distributed solar irradiance prediction, security sensor networks, such as DARPA SIGMA [3], or the Smart Grid for Phasor Measurement Unit (PMU) and Smart Meter data reduction and state estimation. In the general case, on the wire processing will be beneficial to the Internet of Things (IoT) [4]. Finally, the concept can be utilized to add intelligence to the network itself, for example, to handle asynchronous communication over unstable, sporadic, or restricted network domains.

References

1. https://tools.ietf.org/html/rfc7498
2. https://lgscout.com/products/technology/deep-packet-processing/
3. http://www.darpa.mil/program/sigma
4. https://en.wikipedia.org/wiki/Internet_of_Things