A lightweight recommendation system for high-volume network traffic capture retention

Mattijs Jonker

UNIVERSITY OF TWENTE.

Network Abuse

with an ingress perspective

Cloudflare blocks record-breaking 71 million RPS DDoS attack

This weekend, Cloudflare blocked what it describes as the largest volumetric distributed denial-of-service (DDoS) attack to date.

SERGIU GATLAN FEBRUARY 13, 2023 02:50 PM 2

New Linux malware brute-forces SSH servers to breach networks

A new botnet called 'RapperBot' has emerged in the wild since mid-June 2022, focusing on brute-forcing its way into Linux SSH servers and then establishing persistence.

BILL TOULAS AUGUST 04, 2022 12:22 PM 0

DDoS attacks now use new record-breaking amplification vector

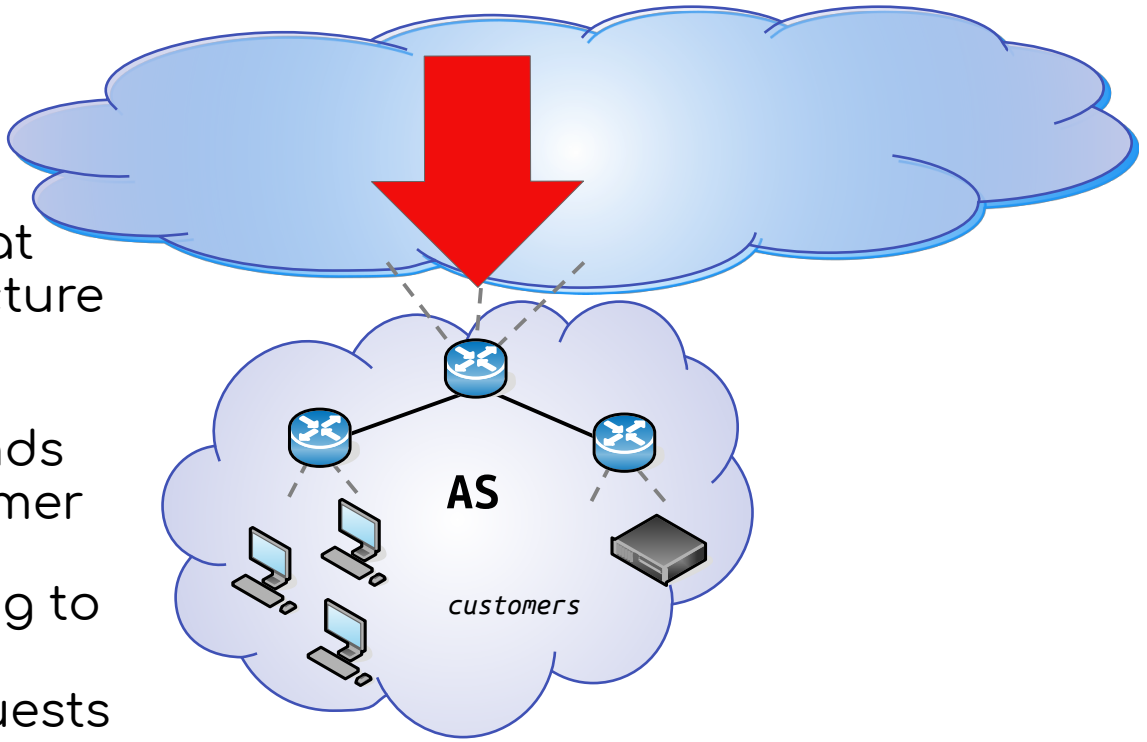
A new reflection/amplification DDoS vector has been spotted in the wild, offering threat actors a record-breaking amplification ratio of almost 4.3 billion to 1.

BILL TOULAS MARCH 08, 2022 10:00 AM 0

Network Abuse

(of the volumetric or high-rate kind)

- (D)DoS attack traffic, targeted at customer or network infrastructure
- Brute-force attacks on authenticated services/backends running in network or at customer
- Malicious ingress traffic, looking to abuse in-network resources as springboard (e.g., spoofed requests to open resolver, ...)
- And so on...



Network Abuse

with an egress perspective

New ‘HinataBot’ botnet could launch massive 3.3 Tbps DDoS attacks

A new malware botnet was discovered targeting Realtek SDK, Huawei routers, and Hadoop YARN servers to recruit devices into DDoS (distributed denial of service) swarm with the potential for massive attacks.

BILL TOULAS MARCH 19, 2023 10:20 AM 1

New GoBruteforcer malware targets phpMyAdmin, MySQL, FTP, Postgres

A newly discovered Golang-based botnet malware scans for and infects web servers running phpMyAdmin, MySQL, FTP, and Postgres services.

SERGIU GATLAN MARCH 10, 2023 02:02 PM 0

New EnemyBot DDoS botnet recruits routers and IoTs into its army

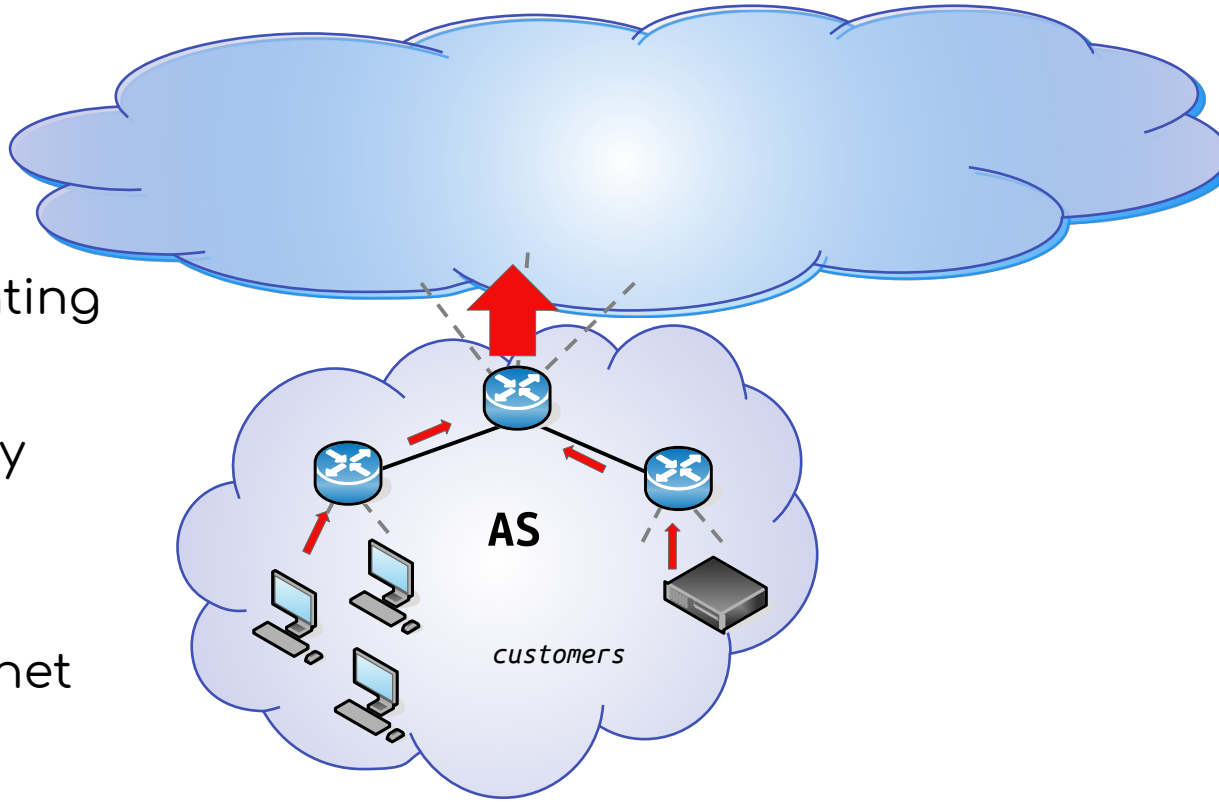
A new Mirai-based botnet malware named Enemybot has been observed growing its army of infected devices through vulnerabilities in modems, routers, and IoT devices, with the threat actor operating it known as Keksec.

BILL TOULAS APRIL 13, 2022 12:00 PM 0

Network Abuse

(of the volumetric or high-rate kind)

- (D)DoS attack traffic, originating from network
  - "Direct" attacks (randomly spoofed)
  - "Indirect" (spoofed)
- Brute-force attacks on Internet services
- Unintentional harmful egress traffic (misconfiguration, packet storms, ...)



Problem

Ideally, you'd retain full network traffic captures for analysis or to later revisit (e.g., forensics)

... however, storage is a challenge

You can store (sampled) flow data (~1/2000x), but you lose payload-related information in the aggregate

Our goal

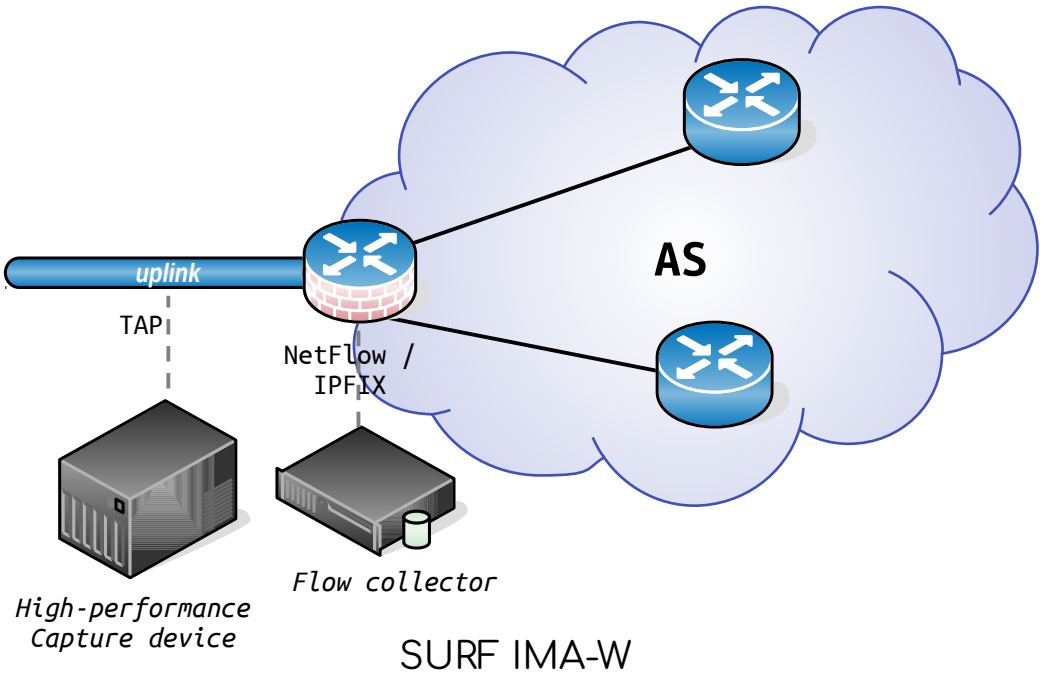
- Build a system that recommends whether to retain or discard full payload capture following moment of capture (closely)
- Aim: help retain bulky data only when needed
  - for later, manual forensic purposes
  - as filter, in a multi-stage system
- And in doing so:
  - Helps address storage challenges
  - Allows for scaling
- Lower costs compared to commercial solutions (e.g., Arista TAP Ag)

System requirements

- Needs to:
  - Operate in real-time (and hence be lightweight)
  - Be sensitive and precise (ideally)
  - Be tunable (balance between precision and recall)
  - Low barrier to deployment (given certain assumptions)
  - Be built on top of open-source and community-driven software, towards an open solution

High-level

- Ingest flow data to make recommendations
- Operate on tumbling windows of configurable duration
- Flows themselves are kept unconditionally



Components [1/5] – Capture Device

- A suitable network capture device that can work at line-rate
  - Professional / commercial
  - Custom built
- Our assumption is that this is (still) feasible in the network and not cost-prohibitive

Components [2/5] – Flow exporter(s)

- Leverage NetFlow v9 / IPFIX capability of edge router(s); or
- Use a dedicated appliance (e.g., Flowmon Probe); or
- In smaller networks, use something such as *softflowd*

Components [3/5] – Flow Collector

- Collector needs to be capable of receiving flow data of a given format
- The collector software needs to be able to produce a feed of collected flows

  ... rather than write to local storage

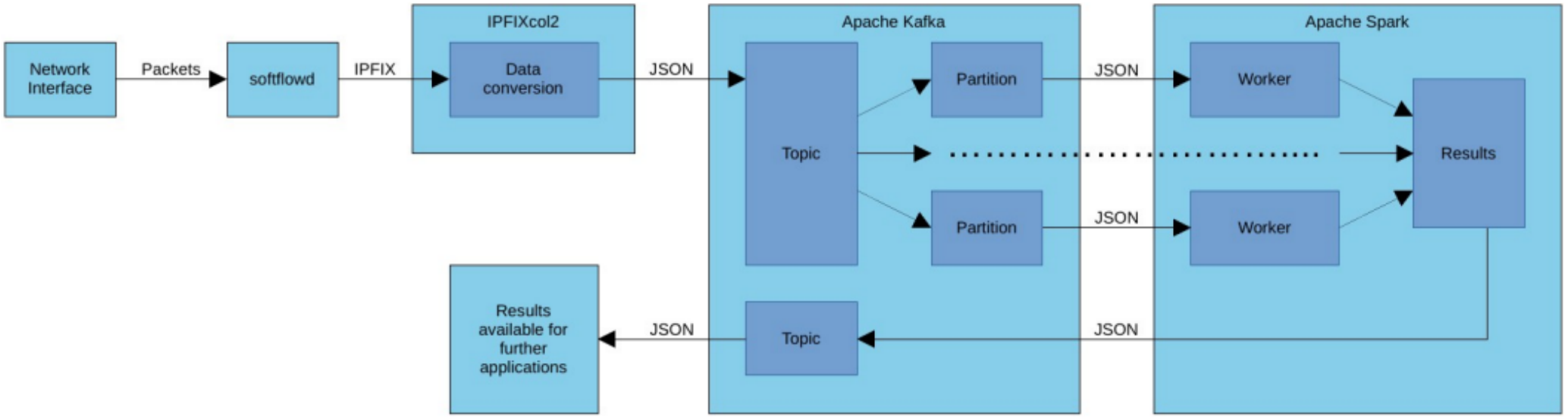
Components [4/5] – Messaging System

- Message System capable of:
  - Carrying the feed of flow data (post-collector)
  - Communicate the binary retain/discard recommendation

Components [5/5] – Analysis Framework

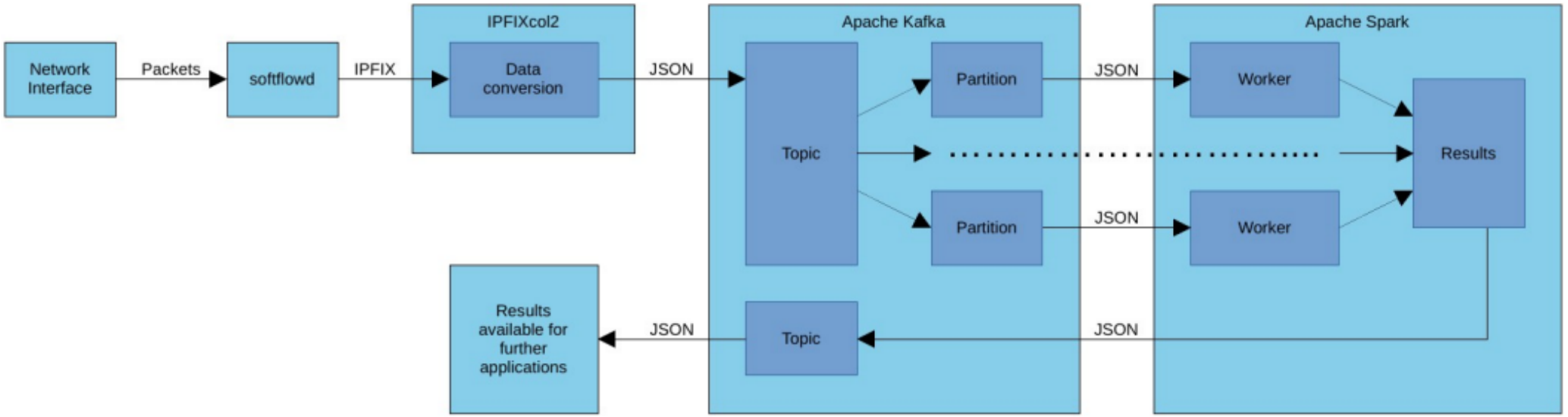
- Needs to be able to consume flow data feed produced by collector and carried by messaging system
- Needs to allow for the recommendation heuristics/logic to be implemented

Our prototype



- Flow Exporter: softflowd
- Flow Collector: IPFIXcol v2
- Message System: Apache Kafka
- Analysis Framework: Apache Spark

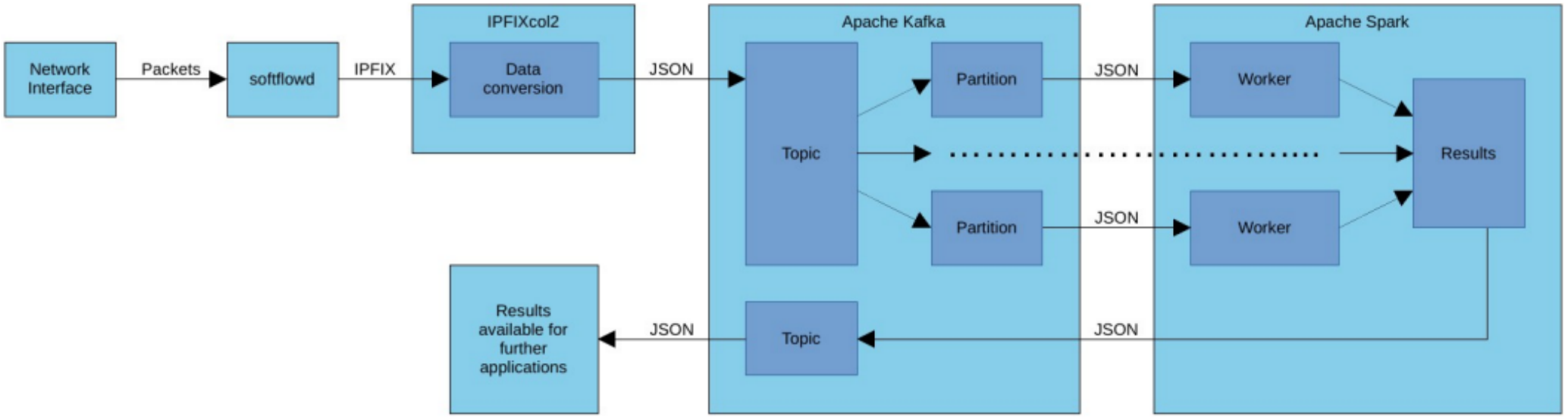
Our prototype



Flow Exporter: softflowd

- Software implementation of flow-based network traffic monitor
- Reads network traffic and keeps track of flow statistics
- Exports under given active/inactive timeouts

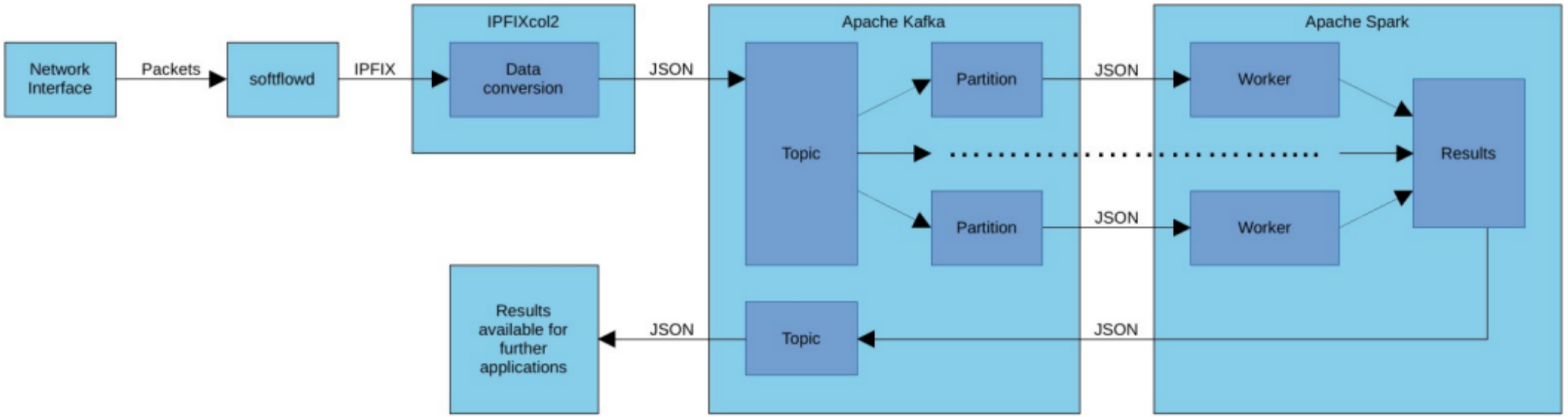
Our prototype



Flow Collector: IPFIXcol v2

- Open-source and well-maintained flow collector (CESNET)
- Listens for IPFIX or NetFlow v5/9 flow data
- Collects and "persists" in given format
  - we have it produce JSON-formatted "messages" to Kafka

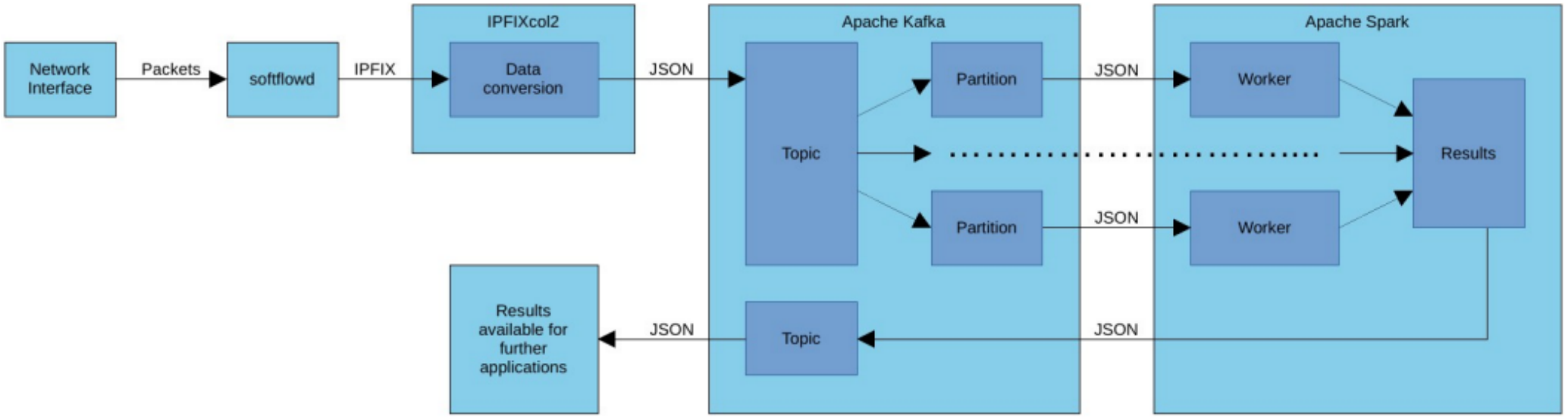
Our prototype



Message System: Apache Kafka

- Widely-used (distributed) event store and stream-processing platform
- Highly scalable
- Can itself “embed” analysis apps (Kafka Streams), but we do not use this

Our prototype



Analysis Framework: Apache Spark

- Widely-used analytics engine for large-scale data processing
  - Distributed, highly scalable
  - Comes with various useful libraries (e.g., Spark ML, Kafka connector)
- We use Spark Structured Streaming to do low-latency (and fault-tolerant) event-processing

A peek under the hood [1/2]

- Calculate entropy over a selection of flow features
  - e.g., source IP entropy changes could signal egress spoofing
  - e.g., ingress flow bytes entropy could signal incoming volumetric attack
  - e.g., increase in ingress #flows and destination IP could signal horizontal scanning
  - Etc.
- Use z-scores (no. std deviations from mean) to find anomalies i.c.w. set tresholds
- Align tumbling window of recommendation with capture
- Built user-defined aggregator to calculate entropy and keep streaming state over triggers

A peek under the hood [2/2]

- Challenges:
  - Flows can outlive tumbling windows
  - Flows have varying duration
  - Flows can be actively timed out

Evaluation (work-in-progress)

- Step 1: synthesize ground truth, layering self-instrumented attacks on benign traffic
  - For certain threat presence ratios (e.g., 1 out of every 5min sees attacks, on avg)
- Step 2: Use precision-recall curves to find optimal thresholds and flow feature selection, under three scenarios:
  - costs of missed attacks and false alarms are equal
  - costs of missed attacks is 10 times higher than that of false alarms
  - costs of false alarms is 10 times higher than that of missed attacks
- Step 3: calculate “gain” in terms of storage

As an enabler

- Leverage architecture to subject flow data to other forms of analyses
- e.g.:
  - Build predictive models for network traffic
  - Detect threats other than scans, brute-force, (D)DoS
  - Decide when to stop capturing (NaWas)

Questions ?