Mind the Cost of Telemetry Data Analysis

Alessandra Fais[1], Gianni Antichi[2], Stefano Giordano[1], Giuseppe Lettieri[1], Gregorio Procissi[1]

[1] Università di Pisa, Italy | [2] Queen Mary University of London, United Kingdom

Introduction

- Stream processing engines **efficiently process continuous amounts (streams) of information**
  - Widely used solutions (Spark Streaming) for a variety of use cases
- Network operators need efficient ways to analyze fine-grained telemetry data
  - In production datacenter networks, hundreds of thousands of switches produce up to millions of reports per second!

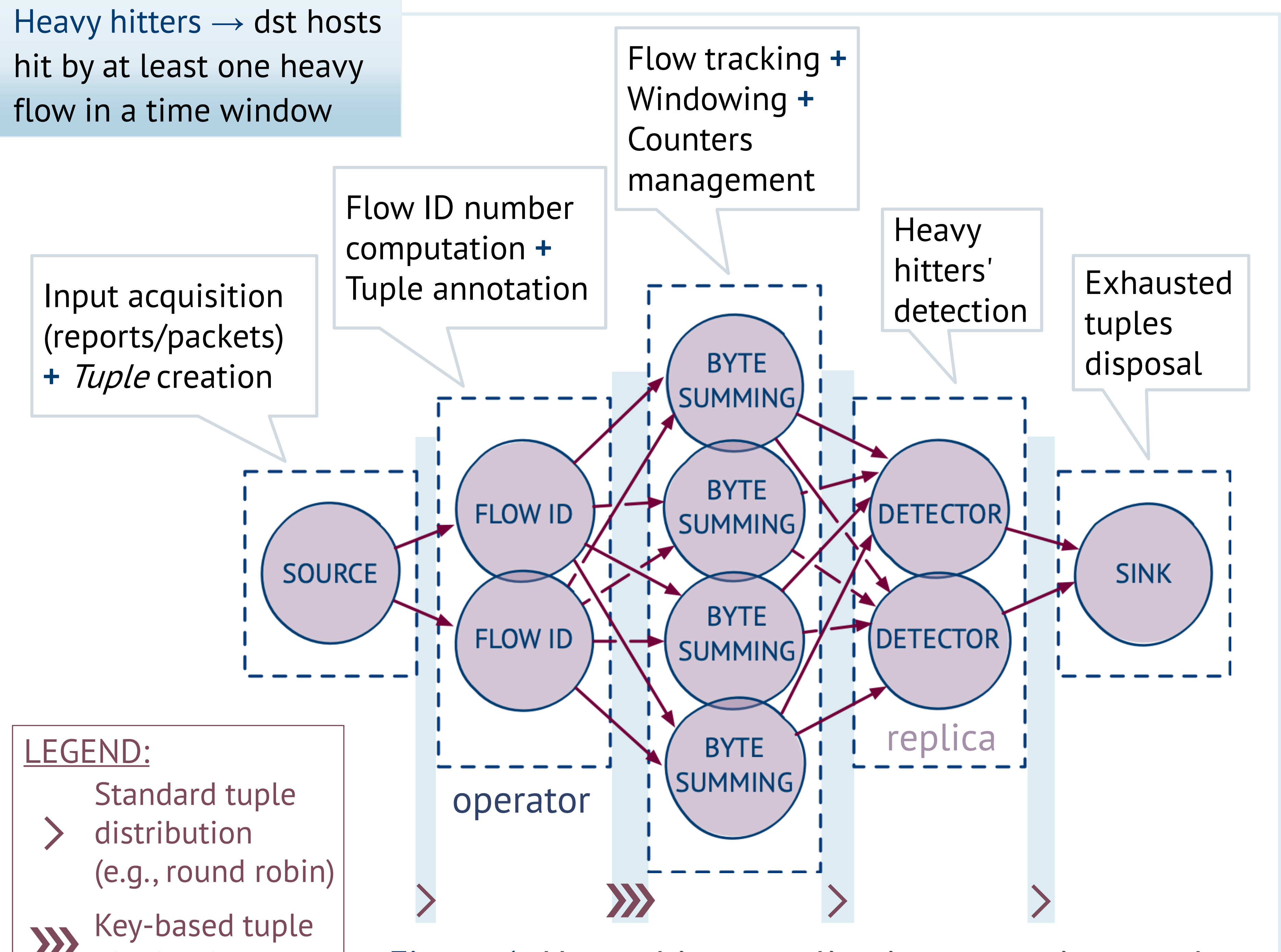
GOAL: What's the best streaming engine for network traffic analysis?

A Qualitative Comparison

	Flink	Spark	Storm	WindFlow
Batching	✗	mandatory	✗	✓
Chaining	✓	✗	✗	✓
Ordering	✗	btw batches	✗	✓
Windows	✓	✓	✗	✓
Event time	✓	partial	✗	✓
Distributed	✓	✓	✓	✗

Findings

- Systems designed for generic data processing over distributed platforms perform poorly with network data
  - Overheads not compensated by the computational burden of the application itself


Heavy hitters → dst hosts hit by at least one heavy flow in a time window

Input acquisition (reports/packets) + *Tuple* creation

Flow ID number computation + Tuple annotation

Flow tracking + Windowing + Counters management

Heavy hitters' detection

Exhausted tuples disposal

LEGEND:

› Standard tuple distribution (e.g., round robin)

››› Key-based tuple distribution

Figure 1: Heavy-hitter application streaming graph

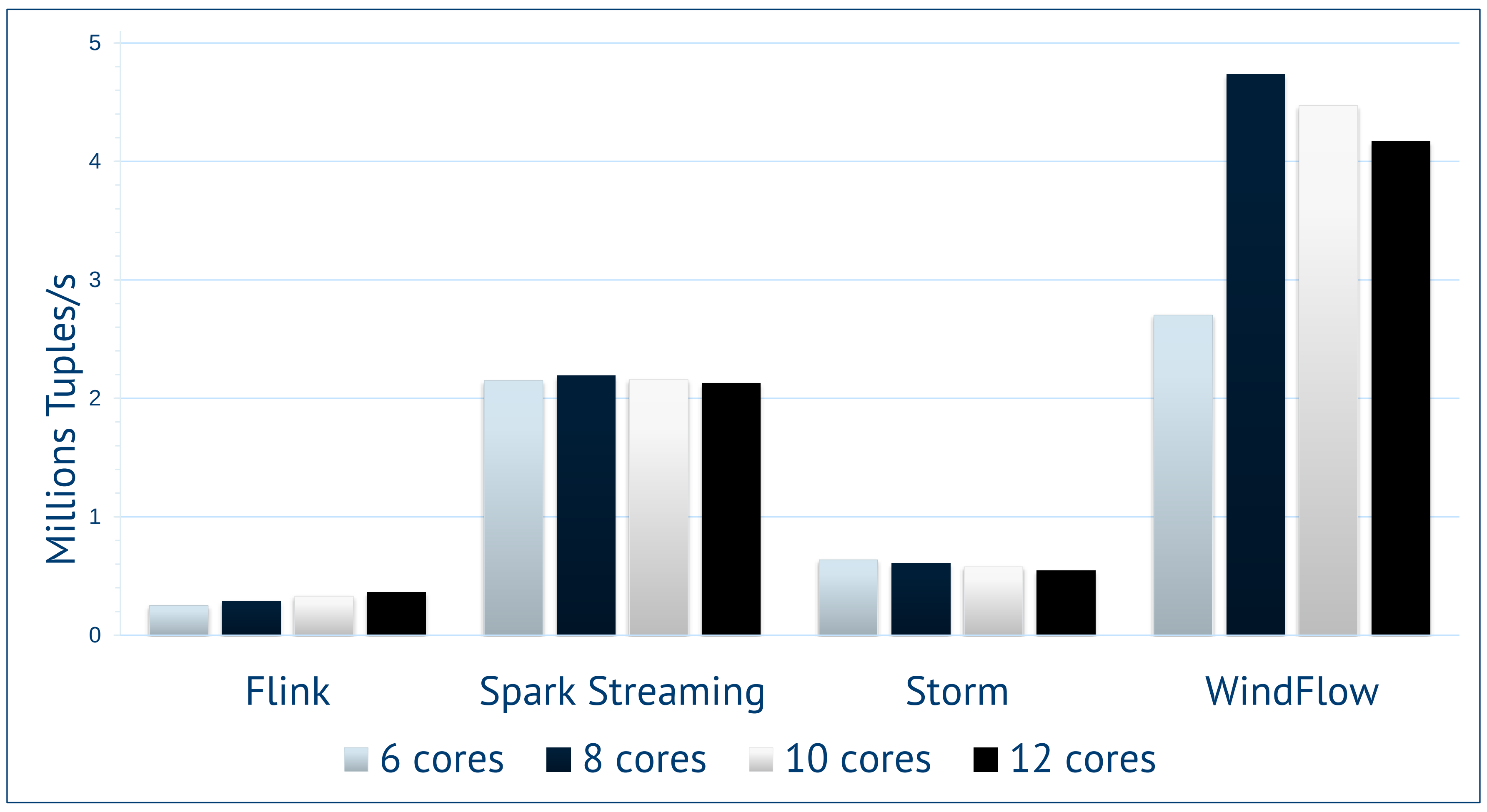


Figure 2: Performance comparison

- WindFlow shows better performance figures
  - More than 2x in most cases than Spark Streaming
  - Around 10x of Flink and Storm
- WindFlow performance scales with n. physical cores
- Other solutions immediately saturate

Resource utilization scenarios
- Physical cores only → number of cores ⩽ 8
- Hyperthreading → 8 < number of cores ⩽ 16

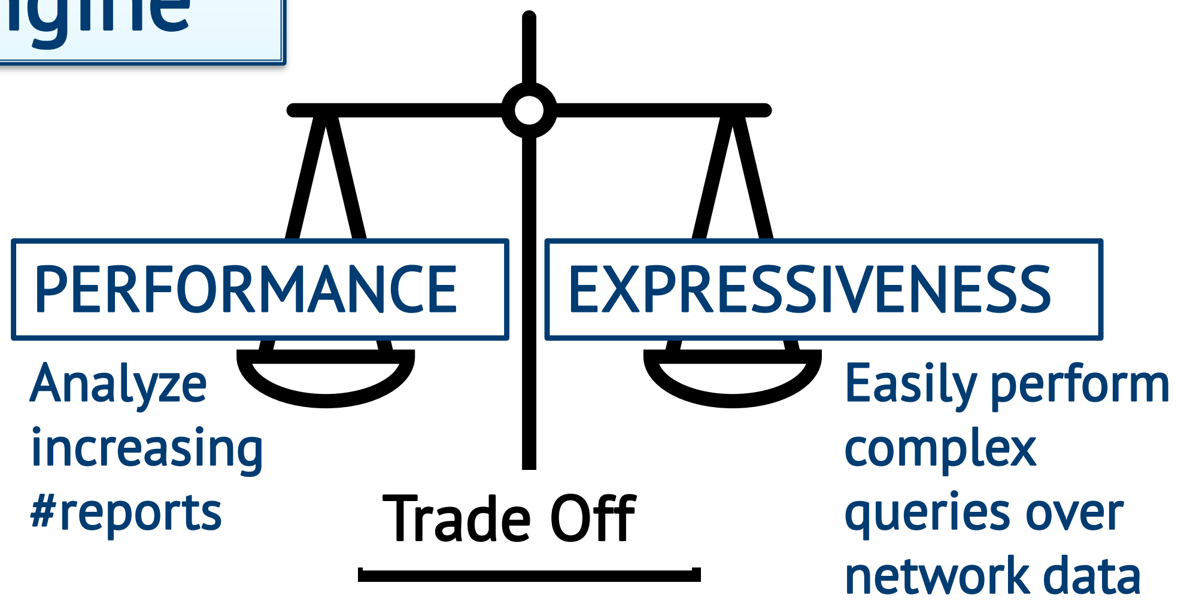
Promising research directions

Design of a networking domain specific streaming engine

Lightweight communication mechanisms → Support for computation distribution over a cluster → Specifically built for network traffic analysis

w/o compromising performance!

Network data analysis traits: sustained input rate + moderate comp. burden


PERFORMANCE

Analyze increasing #reports

EXPRESSIVENESS

Easily perform complex queries over network data

Trade Off

SIGCOMM 2022 Amsterdam | Amsterdam, Netherlands | August 22-26, 2022

Contact information: alessandra.fais@phd.unipi.it | https://for.unipi.it/alessandra_fais/