

# Data Streaming for Network Monitoring

---

Abhishek Kumar  
College of Computing  
Georgia Institute of Technology  
akumar@cc.gatech.edu

# Data Streaming

---

- **Definition:** A data stream is input data arriving at such high rates that it *stresses* the communication, computing and storage infrastructure. In particular, it may not be possible to:
  - Transmit the entire stream.
  - Compute complicated functions over the entire stream.
  - Store the entire stream.
- **Examples:**
  - Continuous data from sensors: radars, telescopes, etc.
  - Aggregated data from multiple streams: credit card transactions, stock trades, etc.
  - Logs of Internet activity, data and traffic.

## Problem Statement – Network Monitoring

---

- The Internet has become a necessity, for businesses, consumers, etc.
- The Internet is also an evolving complex system
- Operators need to monitor their networks to:
  - Infer usage patterns
  - Engineer resource deployments
  - Detect anomalies
- Network attacks add urgency to the problem

## Data Stream Management Systems:

---

- The Stanford STREAM project:
  - Uses the abstract semantics of a stream defined as an *unbounded* bag of tuples and a relation as a *time-varying* bag of tuples.
  - Three classes of operators: *relation-to-relation*, *stream-to-relation*, and *relation-to-stream*.
  - Design of CQL – *Continuous Query Language*
  - Declarative queries translated into *physical query plans*.
  - Query plans support optimizations and fine-grained scheduling.
  - Performance requirements force sharing of state and computations across queries, with graceful degradation in accuracy during overload.

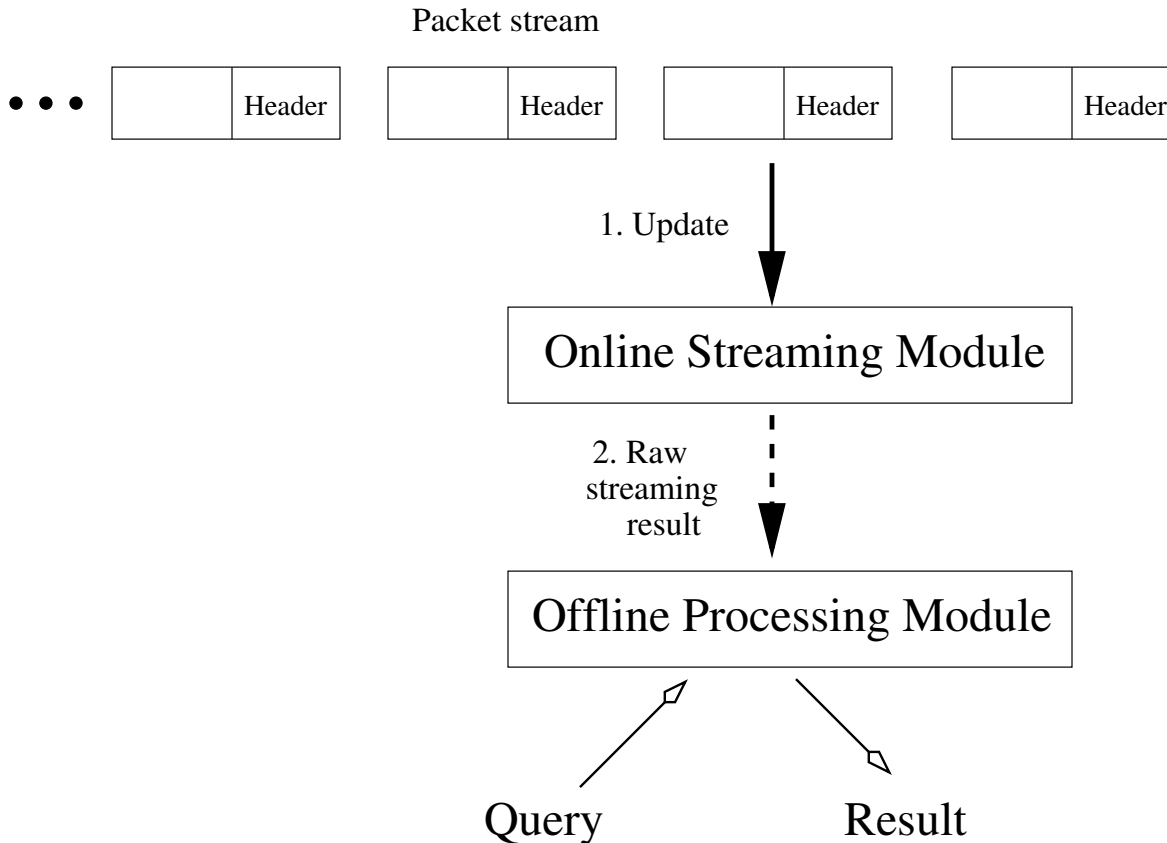
## Data Stream Management Systems:

---

- The AT&T Gigascope:
  - Operational in AT&T's IP Network.
  - Uses the semantics of sliding windows over a stream of timestamped tuples.
  - Design of GSQL – a restriction of SQL.
  - Support for sampling.
  - Gigascope is a GSQL compiler that produces optimized C or C++ code to run on the data stream.
  - Performance requirements force use of low sampling rates, resulting in high inaccuracy of results.

# Data Streaming – Sketch based solutions

---



## Data Streaming – Sketch based solutions

---

- Accurate solutions for a variety of problems, such as:
  - Estimating the total number of flows.
  - Estimating the size of individual flows.
  - Estimating the distribution of flow sizes.
  - Automated detection of common content (worms).
- Suitable model for network monitoring problems
- Meets additional constraints:
  - Constant worst case processing time per item (packet)
  - Constant size of working memory
  - Need to archive observations

## Architecture for Network Data Streaming

---

- Online Streaming Module:
  - Measurement proceeds in epochs (e.g. 100 seconds).
  - Maintain a “digest” data structure in fast memory (SRAM)
  - For each packet, a small number of simple operations
  - Constant computational complexity for average & worst case
  - Data collection is “lossy”, but very fast
  - At the end of the epoch, the digest is paged to disk
- Offline Processing Module:
  - Provide a query interface to users & higher level applications
  - Process the digest to answer queries
  - Processing ranges from simple table lookups to complex statistical estimation algorithms

## Proposal: Data Sketch Management Systems

---

- Identify correct primitives for sketch collection.
- Provide the back-end infrastructure for storing the sketches.
- Support for sketch aggregation and composition.
- Identify correct abstract semantics: Sliding window vs. epochs.
- Flexible support for a broad range of network monitoring applications with superior performance and accuracy.

Thank You !

---