CLOUD COMPUTING APPLICATIONS

Streaming Introduction

Roy Campbell & Reza Farivar

Why Real-Time Stream Processing?

- Real-time data processing at massive scale is becoming a requirement for businesses
  - Real-time search, high frequency trading, social networks
  - Have a stream of events that flow into the system at a given data rate
- The processing system must keep up with the event rate or degrade gracefully by eliminating events. This is typically called load shedding

Why Real-Time Stream Processing?

- MapReduce, Hadoop, etc., store and process data at scale, but not for real-time systems
- There's no hack that will turn Hadoop into a real-time streaming system
  - Fundamentally different set of requirements than batch processing
- Lack of a "Hadoop of real-time" has become the biggest hole in the data processing ecosystem

Cloud Streaming Engines

- Apache Storm
- Twitter Heron
- Apache Flink

- Older non-cloud systems
  - IBM System S
  - Borealis
    - Descendent of Aurora from Brown University. Not active anymore