

Coflow

A Networking Abstraction for Cluster Applications

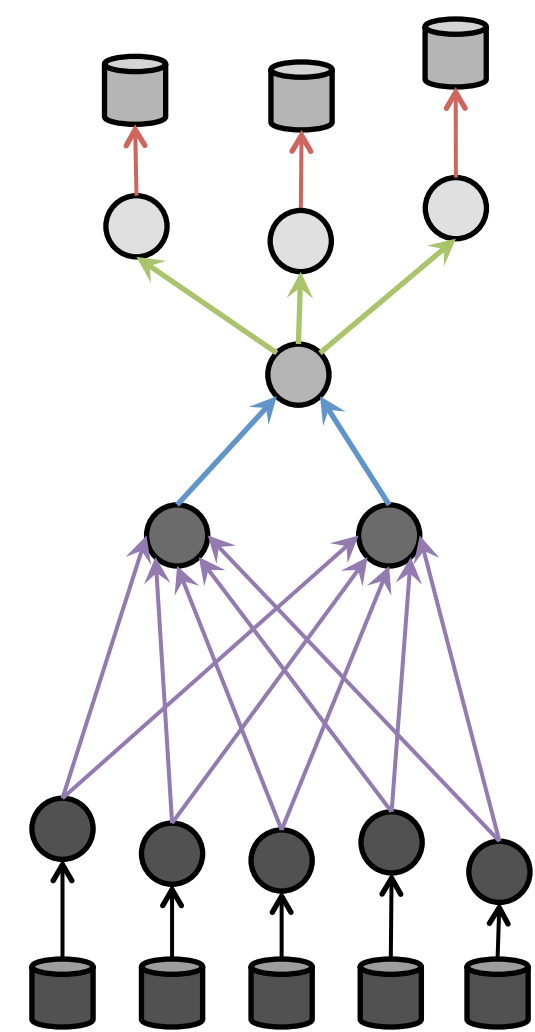
Mosharaf Chowdhury, Gautam Kumar, Sylvia Ratnasamy, Ion Stoica



Cluster Applications

Multi-Stage Data Flows

- » Computation interleaved with communication



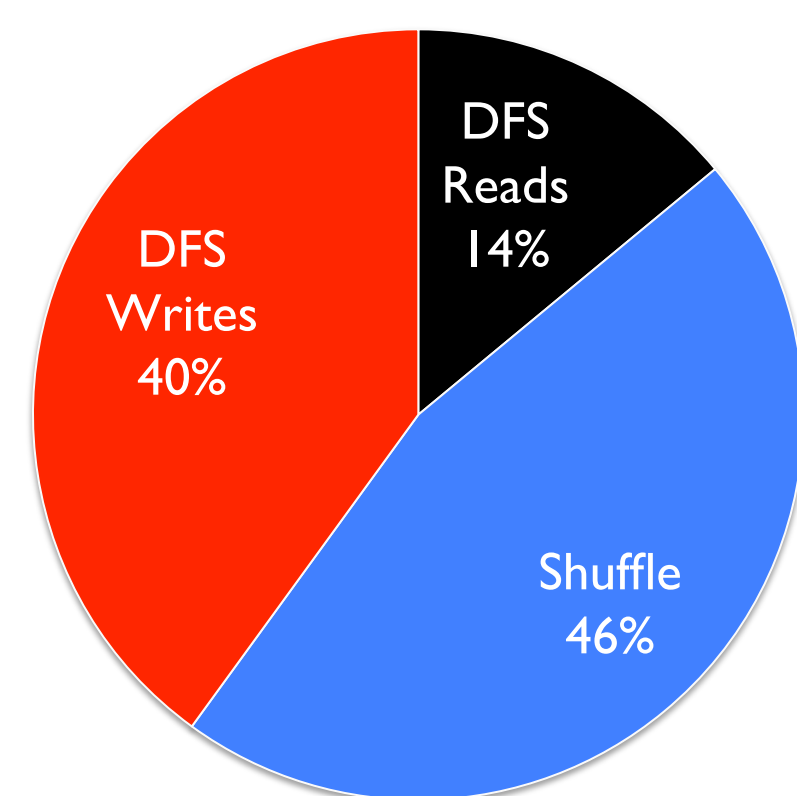
Computation

- » Distributed
- » Runs on many machines

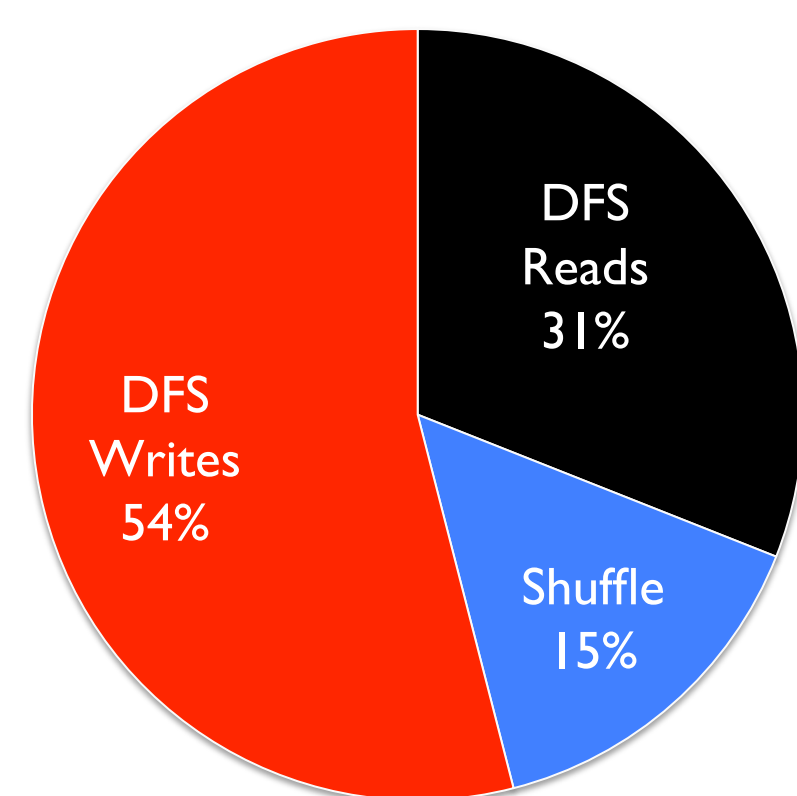
Communication

- » Structured
- » Between machine groups

Data-Intensive Network Traffic



Trace from a 3000-node Hadoop cluster



Trace from a "large" Cosmos cluster

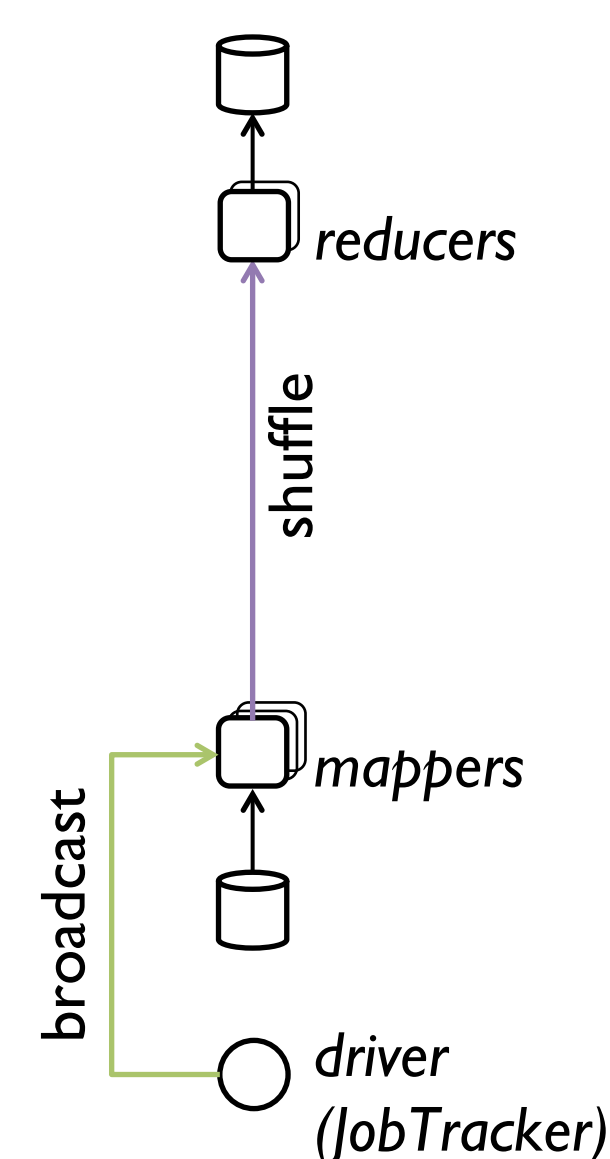
The Flow Abstraction

We get

- » Point-to-point comm.
- » Sequence of packets
- » Independent

We want

- » Multipoint-to-multipoint
- » Collection of flows
- » Shared Objective



The Coflow Abstraction

A semantically-bound collection of flows

Captures and Conveys application intent to the network

- » Performance-centric allocation of the network
- » Greater flexibility in designing applications

A flow is a coflow as well

Examples

Communication Pattern	Coflow	Objective
Intermediate transfers	Many-to-many (Shuffle)	Min completion time
Data dissemination	One-to-many (Broadcast)	Min completion time
Aggregation	Many-to-one (Reduce)	Min completion time
DFS replication	Constrained Anycast	Min completion time
Aggregation	Many-to-one (Incast)	Meet Deadline
Point-to-point	One-to-one	Either

The Coflow API

@driver

```

b ← create(BCAST)
s ← create(SHUFFLE)
...
b.put(id, content)
...
b.terminate()
s.terminate()
    
```

@mapper

```

b.get(id)
...
s.put(ids)
...
    
```

@reducer

```

s.get(ids)
...
    
```

Coflow Scheduler

Input: Diverse coflows arriving over time

- » Some attributes are unknown upon arrival

Output: Allocate resources in near real-time

- » Multi-objective optimization

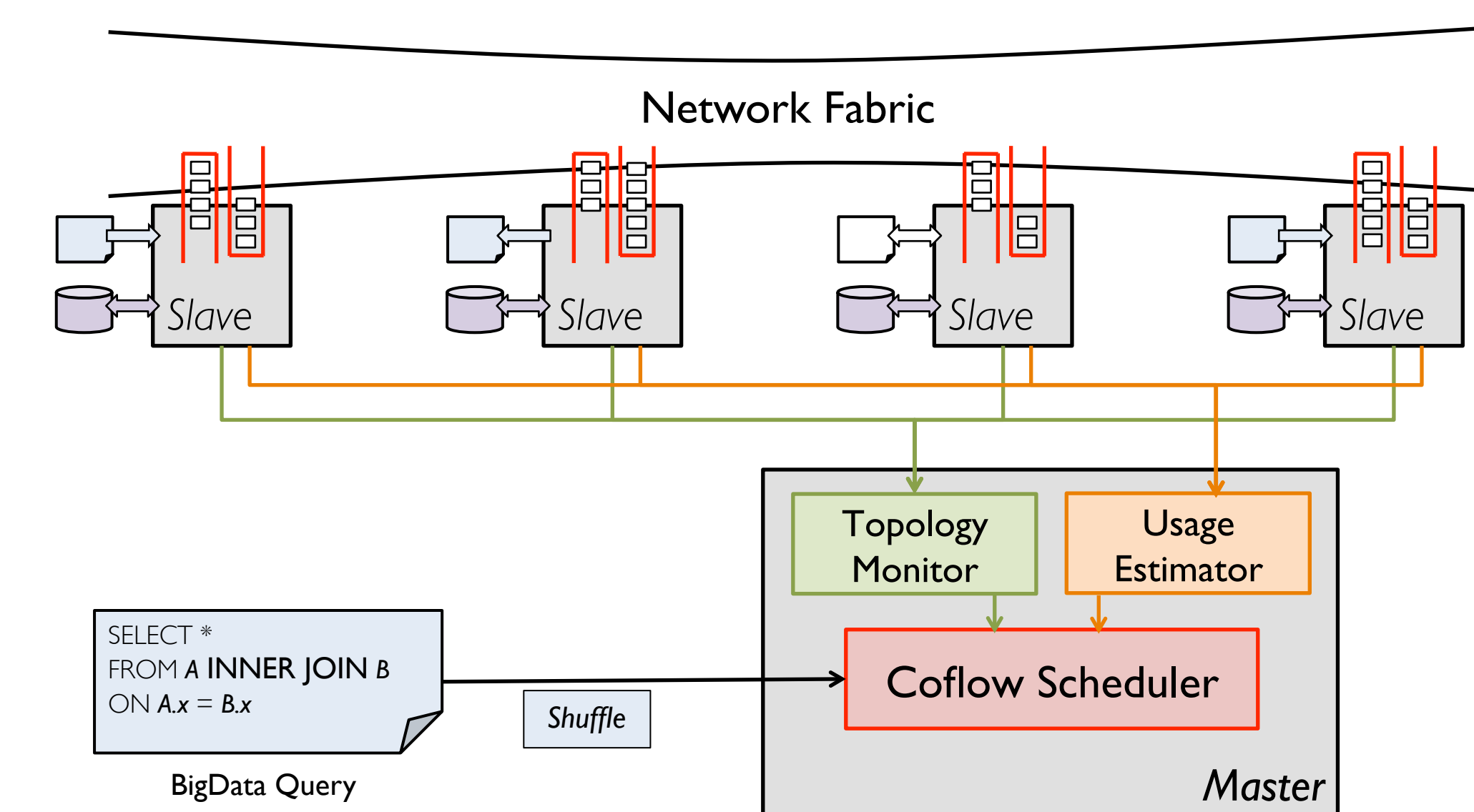
Proven to be NP-hard

- » SRTF et al. heuristics do not work that well
- » LICF (Least-Impact-Coflow-First) is the best so far
- » Uses preemption at the block-level

Enforcement is a major challenge

- » Looking at both app-layer and SDN solutions

System Architecture



Being developed in Scala/Java with a Thrift interface for external applications

Reading List

Overview

- » Coflow: A Networking Abstraction for Cluster Applications – HotNets 2012

Performance Improvements

- » Leveraging Flexibility in Endpoint Placement for a Snappier Network – SIGCOMM 2013 (Submitted)
- » Managing Data Transfers in Computer Clusters with Orchestra – SIGCOMM 2011

Allocation/Sharing

- » FairCloud: Sharing The Network in Cloud Computing – SIGCOMM 2012
- » A Case for Performance-Centric Network Allocation – HotCloud 2012