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

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

<table>
<thead>
<tr>
<th>Communication Pattern</th>
<th>Coflow</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate transfers</td>
<td>Many-to-many (Shuffle)</td>
<td>Min completion time</td>
</tr>
<tr>
<td>Data dissemination</td>
<td>One-to-many (Broadcast)</td>
<td>Min completion time</td>
</tr>
<tr>
<td>Aggregation</td>
<td>Many-to-one (Reduce)</td>
<td>Min completion time</td>
</tr>
<tr>
<td>DFS replication</td>
<td>Constrained Anycast</td>
<td>Min completion time</td>
</tr>
<tr>
<td>Aggregation</td>
<td>Many-to-one (Incast)</td>
<td>Meet Deadline</td>
</tr>
<tr>
<td>Point-to-point</td>
<td>One-to-one</td>
<td>Either</td>
</tr>
</tbody>
</table>

The Coflow API

```java
@driver
b < create(BCAST)
s < create(SHUFFLE)
...
b.put(id, content)
...
b.terminate()
s.terminate()
```

Examples

- 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

- System Architecture

- 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

- 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
    - A Case for Performance-Centric Network Allocation – HotCloud 2012