Traffic Optimization for Hadoop Based on SDN

Siyuan Liu, Jiaxu Zhu, Xinyi Wang

UCSDCSE
Computer Science and Engineering

Problem

Because of congestion from Hadoop application services and other applications sharing the network, also some improper bandwidth distribution for mappers and reducers, Hadoop applications cannot fully take use of the network and distributed systems. More networking problems are caused by properties of Hadoop:

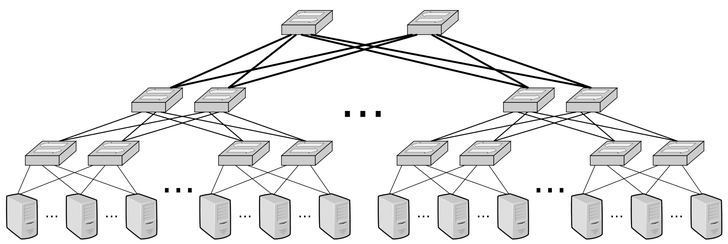
- After a period of map tasks, the reduce tasks start;
- The tasks distribution of Hadoop according to the number of mappers and reducers.

Goal

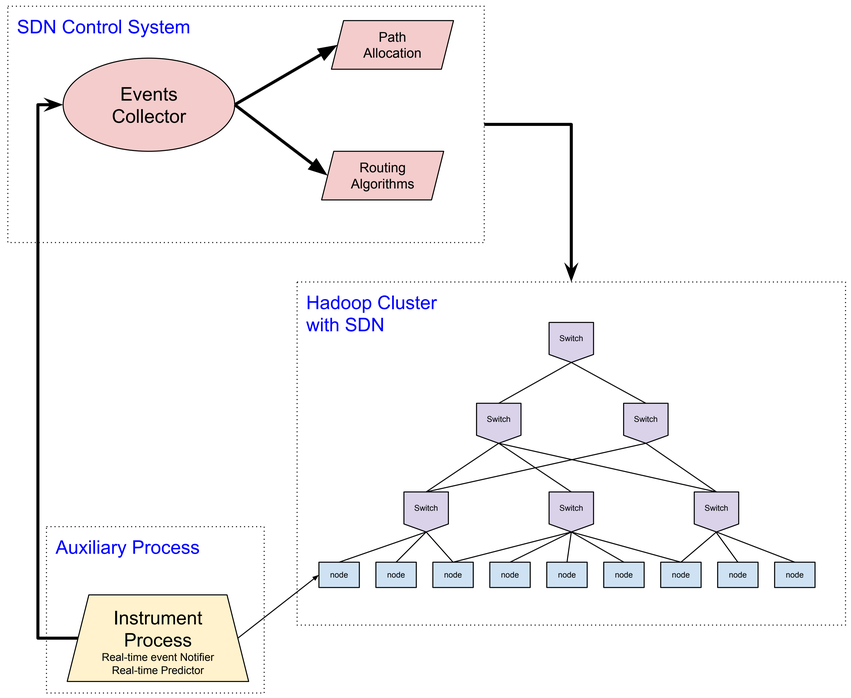
- Design and implement a traffic optimization system based on SDN for improving performance of Hadoop applications.
- Build an SDN control system independently. The SDN system will not influence how the Hadoop works.

Design

- Prototype:
  - Hadoop Cluster Nodes
  - Open Source Switches
  - SDN controller based on OpenFlow
- System Process:
  - Hadoop runs on a network with Fat Tree topology.
  - SDN controller collects information and executes routing computation and flow scheduling.



OpenFlow hadoop

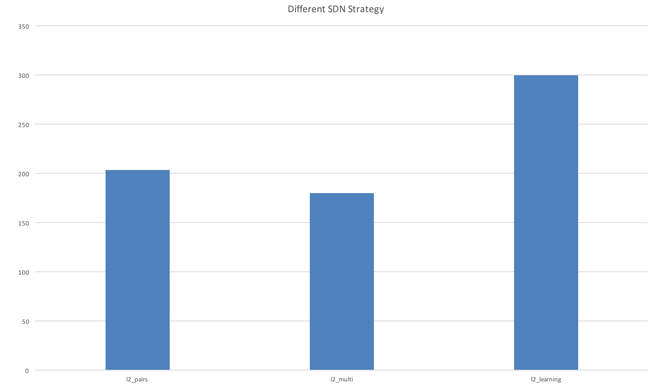


Implementation

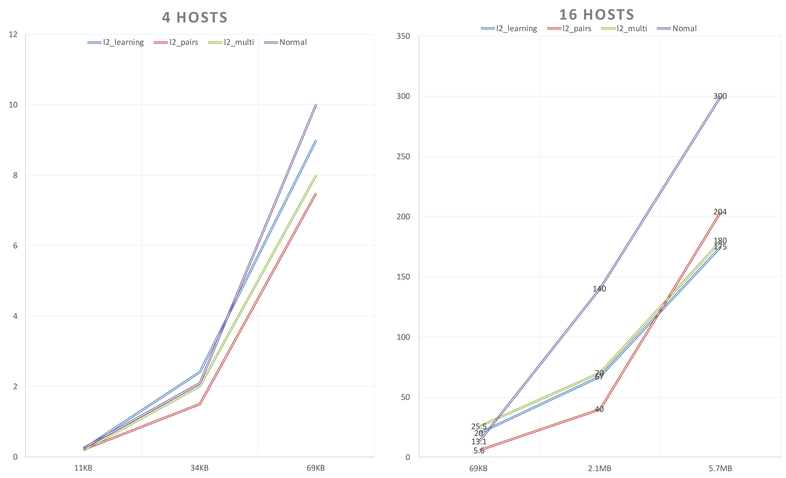
- Set up Hadoop cluster using Vagrant
  - Set up virtual machines using Vagrant.
  - Deploy Hadoop using Cloudera Manager.
- Replace general switches in the network with Open vSwitch
- Build network topology using Mininet
- Build SDN control system with POX

Experiment

- **Question:** Whether or not the performance of Hadoop is improved with SDN?
  - **Experiment**: Hadoop job completion time of general network and SDN
- **Question**: Whether or not the the performance of many kinds of Hadoop applications is improved?
  - **Experiment**: Hadoop job completion time of a sort of Hadoop Benchmarks.



- **Question:** What optimal SDN strategy details truly improve the performance of Hadoop?
  - Experiment: Hadoop job completion time of different SDN strategy adjustment.



Future

- Make experiments for more optimal SDN details
  - fully take use of information collected
- Implement a integral SDN system for Hadoop and other computing application.
- Move to real machines: wired machines and large scaled network
  - deploy SDN system on real network clusters
  - try to make the SDN system work on data centers