Traffic Optimization for Hadoop Based on SDN

Group 7: Siyuan Liu, Jiaxu Zhu, Xinyi Wang

UCSDCSE
Computer Science and Engineering

Problem:

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: Efficient Transmission

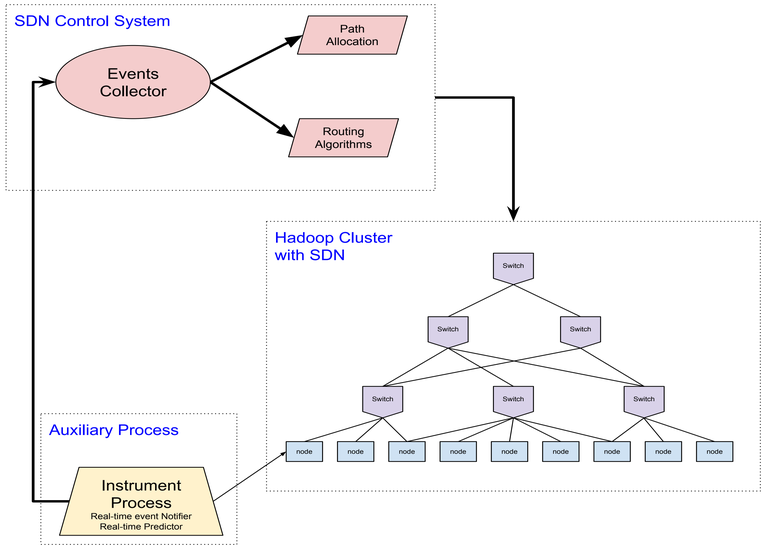
- Design and implement a traffic optimization system based on SDN for improving performance of Hadoop applications.
- Build an SDN control system independently. The SDN system will not influence how the Hadoop works.

Design: System Building

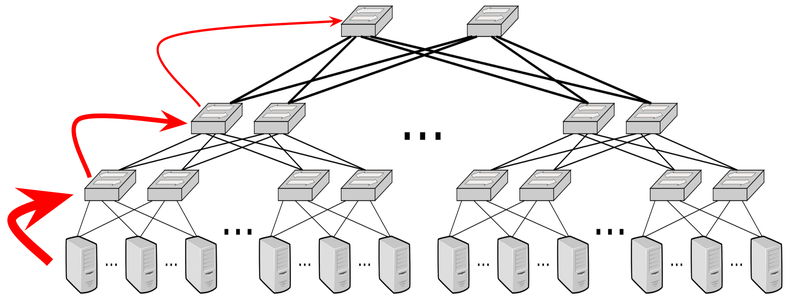
- Prototype:
  - Hadoop Cluster Nodes: **Vagrant** + **Cloudera**
  - Network Topology: **Mininet**
  - Open Source Switches: **Open vSwitch**
  - SDN controller based on OpenFlow: **POX**
- System Process:
  - Hadoop runs on a Fat Tree.
  - SDN controller collects information and executes routing computing and flow scheduling.

Network Scheduling: Modified *Pythia*

- Routing Computing: K-shortest Path
  - Hop-count Based: ensure routing paths on existing network topology.

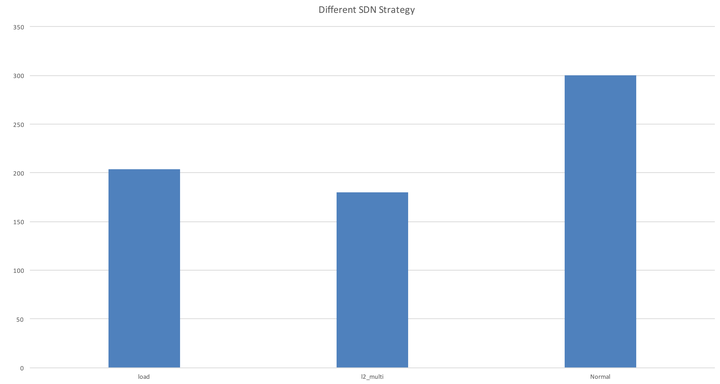


- Flow Scheduling: Load Graph for Scheduling
  - Update a “Load Graph” by information collected from Instrument Processes
  - Use “Load Graph” to distribute flow (Next hops for a switch) on k-shortest paths



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
  - make the SDN system work on data centers