Proposal

Group #7: Siyuan Liu, Xinyi Wang, Jiaxu Zhu

Introduction

Hadoop applications run on compute framework (MapReduce) try to exploit the resource of distributed systems (like distributed file system) to get scalable and reliable computing services. However, 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.

Software Defined Network nowadays separated control plane from data plane, which can be used to monitor, record and control a specific network. It means SDN realizes real-time analysis and enable Hadoop application to be controlled intelligently, which may solve the problem mentioned above and improve the performance of computing.

Our goal is to design and implement a traffic optimization system for Hadoop applications based on SDN. Hadoop applications manage and compute big data on data centers. The inner network of a data center is important to ensure the performance of Hadoop applications. SDN can monitor and control the data flow in real time, and help Hadoop to distribute jobs to devices in the data center and fully take advantage of the whole network.

Approach

We intend to make a change to existing ideas which will be briefly introduced in the next section.
Also, we have a plan about how to realize our goal, which consists of experiments, integration and optimization.

Related Work

In recent yeas, lots of researchers and organizations try to optimize the configuration of network of data centers in order to improve the performance of Hadoop and other computing applications. We find some achievement of researchers at UIUC. They were trying to build a transparent and flexible network framework for big data processing, which was called *FlowComb*. The researchers use logs on switches to realize real-time network management. In detail, they build a control center on the network to predict congestion, reschedule packets and control the interaction between mappers and reducers. Their goal is to fully take use of bandwidth of every path on the network when Hadoop applications exchange data of computation jobs.

There are other researchers citing *FlowComb* to build advanced frameworks or systems. Researchers from HUST try to monitor the the transmission of large data packets, which is the bottleneck in big data applications. The system dynamically forwards flows along multiple equal-cost paths or reschedules transmission, when congestion is detected at certain location. Other researchers from IBM build a system called *Pythia* that has significant overlap with *FlowComb*. They accelerate MapReduce job completion by

Prediction Intelligence which is collected at a central controller and in turn ingested by a chain of network control algorithms (routing, flow scheduling) that optimize network resource allocation.

We plan to firstly validate *FlowComb*, *Pythia* and other current state of the art, then build a new system upon all these work.

Plan

The whole project aims at exploring if SDN can help improve the computation performance of Hadoop systems. The first part is to find state-of-art systems of traffic optimization with SDN. Then we need to implement two or more systems to see the results.

The second step is to analyze the results and extract useful information and factors which do affect the performance of Hadoop system with SDN. Then we need to integrate these factors to design a traffic optimization system for Hadoop applications based on SDN. Through experiment, we make changes to this system and improve the whole performance of the traffic optimization system with SDN for Hadoop computing.

Infrastructures we may use in our experiments:

1. Several workstations (VM provided and laptops of our team members are qualified);
2. SDN-Mininet (Or other OVS) to create a network of hosts and switches, even simulate a data center;
3. Hadoop environment. In fact, we are trying to find a proper tool or some open source that fit in our project.

Schedule

11/5: Finish the first step;
11/24: Finish the second step;
12/3: Draft of the poster;
12/8: Final project report.