Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

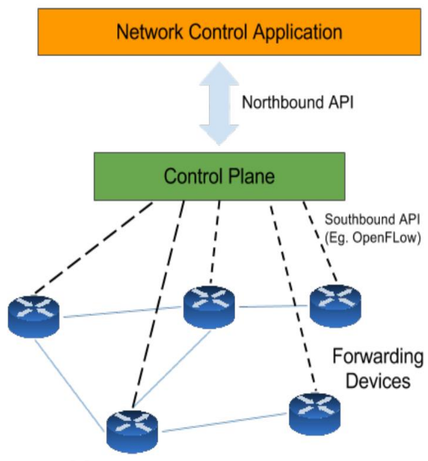
Proactive Configuration of Data Centre Networks for Big Data Processing

Harpreet Singh, Supervisor: Stefan Weber

Introduction

Project Motivation

- Big Data frameworks such as Hadoop need scaling out to thousands of commodity servers as data volumes for analysis grow
  - ➢ Increasing Network Traffic
- Network has been pointed out to be a performance bottleneck in the cloud
- Using SDN, that enables control of the entire network state through a central controller, reactive measures of traffic engineering to optimize a data centre network have been explored
  - ➢ Improves performance
  - ➢ Induces control traffic



Project Aims:

- Device a flow scheduling approach in accordance with big data application patterns, in particular Hadoop, that is proactive in nature
- Determine if there is an increase in network throughput and decrease in application job completion times
- Evaluate the effectiveness of our proactive approach against reactive flow scheduling approaches, namely
  - ECMP
  - Global First Fit

Design

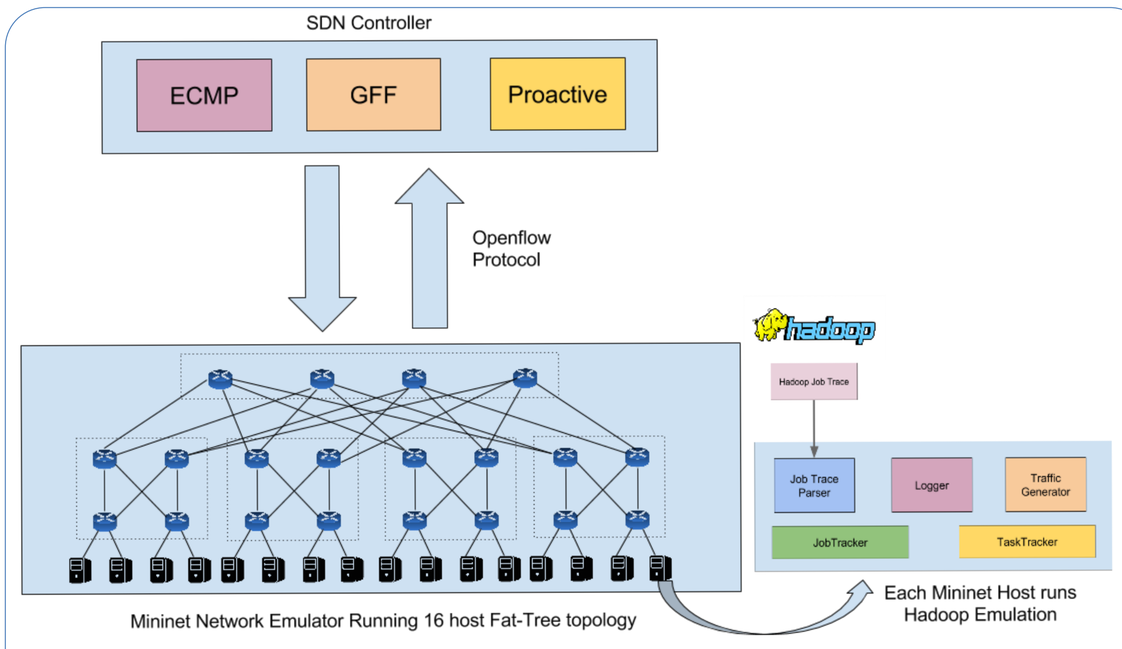


Figure above illustrates the overall high level design of our experiment. It consists of three components

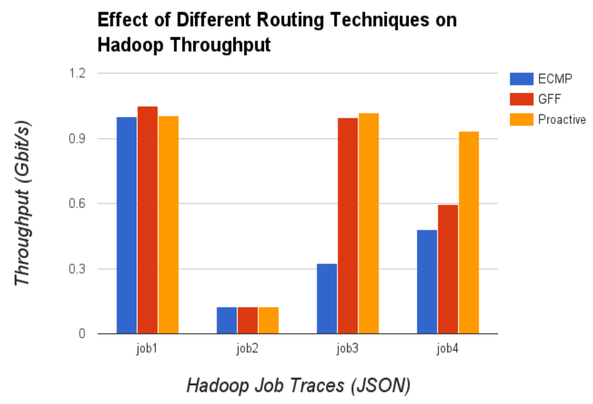
- SDN controller – Controls all switches in the network using OpenFlow protocol, routes traffic via
  - ➢ Reactive Scheduling – ECMP and GFF
  - ➢ Proactive Scheduling
- *Fat-tree* data centre topology with 16 hosts and 20 switches connected to the SDN controller
- Hadoop emulation running on each of the 16 hosts

M.Sc. in Computer Science
(Networks and Distributed Systems)

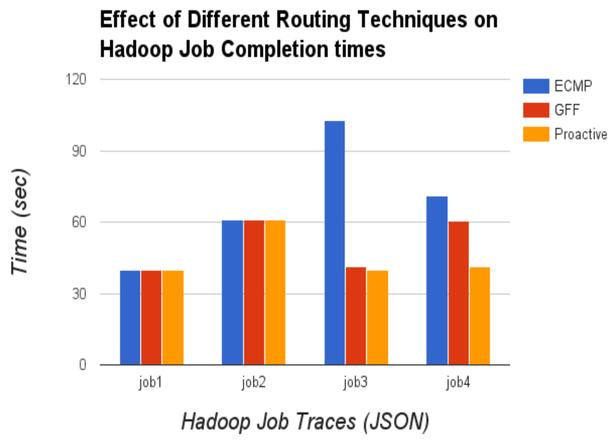
Implementation

- Implemented using Pox SDN Controller with network topology running on Mininet emulator
- Proactive Routing Algorithm
  - ➢ Route Hadoop traffic using GFF (reactive) flow scheduling and log decisions
  - ➢ Use logged decisions from GFF to proactively install flows in the next run of Hadoop, subsequently, default back to GFF
- Measure total bandwidth achieved and Hadoop Job completion times for evaluation

Evaluation



- Effect of flow scheduling on total throughput achieved
- Proactive routing achieves
  - ➢ 59.9% higher throughput than ECMP scheduling
  - ➢ 11.9% higher throughput than GFF scheduling



- Effect of flow scheduling on Hadoop Job Completion times
- Proactive routing achieves
  - ➢ 33.5% faster Job Completion than ECMP
  - ➢ 10% faster Job Completion than GFF

The Hadoop Jobs used for evaluation are traces of Hadoop Jobs forming a part of HiBench application suite (Sort, Nutch, PageRank, Bayers) that have been used by other researchers in evaluations.

Conclusion and Future Work

To optimize data centre network traffic workloads, we introduced a proactive approach of flow scheduling and found it to achieve better performance levels than reactive approaches.
We plan to run the experiment on a real cluster, running Hadoop, and further extend the proactive flow scheduling by adding the ability to automatically generate proactive configuration of the network, based on application patterns.

School of Computer Science & Statistics (SCSS)
Faculty of Engineering, Mathematics and Science