Phurti: Application and Network-Aware Flow Scheduling for MapReduce

Chris Cai, Shayan Saeed, Indranil Gupta, Roy Campbell
BACKGROUND
Software Defined Networking

Applications

Controller

Switches
Network Traffic for MapReduce Jobs

Job 1 Network Traffic
Job 2 Network Traffic

Map Task of Job 1
Map Task of Job 2
Map Task of Job 1
Map Task of Job 2

Node 1
- Reduce Task of Job 1
- Reduce Task of Job 2

Node 2
- Reduce Task of Job 1
- Reduce Task of Job 2

Map Phase
Shuffling Phase
Reduce Phase
MOTIVATION
Application-Awareness in Scheduling

Network Flows for jobs

<table>
<thead>
<tr>
<th>ID</th>
<th>Link</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>$A_2$</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>$A_3$</td>
<td>Z</td>
<td>3</td>
</tr>
<tr>
<td>$B_1$</td>
<td>X</td>
<td>3</td>
</tr>
<tr>
<td>$B_2$</td>
<td>Y</td>
<td>1</td>
</tr>
<tr>
<td>$B_3$</td>
<td>Z</td>
<td>2</td>
</tr>
<tr>
<td>$C_1$</td>
<td>X</td>
<td>2</td>
</tr>
</tbody>
</table>

Fair Sharing

Application-aware Scheduling
Network-Awareness in Scheduling

Network-Agnostic Scheduling

Network-Aware Scheduling
ARCHITECTURE
Phurti: Detecting Flow Interference

OpenFlow switch

N1
N2
N3
N4

Flow 1
Flow 2
Flow 3
ALGORITHM
Smallest Maximum Sequential-traffic First (SMSF)

- Sequential-traffic $T_{ij}$ of a MapReduce job: the traffic a job needs to transmit from host $i$ to host $j$

Intuition behind SMSF: the size of maximum sequential-traffic of a job will likely determine its shuffle completion time

Maximum Sequential-traffic of Job 1: 300MB  Maximum Sequential-traffic of Job 2: 1G
Smallest Maximum Sequential-traffic First (SMSF)

Maximum Sequential-traffic of Job 1: 300MB

Maximum Sequential-traffic of Job 2: 1G

1Gbps Link

Job1: $(1-f) \times 1$ Gbps

Job2: $f \times 1$ Gbps 
$(0 < f < 1)$
EVALUATION
Evaluation

- Testbed: 6 nodes, 2 HP SDN switches
- SWIM workload: workload generated from Facebook Hadoop trace

<table>
<thead>
<tr>
<th>Job Size Bin</th>
<th>% of total jobs</th>
<th>% of total bytes in shuffled data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>62%</td>
<td>5.5%</td>
</tr>
<tr>
<td>Medium</td>
<td>16%</td>
<td>10.3%</td>
</tr>
<tr>
<td>Large</td>
<td>22%</td>
<td>84.2%</td>
</tr>
</tbody>
</table>
Job Completion Time: SMSF VS Fair Sharing

Difference in Job Completion Time between SMSF and Fair Sharing (sec)
## Job Completion Time Improvement By Job Types

<table>
<thead>
<tr>
<th>Job Type</th>
<th>Average Fractional Improvement</th>
<th>95th Percentile Fractional Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>Small</td>
<td>0.22</td>
<td>0.24</td>
</tr>
<tr>
<td>Medium</td>
<td>0.14</td>
<td>0.16</td>
</tr>
<tr>
<td>Large</td>
<td>0.16</td>
<td>0.17</td>
</tr>
</tbody>
</table>

- **Overall** shows the lowest improvement, with an average of 0.20 and a 95th percentile of 0.25.
- **Small** has the highest improvement, with an average of 0.22 and a 95th percentile of 0.24.
- **Medium** and **Large** fall in between, with average improvements of 0.14 and 0.16, respectively, and 95th percentiles of 0.16 and 0.17, respectively.

The bar chart illustrates these improvements, showing the range of fractional improvement across different job types.
Questions
ACC Demo
Rescue Coordination Center

Local Rescue Agency

Phurti: Assurance for mission-critical traffic

Rescue Coordination Center

Dedicated Network