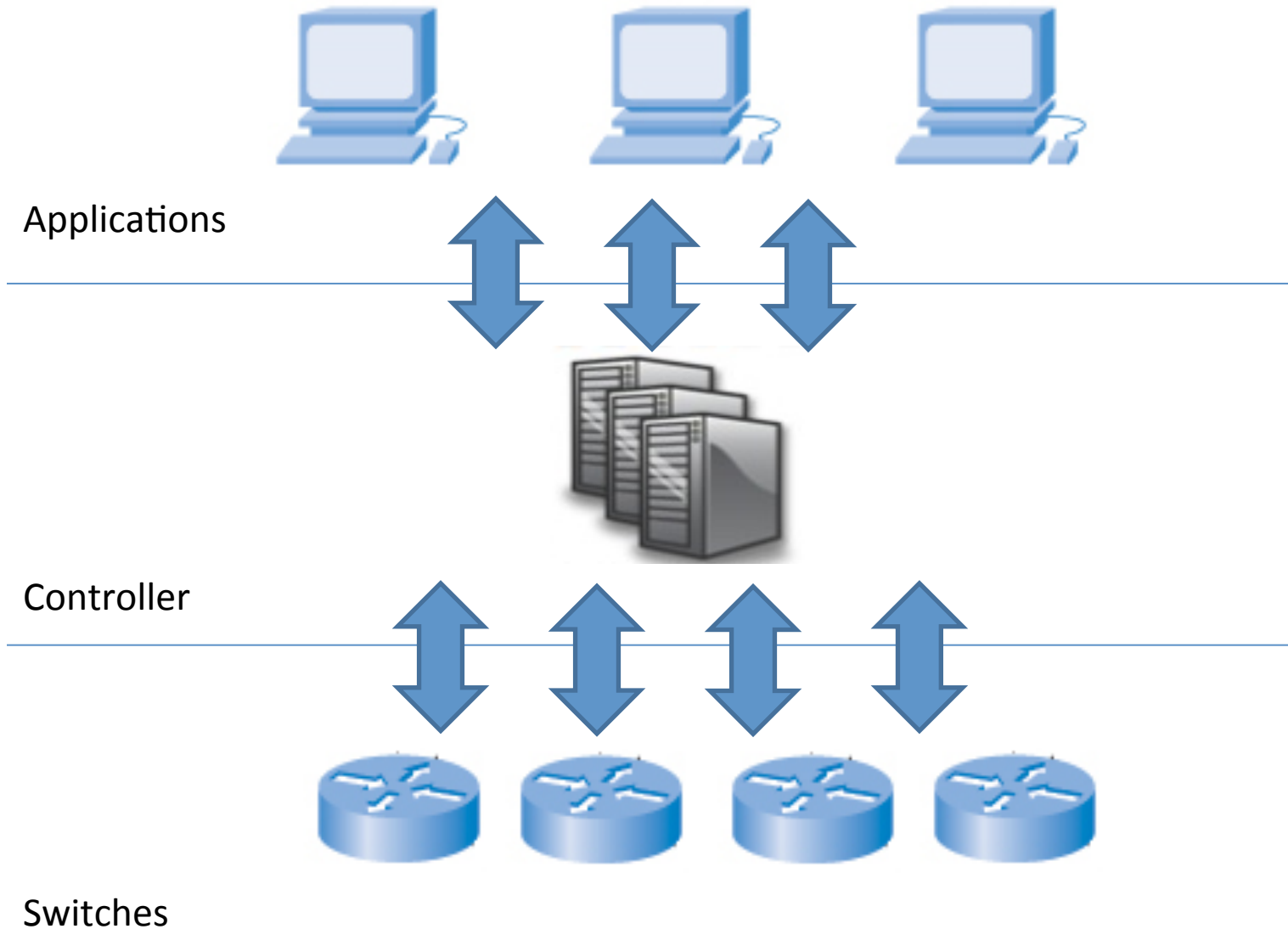


Phurti: Application and Network-Aware Flow Scheduling for MapReduce

Chris Cai, Shayan Saeed, Indranil
Gupta, Roy Campbell

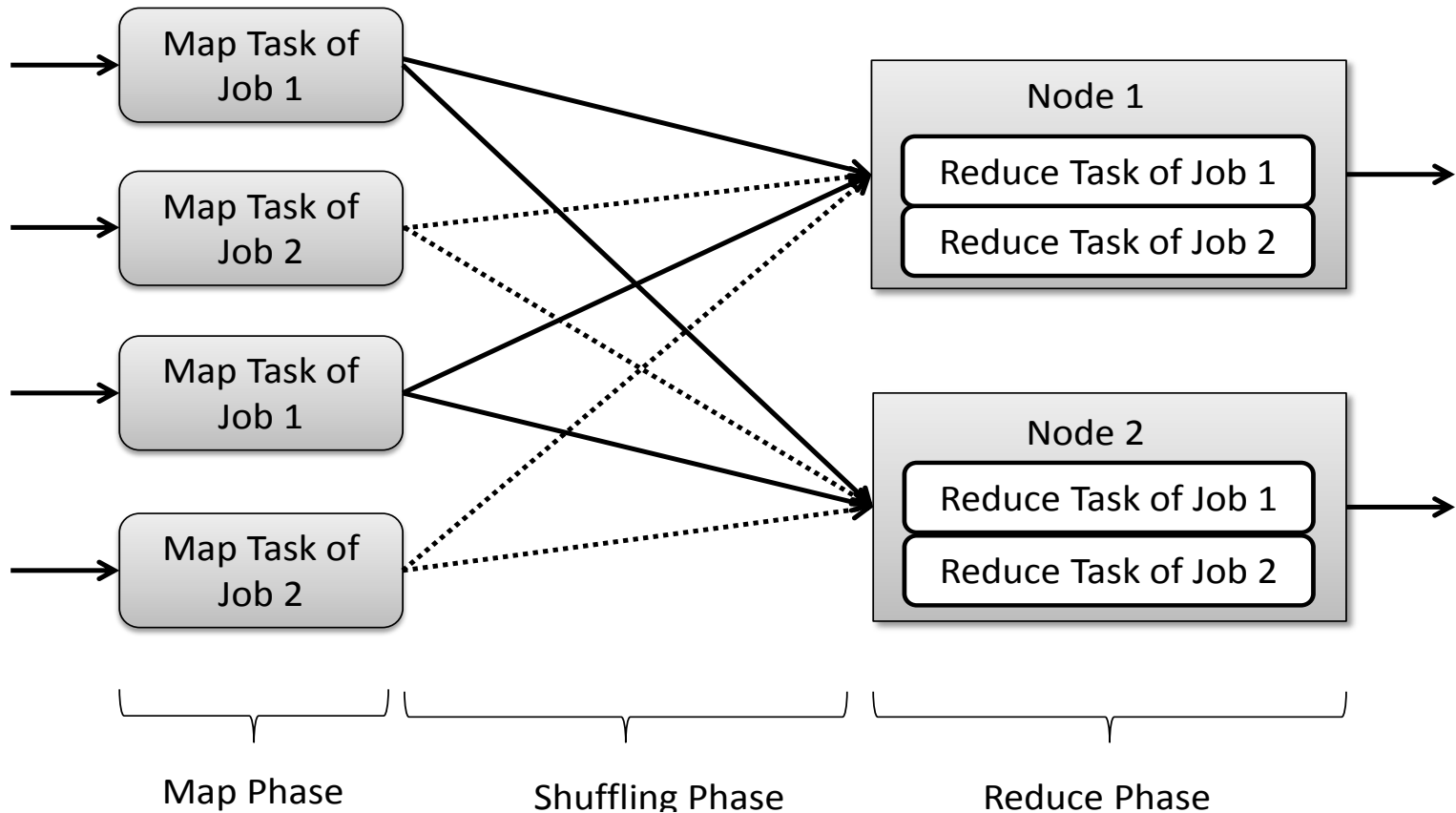
BACKGROUND

Software Defined Networking



Network Traffic for MapReduce Jobs

Job 1 Network Traffic —————>
Job 2 Network Traffic ·········>

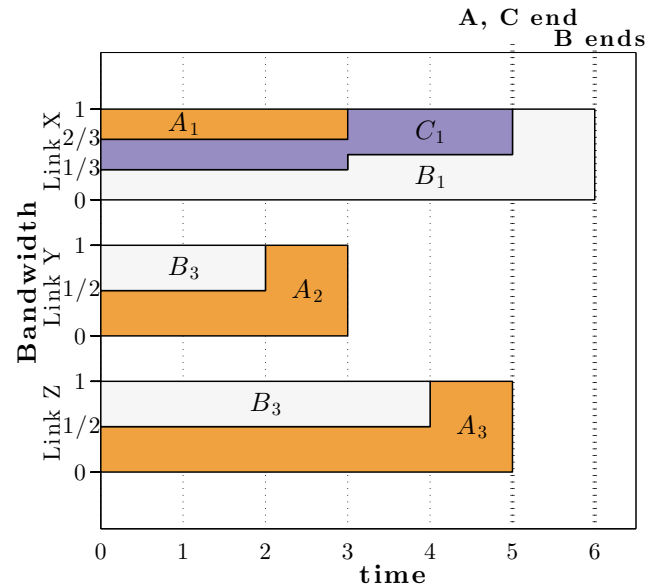


MOTIVATION

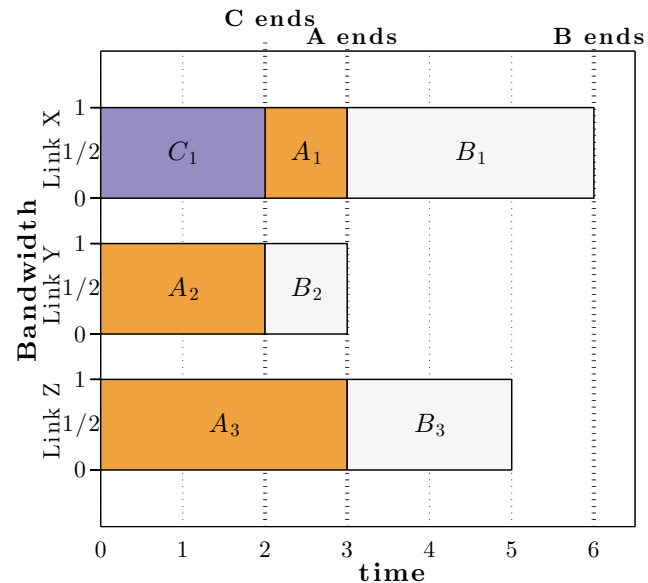
Application-Awareness in Scheduling

ID	Link	Size
A_1	X	1
A_2	Y	2
A_3	Z	3
B_1	X	3
B_2	Y	1
B_3	Z	2
C_1	X	2

Network Flows for jobs

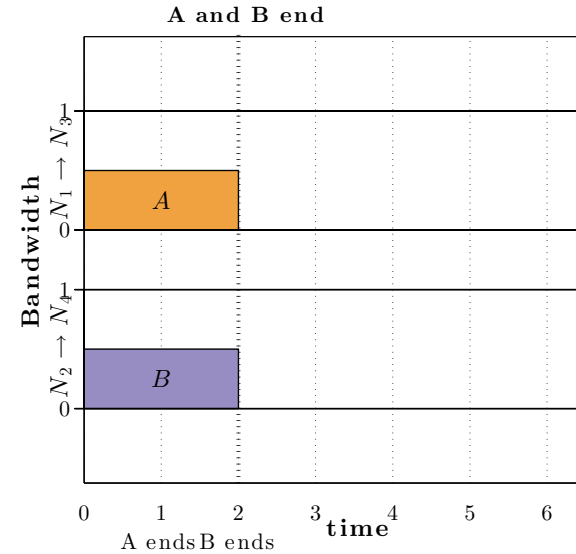
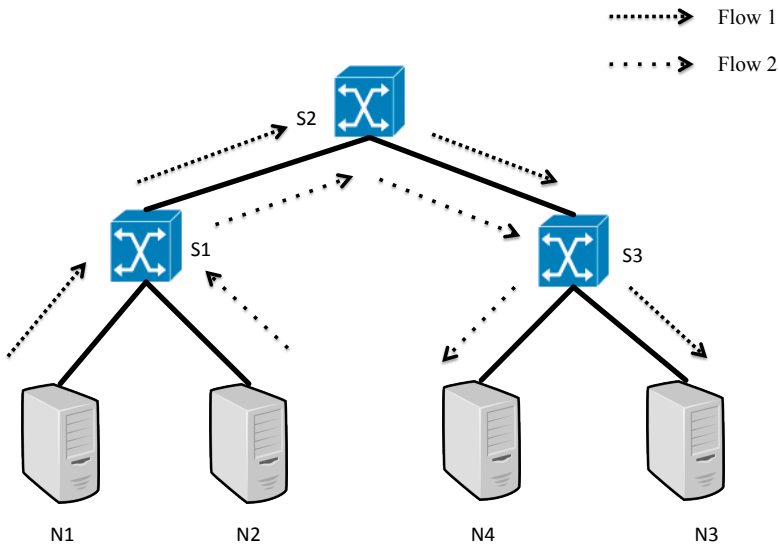


Fair Sharing

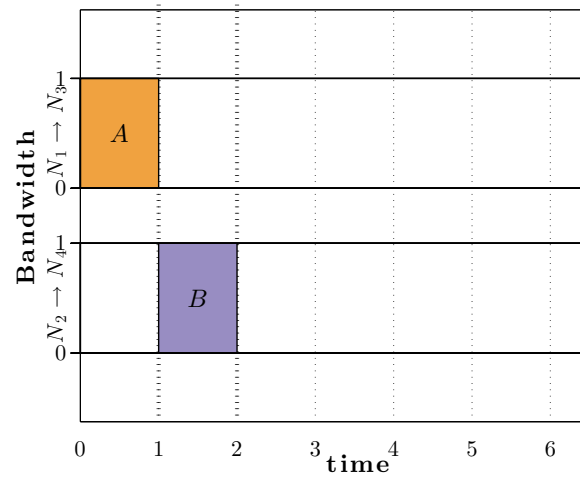


Application-aware Scheduling

Network-Awareness in Scheduling



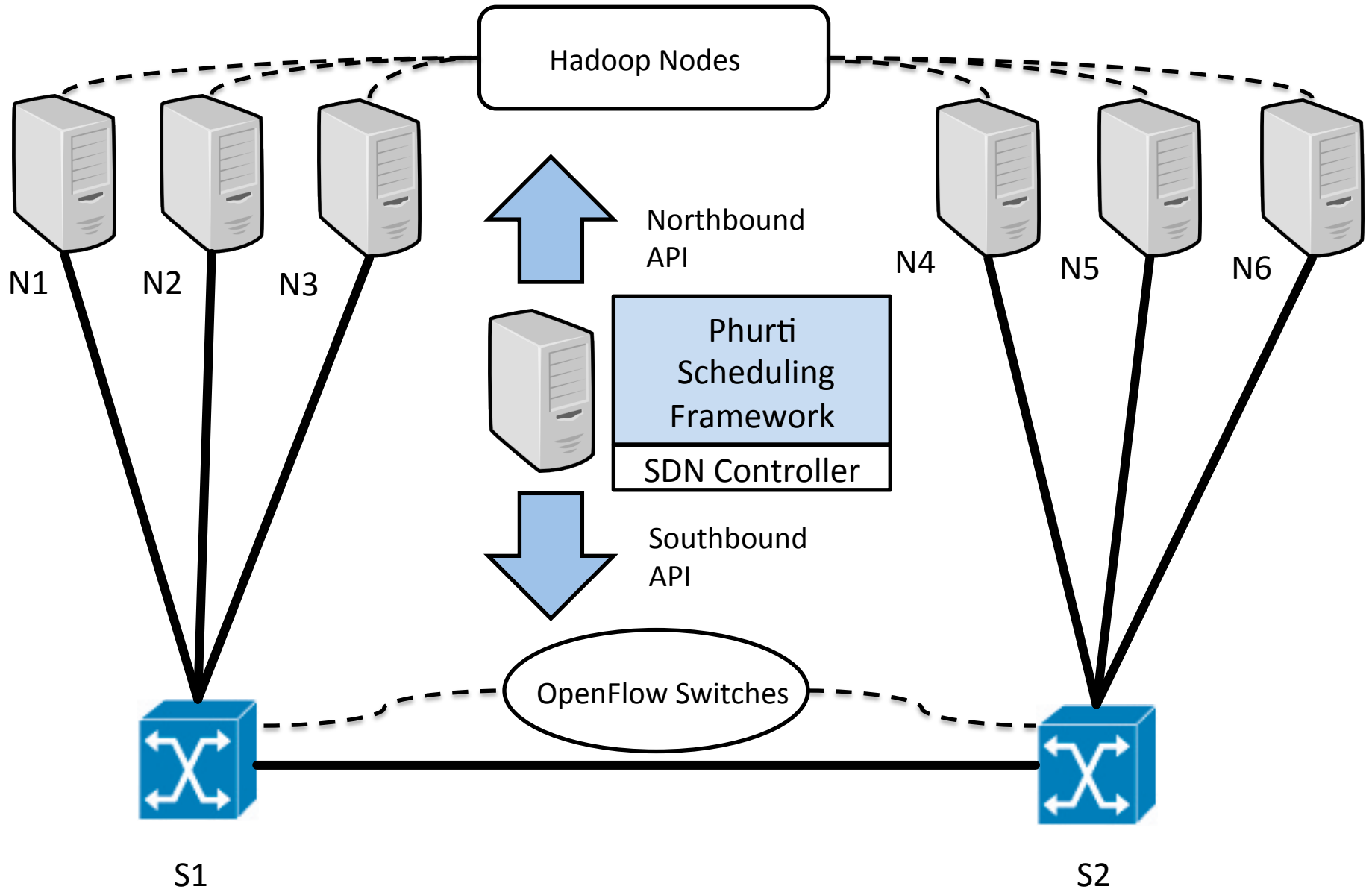
Network-Agnostic Scheduling



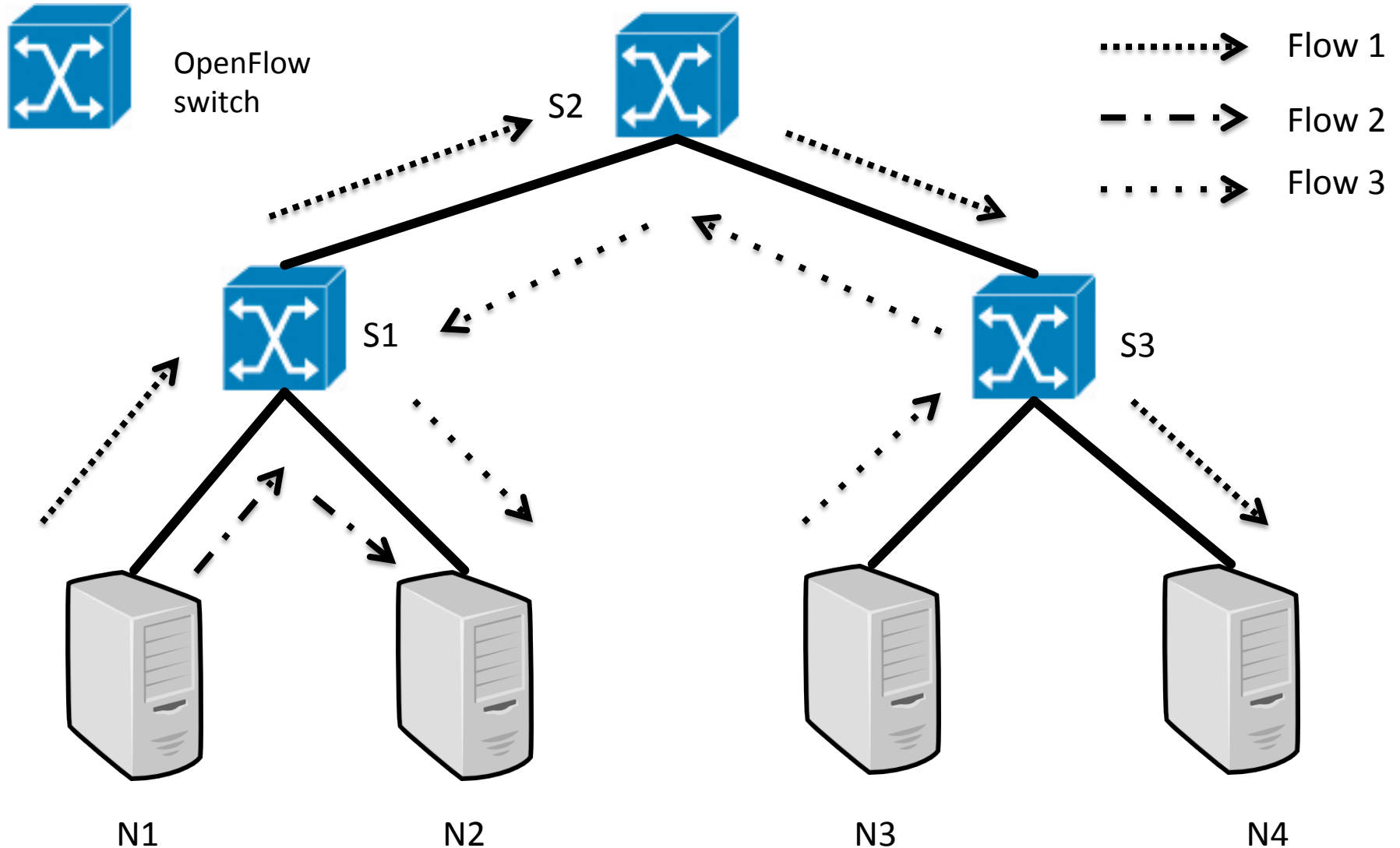
Network-Aware Scheduling

ARCHITECTURE

Phurti Architecture



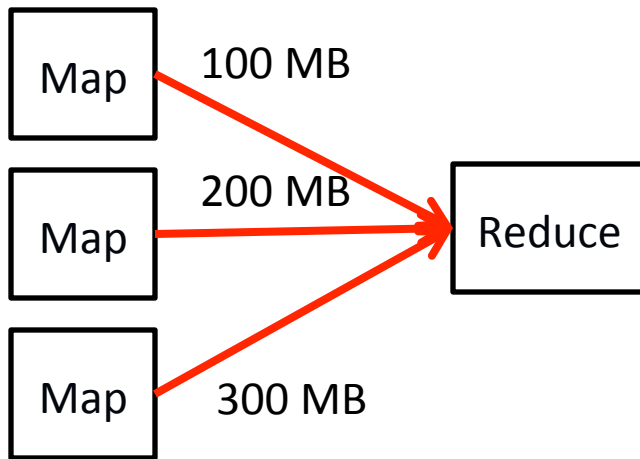
Phurti: Detecting Flow Interference



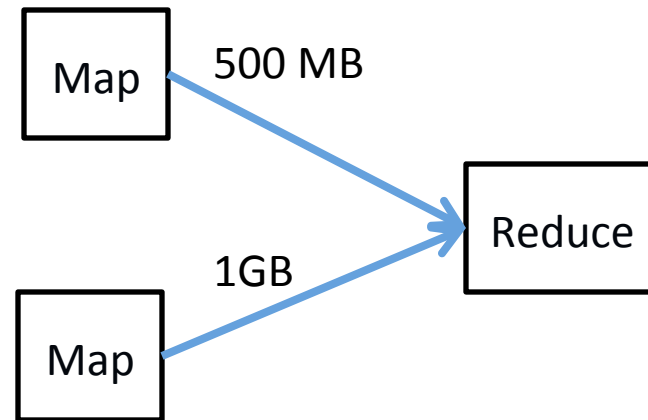
ALGORITHM

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



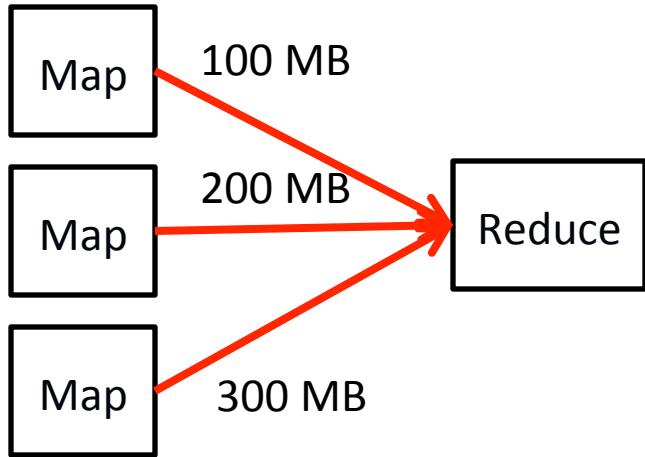
Maximum Sequential-traffic of Job 1: 300MB



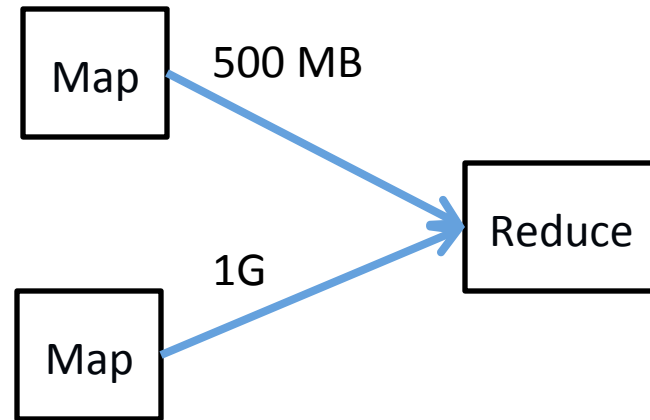
Maximum Sequential-traffic of Job 2: 1G

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

Smallest Maximum Sequential-traffic First (SMSF)

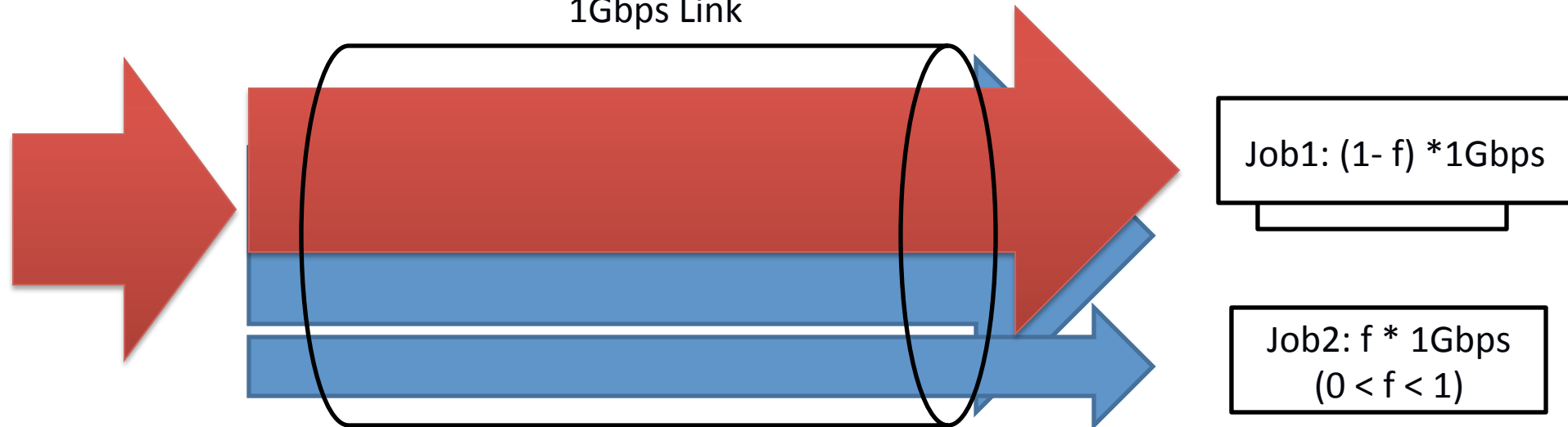


Maximum Sequential-traffic of Job 1: 300MB



Maximum Sequential-traffic of Job 2: 1G

1Gbps Link



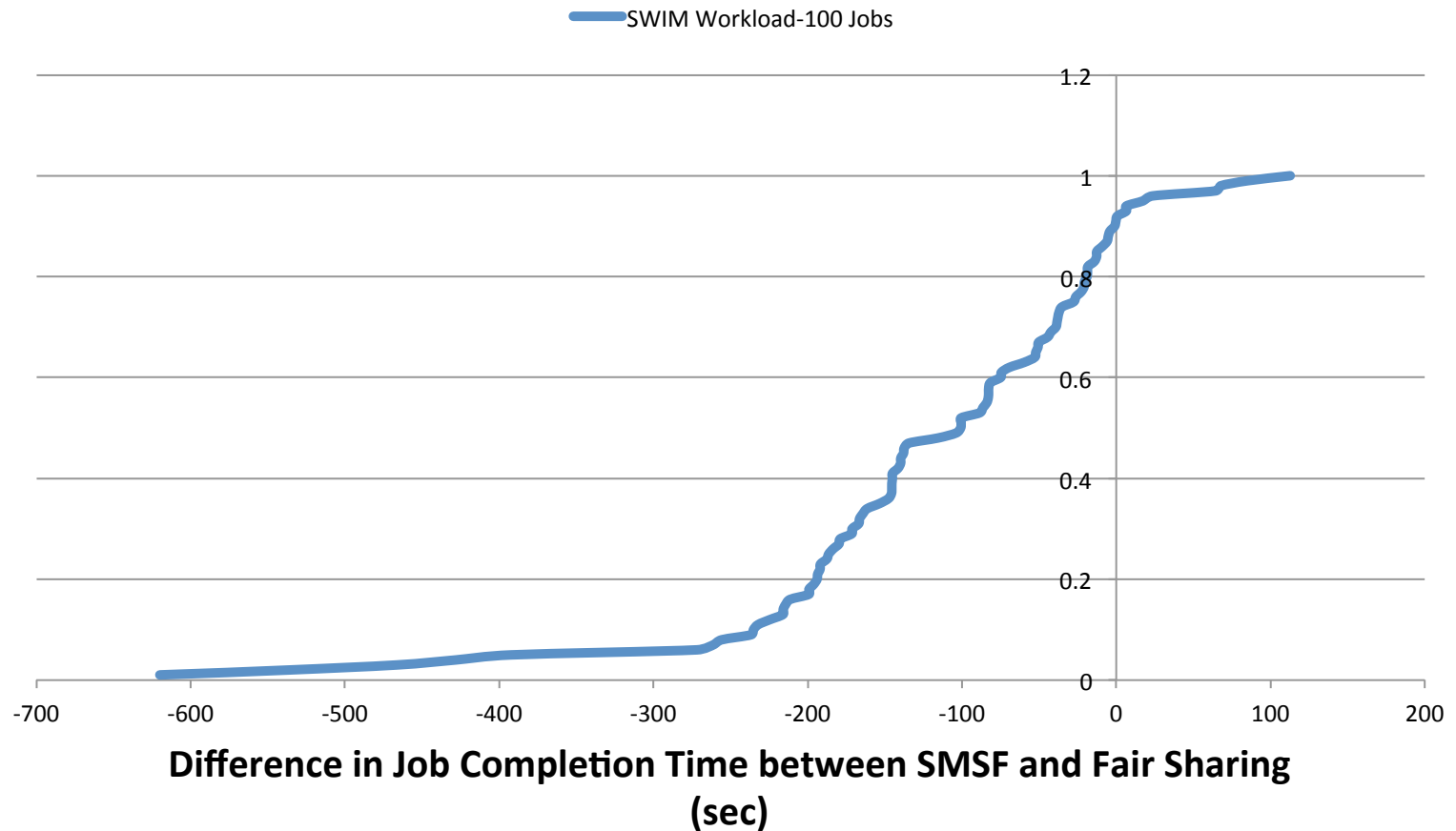
EVALUATION

Evaluation

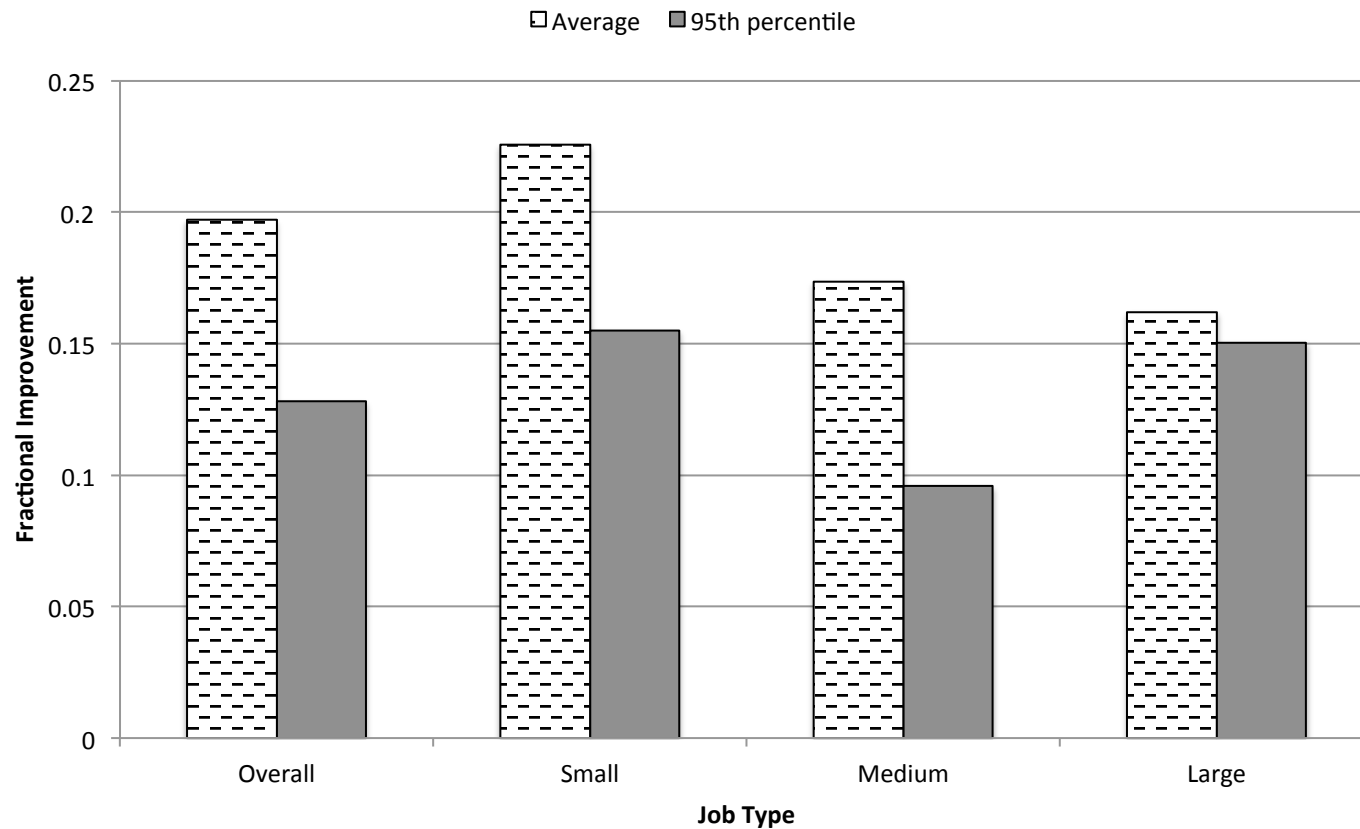
- Testbed: 6 nodes, 2 HP SDN switches
- SWIM workload: workload generated from Facebook Hadoop trace

Job Size Bin	% of total jobs	% of total bytes in shuffled data
Small	62%	5.5%
Medium	16%	10.3%
Large	22%	84.2%

Job Completion Time: SMSF VS Fair Sharing

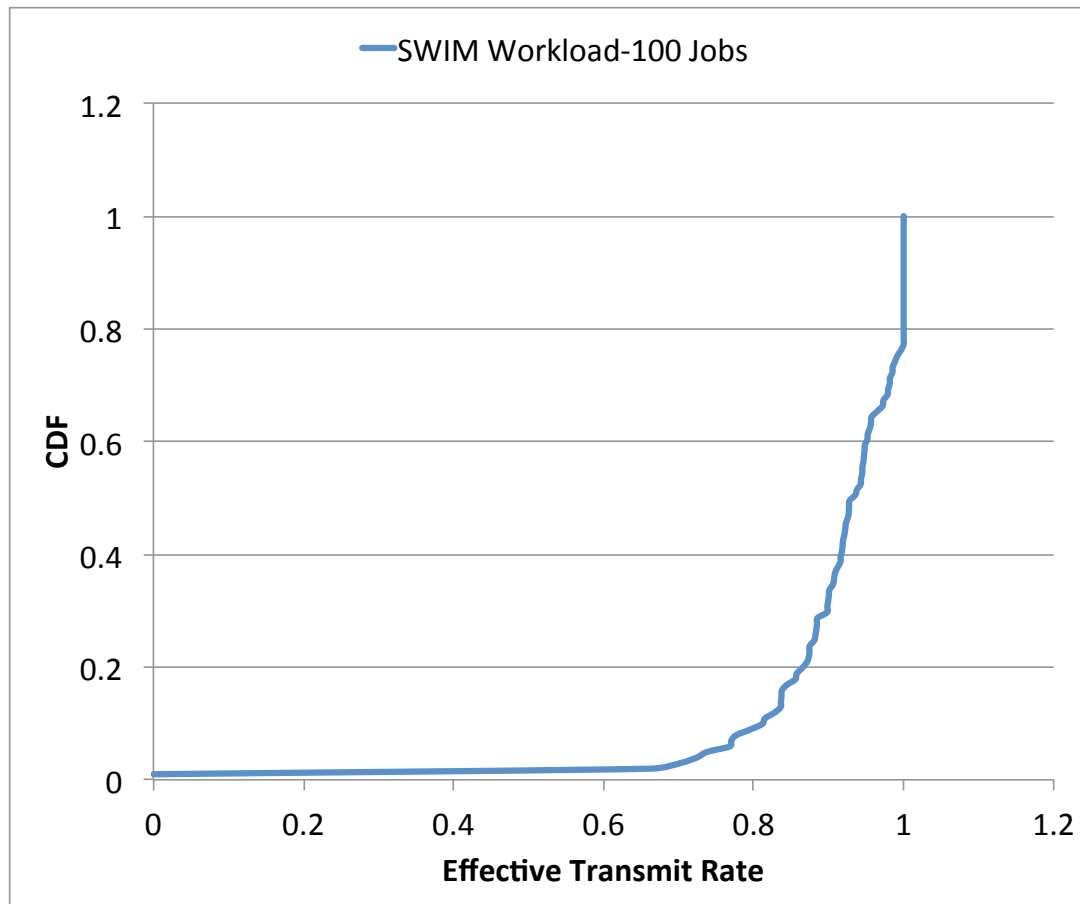


Job Completion Time Improvement By Job Types

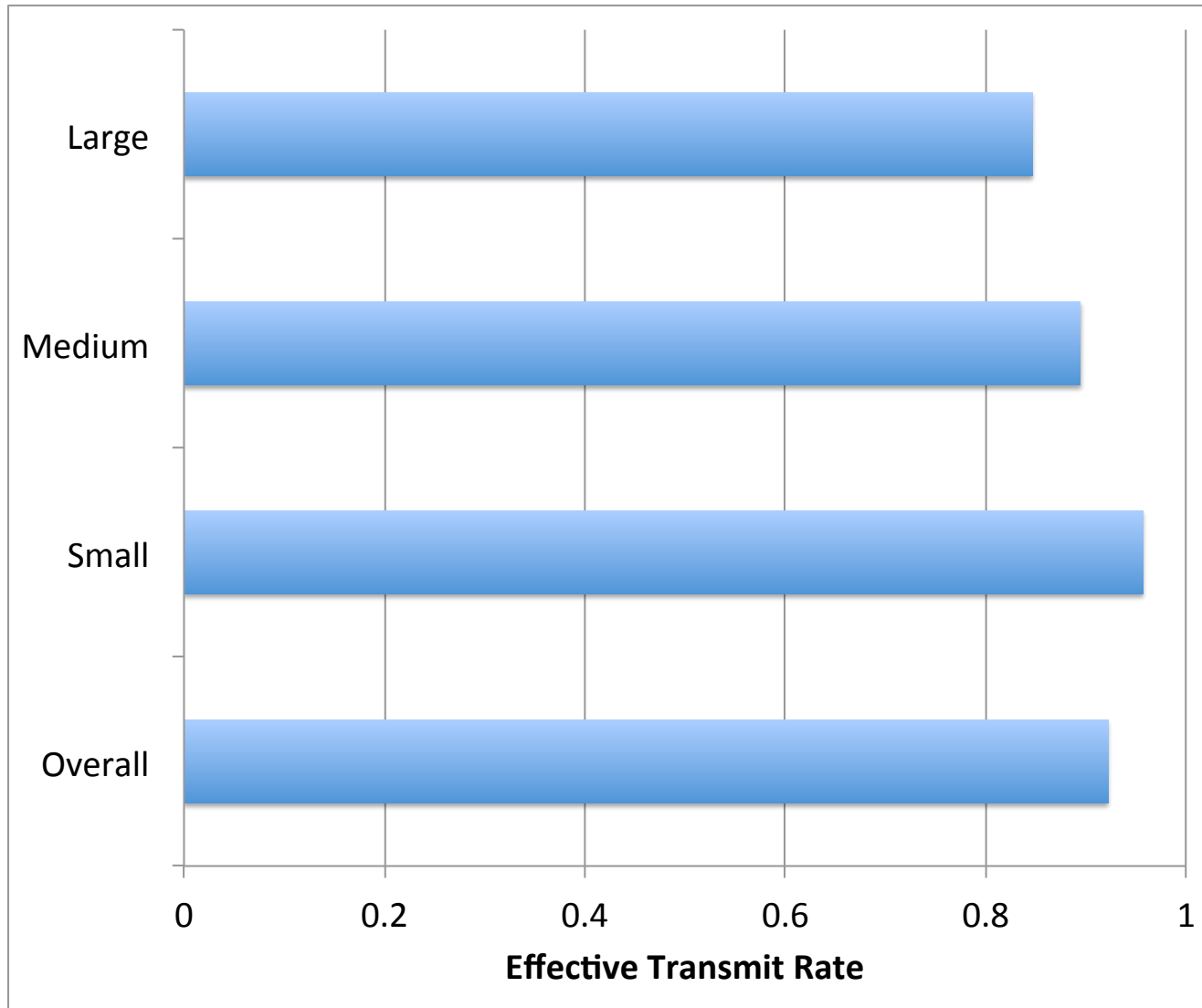


Effective Transmit Rate

Effective Transmit Rate: percentage of time a flow spends in “Transmit” state during its entire transmission

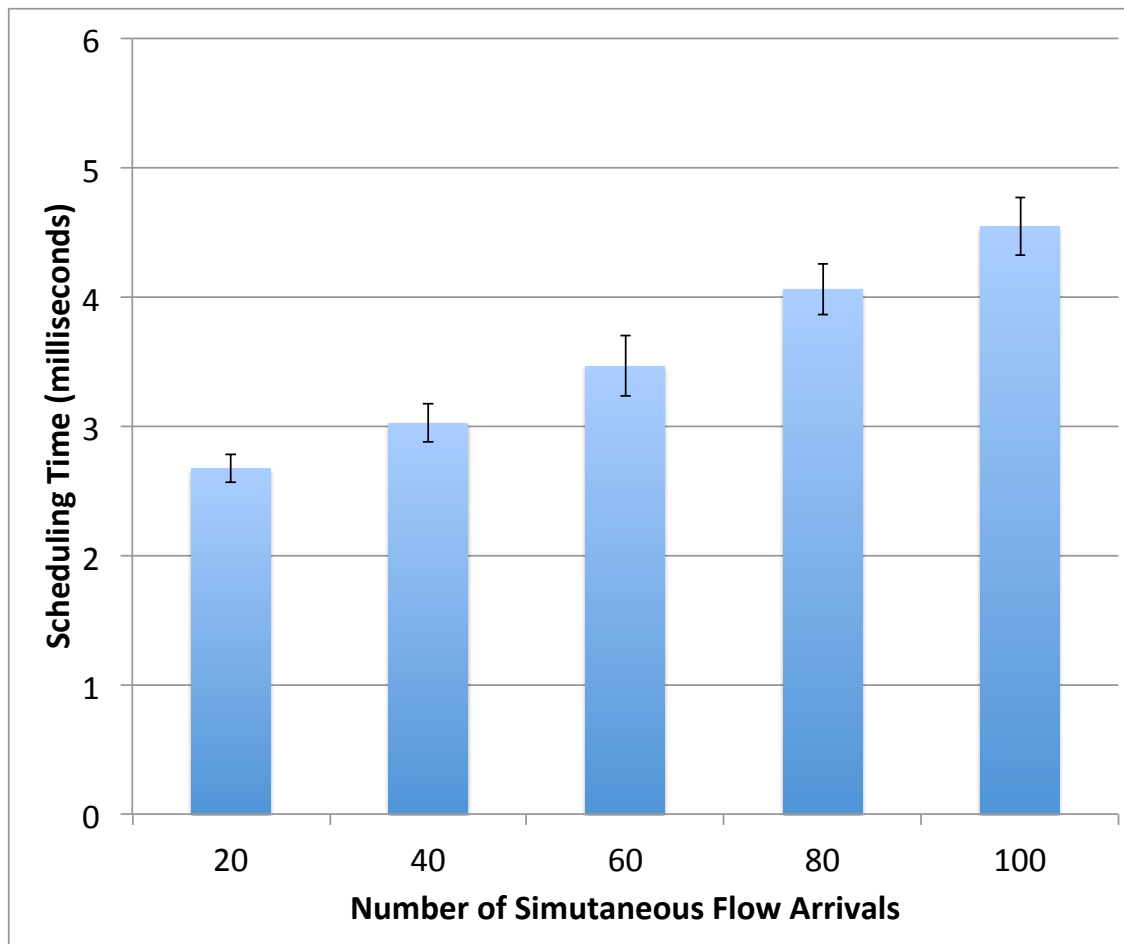


Effective Transmit Rate By Job Size



Flow Scheduling Overhead

Simulate a fat-tree topology with 128 hosts and launch multiple flows simultaneously



Conclusion

- We present Phurti, which is an application and network-aware flow scheduling framework for multi-tenant MapReduce cluster.
- We design Smallest Maximum Sequential-traffic First (SMSF) to optimizing shuffle completion time of MapReduce jobs.
- We evaluate Phurti via real testbed and workload and demonstrate Phurti can improve average job completion by nearly 20% with low overhead.

Questions