Ming Hsieh Department of Electrical Engineering

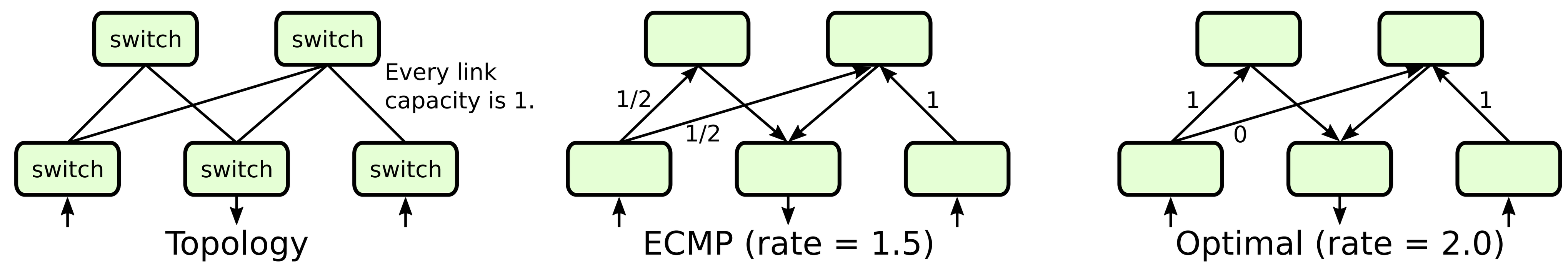
USC Viterbi School of Engineering

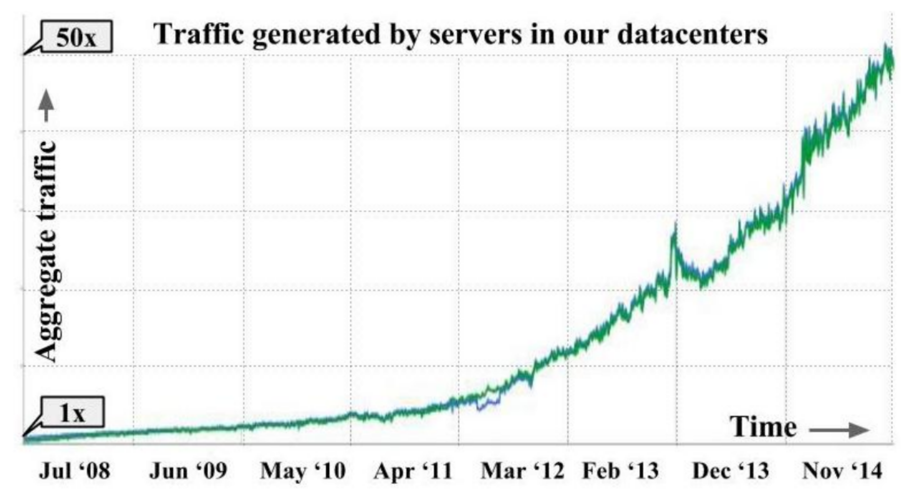
Low-Latency Throughput-Optimal Load Balancing for Datacenter Networks

Sucha Supittayapornpong and Michael J. Neely

Introduction

- Traffic load-balancing in datacenters improves network utilization.
- Conventional algorithms, including ECMP, are suboptimal.





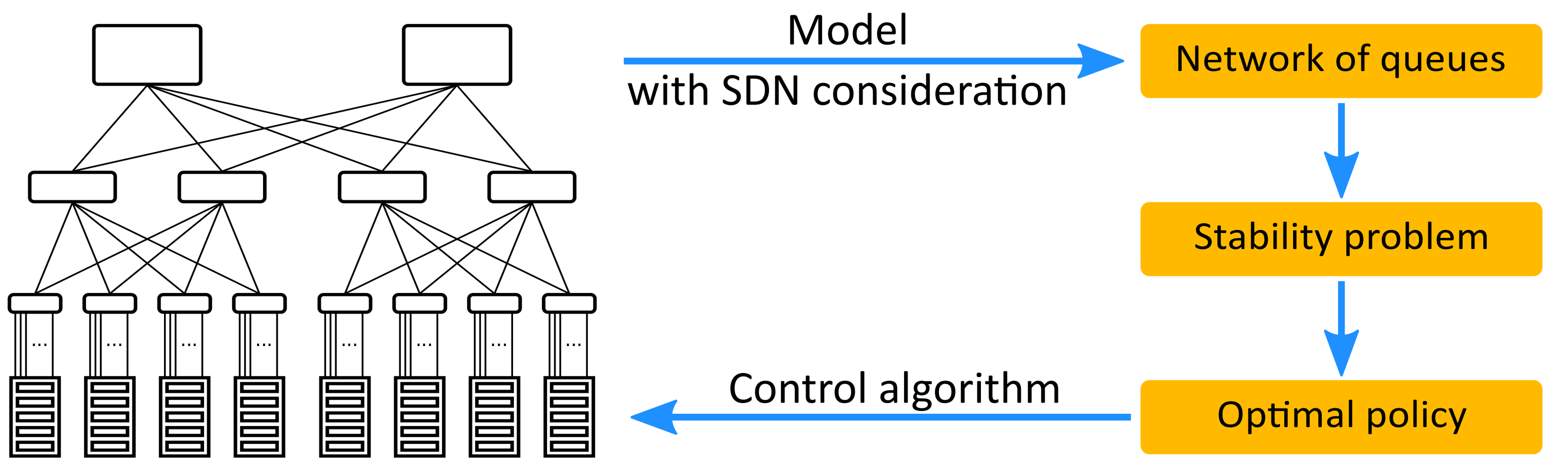
Google's datacenter traffic grows exponentially.

[SIGCOMM'15] Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network

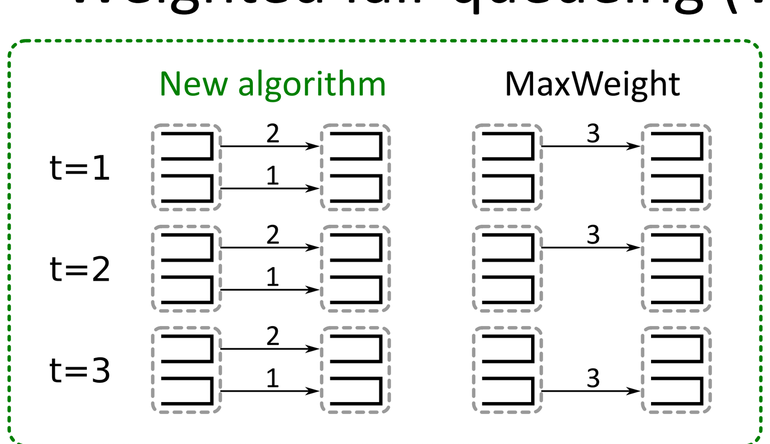
- This work develops a new throughput-optimal load-balancing algorithm that can be implemented in software-defined networking (SDN) switches.
- The algorithm works gracefully with TCP flows and outperforms the ECMP algorithm in simulation.

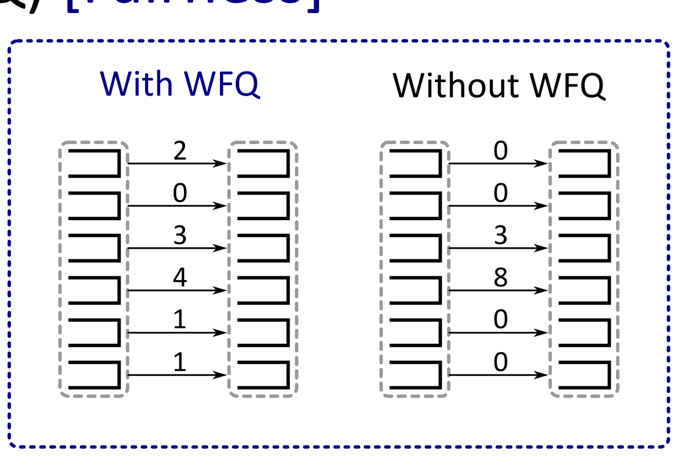
Algorithm Design

- The load-balancing algorithm is developed from an optimal policy that stabilizes a network of queues.



- The optimal policy has two special properties:
  - Link capacity sharing [Low Latency]
  - Weighted fair queueing (WFQ) [Fairness]

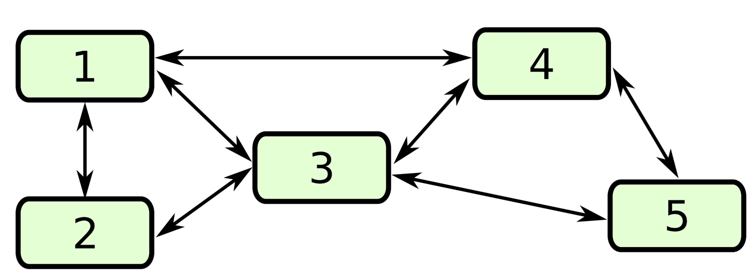




- **Theorem (Throughput optimality):**
For any arrival rates inside a network capacity region, the load-balancing algorithm stabilizes all queues in a network.

Load-Balancing Algorithm

- Exchange of queue information:



Example:
Local queue information $\{Q_1^d(t) : d \in \mathcal{D}\}$ at switch 1 is sent to switches 2, 3, 4.
Local queue information $\{Q_5^d(t) : d \in \mathcal{D}\}$ at switch 5 is sent to switches 3, 4.

$Q_{ij}^d(t)$ No. of group-d packets at switch i's interface j

$Q_i^d(t) = \sum_j Q_{ij}^d(t)$ No. of group-d packets at switch i

- Weights of WFQ:

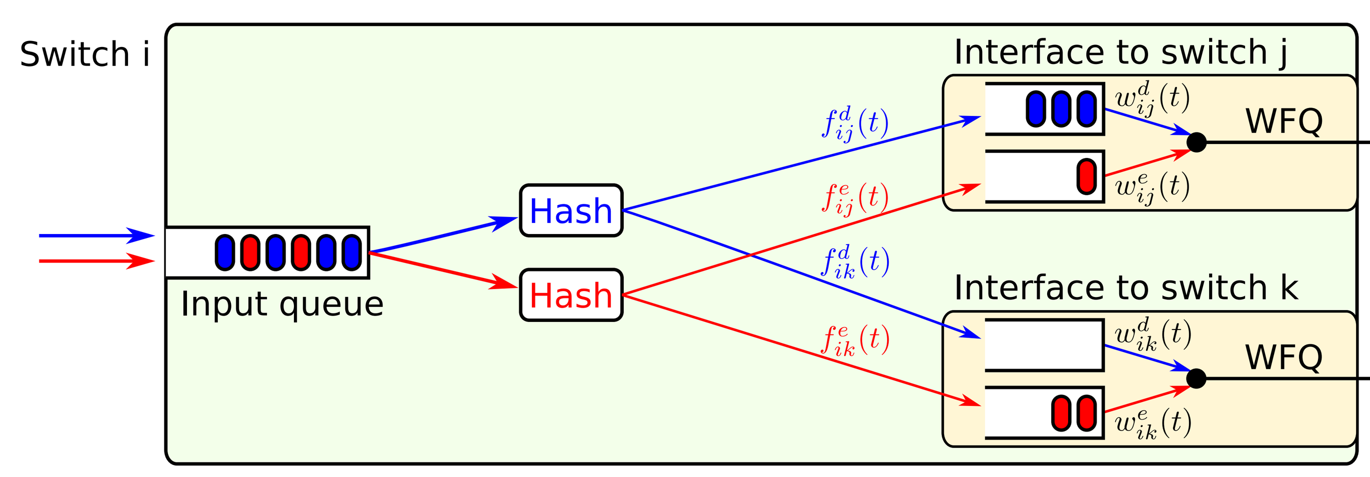
$$w_{ij}^d(t) = \max[1, Q_i^d(t) - Q_j^d(t) + r_{ij}^d(t-T)/\alpha]$$

$w_{ij}^d(t)$ Weight of WFQ for group-d queue at link (i,j)

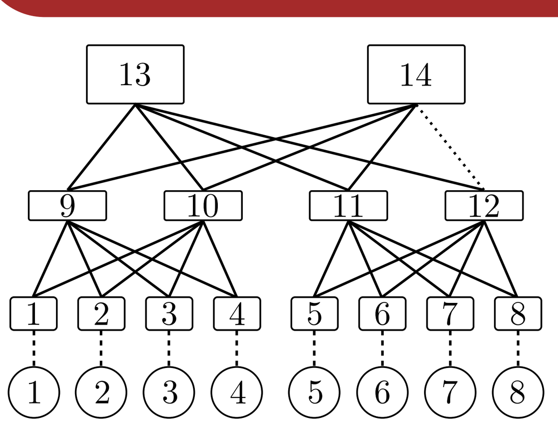
$r_{ij}^d(t-T)$ Previous measured group-d traffic over link (i,j)

- Hash ratio:

$f_{ij}^d(t)$ Hash ratio of group-d traffic to link (i,j)
Its value is a function of $\{Q_{ij}^d(t), r_{ij}^d(t-T) : d \in \mathcal{D}\}$.

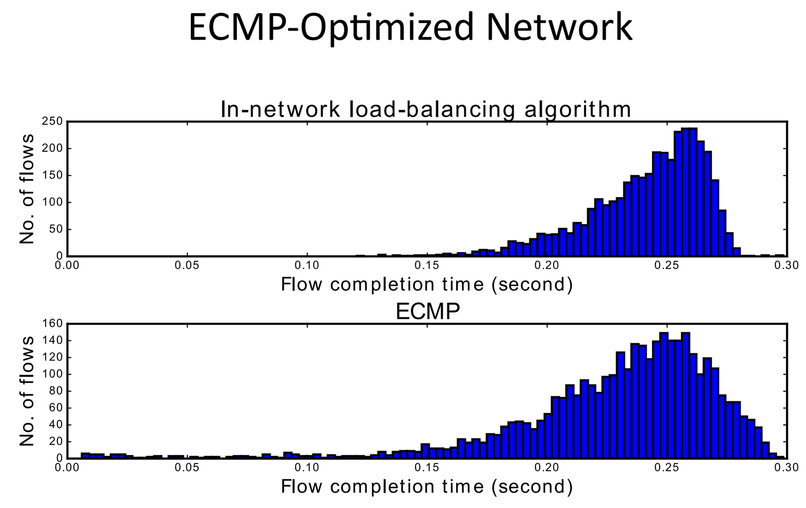


Simulation Results

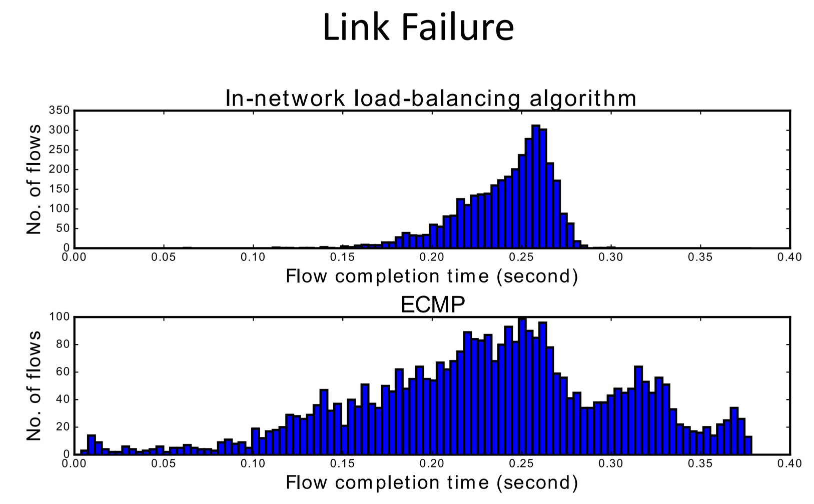


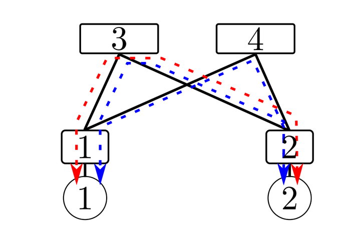
Maximum queue size is 200 pkts.
All to all TCP flows
New Reno TCP
A flow transfers 1MB of data.
The total number of flows is 3578.

ECMP-Optimized Network



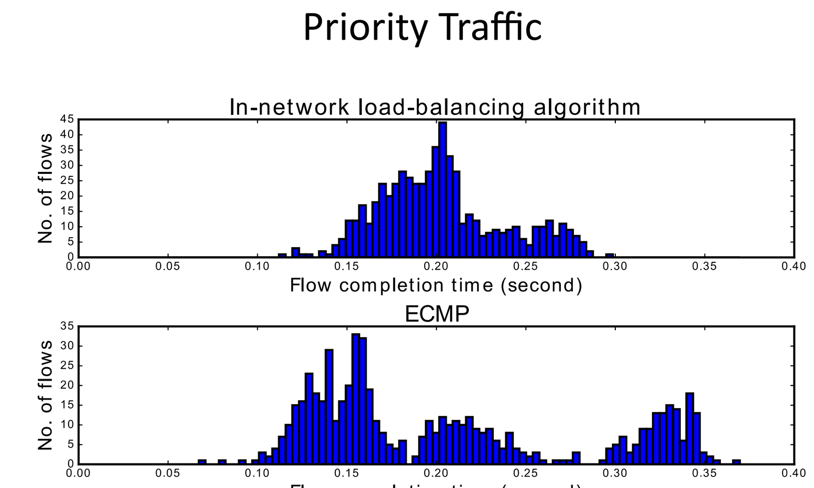
Link Failure





Priority flow
1 always-on flow
Normal flow
300 flows

Priority Traffic



Big Picture & Future Work

- This work illustrates the practicality of theoretical network optimization to solve a load-balancing problem.
- It is part of the exploration of mathematical network optimization and programmable networking:
  - Load-balancing algorithm [1]
  - Finite-buffer algorithm [2]
  - Convergence analysis [3][4]
  - Quality of information [5][6]

[1] S. Supittayapornpong and M. J. Neely, "Throughput-Optimal Load Balancing for Datacenter Networks." -- Under review for IEEE INFOCOM 2017
[2] S. Supittayapornpong and M. J. Neely, "Achieving Utility-Delay-Reliability Tradeoff in Stochastic Network Optimization with Finite Buffers," Proc. IEEE INFOCOM, Apr. 2015.
[3] S. Supittayapornpong and M. J. Neely, "Time-Average Stochastic Optimization with Non-convex Decision Set and its Convergence," Proc. IEEE WiOpt, May 2015. -- Best paper award
[4] S. Supittayapornpong, L. Huang, and M. J. Neely, "Time-Average Optimization with Nonconvex Decision Set and its Convergence," Proc. IEEE CDC, Dec. 2014.
[5] S. Supittayapornpong and M. J. Neely, "Quality of Information Maximization for Wireless Networks via a Fully Separable Quadratic Policy," IEEE/ACM TON, Apr. 2015.
[6] S. Supittayapornpong and M. J. Neely, "Quality of Information Maximization in Two-Hop Wireless Networks," Proc. IEEE ICC, Jun. 2012.

"Practical Network Optimization" | supittay@usc.edu | http://www-scf.usc.edu/~supittay/