Improving Valiant Routing for Slim Fly Networks

Deyu Han

Zhaofeng Wang

David P. Bunde

Partially supported by National Science Foundation grant CNS-1423414 and the Paul K. and Evalyn Elizabeth Richter Memorial Funds.

Designing new HPC topologies

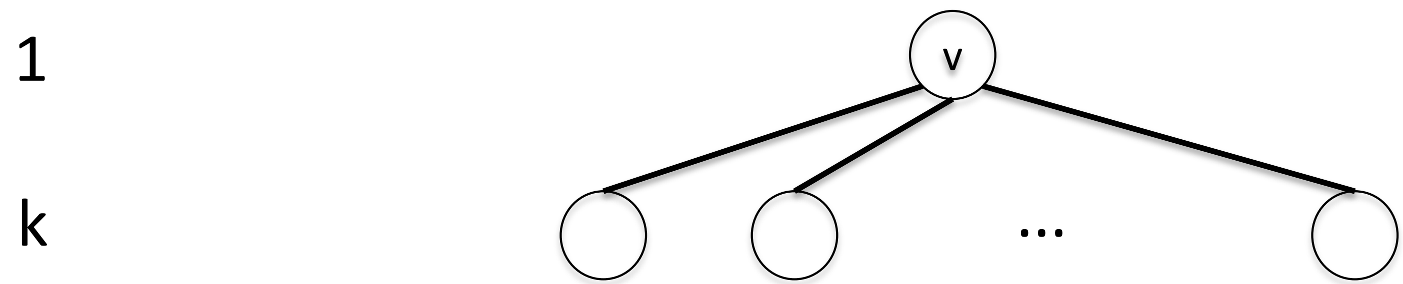
- Minimizing system diameter
  - low latency to support fine-grained parallelism
  - reduces power per message
  - less opportunity for inter-packet interference

Moore bound

- How many vertices of degree k can be within distance D?

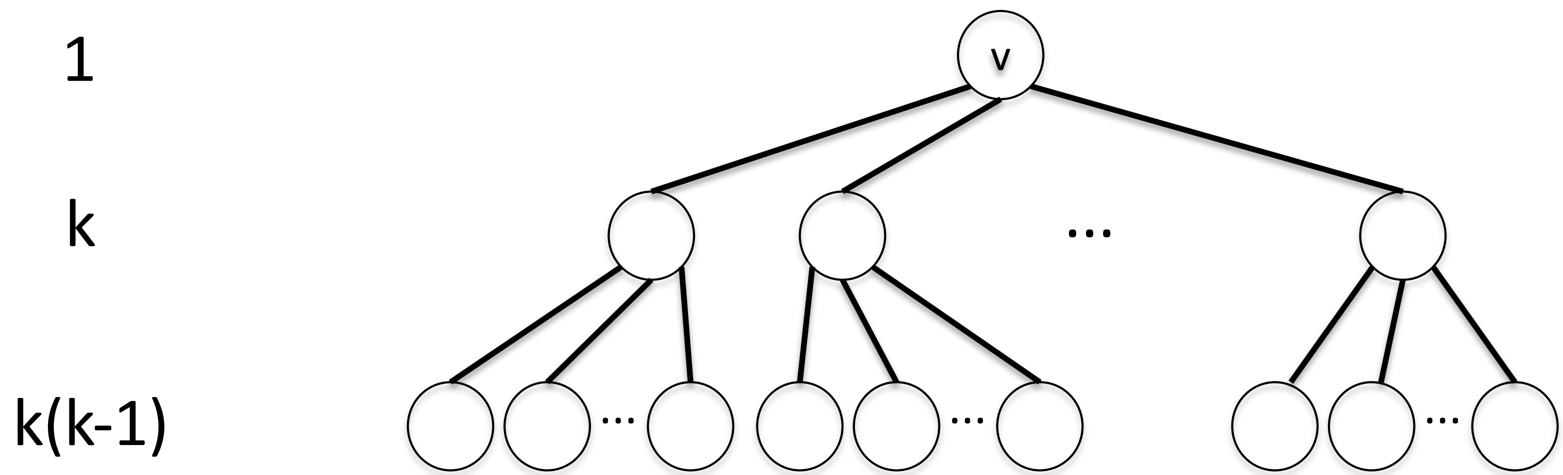
Moore bound

- How many vertices of degree k can be within distance D?



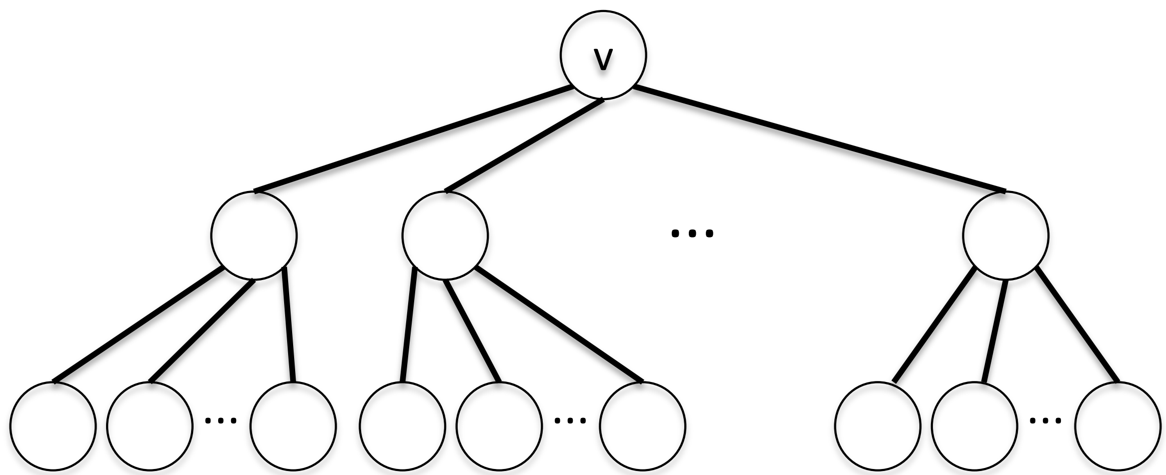
Moore bound

- How many vertices of degree k can be within distance D?



Moore bound

- How many vertices of degree k can be within distance D?



1

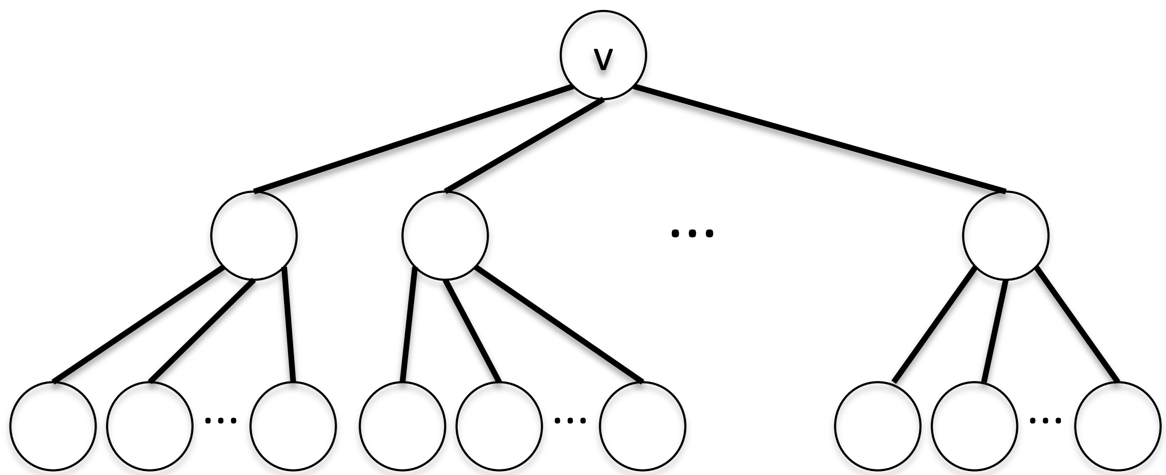
k

$k(k-1)$

$k(k-1)^2$

Moore bound

- How many vertices of degree k can be within distance D?



1

k

$k(k-1)$

$k(k-1)^2$

max vertices: $1 + k\sum_{i=0}^{D-1}(k-1)^i$

Slim Fly

- Algebraically-specified family of graphs
- Based on MMS graphs [McKay, Miller, Širán, 1998]
  - Diameter 2
  - close to Moore bound (within 12% for 8,192 vertices)
- [Besta and Hoefler, 2014] developed as network topology
  - high performance
  - cheaper to build
  - resilient to link failures

Slim Fly

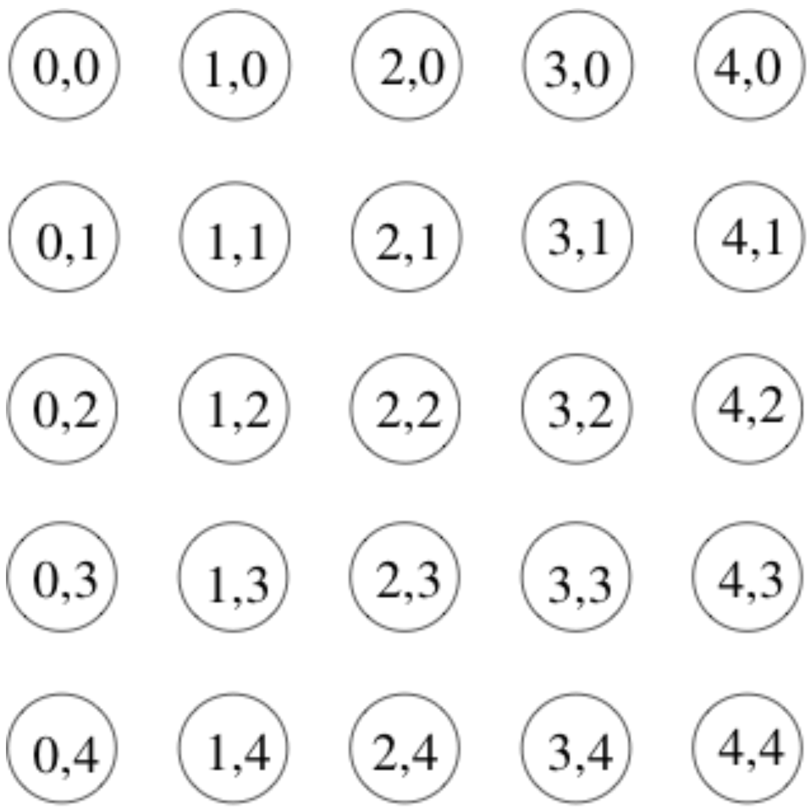
- Choose prime power q not congruent to 2 mod 4
- Find ξ that generates F_q
- Select sets X and X' based on q mod 4

  For q = 1 mod 4,

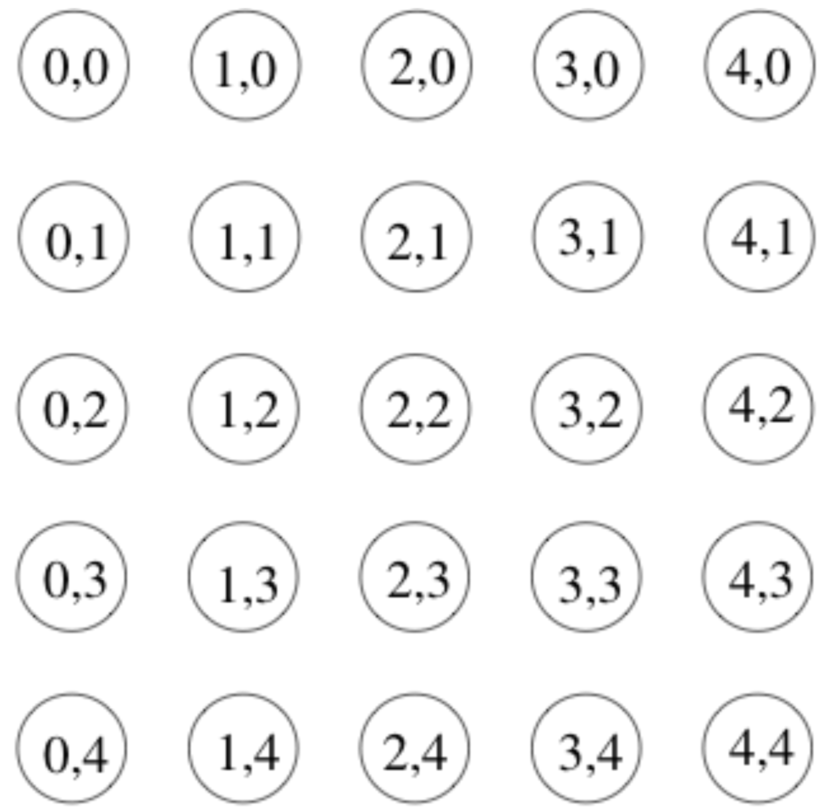
  $X = \{ 1, \xi^2, \xi^4, \ldots, \xi^{q-3} \}$

  $X' = \{ \xi, \xi^3, \xi^5, \ldots, \xi^{q-2} \}$

Slim Fly



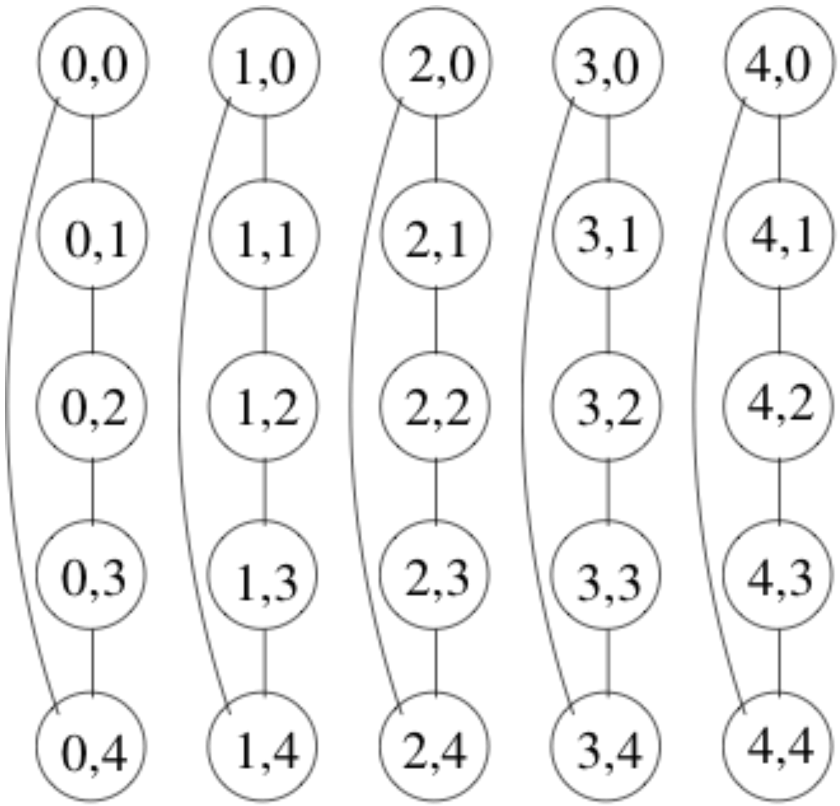
(0,x,y) switches



(1,m,c) switches

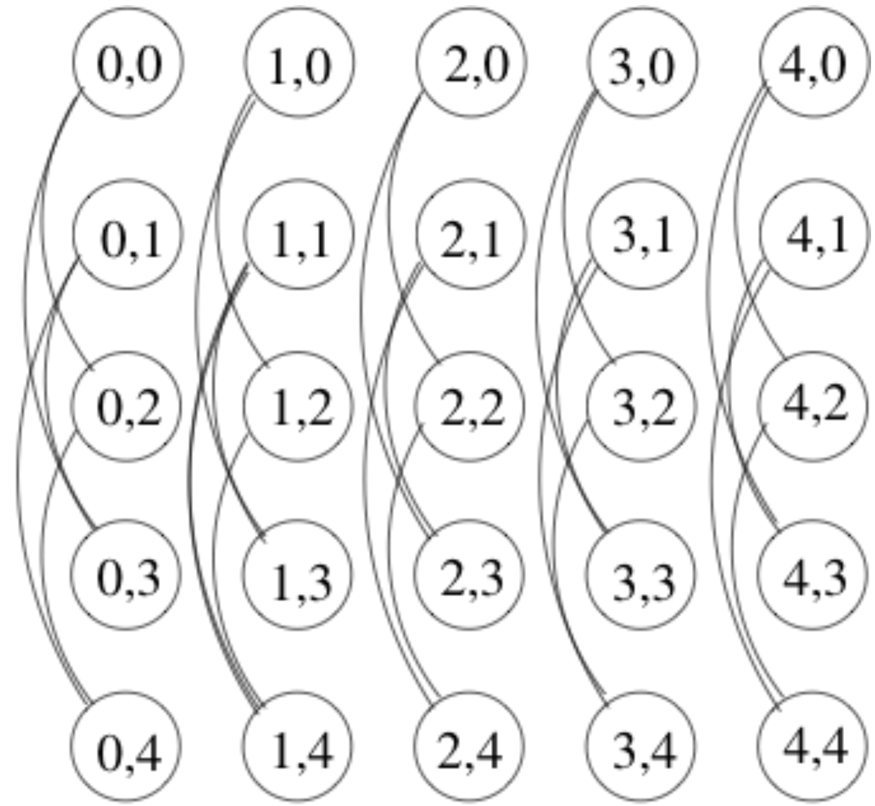
Each switch has attached compute nodes

Slim Fly



(0,x,y) switches

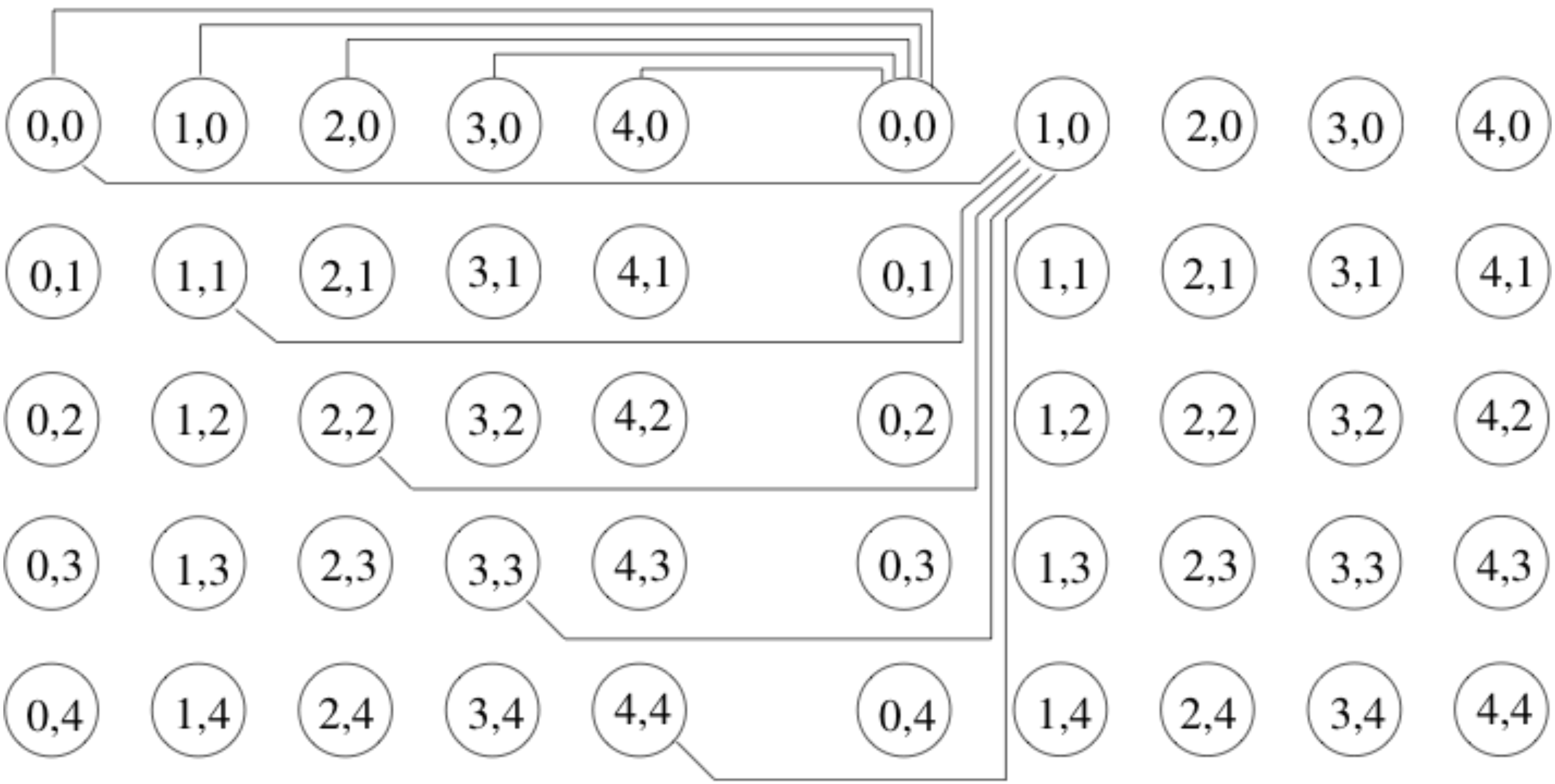
(0,x,y) and (0,x,y’) connect iff y-y’ is in X



(1,m,c) switches

(1,m,c) and (1,m,c’) connect iff c-c’ is in X’

Slim Fly



Edges between (0,x,y) and (1,m,c) switches

(0,x,y) and (1,m,c) connect iff y = mx+c

Valiant routing

- Shortest path/minimal routing can deterministically cause hot-spots for some communication patterns
- Instead, each packet randomly chooses an intermediate node and goes to it before heading to destination

Adaptive routing

- Valiant routing avoids worst-case behavior

  ...but not good when traffic already distributed

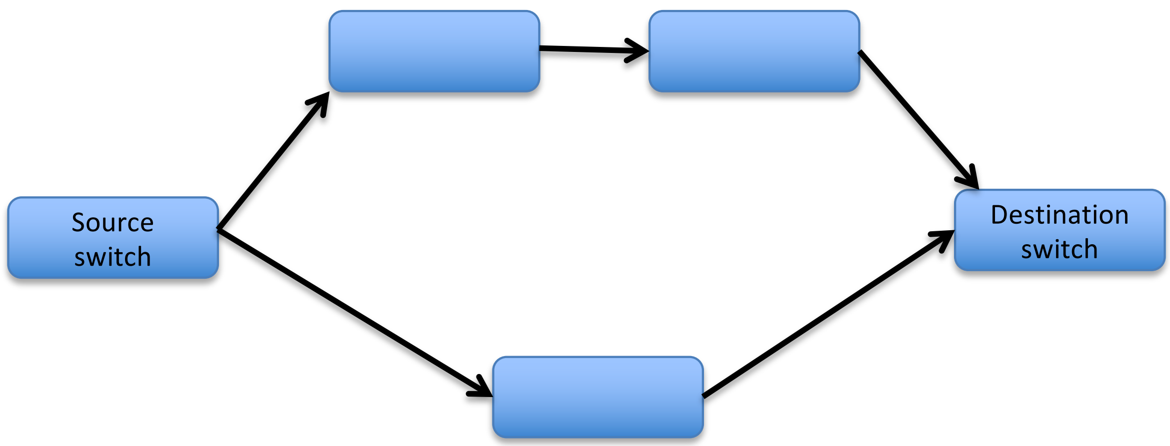
Adaptive routing

- Valiant routing avoids worst-case behavior

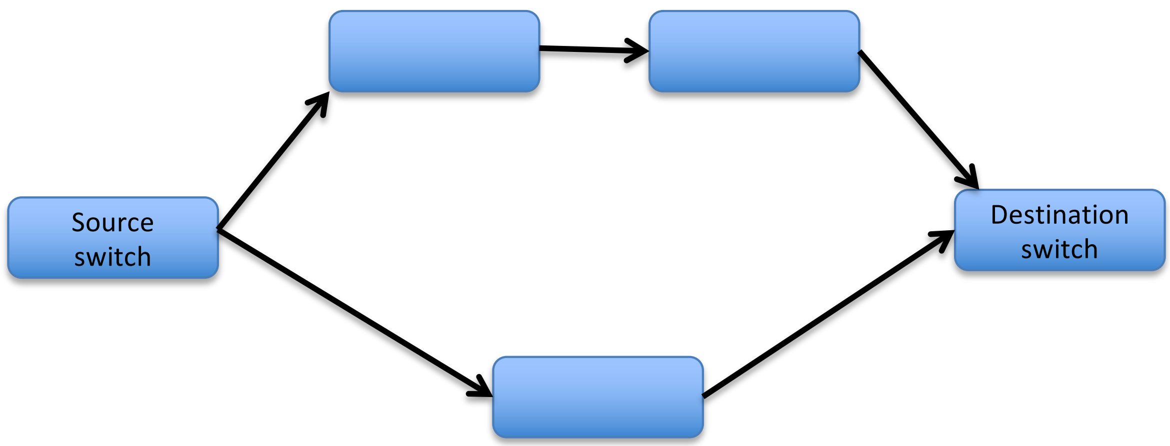
  ...but not good when traffic already distributed

- Idea: Use minimal routing unless hot-spots develop, in which case switch to Valiant

UGAL



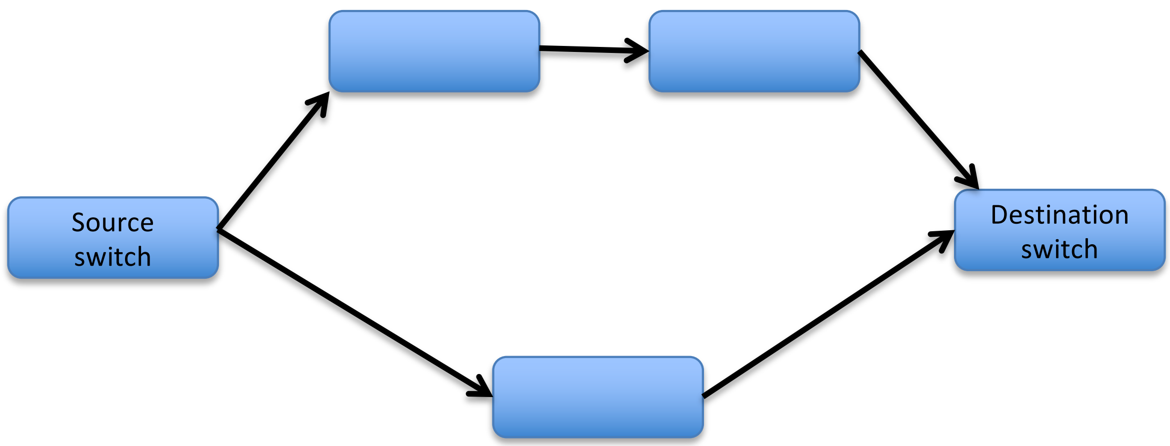
UGAL



Estimated delivery time for each path:

UGAL-G: sum of length of message queues along path

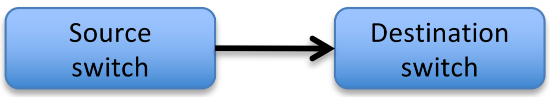
UGAL



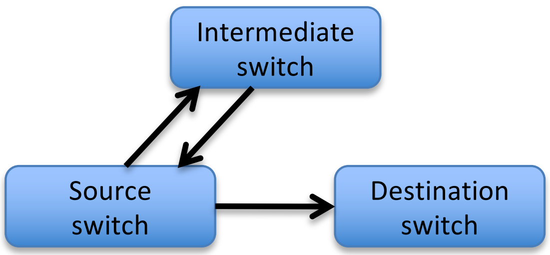
Estimated delivery time for each path:

- UGAL-G: sum of length of message queues along path
- UGAL-L: length of first queue × path length

Problem with Valiant on Slim Fly

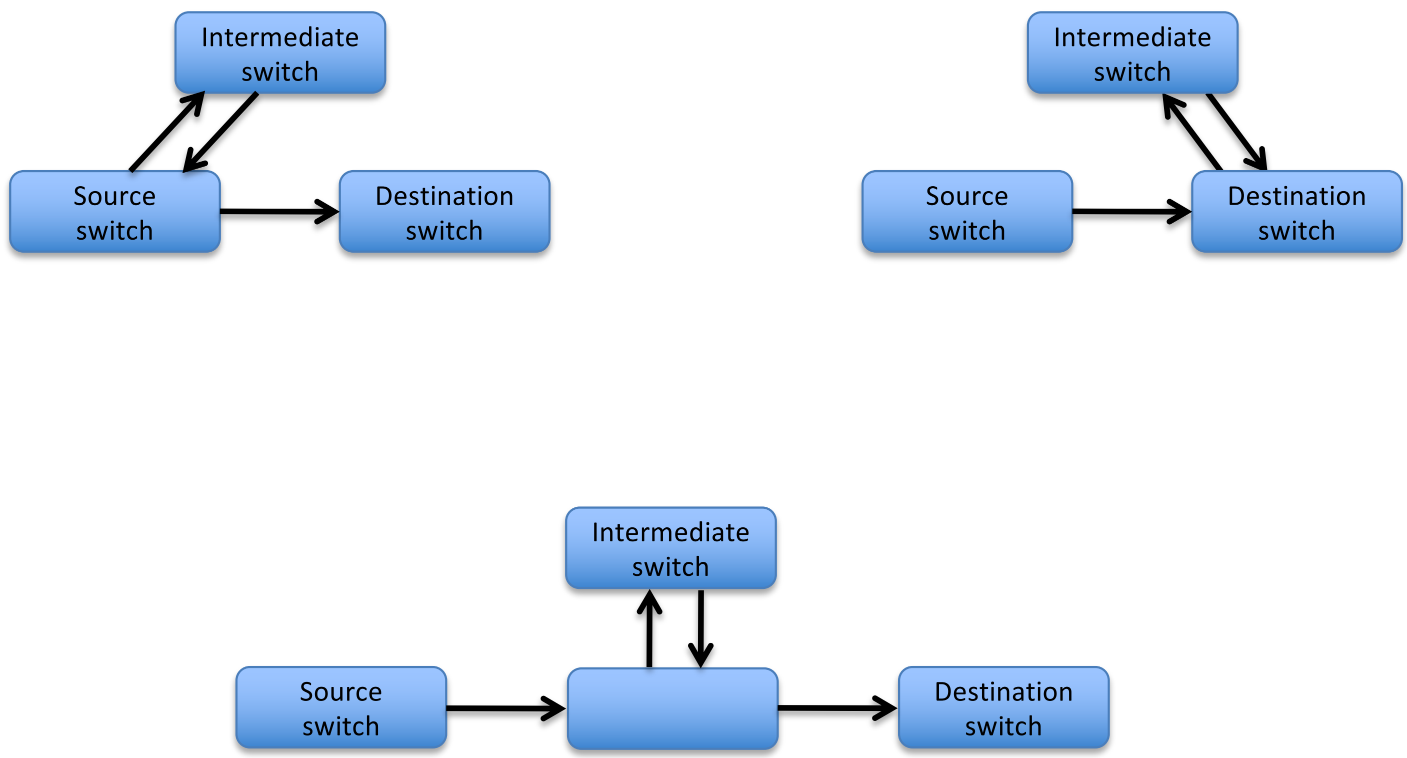


Problem with Valiant on Slim Fly



Problem with Valiant on Slim Fly

Problem with Valiant on Slim Fly



Our idea

- Valiant routing: Choose intermediate switch randomly *among those that don't cause a loop*

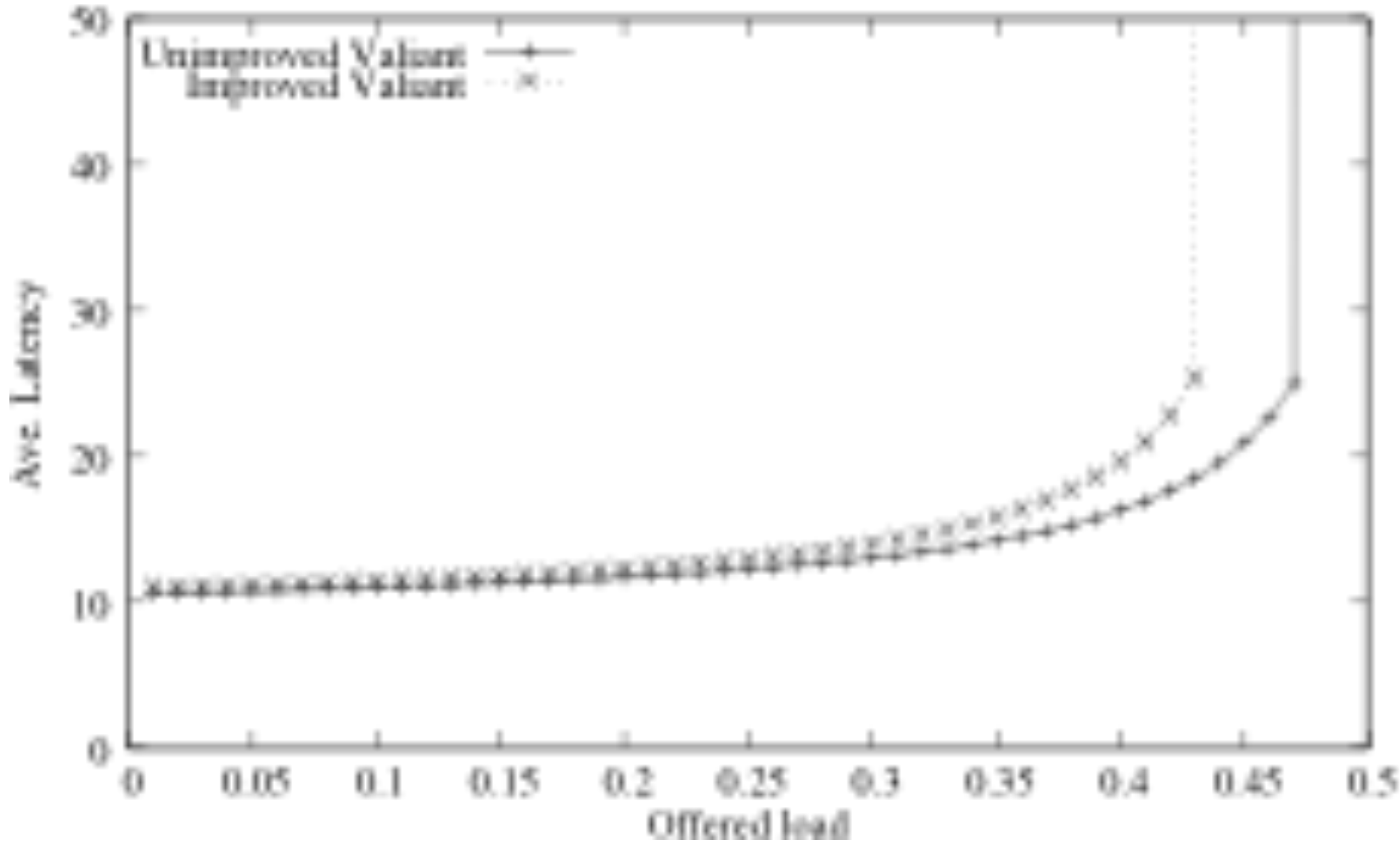
- UGAL-L: Use this improved version of Valiant routing when selecting an indirect path

Experimental setup

- Packet-level simulation
- q = 5..13, nodes as needed to balance network
- "Worst case" communication pattern with many hot spots
  - Divide system into chains of switches
  - Each node sends to randomly-chosen node on next switch

Performance for Valiant routing

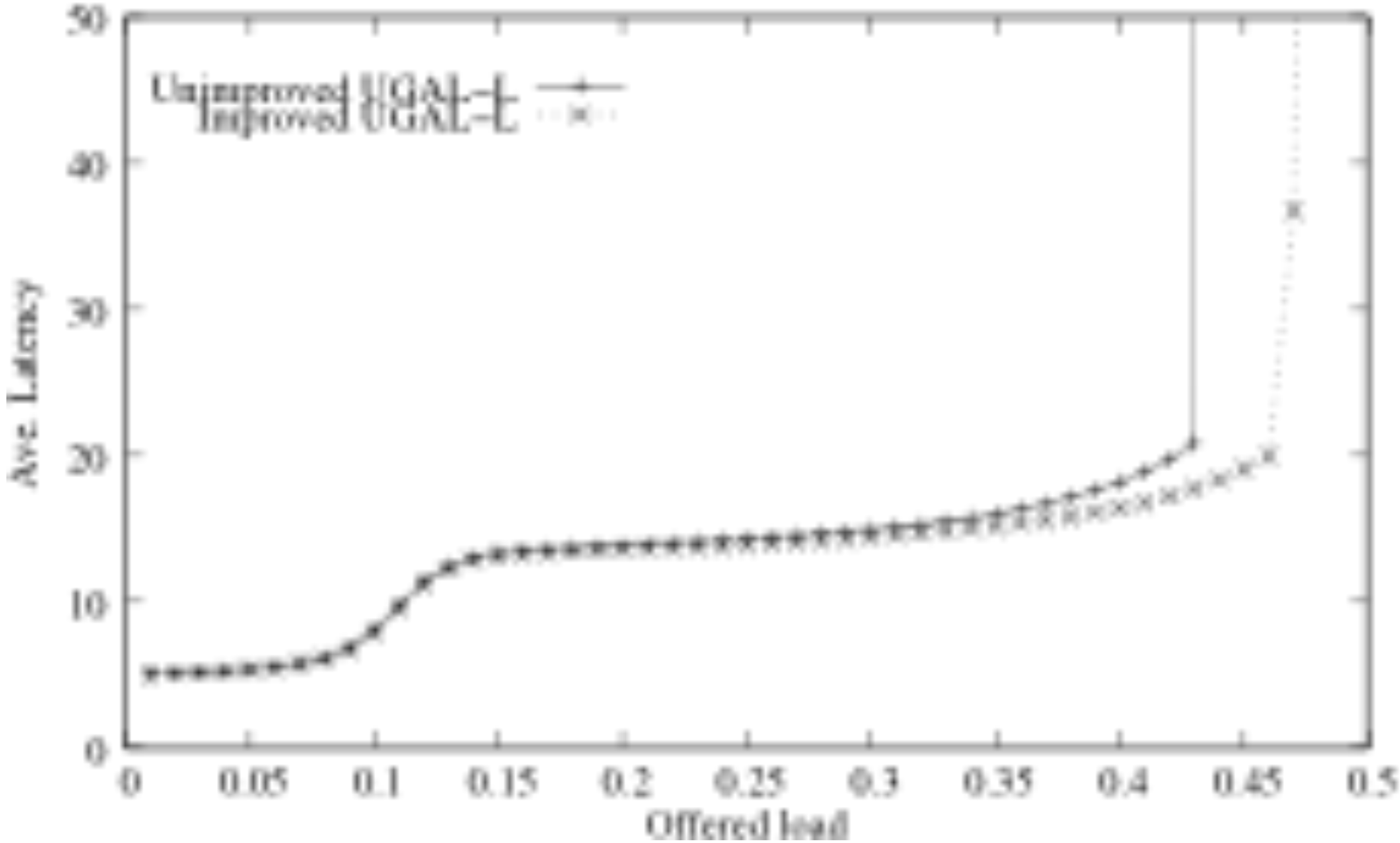
(q=13)



Likely explanation: The improved algorithm has longer path lengths (6.0 vs 5.8)

Performance for Adaptive routing (UGAL-L)

(q=13)



Relationship to system size

Relationship to system size

Larger system means

- Larger value of k so fewer loops

  On a diameter-2 Moore graph with degree k and uniform traffic, only 1/(k+1) of the packets loop

Relationship to system size

Larger system means

- Larger value of k so fewer loops
- More nodes per switch so hotter hot-spots

Relationship to system size

Larger system means

- Larger value of k so fewer loops
- More nodes per switch so hotter hot-spots

q	# nodes	baseline	improved	% diff
5	150	0.62	0.66	6.4
7	490	0.48	0.53	10.4
11	1,936	0.42	0.47	11.9
13	3,042	0.43	0.47	9.3

Future Work

- Scaling of improvement to larger systems
- Effect on other communication patterns
- Effect on other adaptive routing algorithms
- Applications to other topologies

Thanks!

dbunde@knox.edu