Shadow MACs: Scalable Label-switching for Commodity Ethernet

Kanak Agarwal, Colin Dixon*, Eric Rozner, John Carter
IBM Research, Austin, TX

* now at Brocade

SDN: The Future!

- Rose-colored glasses:
  Fine-grained, dynamic control of the network
- Supported by:
  - Flow mod's based on diverse set of pkt hdr fields
  - Network measurements obtained in milliseconds[1]
  - Flow mods installed hundreds of times a second[2]

[1] Rasley, et al. Planck: Millisecond-scale Monitoring and Control for Commodity Networks. SIGCOMM'14.

[2] Rostos et al. OFLOPS: An Open Framework for OpenFlow Switch Evaluation. PAM'12.

SDN: The Future!

- Rose-colored glasses:

  Fine-grained, dynamic control of the network

- Supported by:
  - Flov fields
  - Netw nds[1]
  - Flow ed hundreds of times a second[2]

Most SDN deployments limited to overlays or small production environments

[1] Rasley, et al. Planck: Millisecond-scale Monitoring and Control for Commodity Networks. SIGCOMM'14.

[2] Rostos et al. OFLOPS: An Open Framework for OpenFlow Switch Evaluation. PAM'12.

SDN: The Future?

- Significant issues can arise at scale!

- Flow mod’s based on diverse set of pkt hdr fields

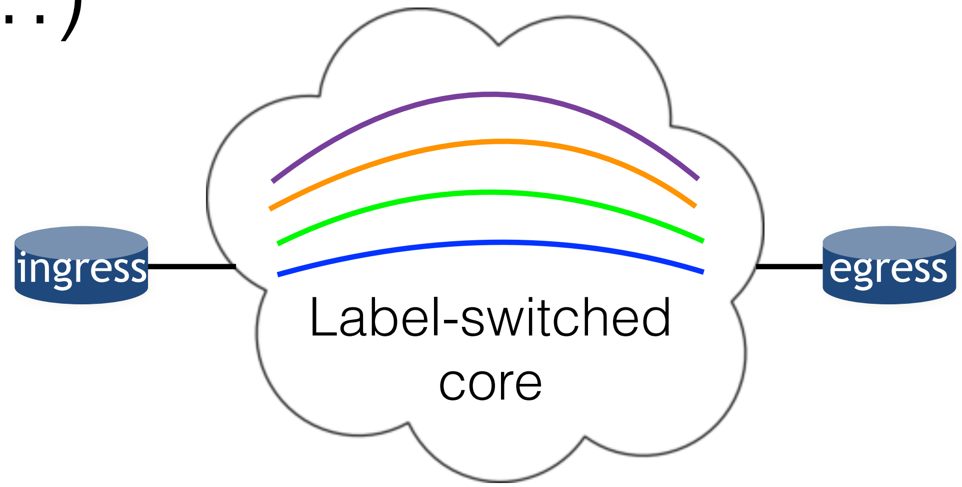
TCAMs expensive, only few 1,000 rules supported

- Network measurements obtained in milliseconds
- Flow mods installed hundreds of times a second

Consistent network updates are hard!

Label Switching to the Rescue!

- Label switching common forwarding mechanism (Frame Relay, ATM, MPLS, …)



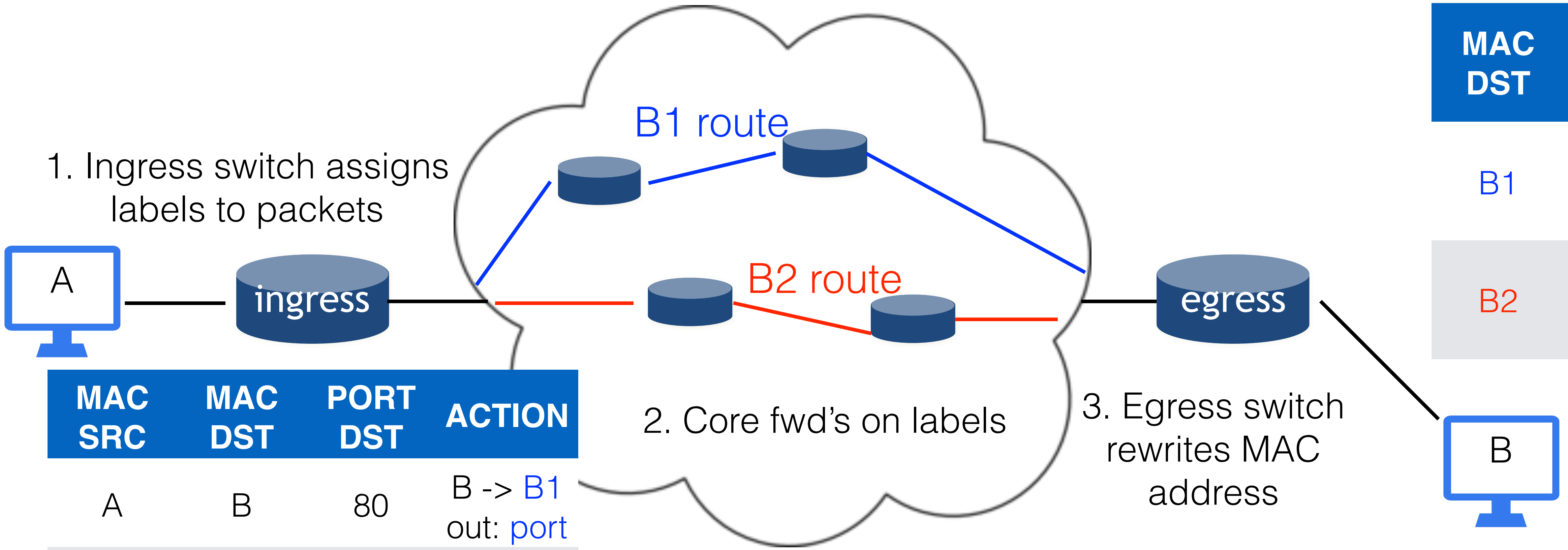
- We'll borrow:

- Label-switched core: fixed-width, exact-match lookups map easily into large forwarding tables

- Opaque labels: not assoc to physical endpoint in n/w

Our solution: Shadow MACs

- Opaque forwarding label: Destination MAC address
  - Fast, cheap and large fwd'ing tables already in switch!
  - OpenFlow flow mods on ingress/egress guide onto paths



1. Ingress switch assigns labels to packets

2. Core fwd's on labels

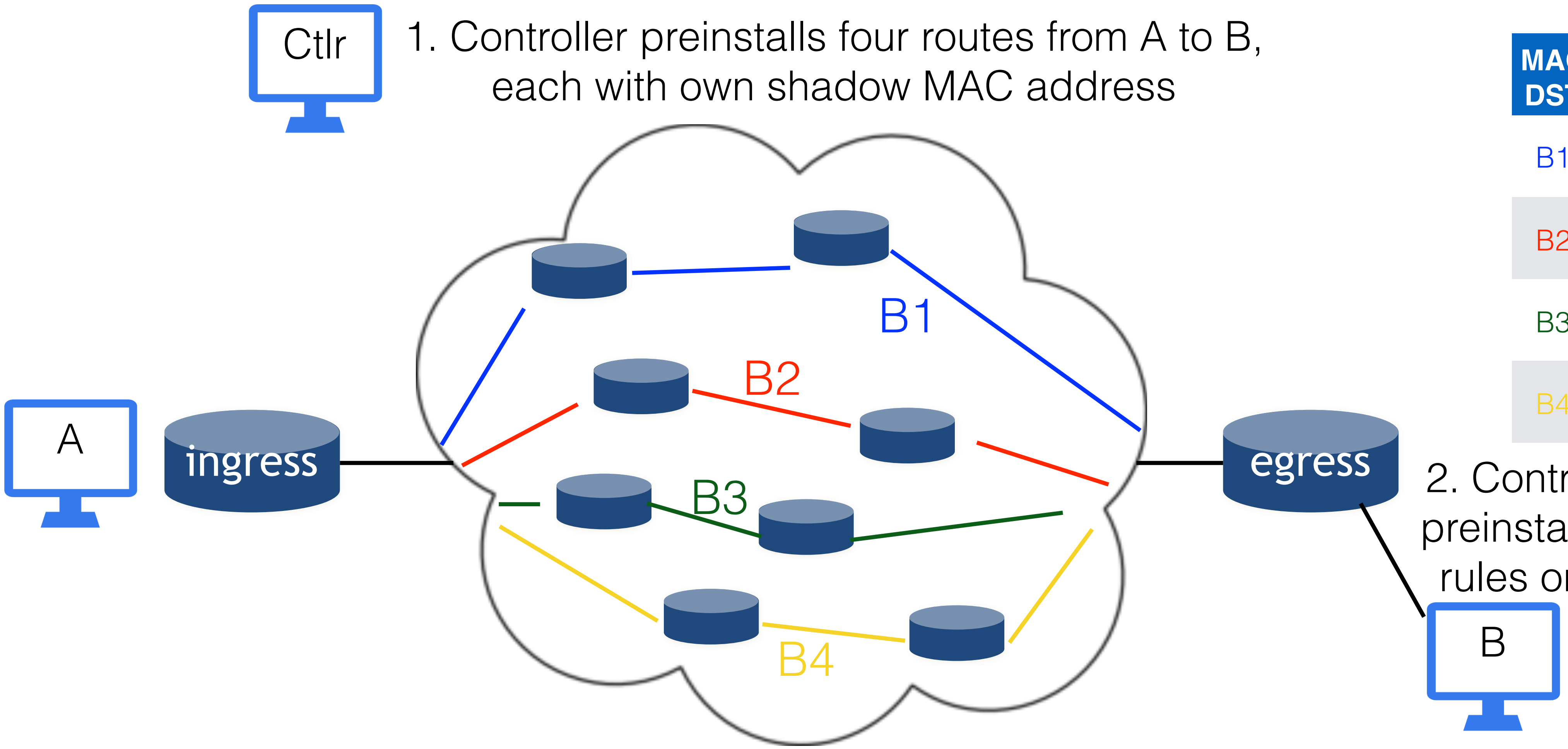
3. Egress switch rewrites MAC address

MAC SRC	MAC DST	PORT DST	ACTION
A	B	80	B -> B1 out: port
A	B	*	B -> B2 out: port

MAC DST	ACTION
B1	B1 -> B out: port
B2	B2 -> B out: port

Shadow MACs: Rerouting

- Opaque labels: no physical host → preinstall routes
- Ingress guiding: Changing routes now an atomic action!



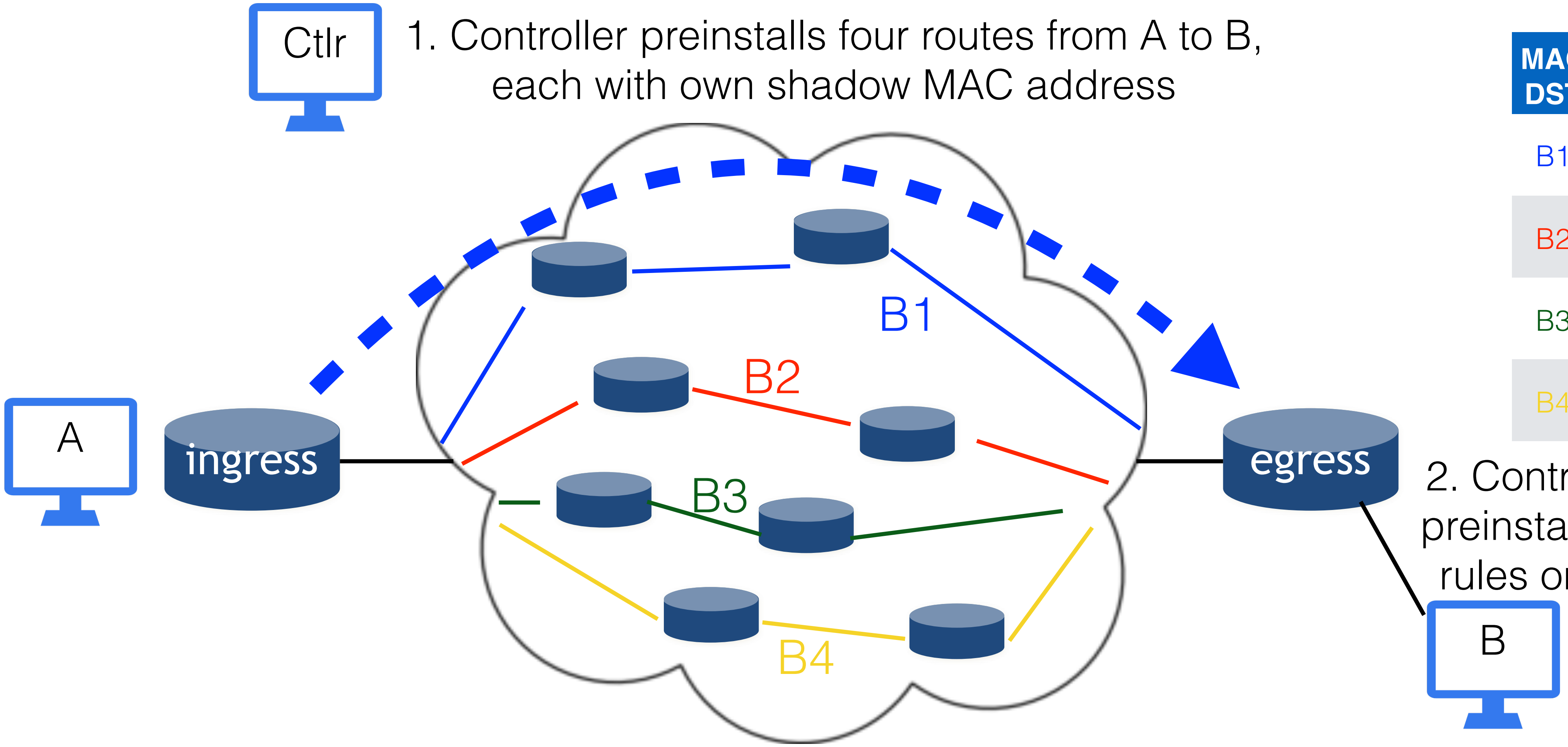
1. Controller preinstalls four routes from A to B, each with own shadow MAC address

2. Controller also preinstalls rewrite rules on egress

MAC DST	ACTION
B1	B1 -> B out: port
B2	B2 -> B out: port
B3	B3 -> B out: port
B4	B4 -> B out: port

Shadow MACs: Rerouting

- Opaque labels: no physical host → preinstall routes
- Ingress guiding: Changing routes now an atomic action!



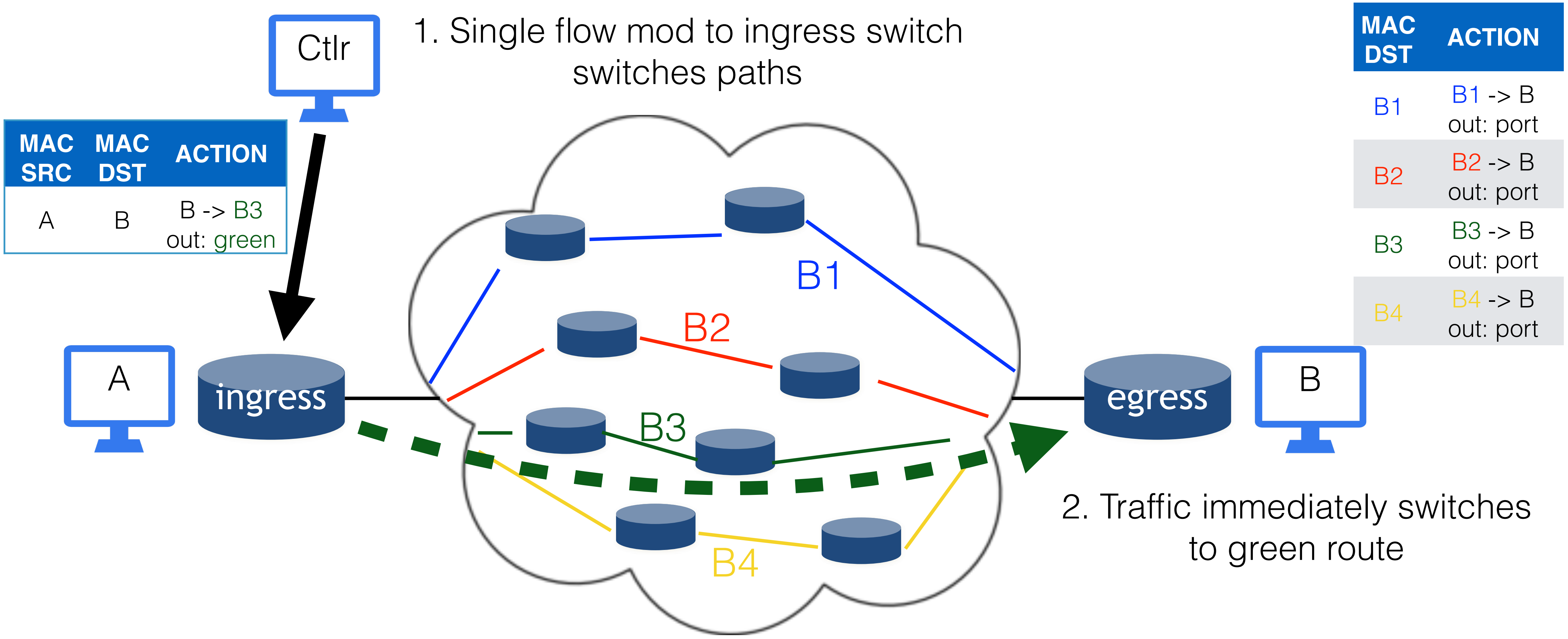
1. Controller preinstalls four routes from A to B, each with own shadow MAC address

2. Controller also preinstalls rewrite rules on egress

MAC DST	ACTION
B1	B1 -> B out: port
B2	B2 -> B out: port
B3	B3 -> B out: port
B4	B4 -> B out: port

Shadow MACs: Rerouting

- Opaque labels: no physical host → preinstall routes
- Ingress guiding: Changing routes now an atomic action!



1. Single flow mod to ingress switch switches paths

2. Traffic immediately switches to green route

MAC SRC	MAC DST	ACTION
A	B	B -> B3 out: green

MAC DST	ACTION
B1	B1 -> B out: port
B2	B2 -> B out: port
B3	B3 -> B out: port
B4	B4 -> B out: port

Benefits

- Controller guides pkts onto intelligently selected paths
  - Load balancing, link fail-over, route via middleboxes, differentiated services, …
- Decouples network edge from core
  - Consistent n/w updates, fast rerouting, multi-pathing, …
- Maps fine-grained matching to fixed destination-based rules
  - Pushes TCAM rules to FDB, limits TCAM usage in core
- Implementable today!

TCAM Usage

- TCAM usage:
  - Core switches use little/no TCAM rules
  - TCAM rules limited to edges, best case (OVS) uses no TCAM
- L2 forwarding tables are typically largest tables in switches
  - Scales better (up to 124x more L2 entries than TCAM)

	Broadcom Trident	IBM Rackswitch	HP ProVision	Intel FM6000	Mellanox SwitchX
TCAM	~4K	1K	1500	24K	0?
L2/Eth	~100K	~124K	~64K	64K	48K
X more L2	~25x	~124x	~42x	~2.6x	∞

10Gbps Ethernet Switch Table Sizes (# entries) [1]

[1] B. Stephens, et al. PAST: Scalable ethernet for data centers. *CoNEXT*, 2012.

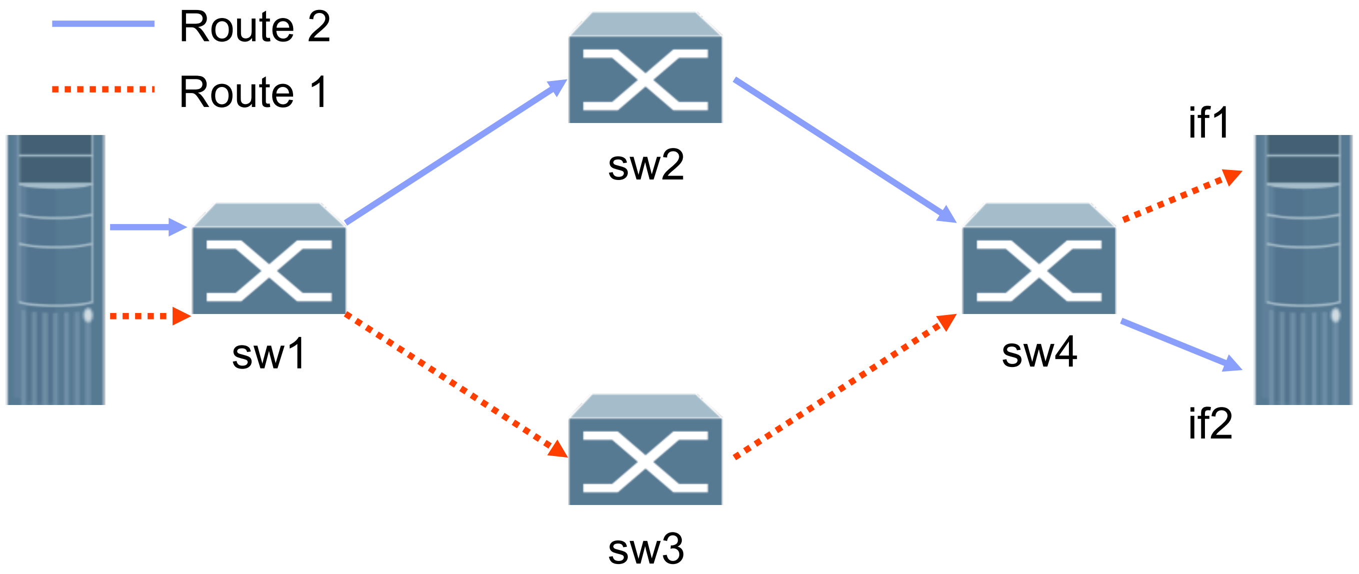
Fast, Consistent Updates

- Consistent Route updates:
  - SDN controller can pre-install routes
  - Atomic reroute: single flow-mod at ingress switch
- Two ways to achieve:
  - MAC address rewriting (OpenFlow)
  - ARP spoof (SDN controller sends GARP response)

E2E Multi-pathing

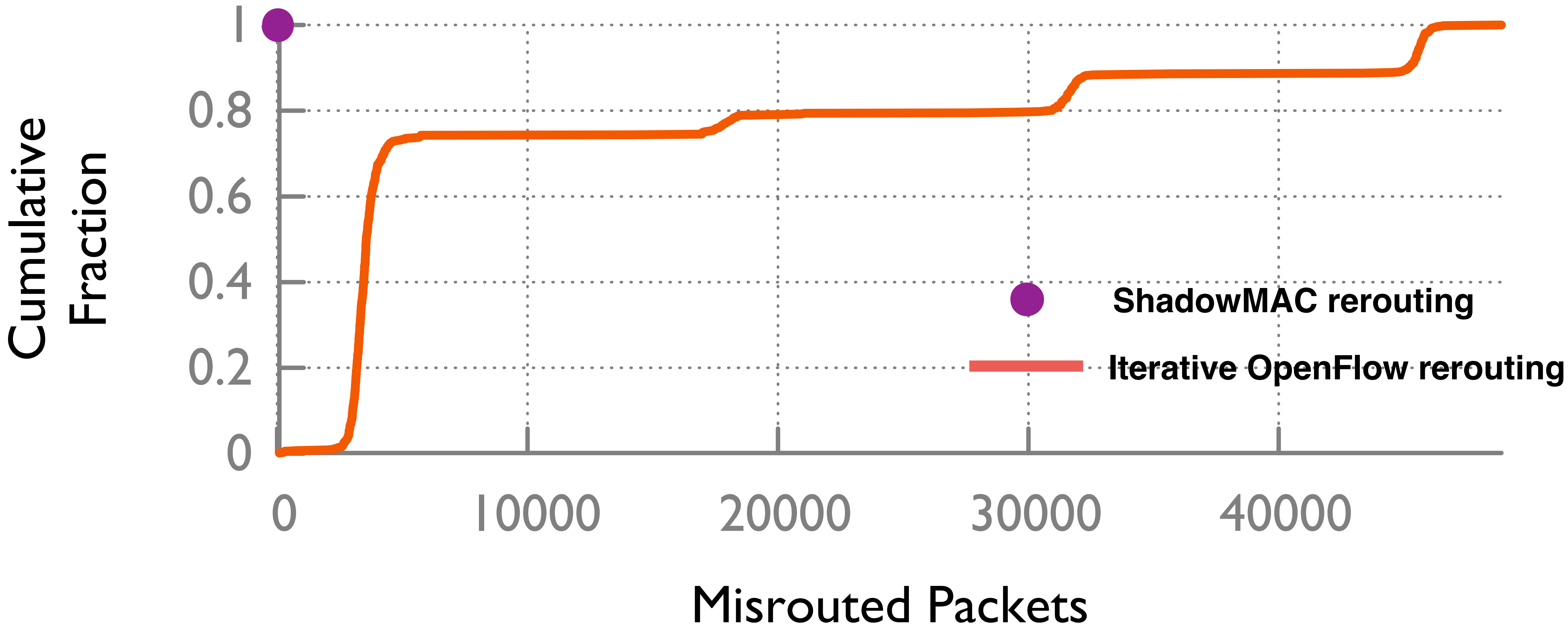
- SDN controller can allocate multiple distinct paths (shadow MACs) per destination
- OVS can allocate flows in round-robin fashion
- Benefits over ECMP
    - True L2 solution (ECMP is L3)
    - More control: per-path, instead of per-hop

Testbed Methodology



- UDP pkts start on Route 1, switch to Route 2
- Goal: measure # times per-pkt consistency violated, compare:
    - Shadow MAC rerouting
    - Traditional, iterative OpenFlow (order: sw4, sw2, sw1)
        - Uses Static Flow Pusher (barrier msg's not implemented)

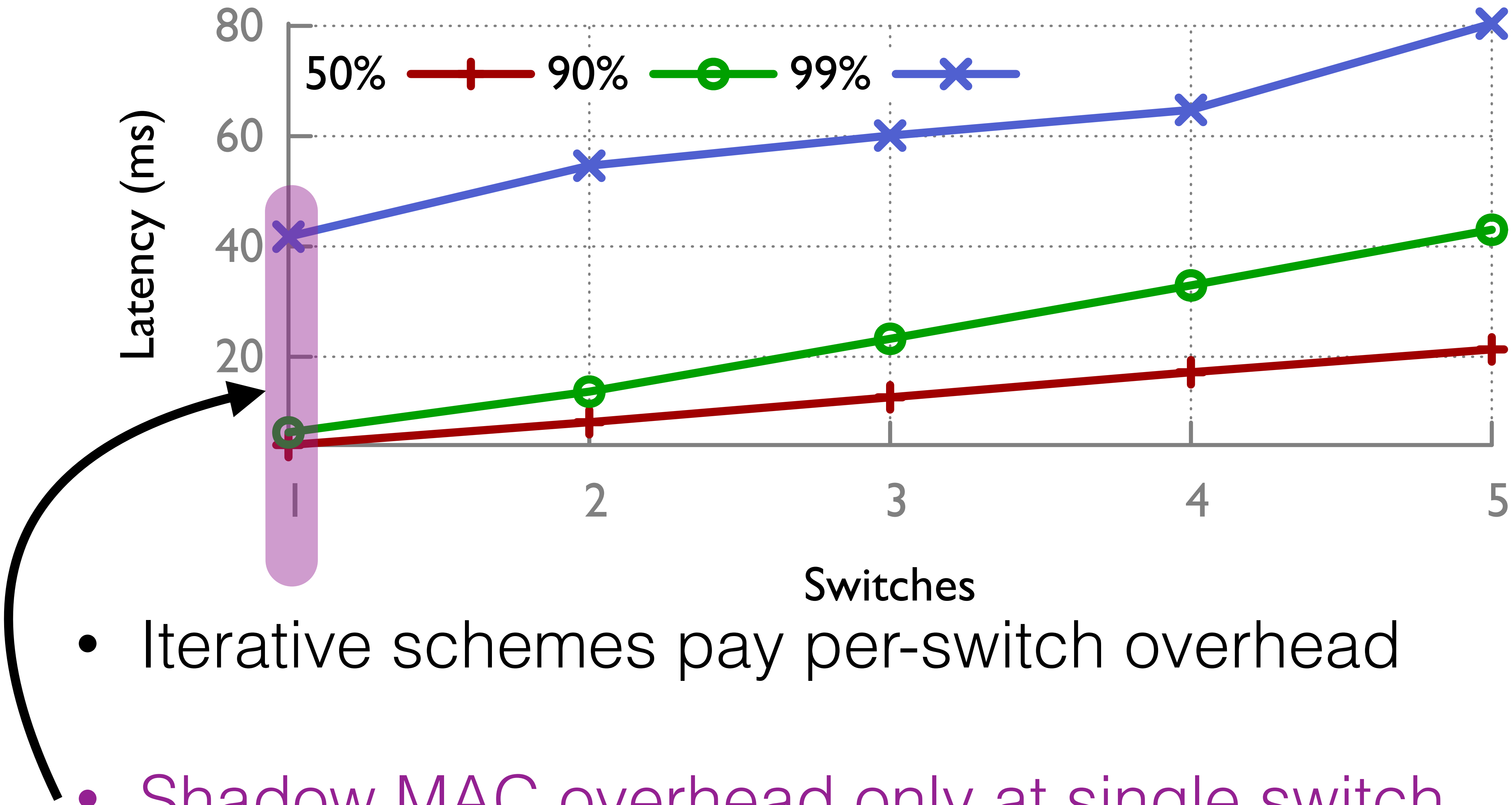
Per-Pkt Consistency



- CDF over 700 runs: at least 1 pkt misrouted every time
- Loss in ~5% of cases
- ShadowMACs: no inconsistency & no loss!

Per-pkt consistency violated

Iterative Flowmod Overhead

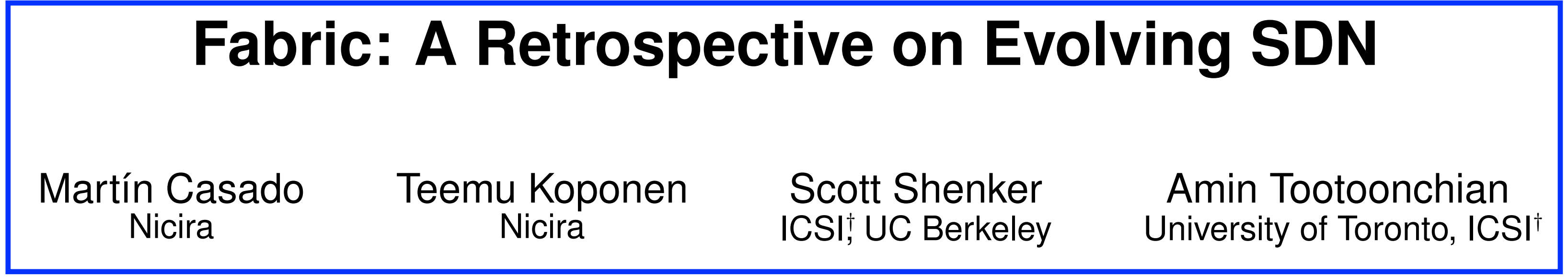


- Iterative schemes pay per-switch overhead

- Shadow MAC overhead only at single switch

  - 20-40 ms faster than traditional schemes

Related Work

- Have we seen this before?
  - Label-switching common
- 

  Fabric: A Retrospective on Evolving SDN

  Martín Casado, Nicira
  Teemu Koponen, Nicira
  Scott Shenker, ICSI[†], UC Berkeley
  Amin Tootoonchian, University of Toronto, ICSI[†]

  HotSDN ‘12
  - Motivated by separate, clean host-network, operator-network and packet-switch interfaces
  - MPLS: Little support in switches
- Consistent route updates [Reitblatt12, Jin14, …]

Summary

- SDN networks have issues at scale
  - Dynamic, fine-grained control of the network is challenging
- Label-switching using Shadow MACs is promising
  - Flexible edge steers traffic via OVS
  - Opaque labels (destination MAC) allow pre-installation of routes
  - Very practical: DMAC tables are widespread, large and fast
- Shadow MACs is a flexible architecture
  - Enable fast, atomic route updates, straight-forward mechanisms to implement multi-path, differentiated services, load-balancing, etc

Questions?

- Eric Rozner
  erozner@us.ibm.com

We are hiring at
IBM Research in Austin!

- All areas
- All experience-levels

- Co-authors:
  Kanak Agarwal, Colin Dixon, John Carter