In-Network Address Caching for Virtual Networks

Lior Zeno, Ang Chen, Mark Silberstein

Virtual networks *enable* the public cloud, but they are also *difficult* to manage.

Virtual-to-physical IP translation is *challenging*, *but necessary* for packet forwarding.

Sailfish: Accelerating Cloud-Scale Multi-Tenant Multi-Service Gateways with Programmable Switches

Bluebird: High-performance SDN for Bare-metal Cloud Services

ABST
The clo
hub of
gatewa

Manikandan Arumugam[1] Deepak Bansal[3] Navdeep Bhatia[1] James Boerner[3] Simon Capper[1]

Andromeda: Performance, Isolation, and Velocity at Scale in Cloud Network Virtualization

The bare
as a Ser
tomers a

Mic
Dima

Nil

***Achelous*: Enabling Programmability, Elasticity, and Reliability in Hyperscale Cloud Networks**

Chengkun Wei[†*], Xing Li[†§*], Ye Yang[§†*], Xiaochong Jiang[†], Tianyu Xu[†], Bowen Yang[§], Taotao Wu[§], Chao Xu[§], Yilong Lv[§], Haifeng Gao[§], Zhentao Zhang[§], Zikang Chen[§], Zeke Wang[†], Zihui Zhang[†], Shunmin Zhu[‡§□], Wenzhi Chen[†□]

[†]Zhejiang University [§]Alibaba Cloud [‡]Tsinghua University

ABSTRACT **Abstract**

system together end to end. We developed *Andromeda*,

What

In-network IP address caching

Why

Need efficient updates & fast routing

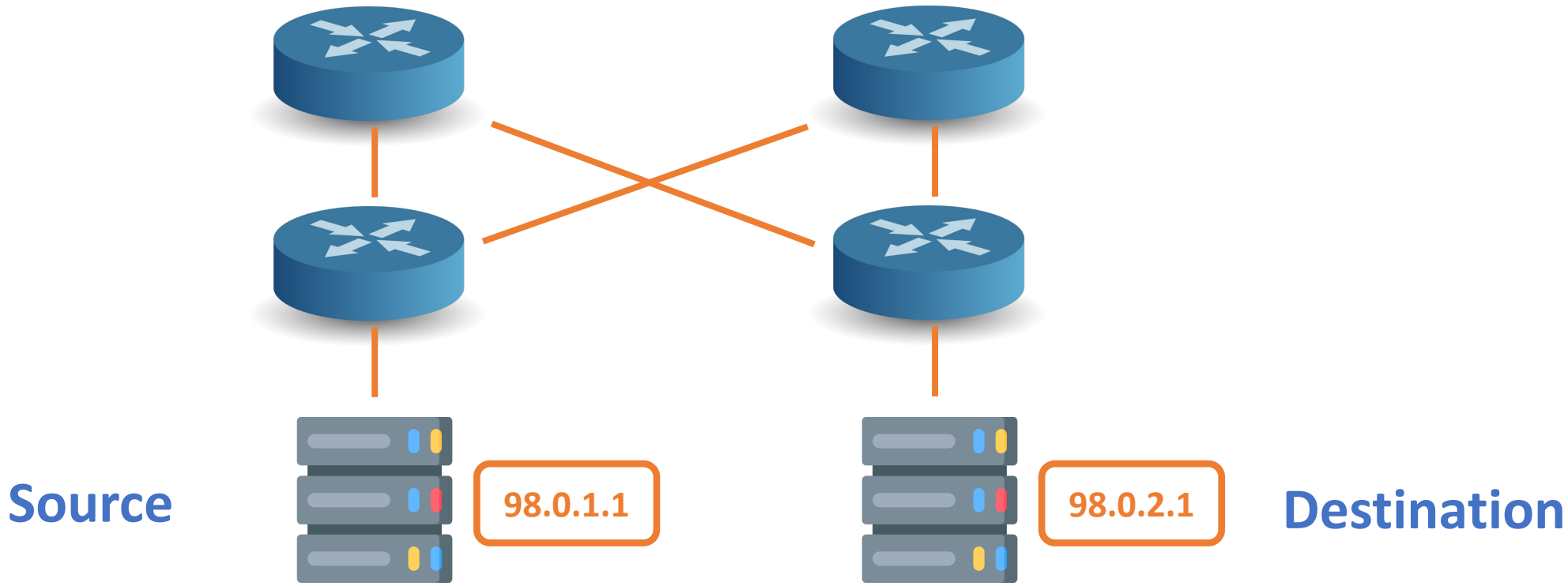
How

Systematic approach backed by distributed computing results

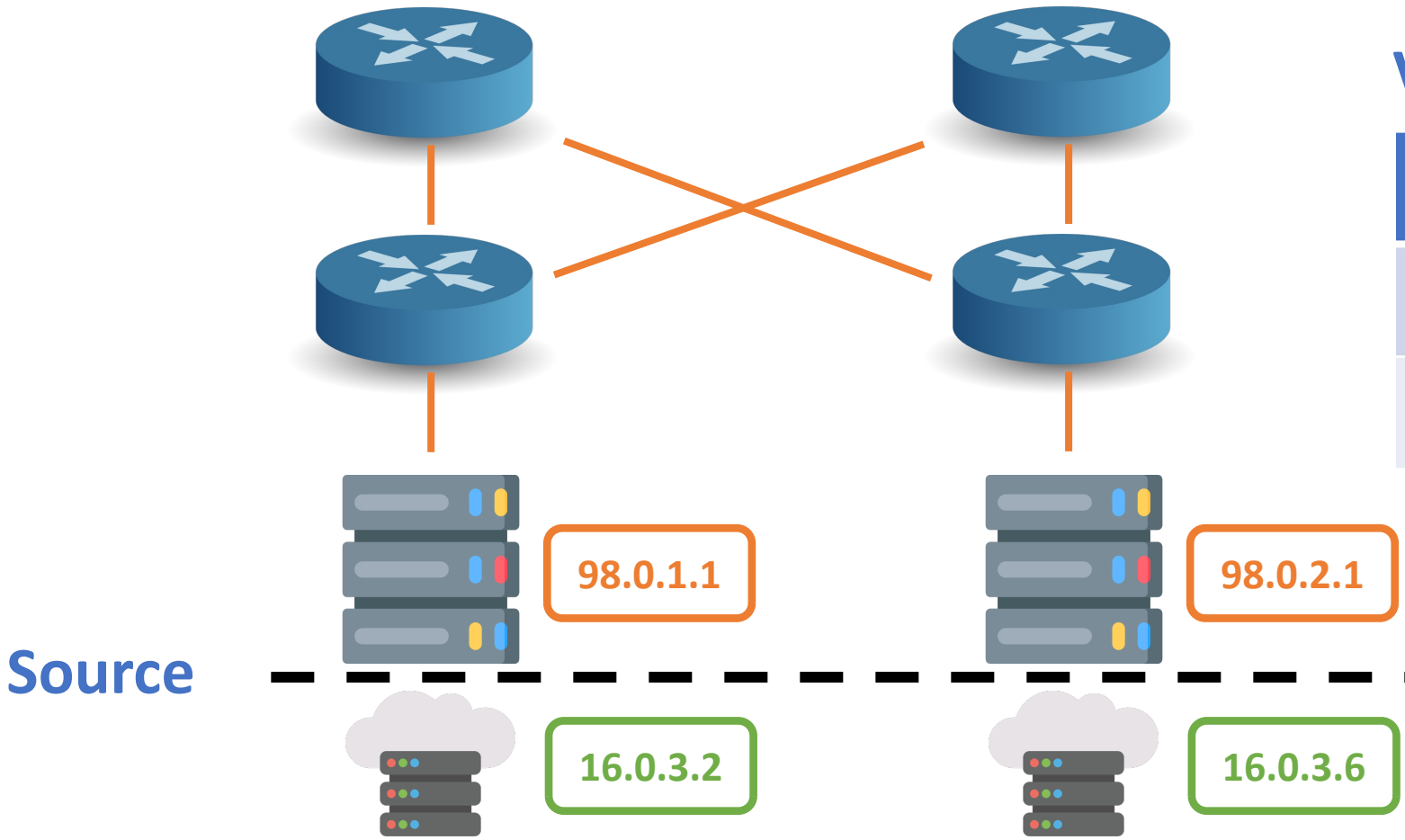
Results

Better	7.8×	4.3×	6.1×
	FCT	First-packet	Bandwidth overhead

Background:
Packet Forwarding in Physical Networks



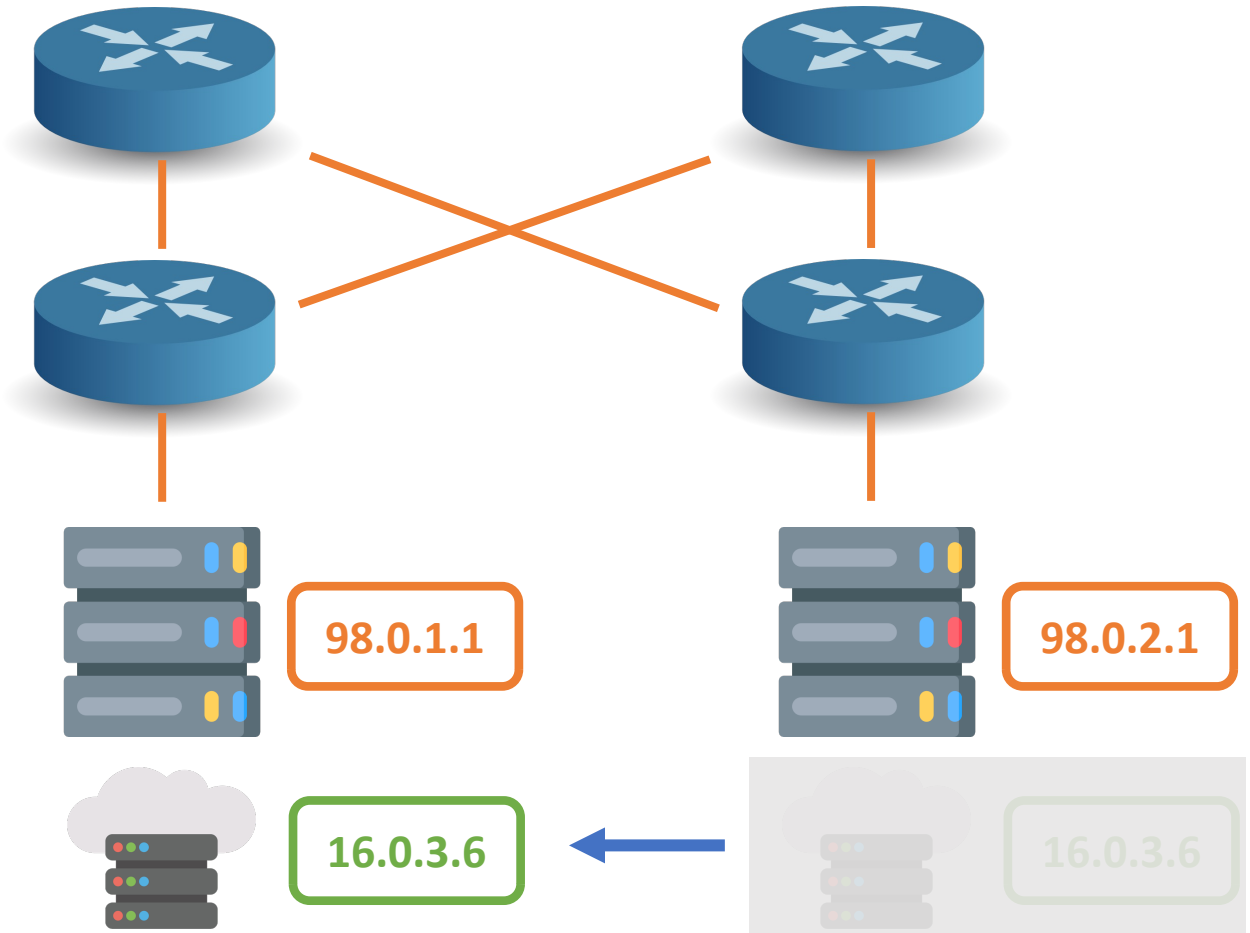
Background:
Packet Forwarding in Virtual Networks



V2P Mappings

Virtual	Physical
16.0.3.2	98.0.1.1
16.0.3.6	98.0.2.1

Background: VM Roaming



V2P Mappings

Virtual	Physical
16.0.3.6	98.0.2.1

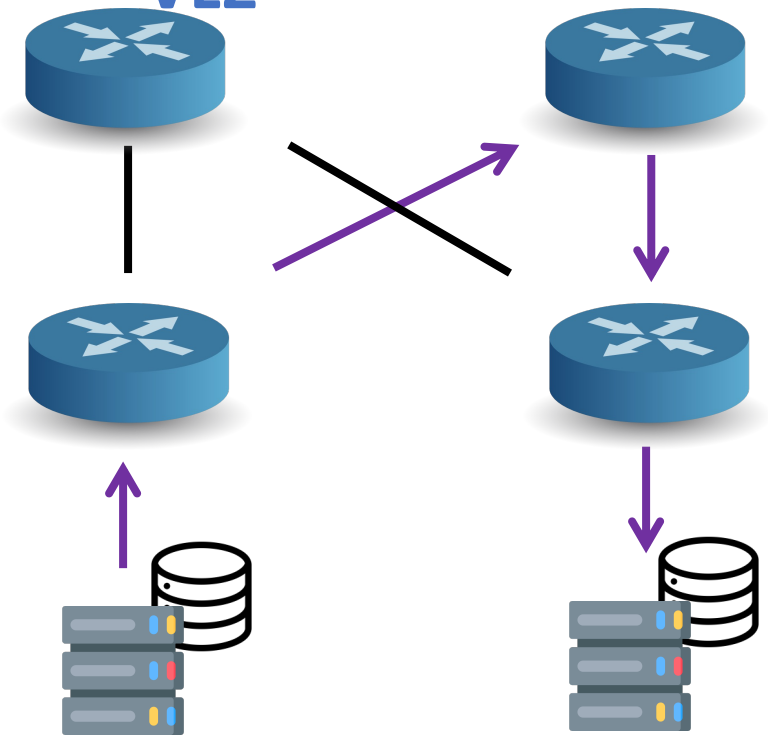
Virtual	Physical
16.0.3.6	98.0.1.1

Where to Perform Address Translation?

Host-Driven Design

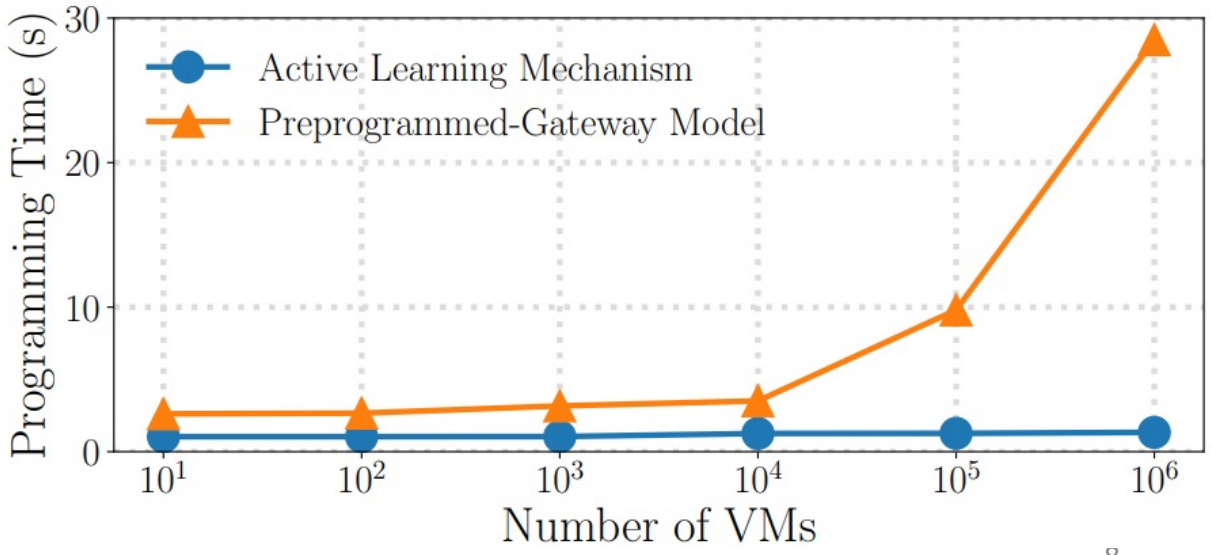
AccelNet VL2

Achelous



Fast packet forwarding

Slow updates: large networks – seconds, single server – 10s ms



Achelous: Enabling Programmability, Elasticity, and Reliability in Hyperscale Cloud Networks, SIGCOMM'23.

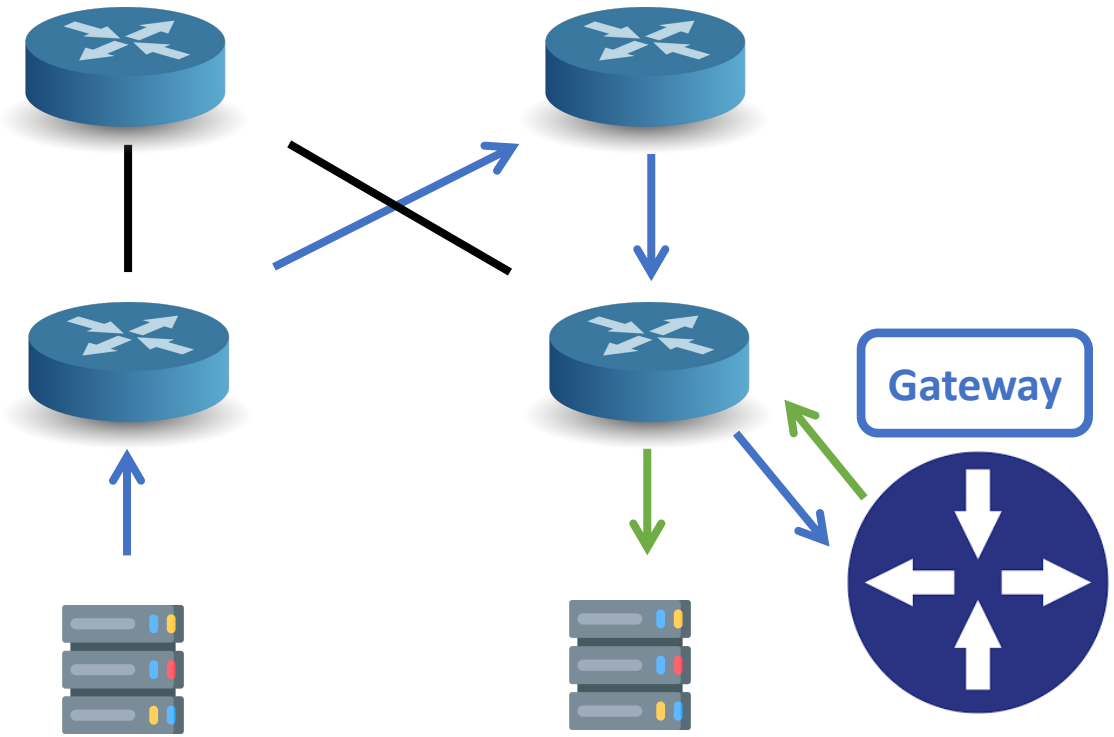
Gateway-Driven Designs

Bluebird

Sailfish

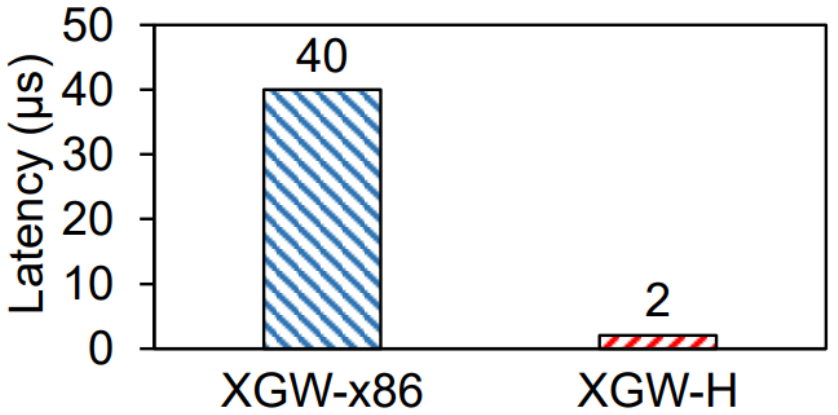
Andromeda

Zeta



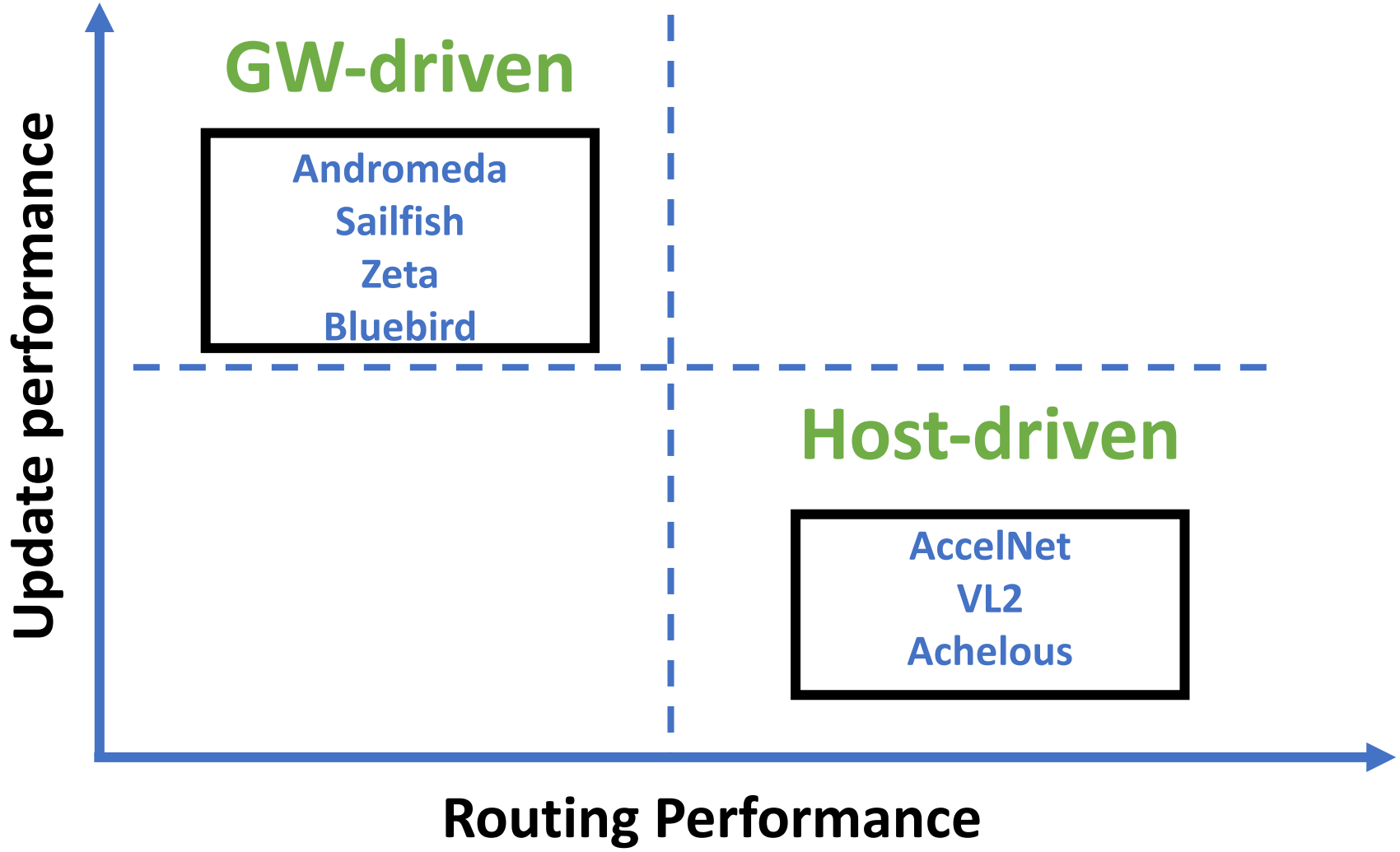
Fast updates

Slow packet forwarding (40µs processing overhead), network bandwidth overhead

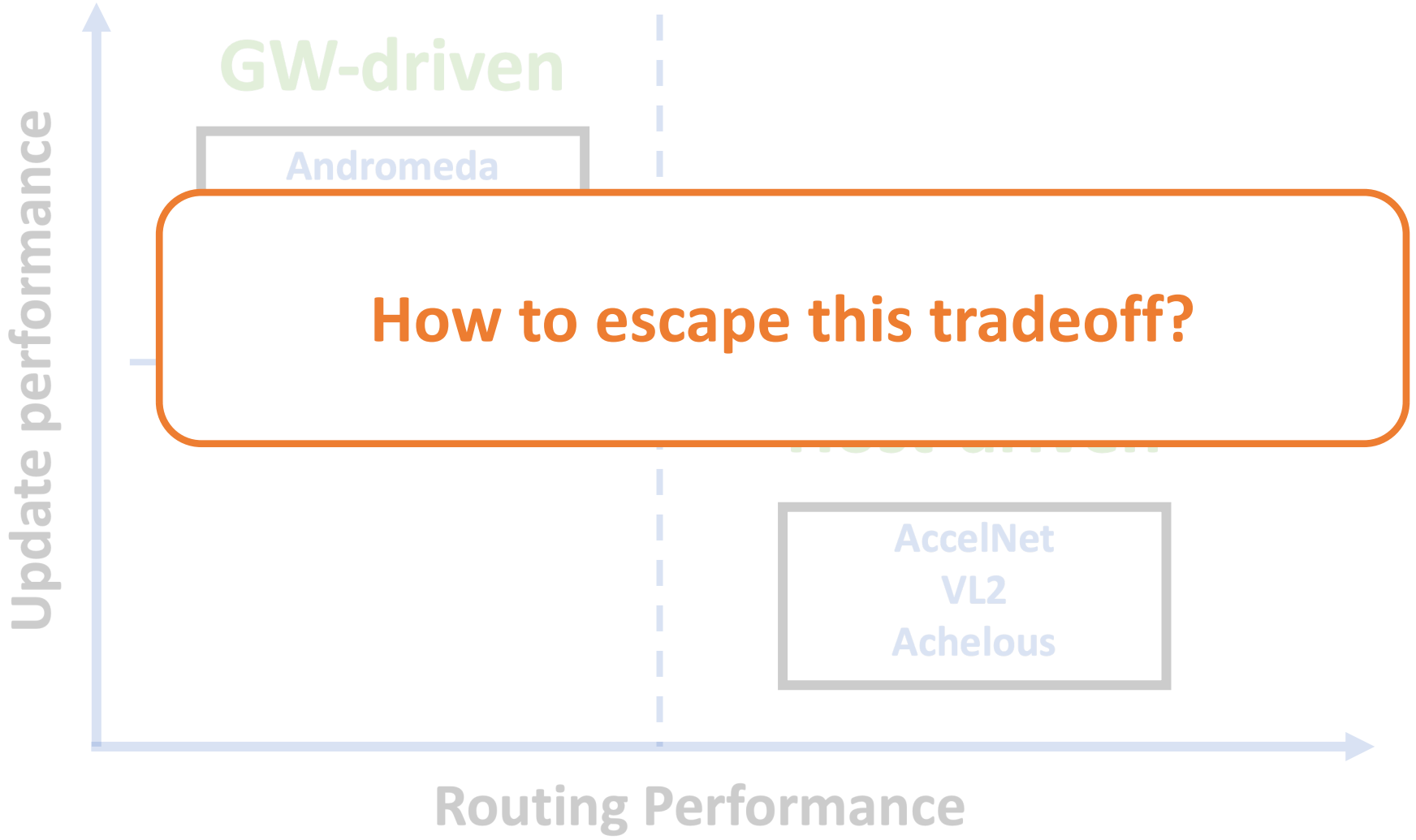


Sailfish: Accelerating Cloud-Scale Multi-Tenant Multi-Service Gateways with Programmable Switches, SIGCOMM'21.

The Tradeoff



The Lookup-Update Tradeoff



The Read (Lookup)-Write (Update) Tradeoff

Limiting the replication factor balances read and write costs

The Dangers of Replication and a Solution

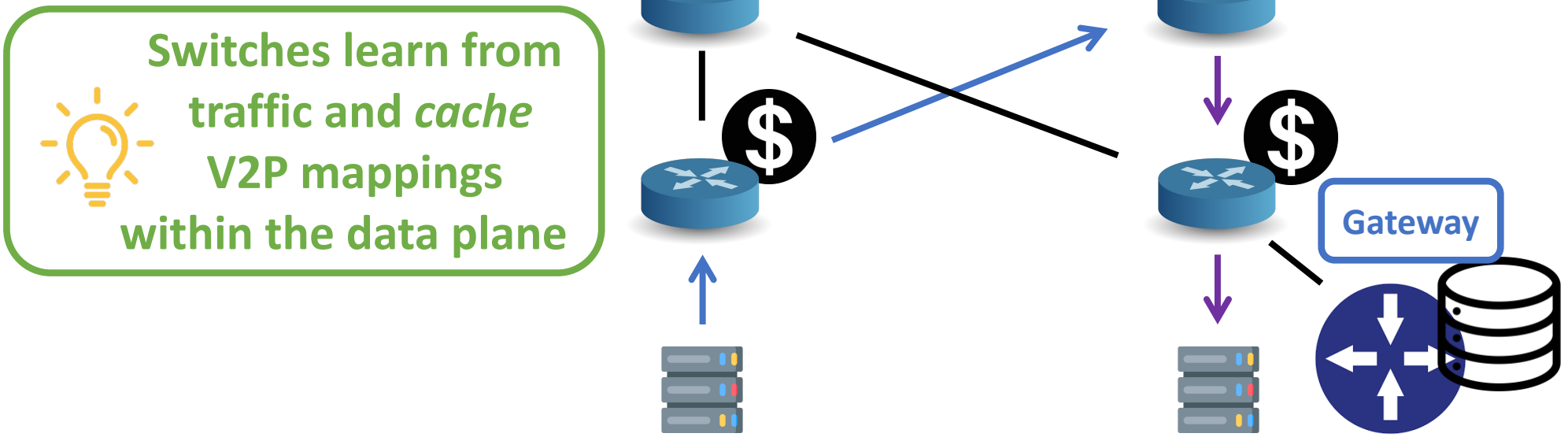
Jim Gray (Gray@Microsoft.com)

THE COST OF DATA REPLICATION +

Hector Garcia-Molina
Daniel Barbara

Department of Electrical Engineering and Computer Science
Princeton University
Princeton, New Jersey 08544

Our Idea: In-Network Address Caching

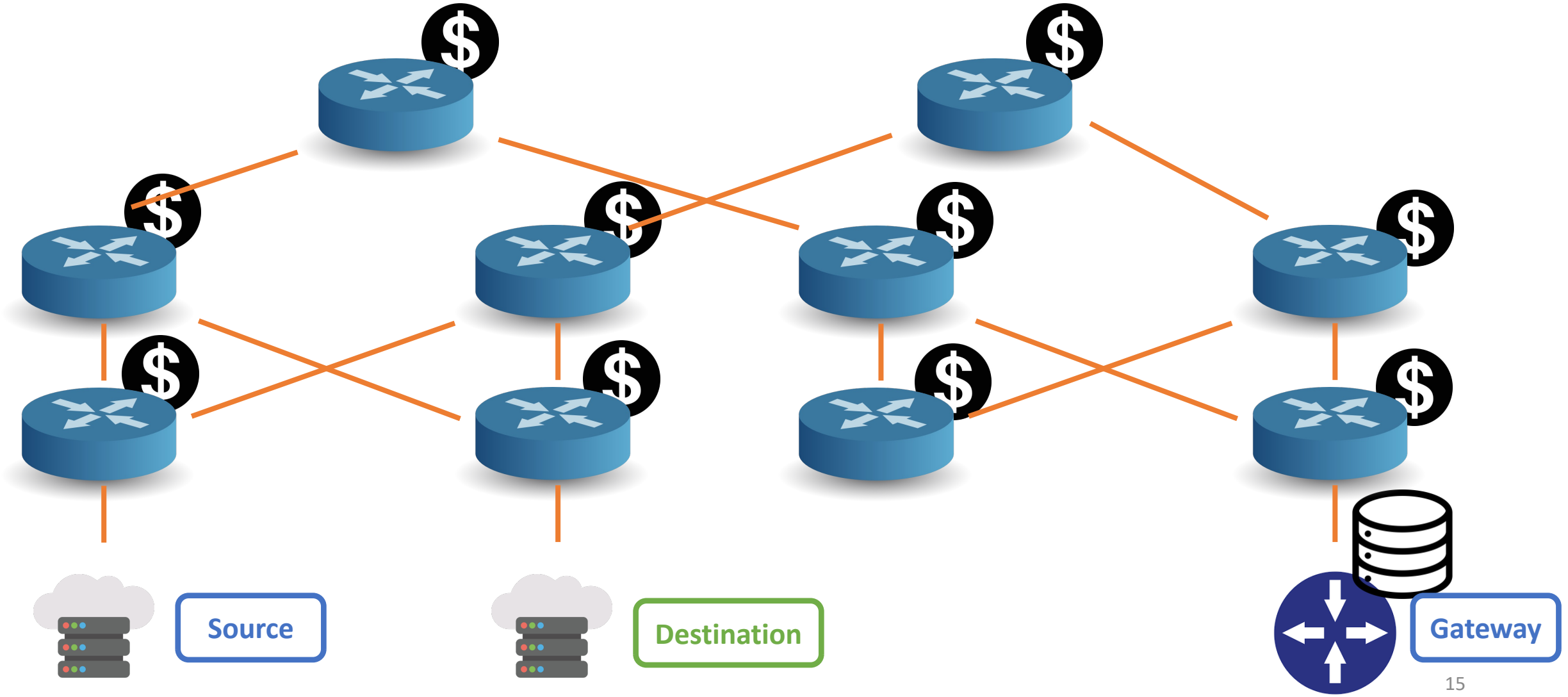


Fast routing
Fast updates
Reduced gateway costs
No routing changes
Incrementally deployable
Compatible with existing switches

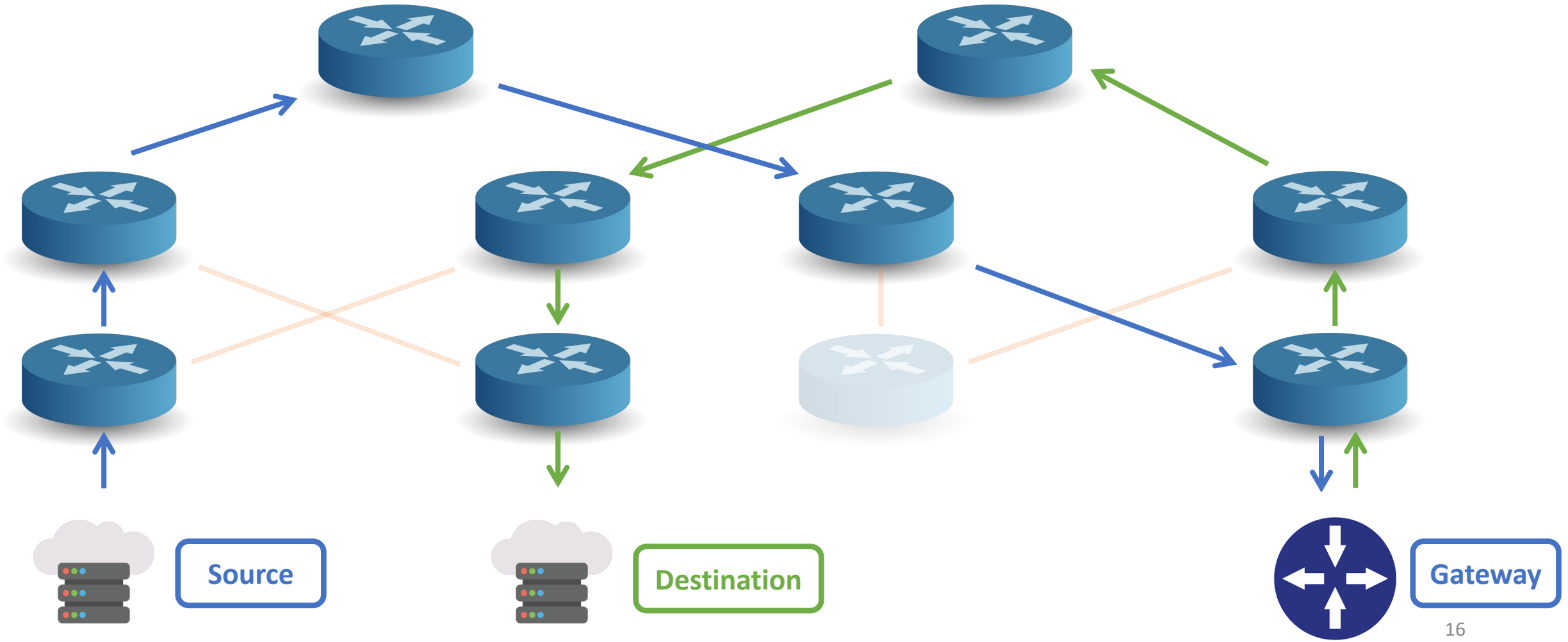
Agenda

- Background
- SwitchV2P: overview
- Challenges
- Design
- Experimental results

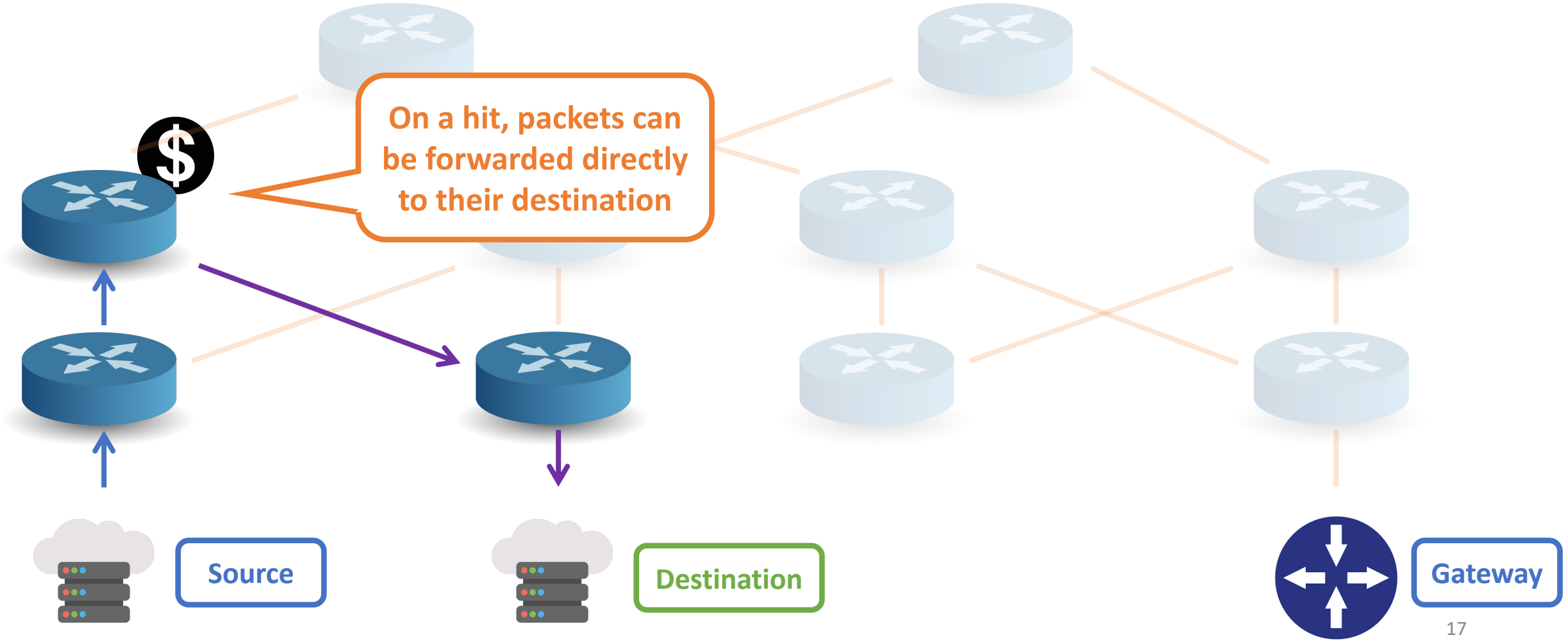
In-Network Address Translation



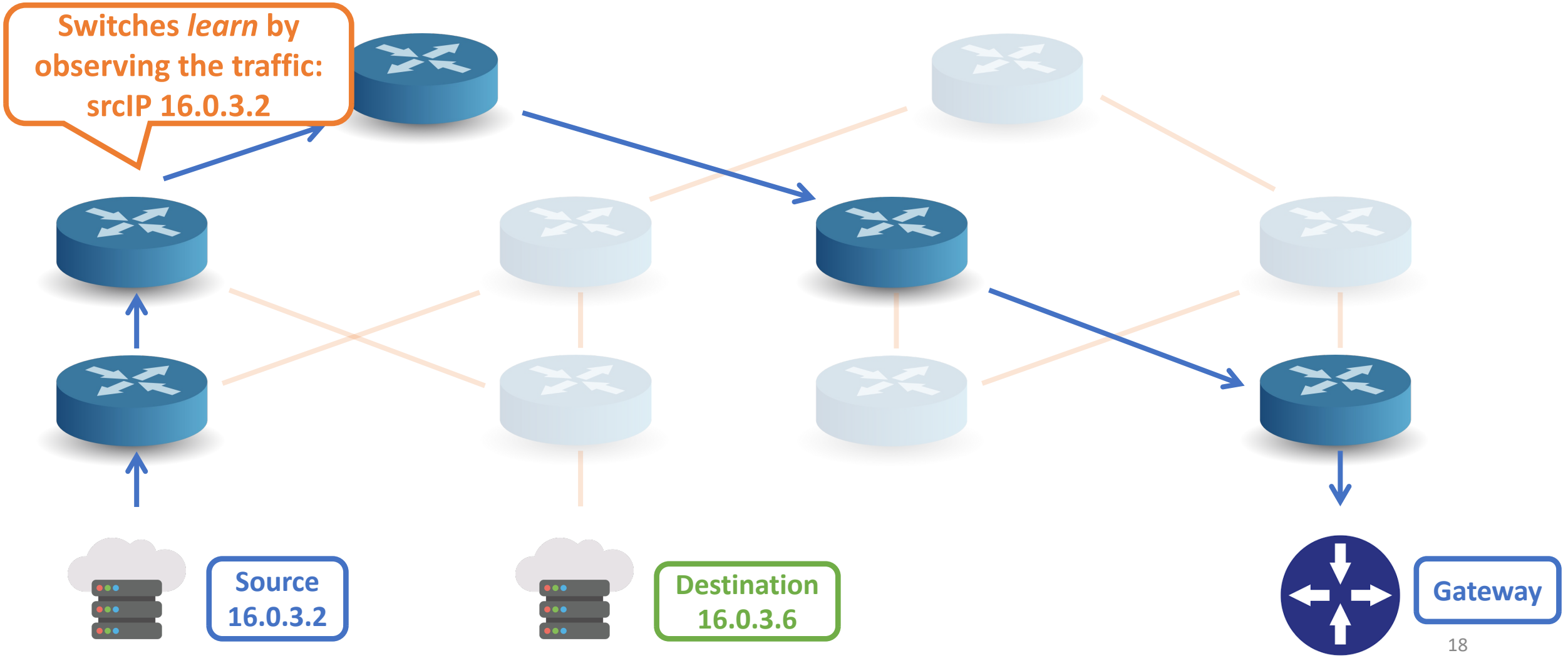
Cache Miss



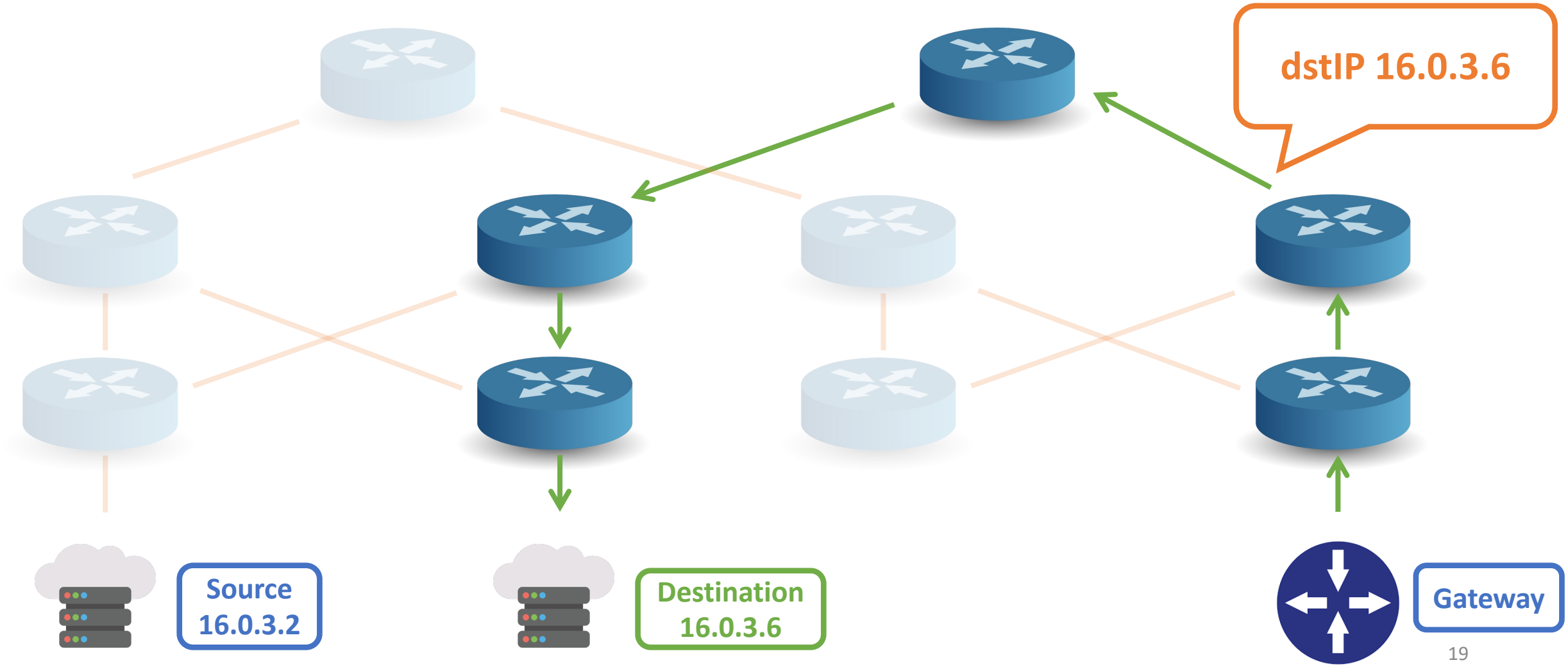
Cache Hit: Source → Destination



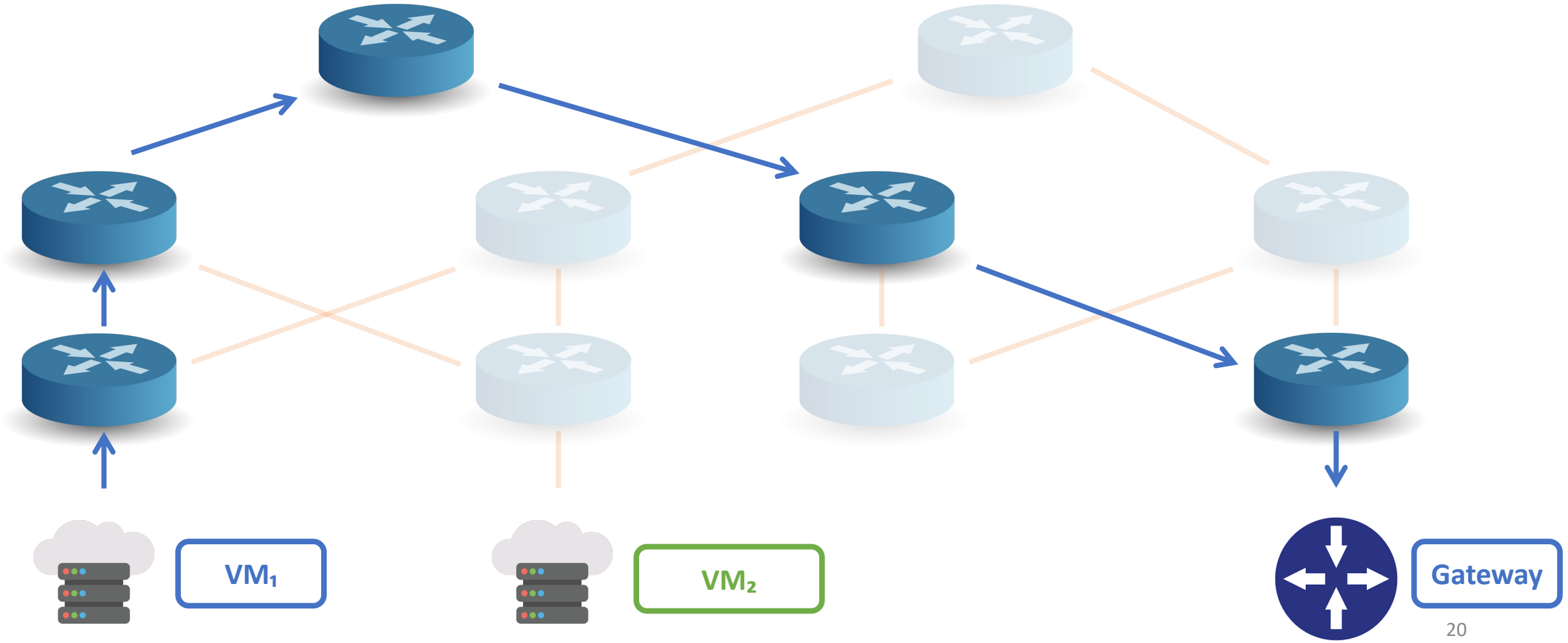
Caching by Learning



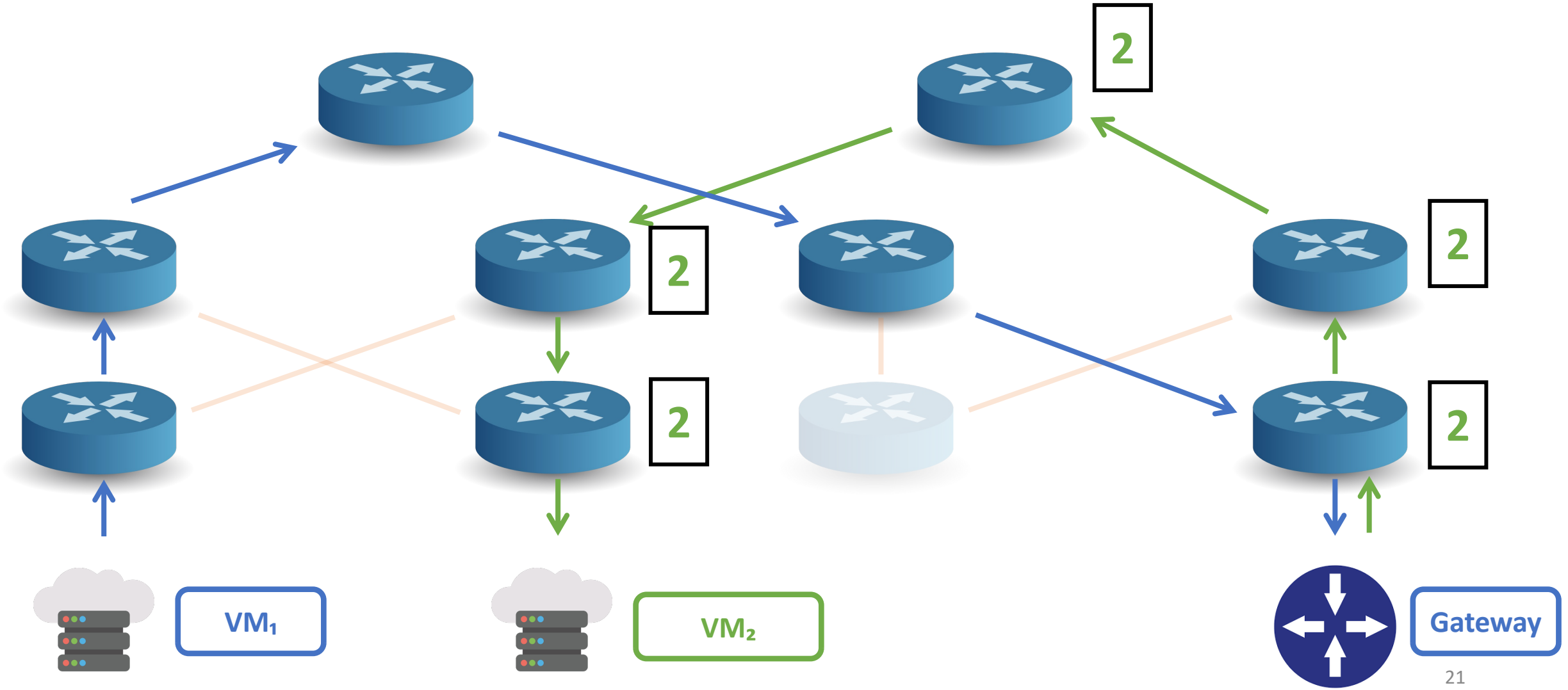
Caching by Learning



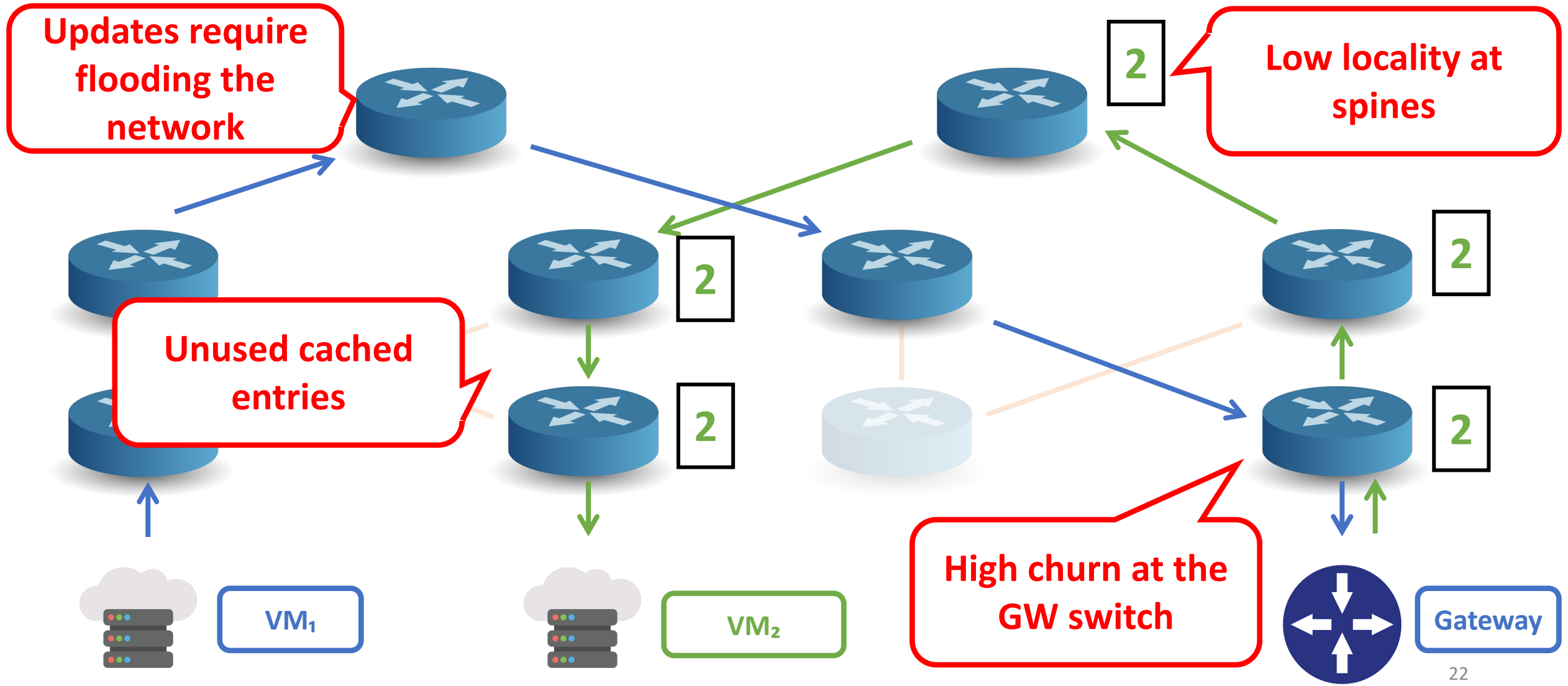
The Greedy Approach



The Greedy Approach



Local Decisions are not Enough

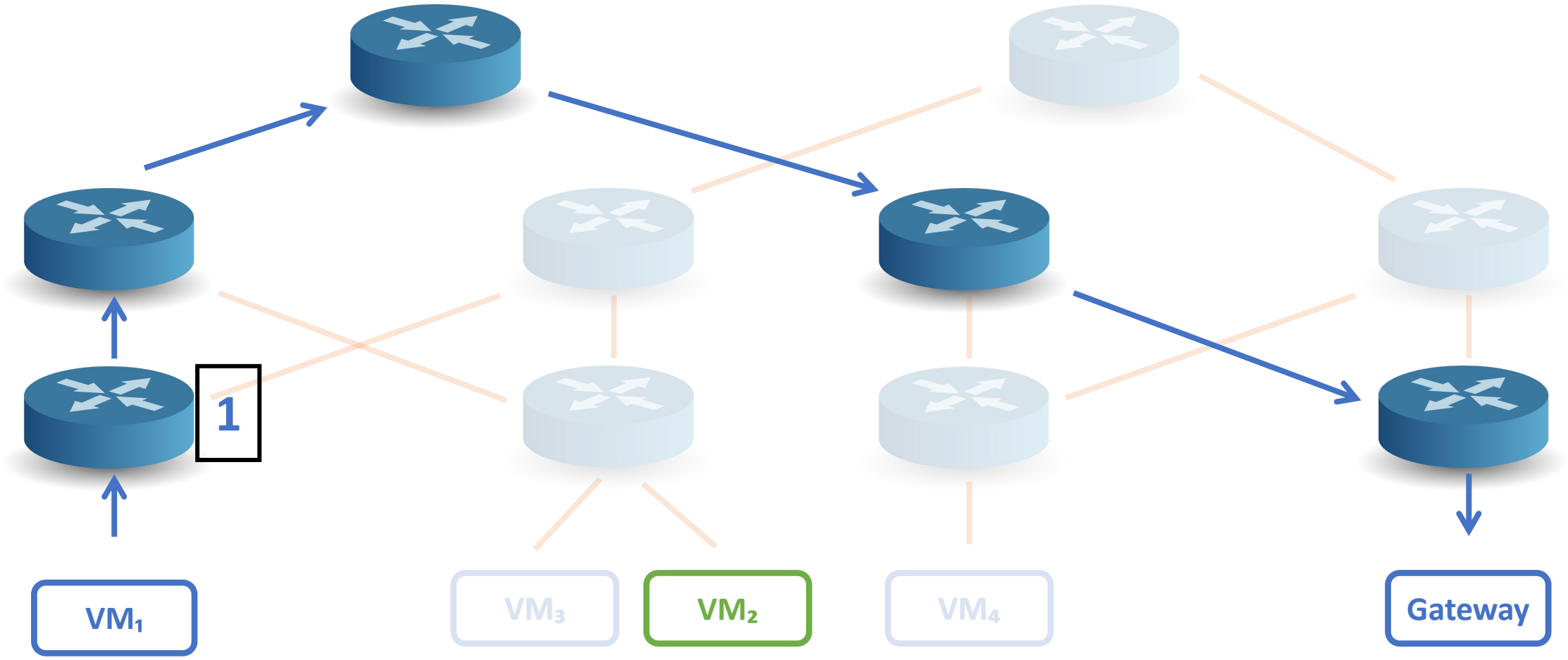


Topology-Aware Caching

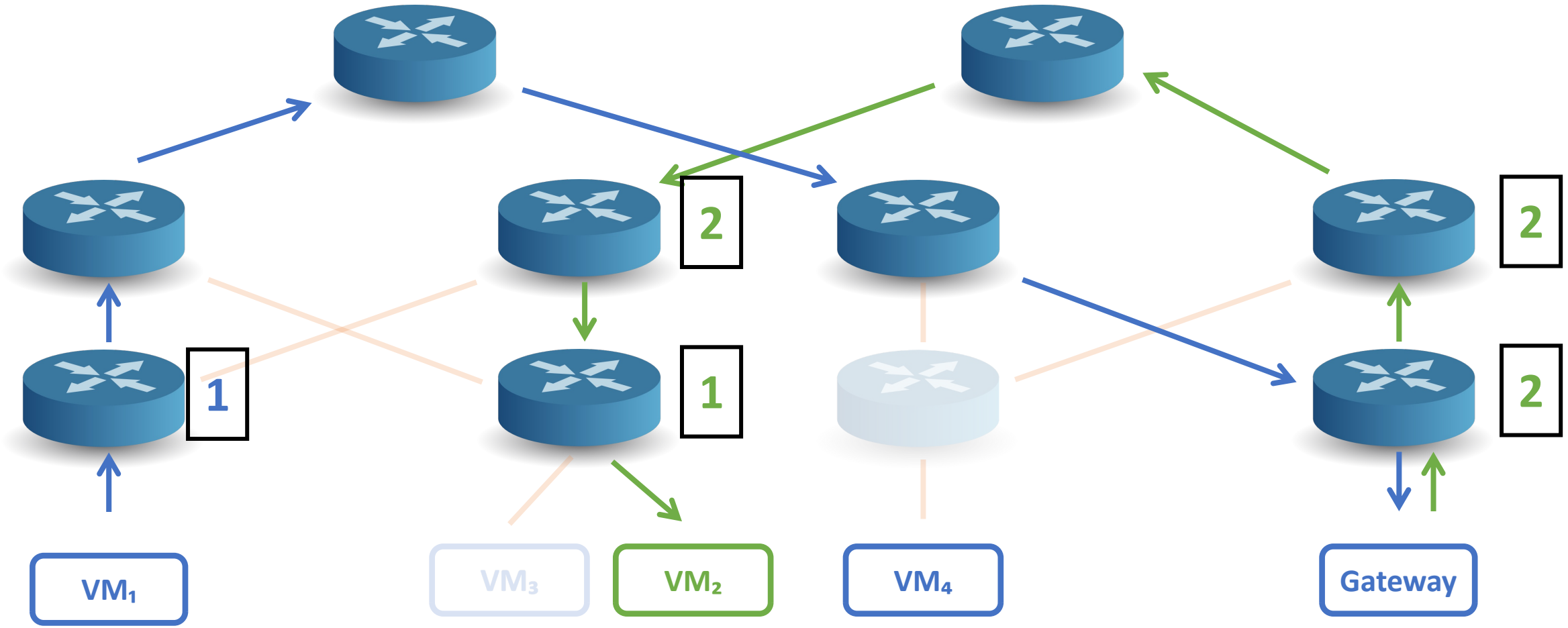
- Direct-mapped cache with small metadata (1 bit)
- ToRs learn source addresses
- GW ToRs learn destination addresses
- Evicted entries are spilled to other switches
- Popular entries are promoted to upper levels
- Move mappings to the traffic

Please see the paper for more details!

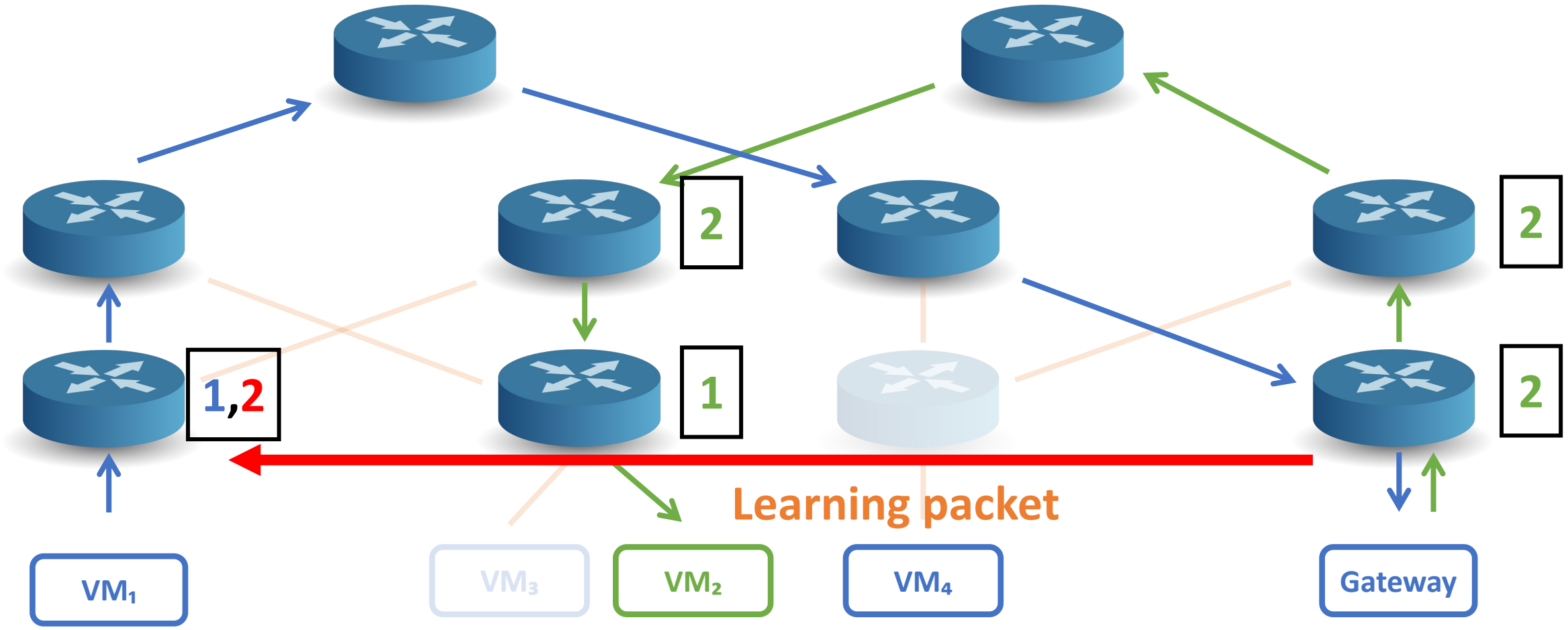
Example: $VM_1 \rightarrow VM_2$



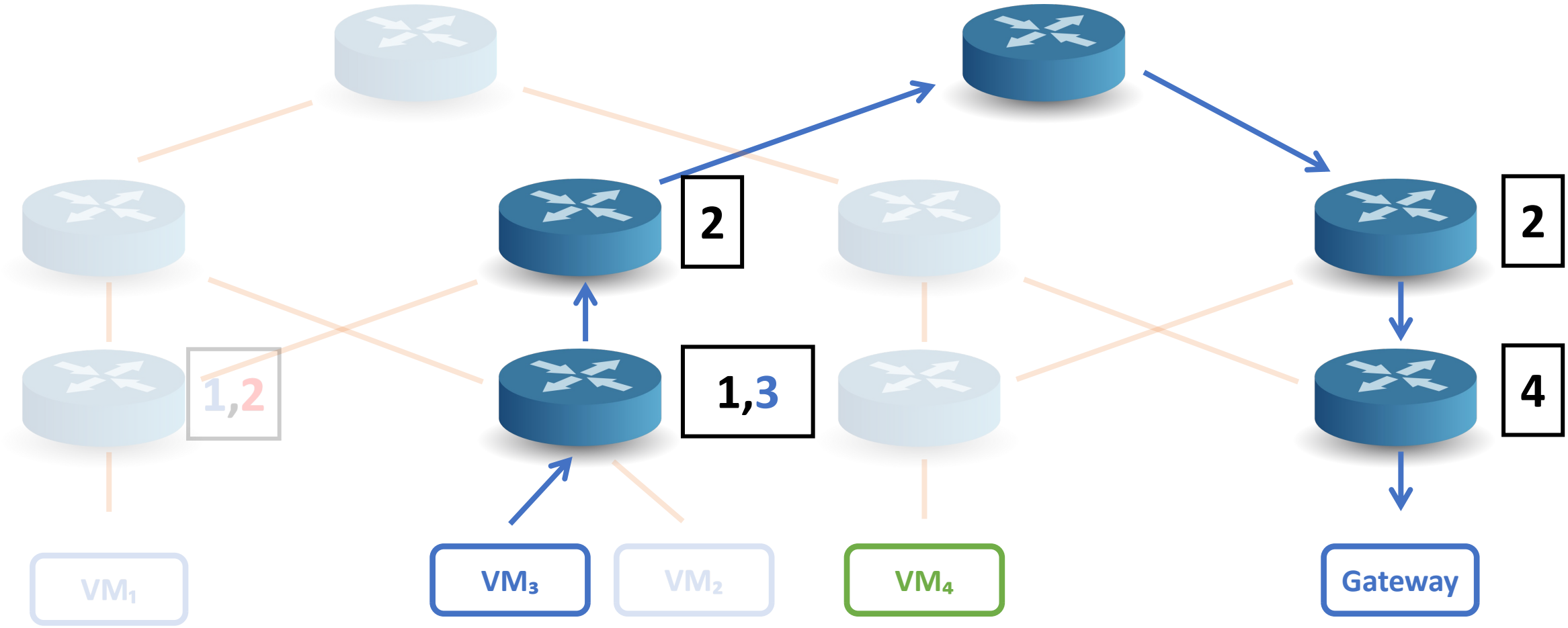
Example: $VM_1 \rightarrow VM_2$



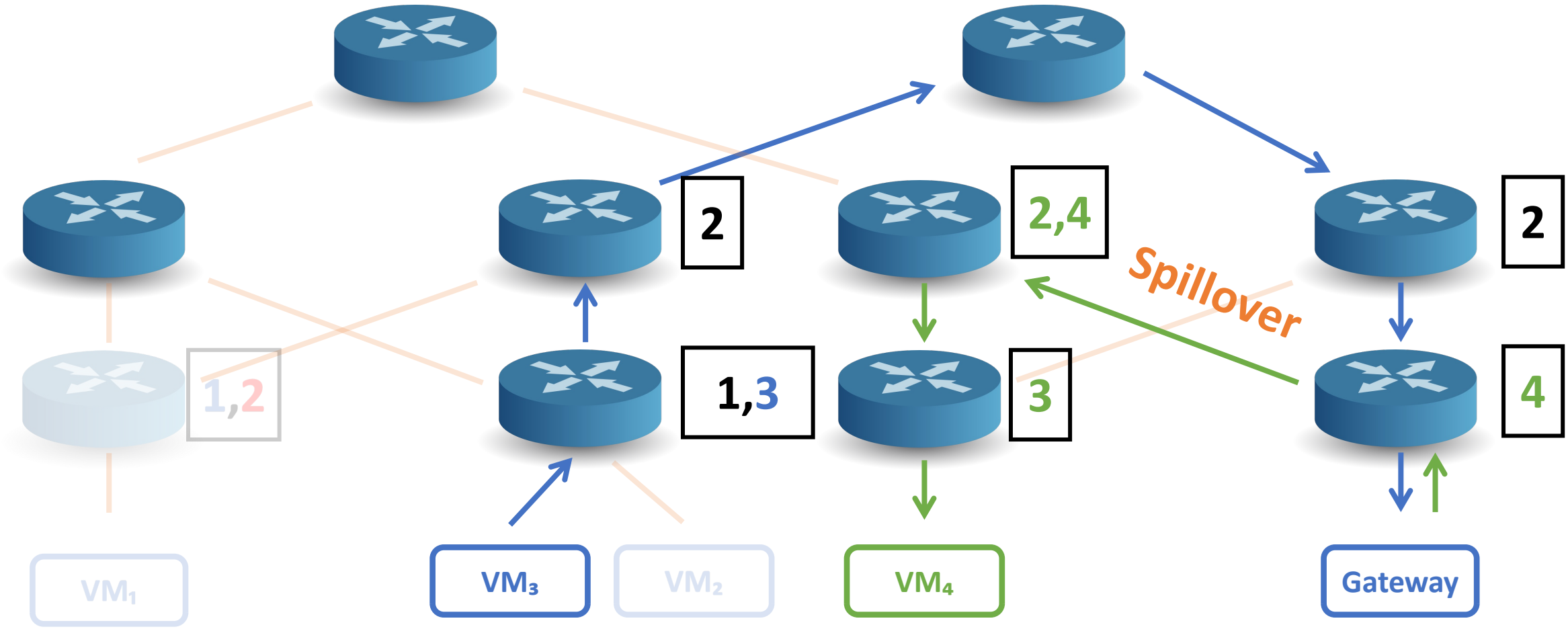
Example: $VM_1 \rightarrow VM_2$



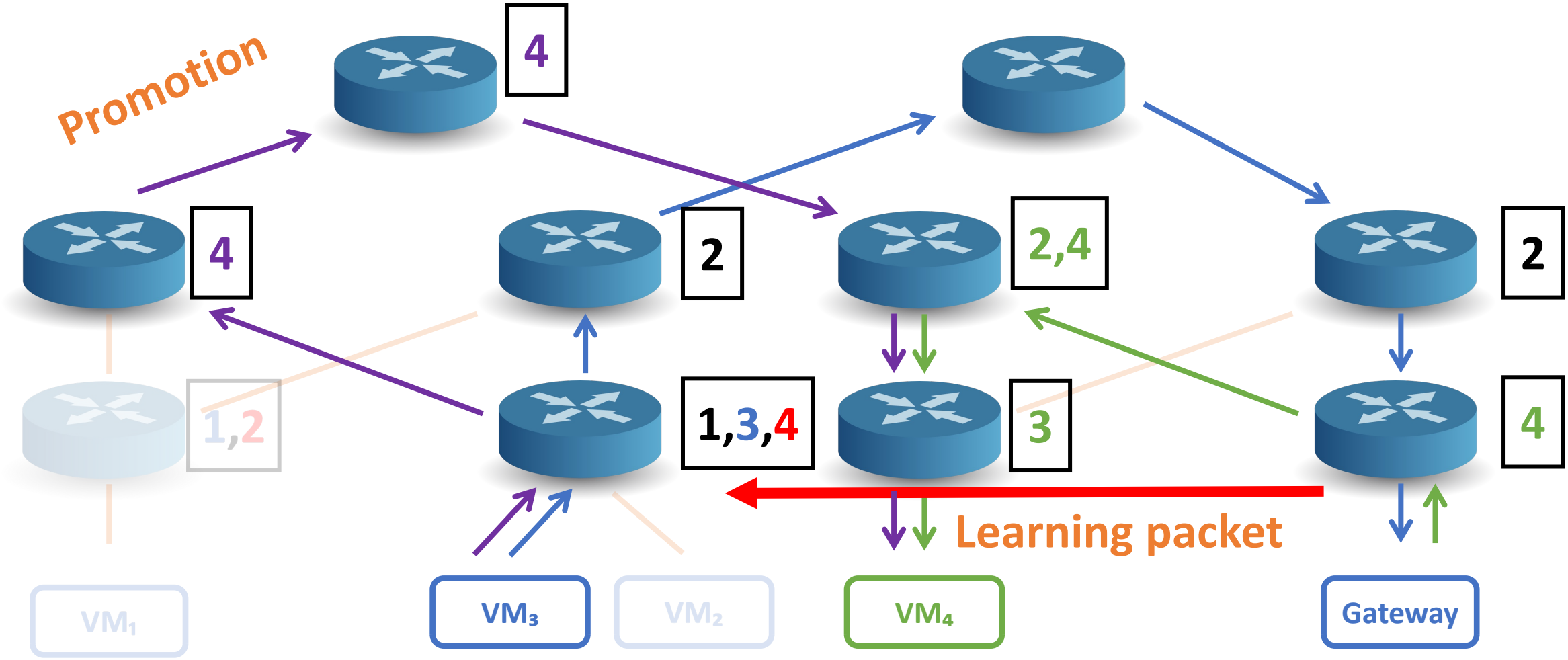
Example: $VM_3 \rightarrow VM_4$



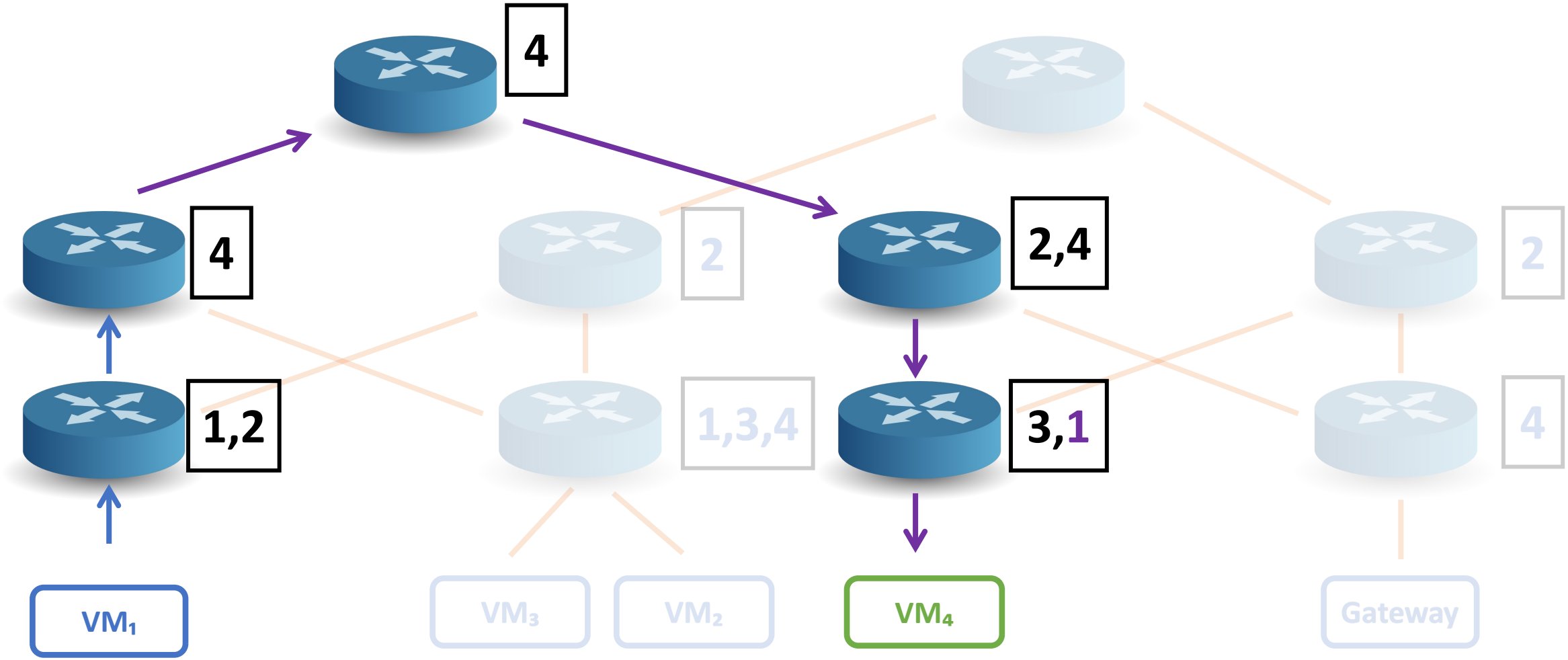
Example: $VM_3 \rightarrow VM_4$



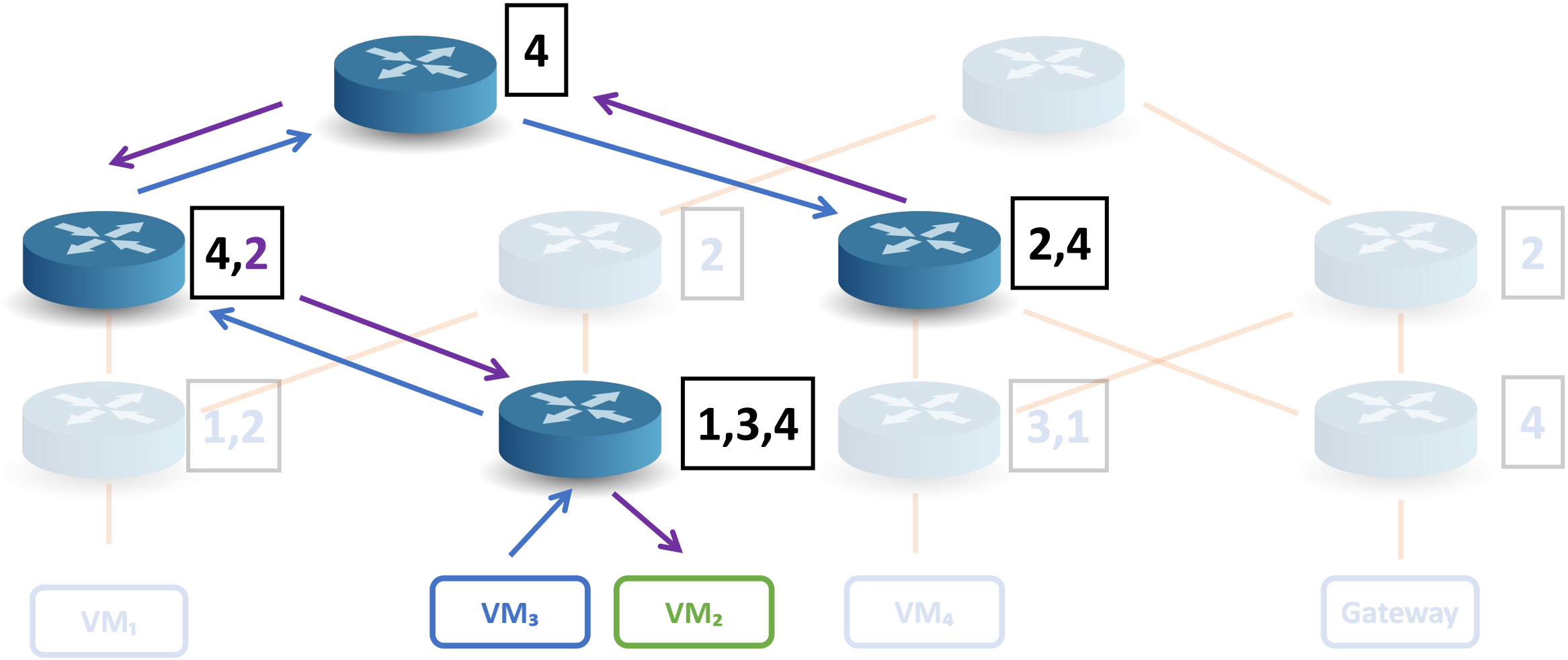
Example: $VM_3 \rightarrow VM_4$



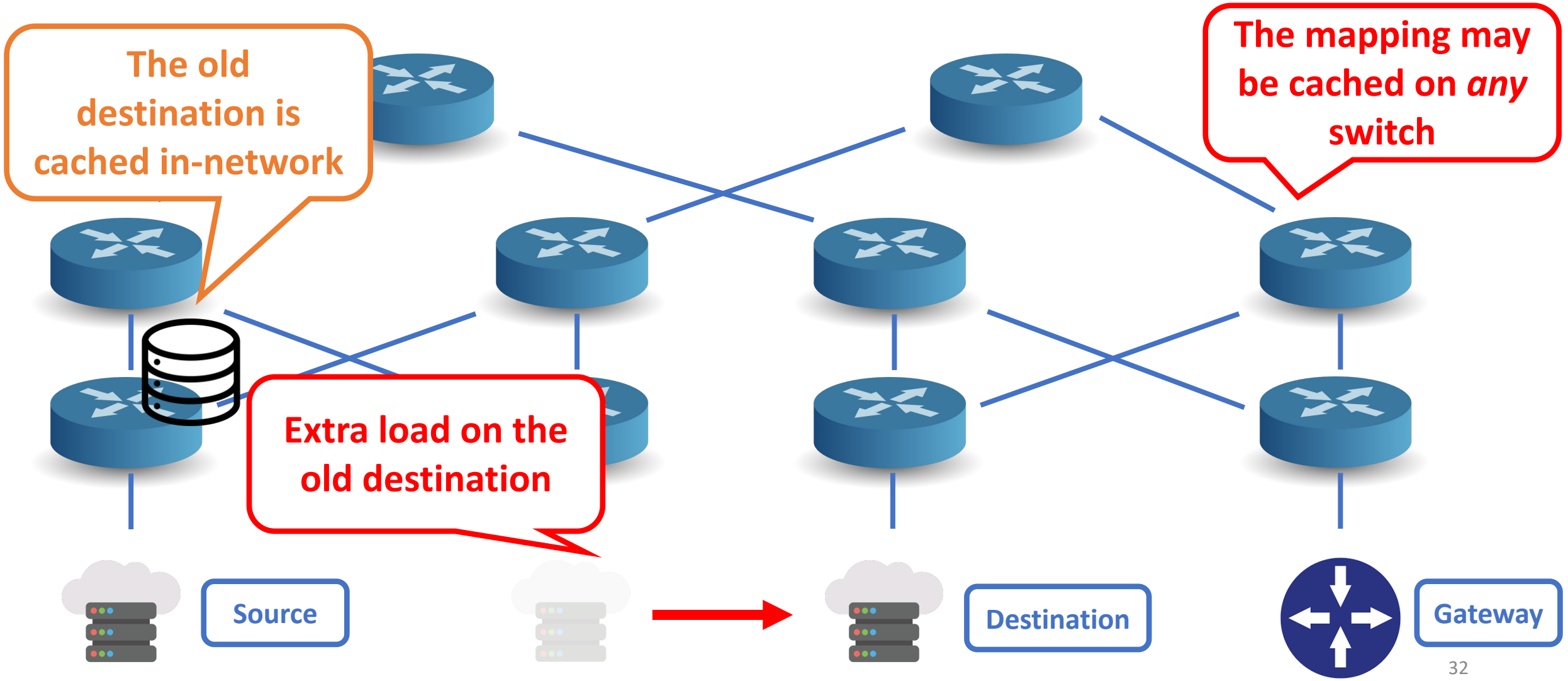
Example: $VM_1 \rightarrow VM_4$



Example: $VM_3 \rightarrow VM_2$



Updates


The old destination is cached in-network
The mapping may be cached on *any* switch
Extra load on the old destination
Source
Destination
Gateway

Updates

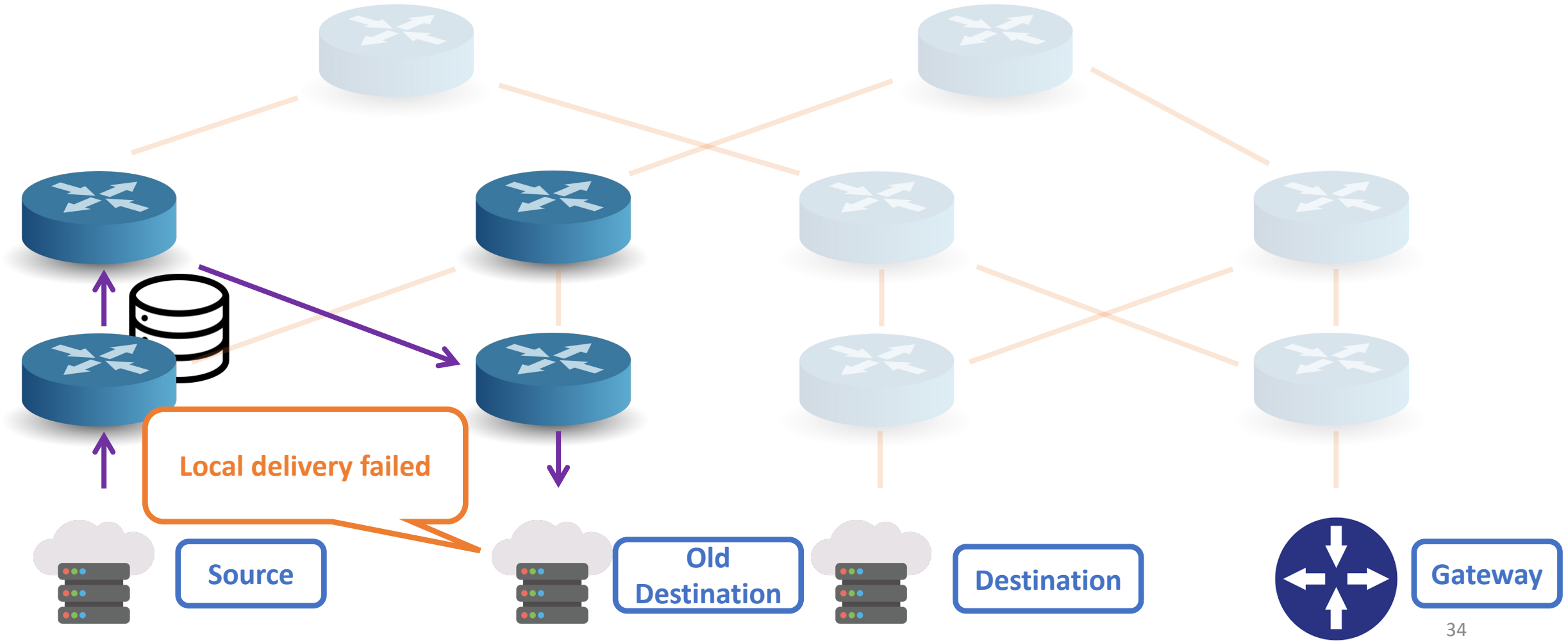
Goal

Minimize the number of misdelivered and invalidation messages

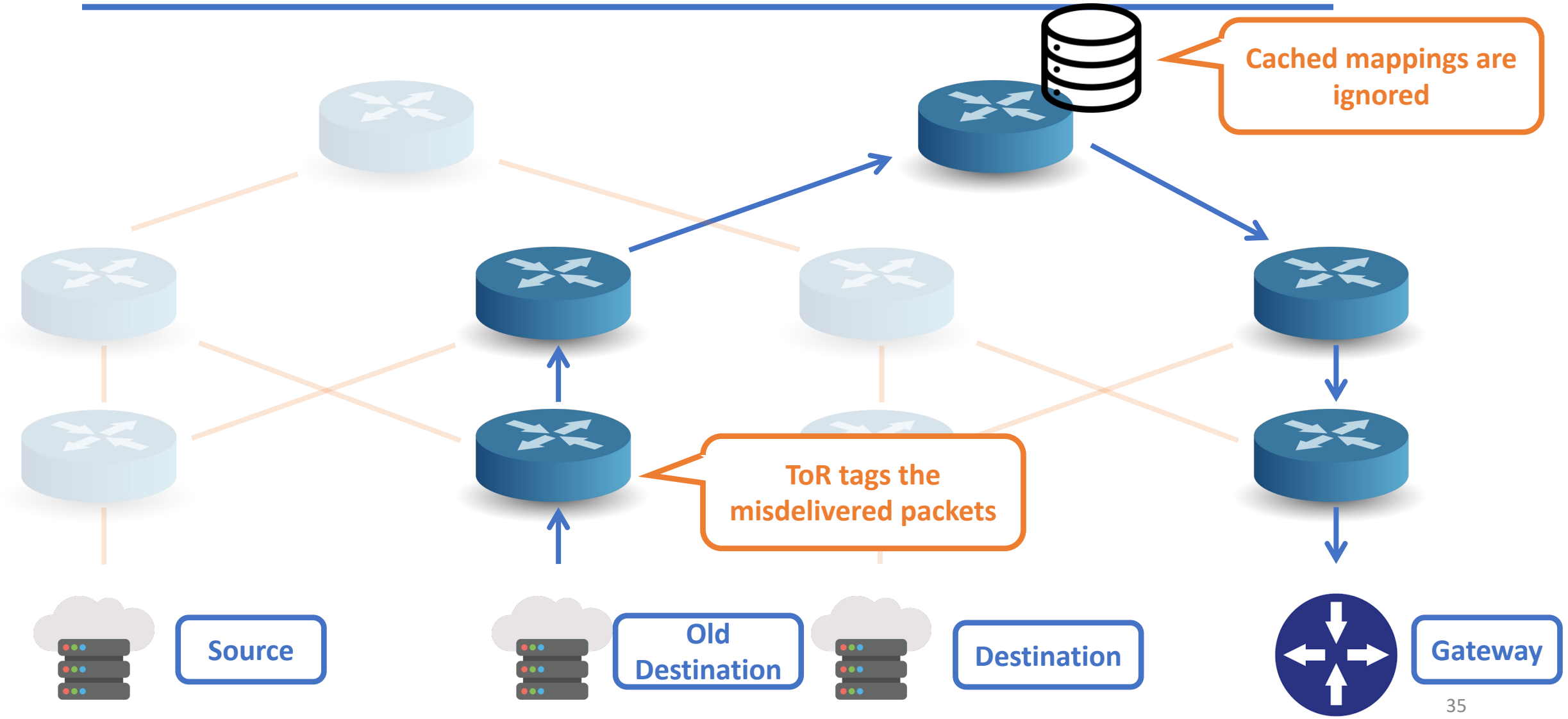
Key idea

Cache coherence is not necessary

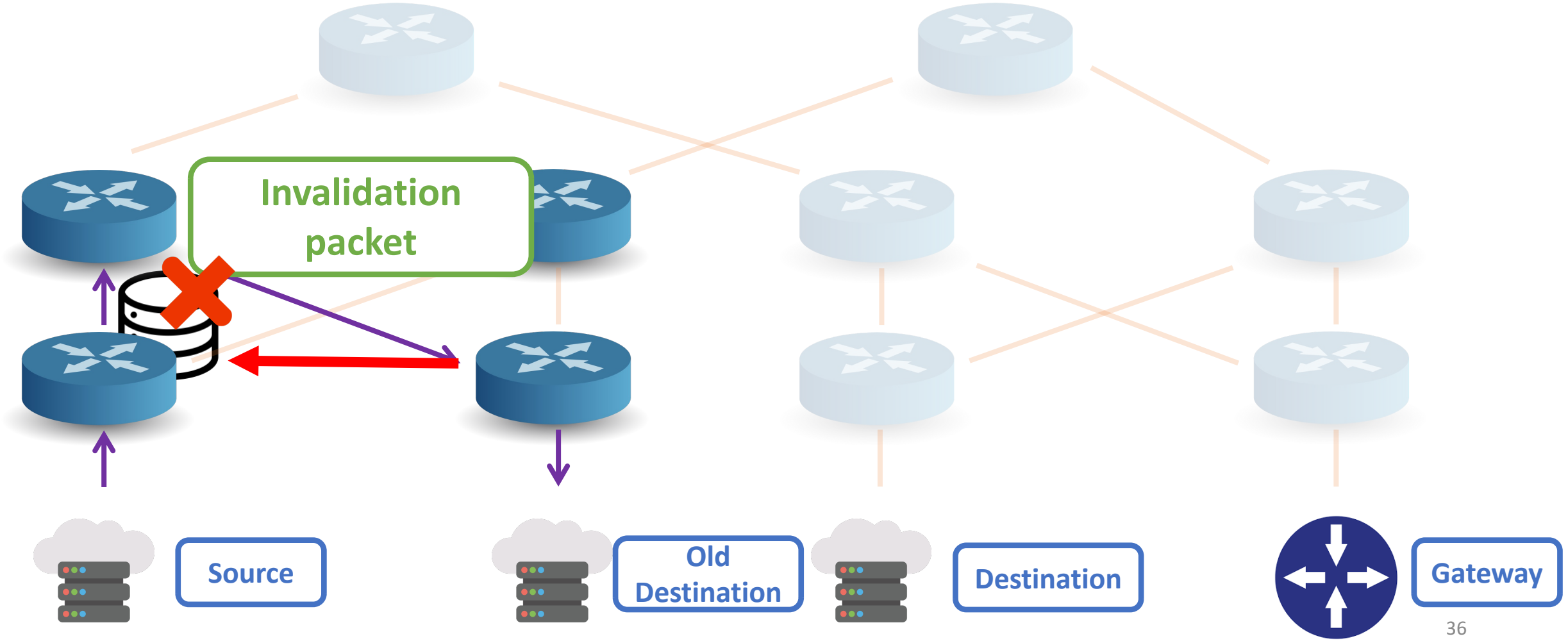
Keeping Correct Forwarding



Keeping Correct Forwarding



Lazy Invalidations



Simulations

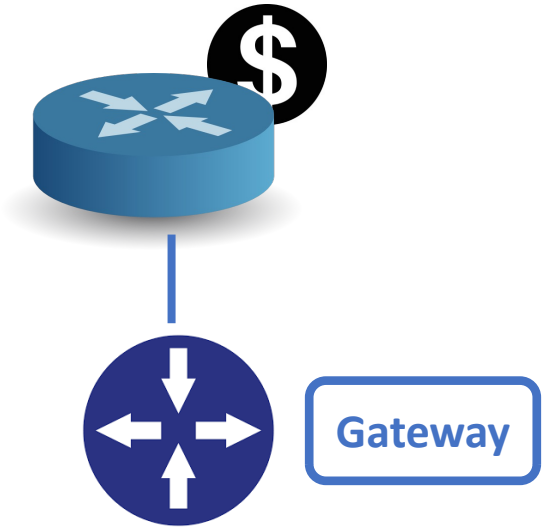
- Large network topologies: 10K VMs, 128 servers, 80 switches (>800 switches for Alibaba)
- Traces: Hadoop, WebSearch, Alibaba RPC, Microbursts, Video
- Network- and application-level metrics

- Main results:
  - Up to 7.8× reduction in FCT and 4.3× reduction in first packet latency
  - Low miss rates (below 1%) - same performance with an order-of-magnitude fewer gateways
  - Reduced network load
  - Low migration costs

Baselines

Andromeda (w/o offloading)

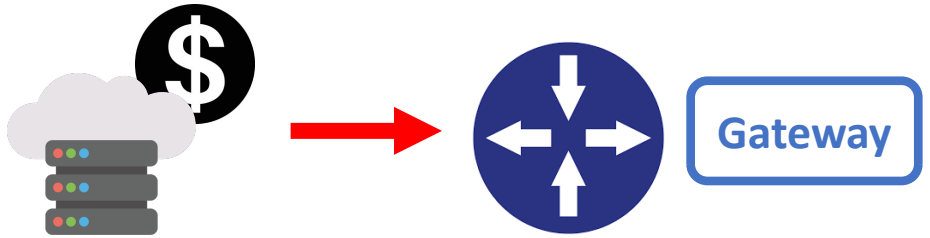
Sailfish



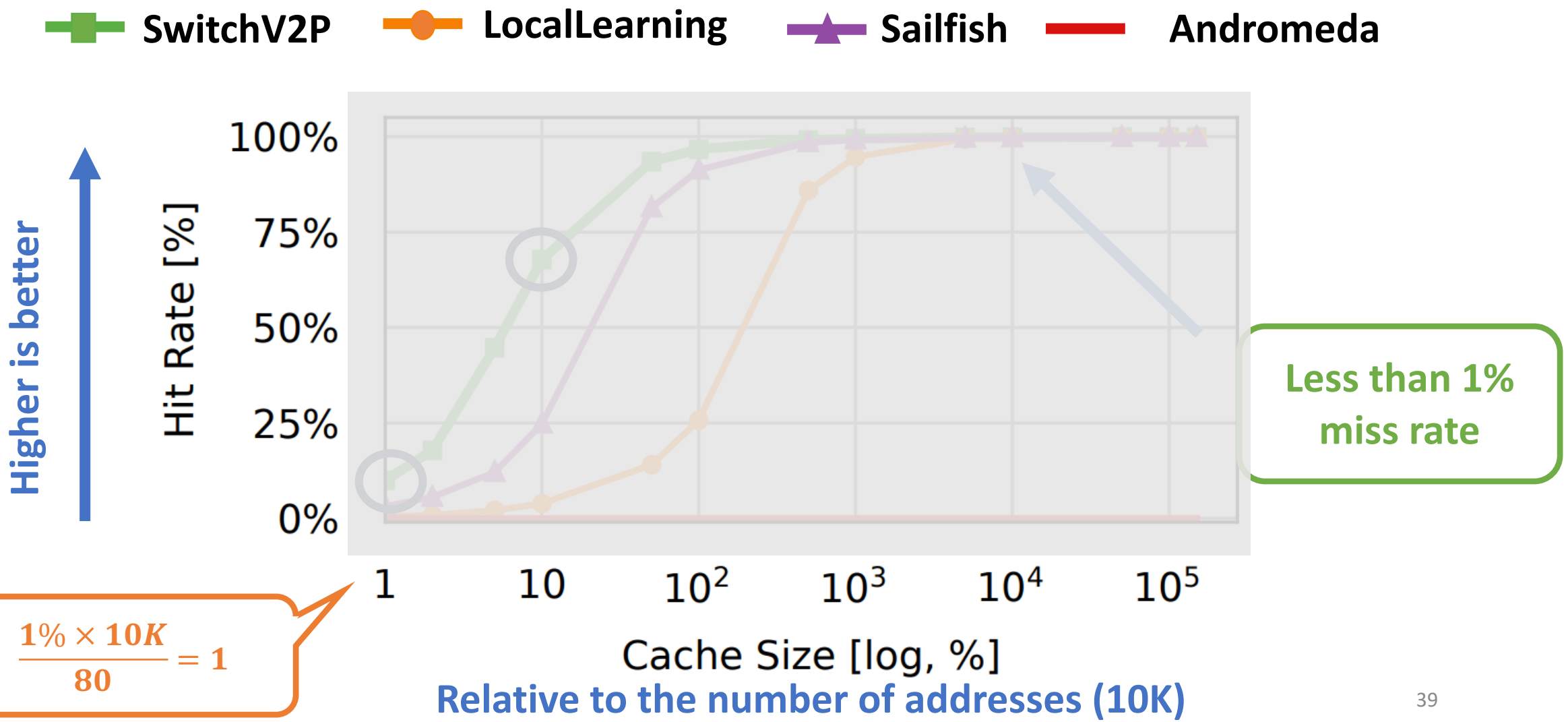
LocalLearning

AccelNet

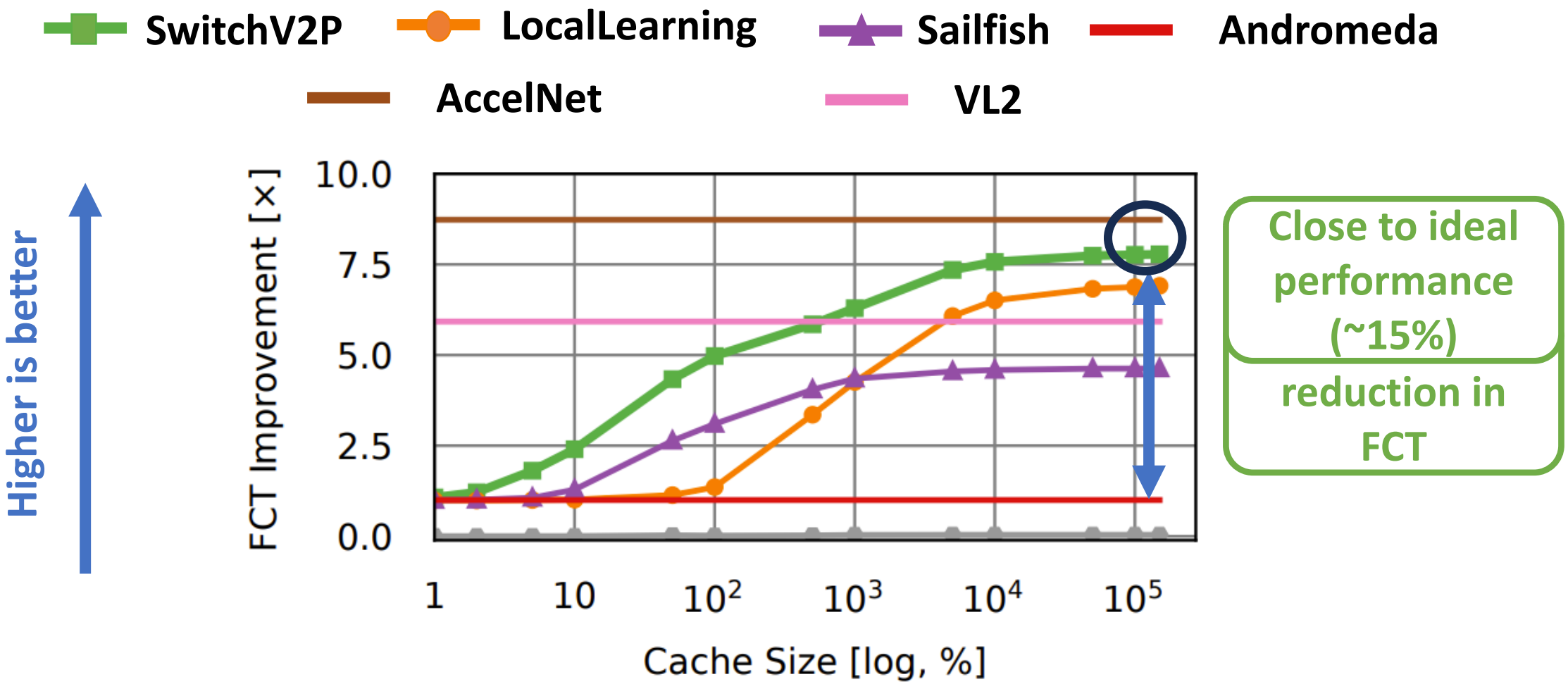
VL2



Hadoop: Hit Rate



Hadoop: FCT



Relative to the number of addresses (10K)

Updates: Results

	Avg. Packet Latency	Gateway Packets	Misdelivered Packets	Total Invalidation Packets
NoCache	1×	100%	1×	
OnDemand	0.25×	0%	11×	
SwitchV2P	0.25×	8.7%	1.2×	24

SwitchV2P reduces the load on the stale destination with a small number of invalidation packets

Conclusions

- Give the power to the switches!
- In-network address translation is practical and efficient
- Key ideas: topology-aware caching, move mappings to the traffic, and lazy invalidations
- Up to 7.8× reduction in FCT and 4.3× reduction in first packet latency
- Up to 6.1× reduction in bandwidth overheads

Thank you!
Questions?

liorz@campus.technion.ac.il