Duet: Cloud Scale Load Balancing with Hardware and Software

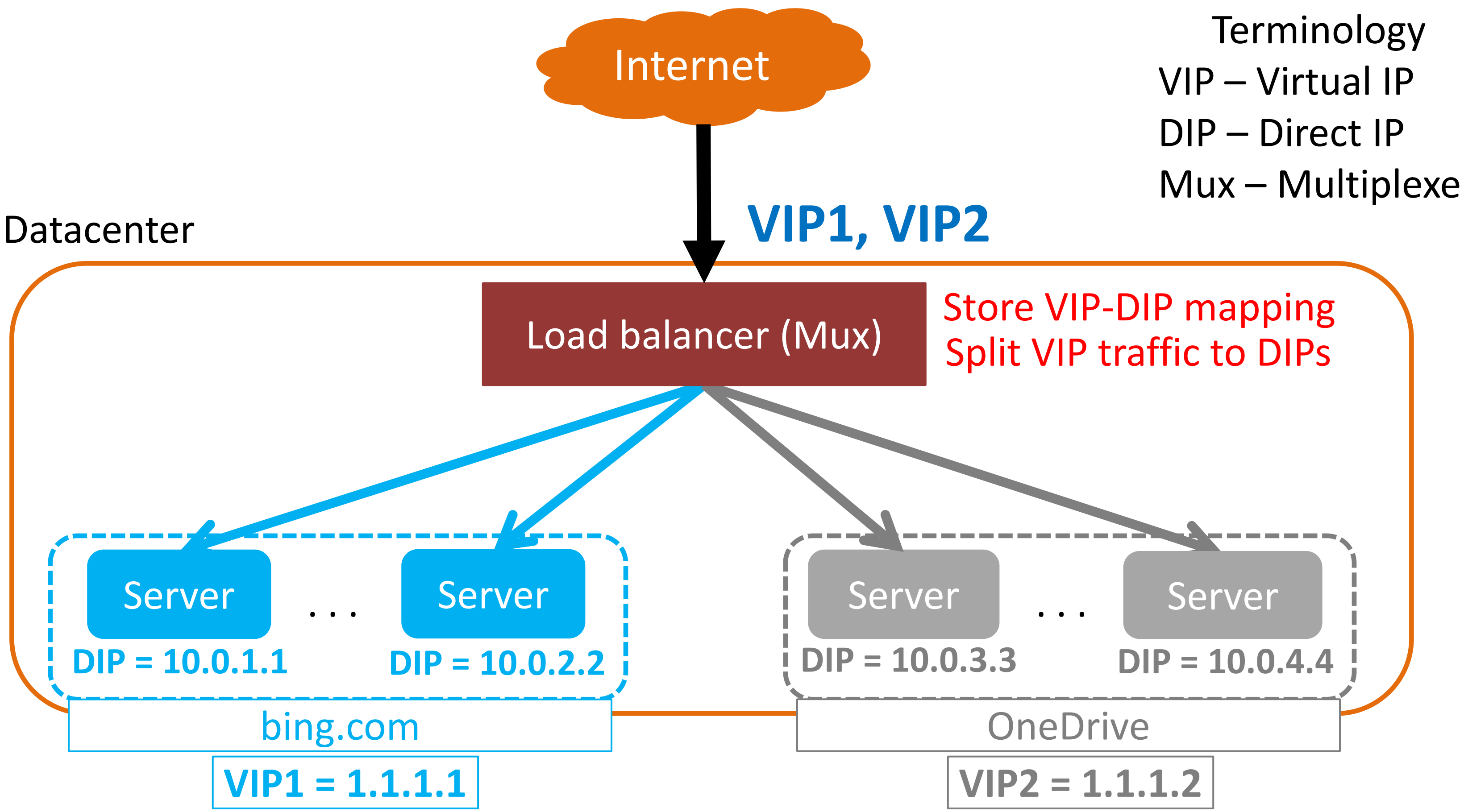
Rohan Gandhi Hongqiang Harry Liu Guohan Lu Jitu Padhye

Y. Charlie Hu Lihua Yuan Ming Zhang

PURDUE UNIVERSITY

Microsoft

Load Balancer is Critical For Online Services



Load balancer provides high availability and scalability

Existing LBs Have Limitations

Specialized Hardware LBs

Too costly

$100+ million for 15 Tbps

Poor robustness

1+1 redundancy

Software LBs

Scale with demand

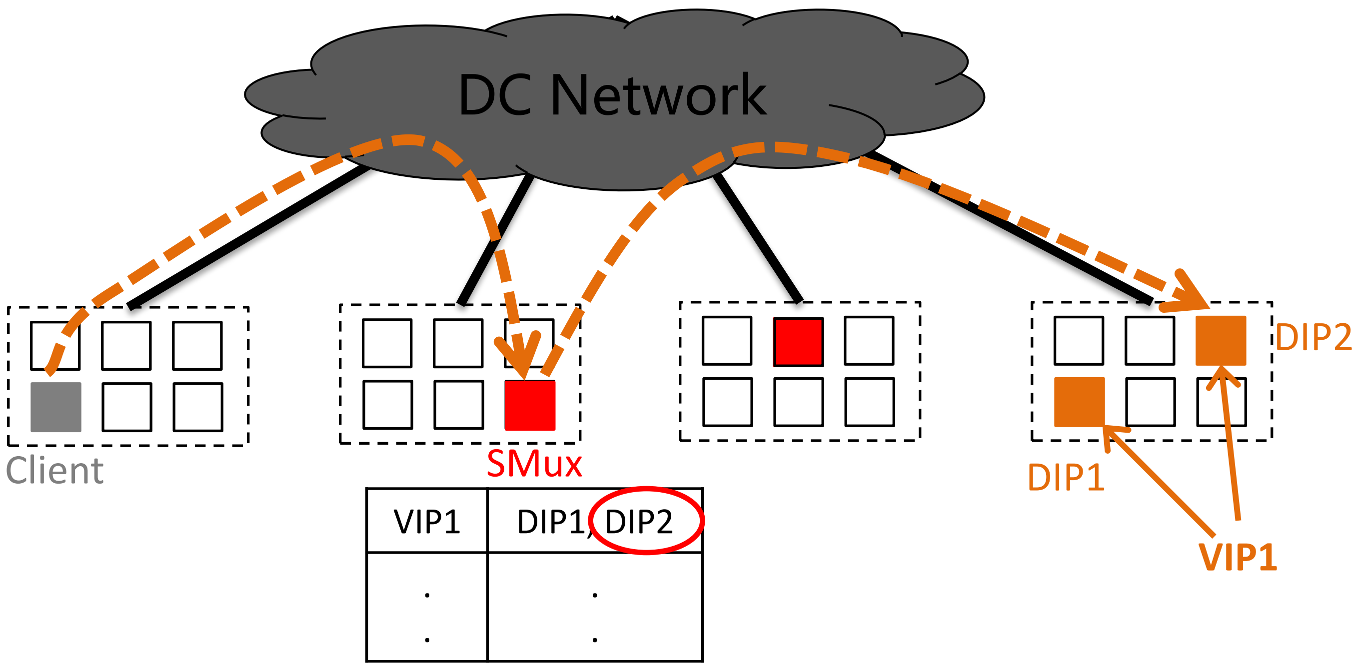
Scale up/down according to VIP traffic

High robustness

n+1 redundancy

Software LBs have cost and performance limitations

Software LB Design

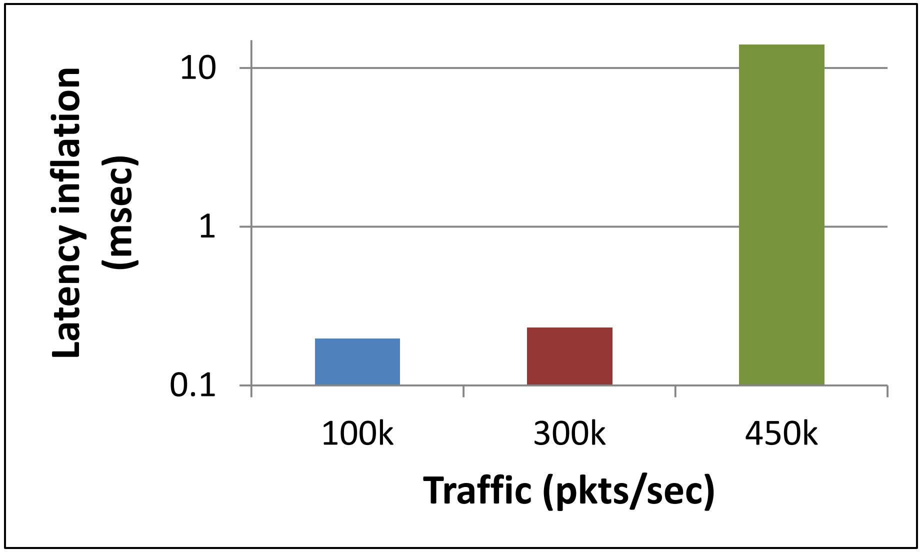


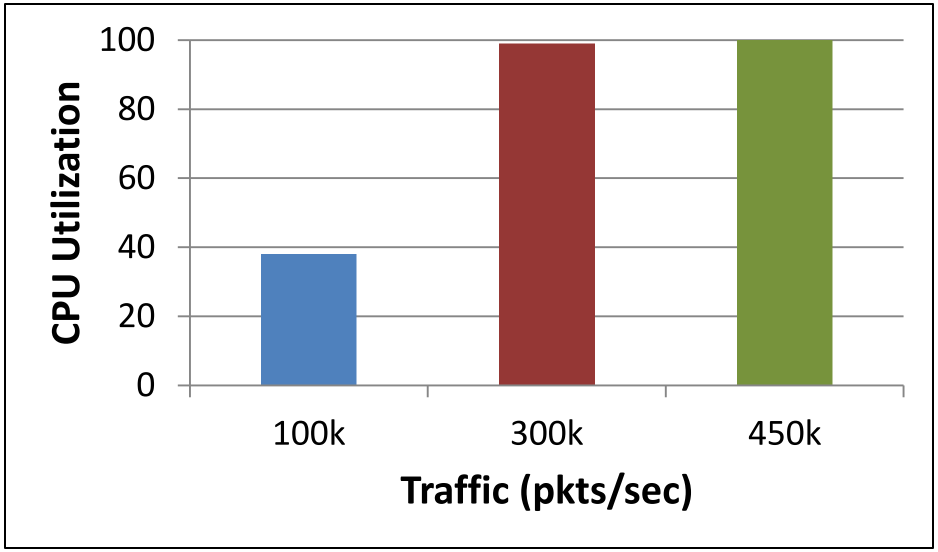
Software LB Benefits:

- High robustness
- Scale with demand

Ananta: Cloud-scale load balancing [SIGCOMM'13]

Software LB has Limitations





High latency inflation: 200 usec

Low capacity: 300k pkts/sec

5k SMuxes needed at 15Tbps traffic in 50k server DC

How can we build high performance, low cost and robust load balancer?

Duet ideas

- Use commodity switches as hardware Muxes
- Use software Muxes as a backstop

Can Switch Act As a Mux

Switches offer:

- High capacity (500+ million pkts/sec)
- Low latency inflation (1 usec)

Mux functionalities

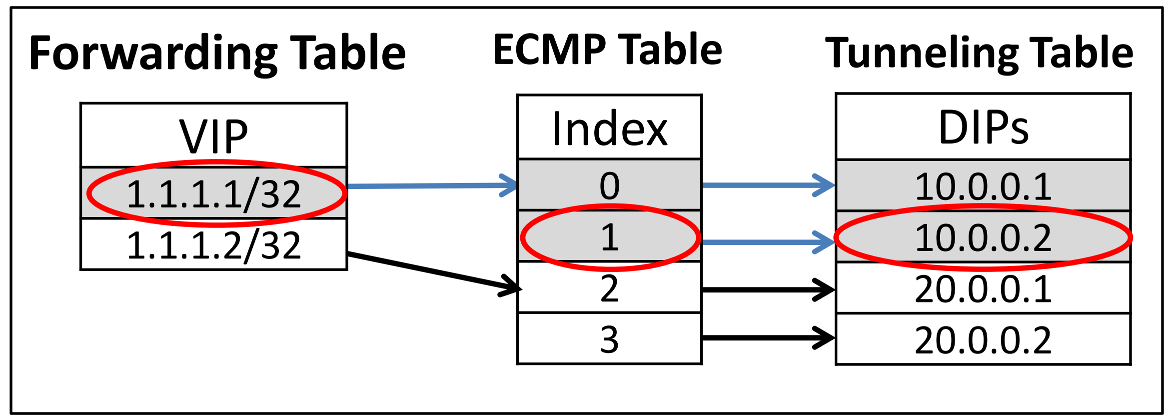
- Split VIP traffic across DIPs
- Forward VIP traffic to DIPs

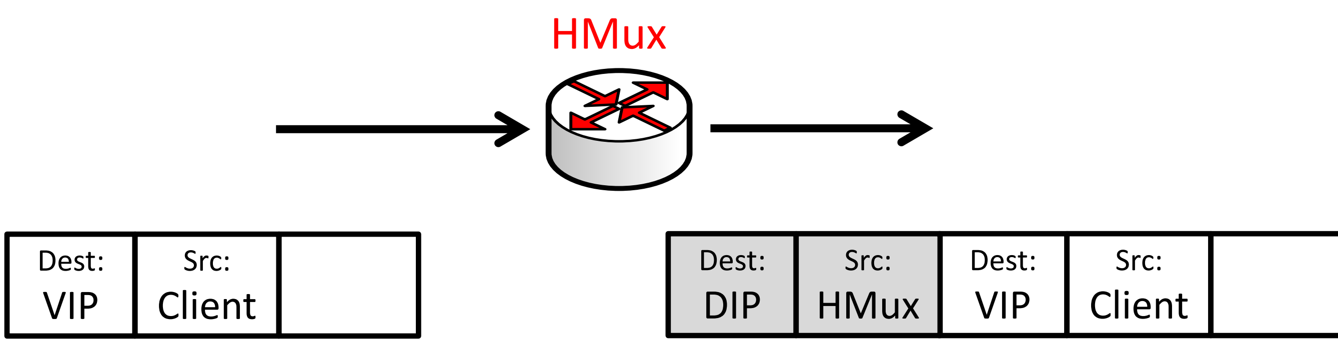
Switch resources

- ECMP
- Tunneling

Implementing HMux on Switch

VIP	DIPs
1.1.1.1/32	10.0.0.1 10.0.0.2
1.1.1.2/32	20.0.0.1 20.0.0.2





Key Design Challenges

- Limited switch memory
- High failure robustness
- VIP assignment
- VIP migration

Challenge 1: Switches have Limited Memory

Workload: 100k+ VIPs and 1+ millions DIPs

Single HMux cannot store all VIPs and DIPs

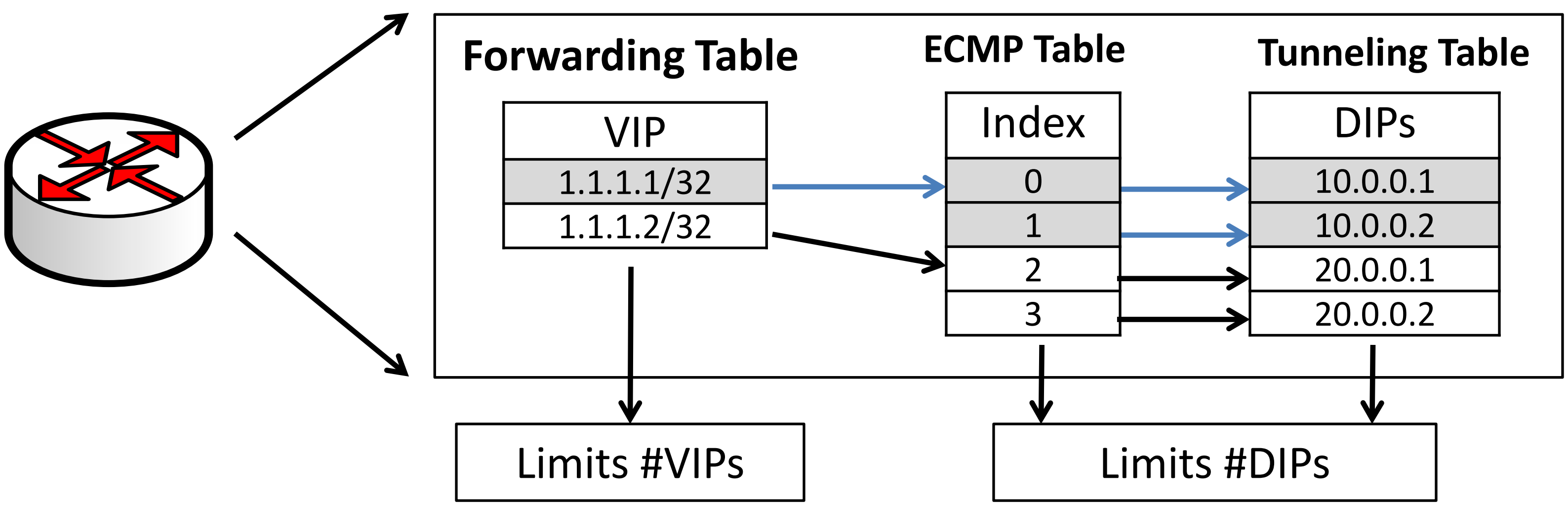


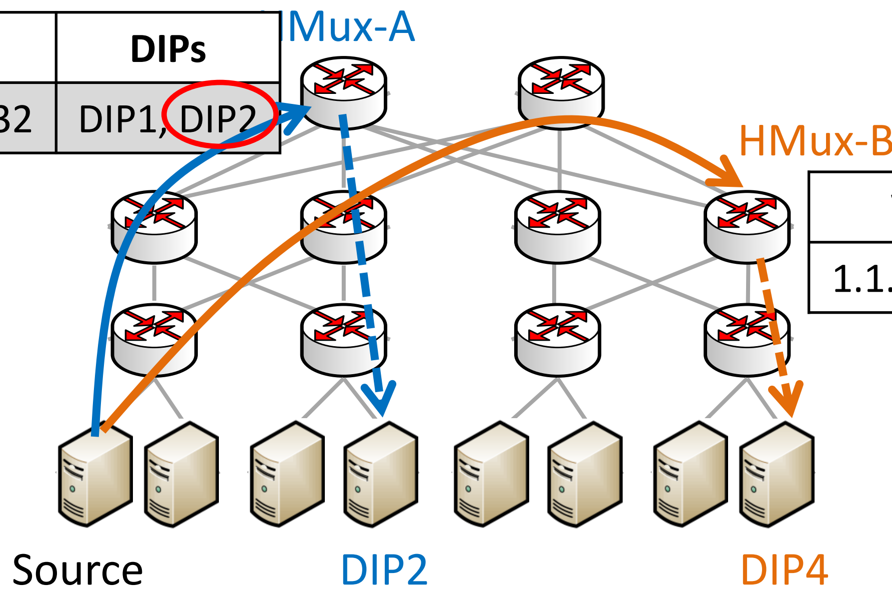
Table	Forwarding	ECMP	Tunneling
Max. size	16k	4k	512

Max VIPs

Max DIPs

Solution: Partitioning VIPs across HMuxes

VIP	DIPs
1.1.1.1/32	DIP1, DIP2



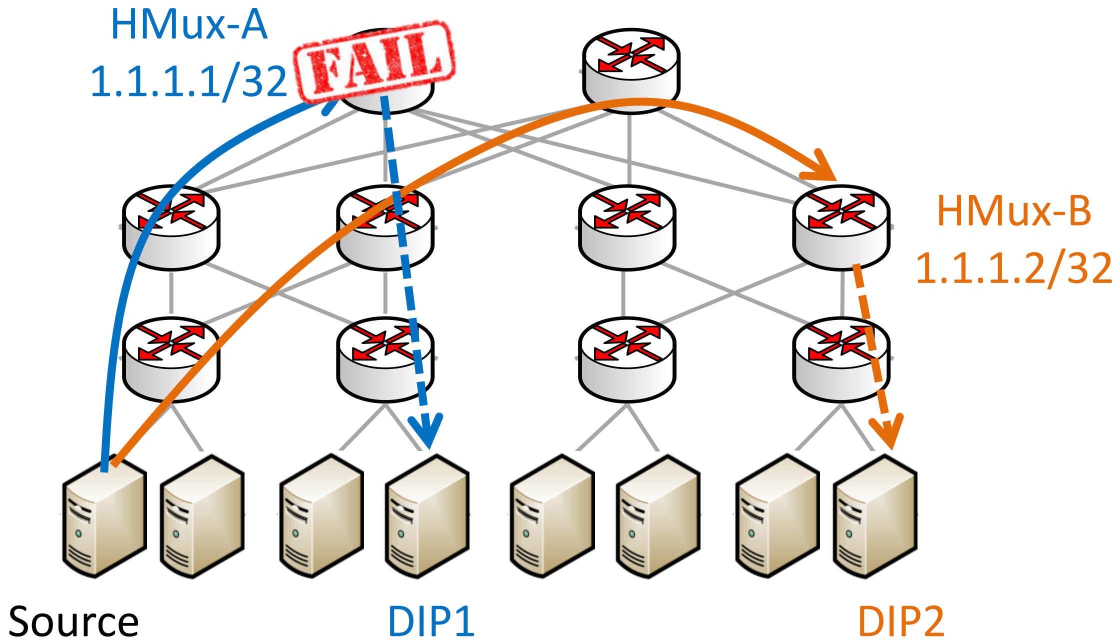
VIP	DIPs
1.1.1.2/32	DIP3, DIP4

Capacity	VIPs	DIPs
Single HMux	16k	512
All HMuxes	16k	512 * 2k = 1M

Fixed

Scales with #DIPs

Challenge 2: High Robustness



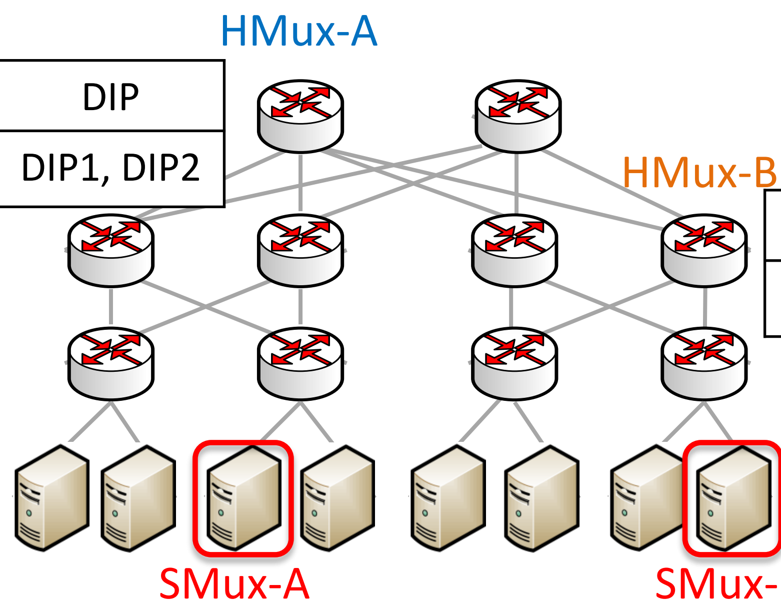
- Availability during failure?
- Large number of VIPs?

Idea: Integrate SMux with HMux

	HMuxes	SMuxes	Duet
Low latency	✓	✗	✓
High capacity	✓	✗	✓
High availability	✗	✓	✓
Scale to large #VIPs	✗	✓	✓

?

Solution: Use SMuxes As a Backstop



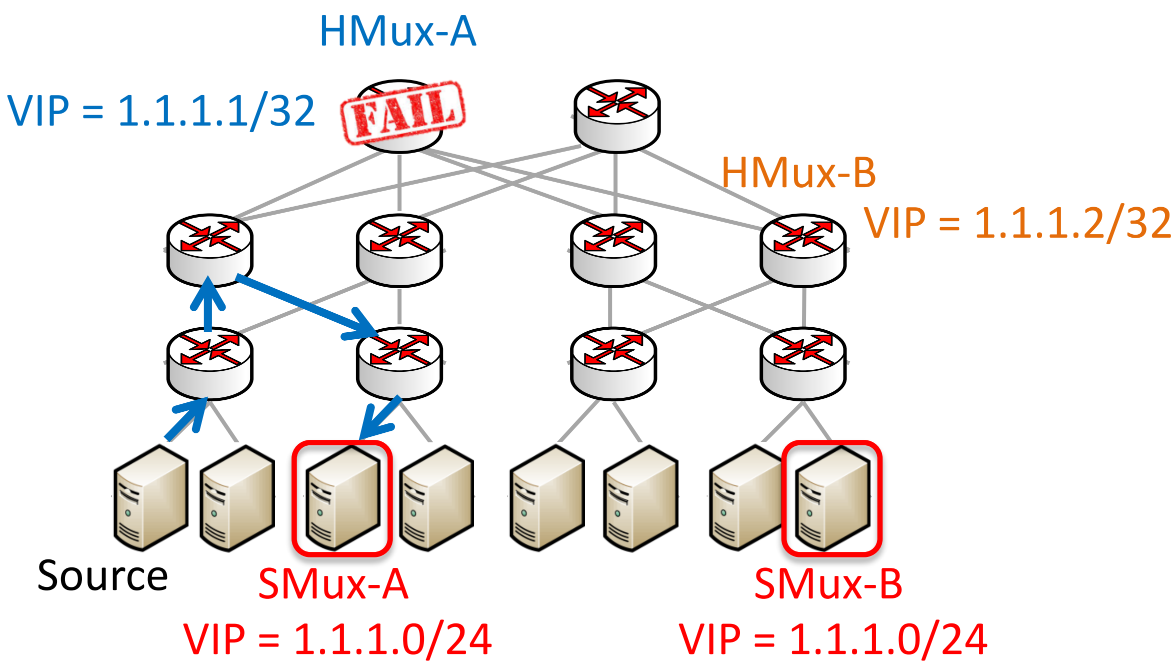
VIP	DIP
1.1.1.1/32	DIP1, DIP2

VIP	DIP
1.1.1.2/32	DIP3, DIP4

VIP	DIP
1.1.1.1/32	DIP1, DIP2
1.1.1.2/32	DIP3, DIP4

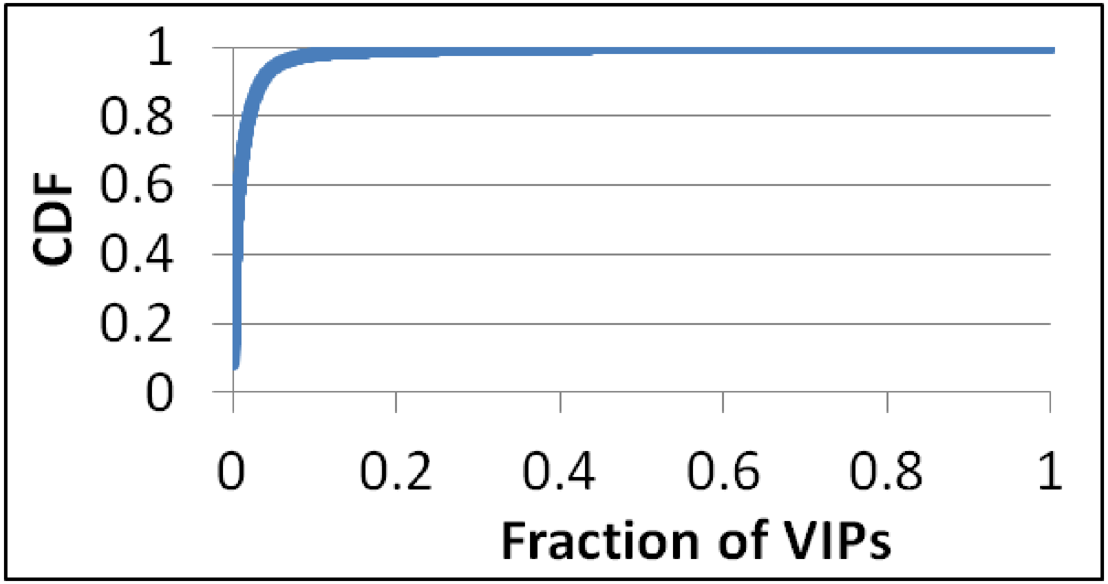
VIP	DIP
1.1.1.1/32	DIP1, DIP2
1.1.1.2/32	DIP3, DIP4

Solution: Use SMuxes As a Backstop



- High availability during failure
- Scale to large #VIPs

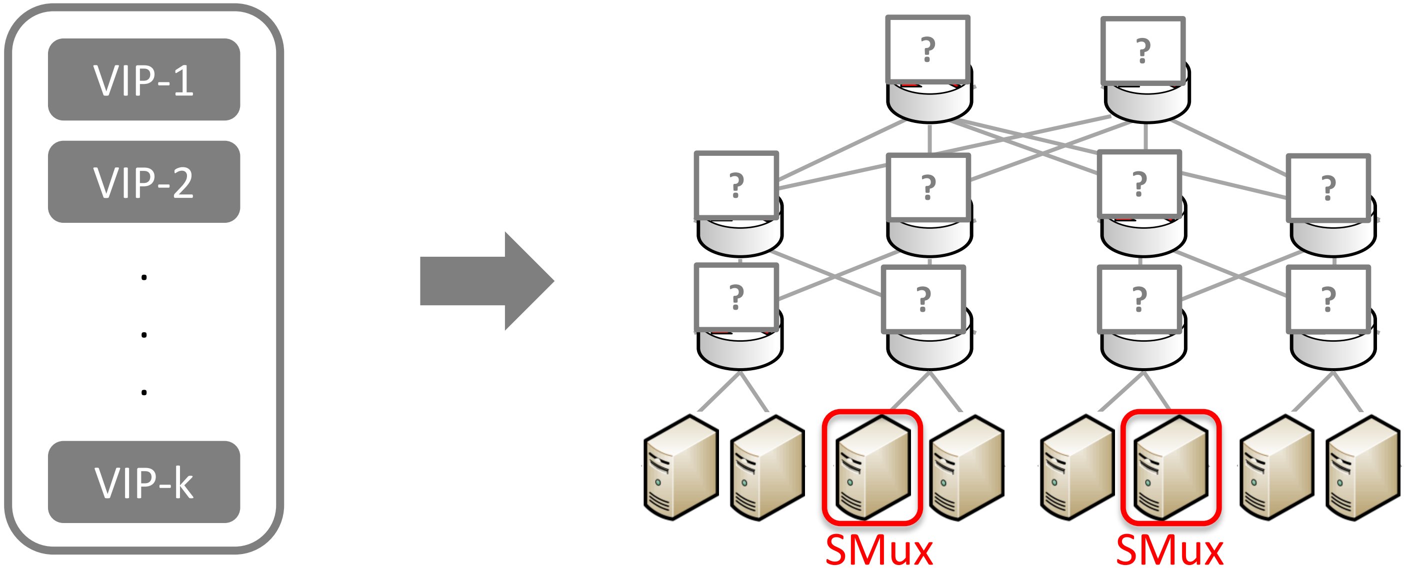
VIP Traffic Distribution is Highly Skewed



Top 10% VIPs carry 99% traffic

Duet handles 86-99.9% traffic using HMuxes

Challenge 3: How to Assign VIPs?

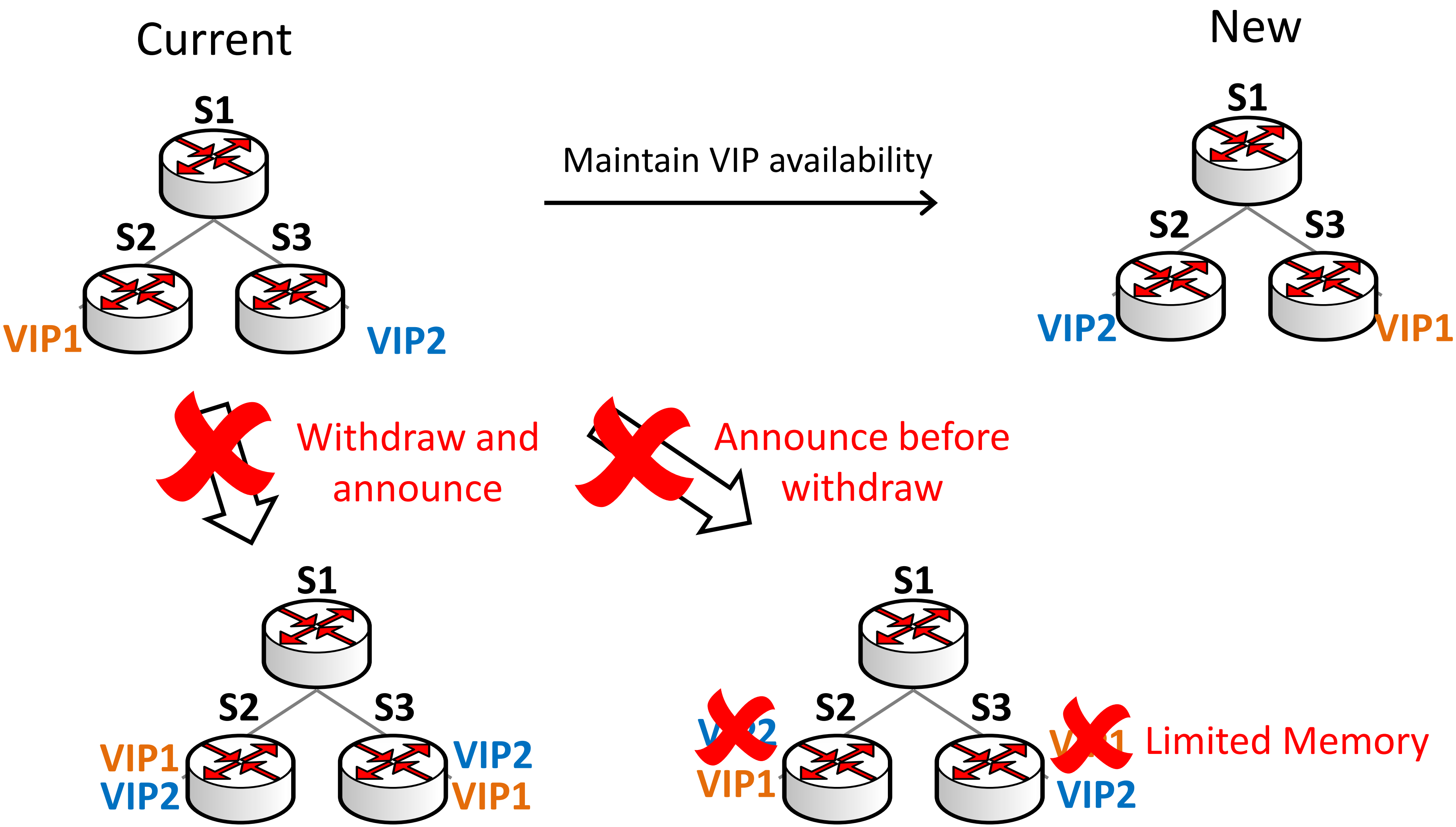


Objective: Maximize traffic handled by HMuxes

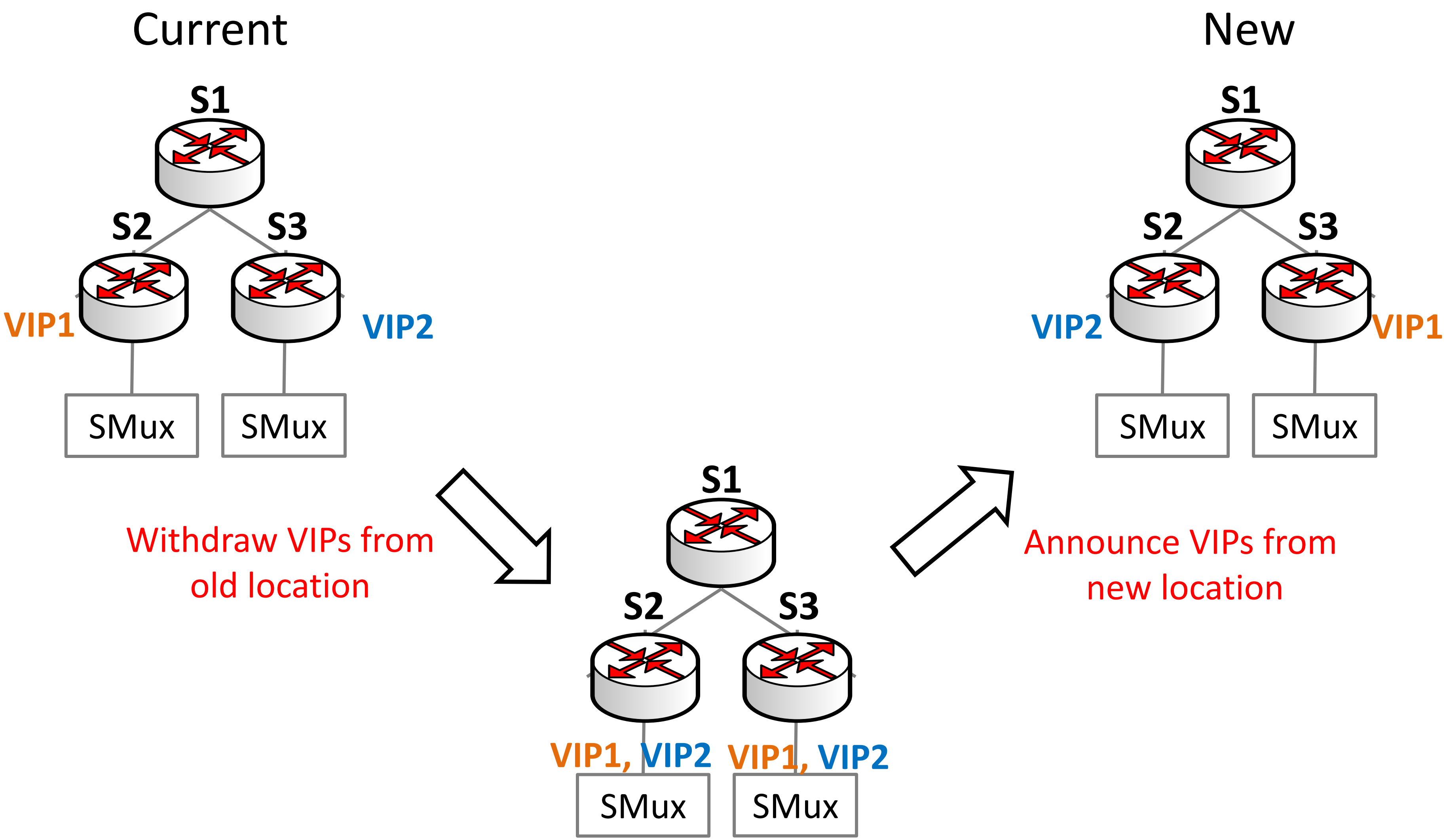
Input:
VIP traffic, DIP locations
Topology

Constraints:
Switch memory
Link capacity

Challenge 4: How to Migrate VIPs?



Solution: Migrate VIPs through SMuxes



Fast and maintains VIP availability

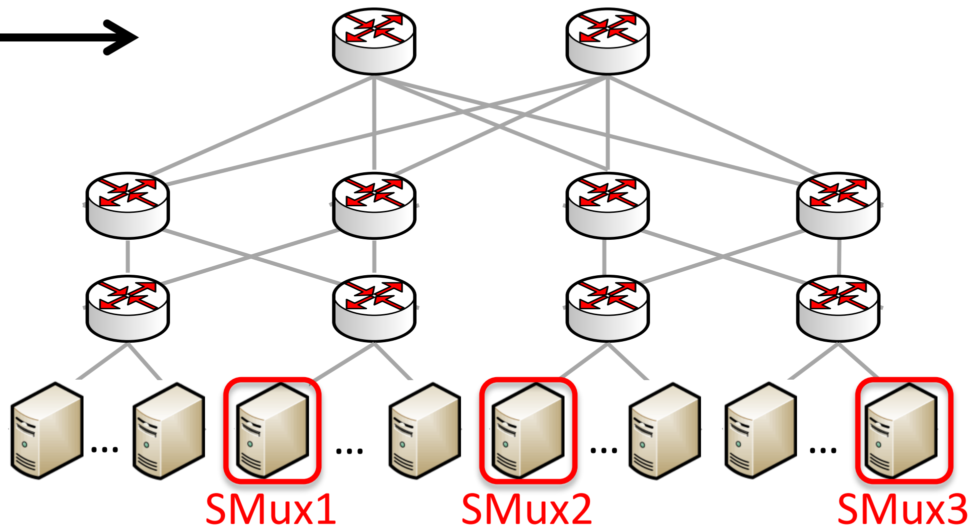
Duet Extensions

- ☑ SNAT
- ☑ Support VIPs with 512+ DIPs
- ☑ Port based load balancing
- ☑ Load balancing in virtualized networks

Experimental Setup

Testbed

- 10 switches, 3 SMuxes
- 10 VIPs, 34 DIPs

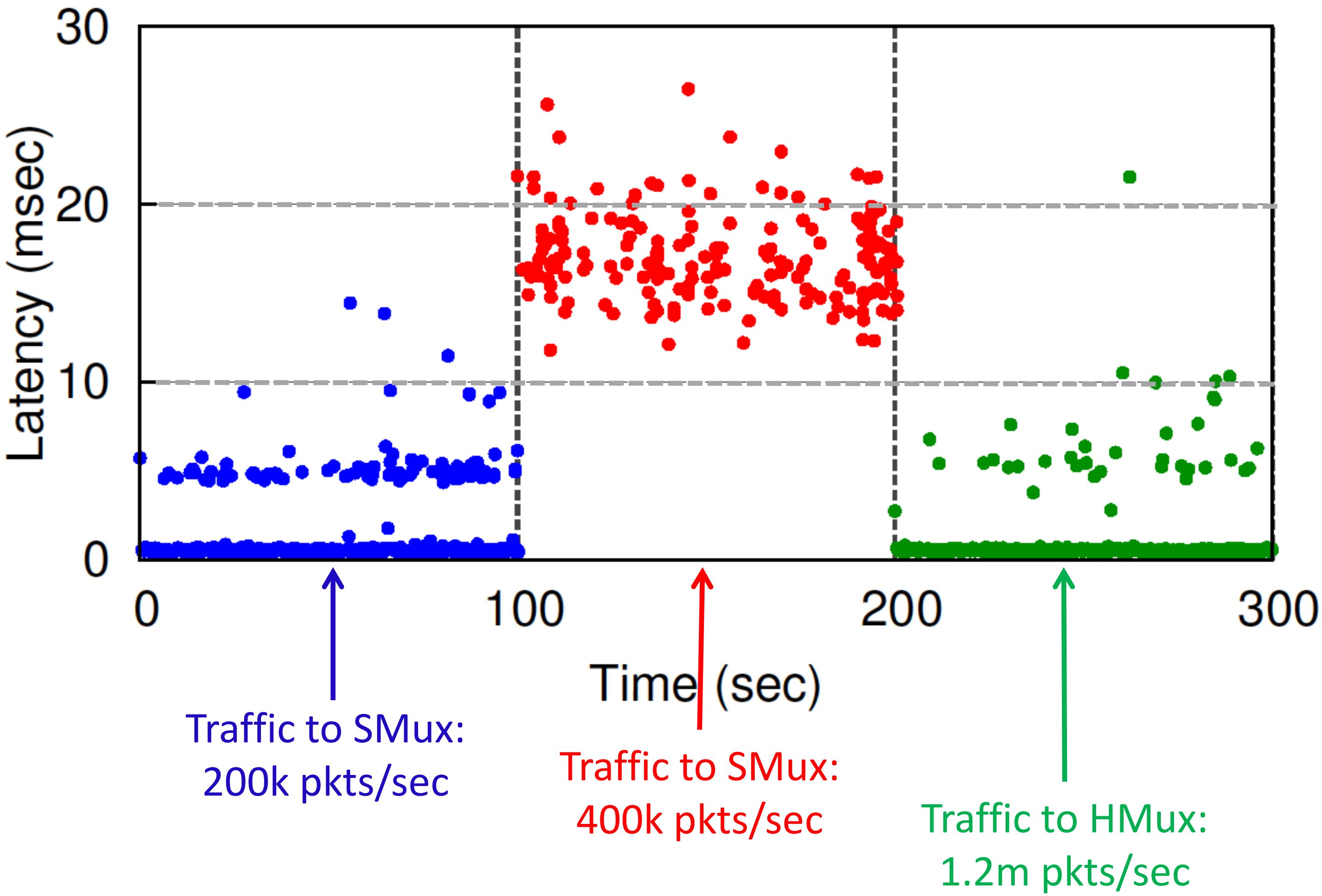


Simulation

- Topology and traffic trace from Azure DC

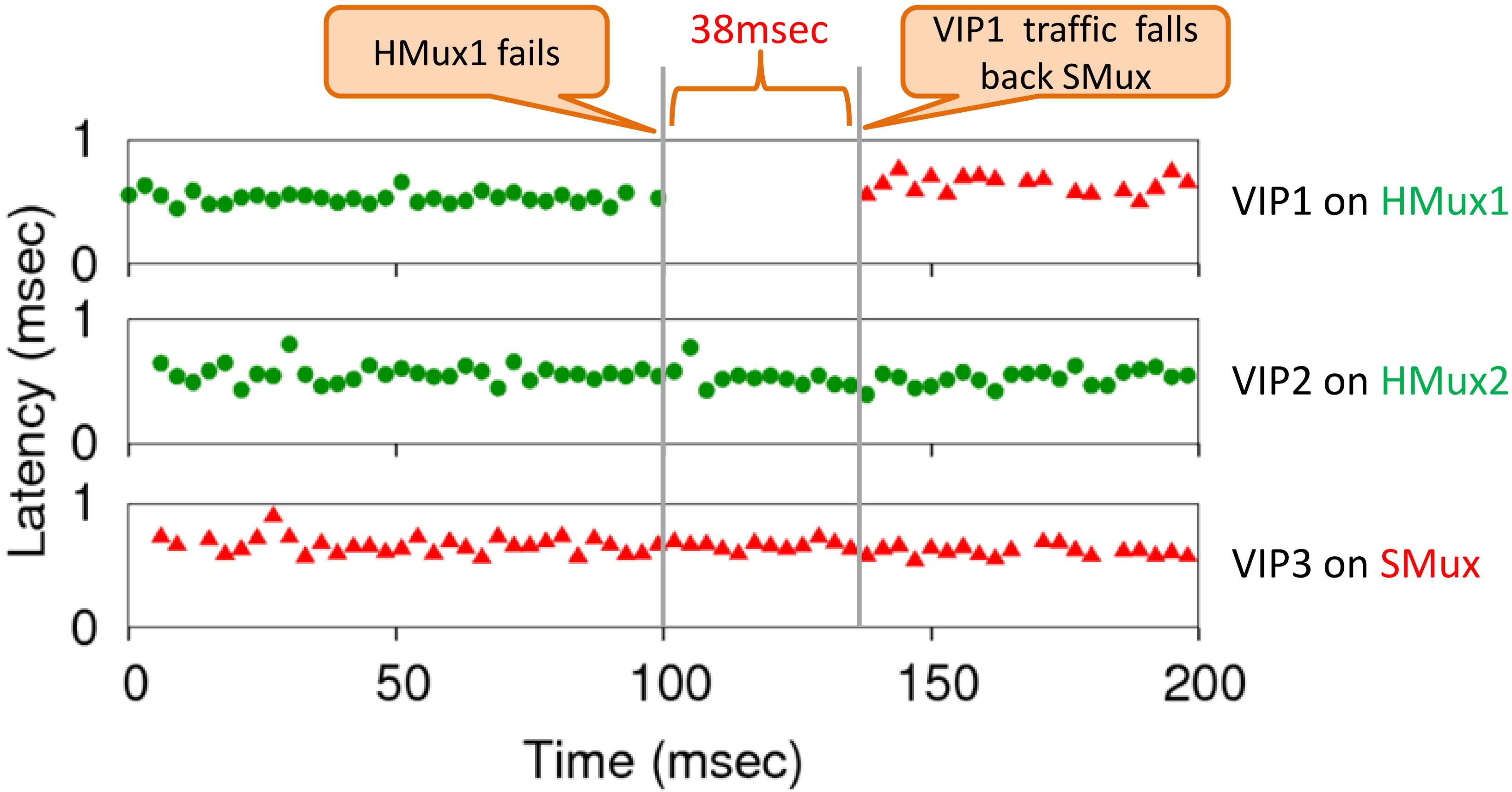
- High capacity
- High availability
- Low cost

Duet Provides High Capacity

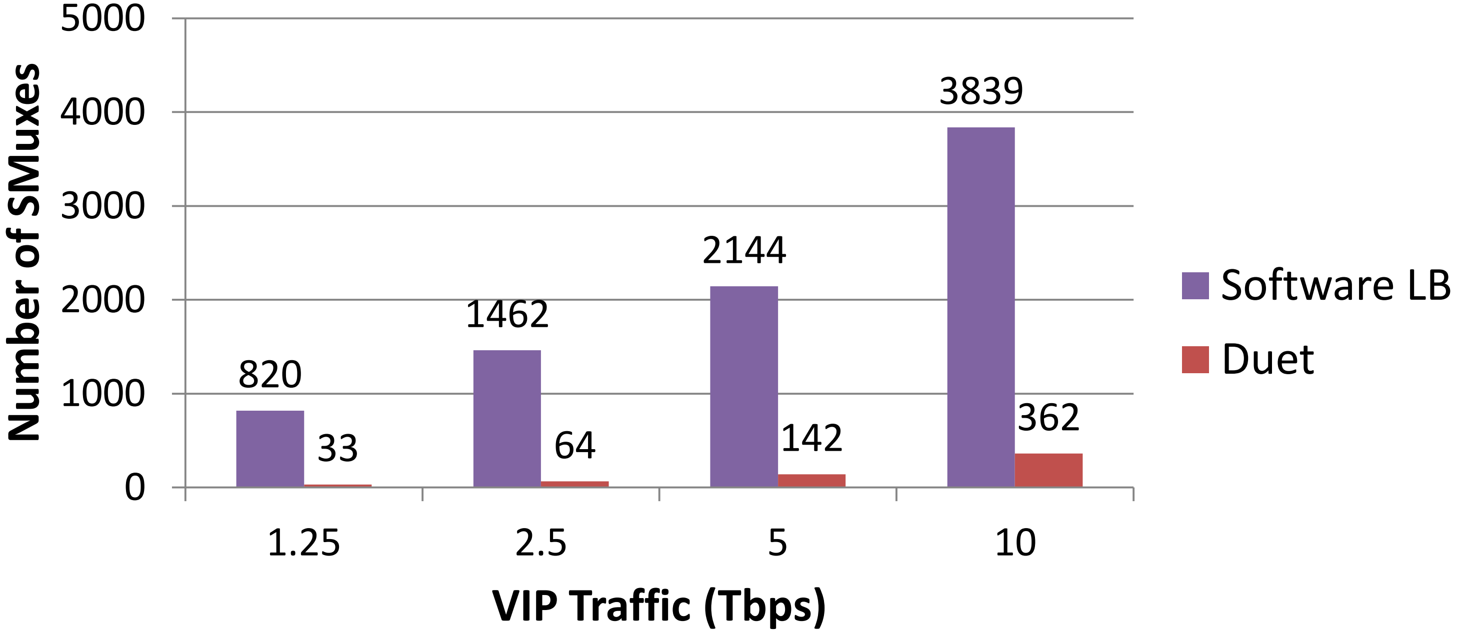


HMux has larger capacity and lower latency than SMux

Duet Provides High Availability



Duet Reduces Cost



Duet reduces cost by 10-24x

Summary

- Specialized and software LBs have cost and performance problems
- Duet key ideas:
  - Use commodity switches as HMuxes
  - Use small number of SMuxes as backstop
- Benefits:
  - Low latency
  - High capacity
  - High robustness
  - Low cost