PRINCETON
UNIVERSITY

HULA: Scalable Load Balancing Using Programmable Data Planes

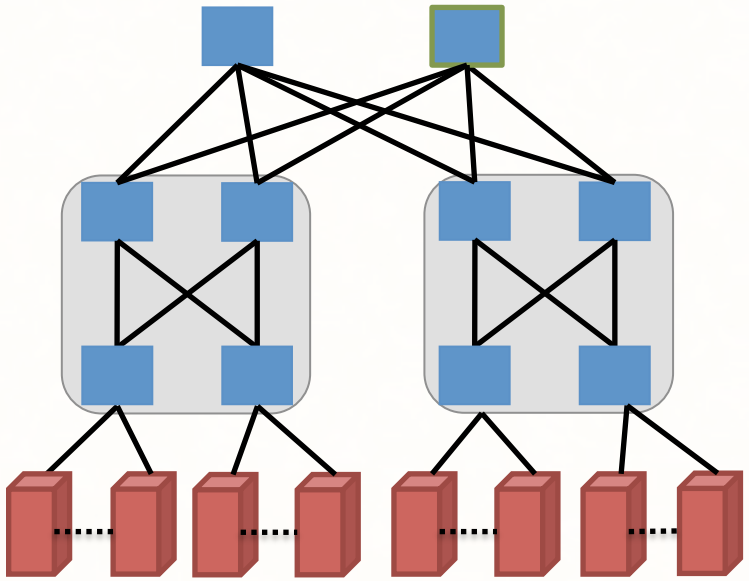
Naga Katta[1]

Mukesh Hira[2], Changhoon Kim[3], Anirudh Sivaraman[4], Jennifer Rexford[1]

1.Princeton 2.VMware 3.Barefoot Networks 4.MIT

Datacenter Load Balancing

- Multiple network paths
- High bisection bandwidth
- Volatile traffic
- Multiple tenants



A Good Load Balancer

- Multiple network paths
  - Track path performance
  - Choose best path
- High bisection bandwidth
- Volatile traffic
- Multiple tenants

A Good Load Balancer

- Multiple network paths
  - Track path performance
  - Choose best path
- High bisection bandwidth
  - Fine grained load balancing
- Volatile traffic
- Multiple tenants

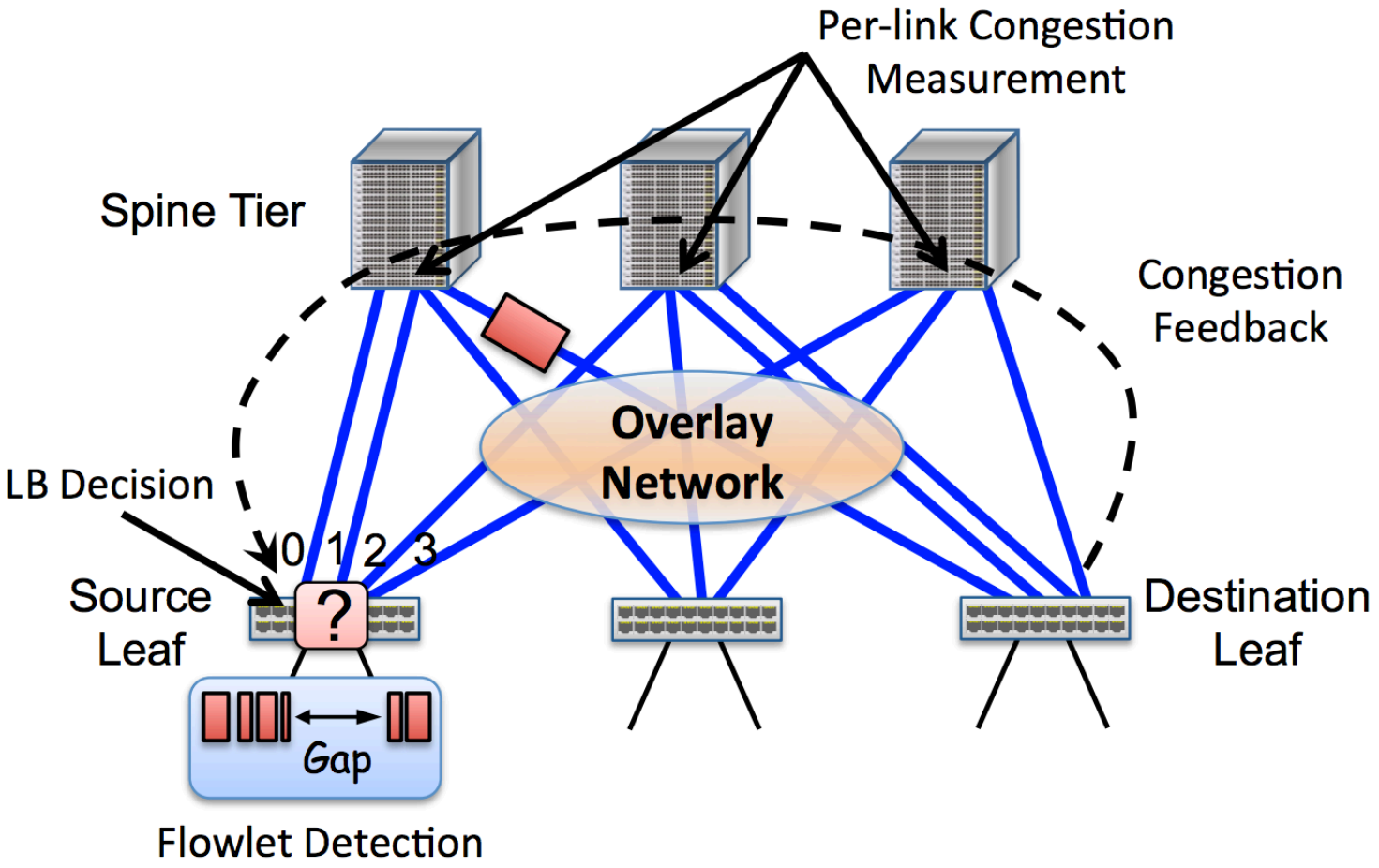
A Good Load Balancer

- Multiple network paths
  - Track path performance
  - Choose best path
- High bisection bandwidth
  - Fine grained load balancing
- Volatile traffic
  - In-dataplane
- Multiple tenants

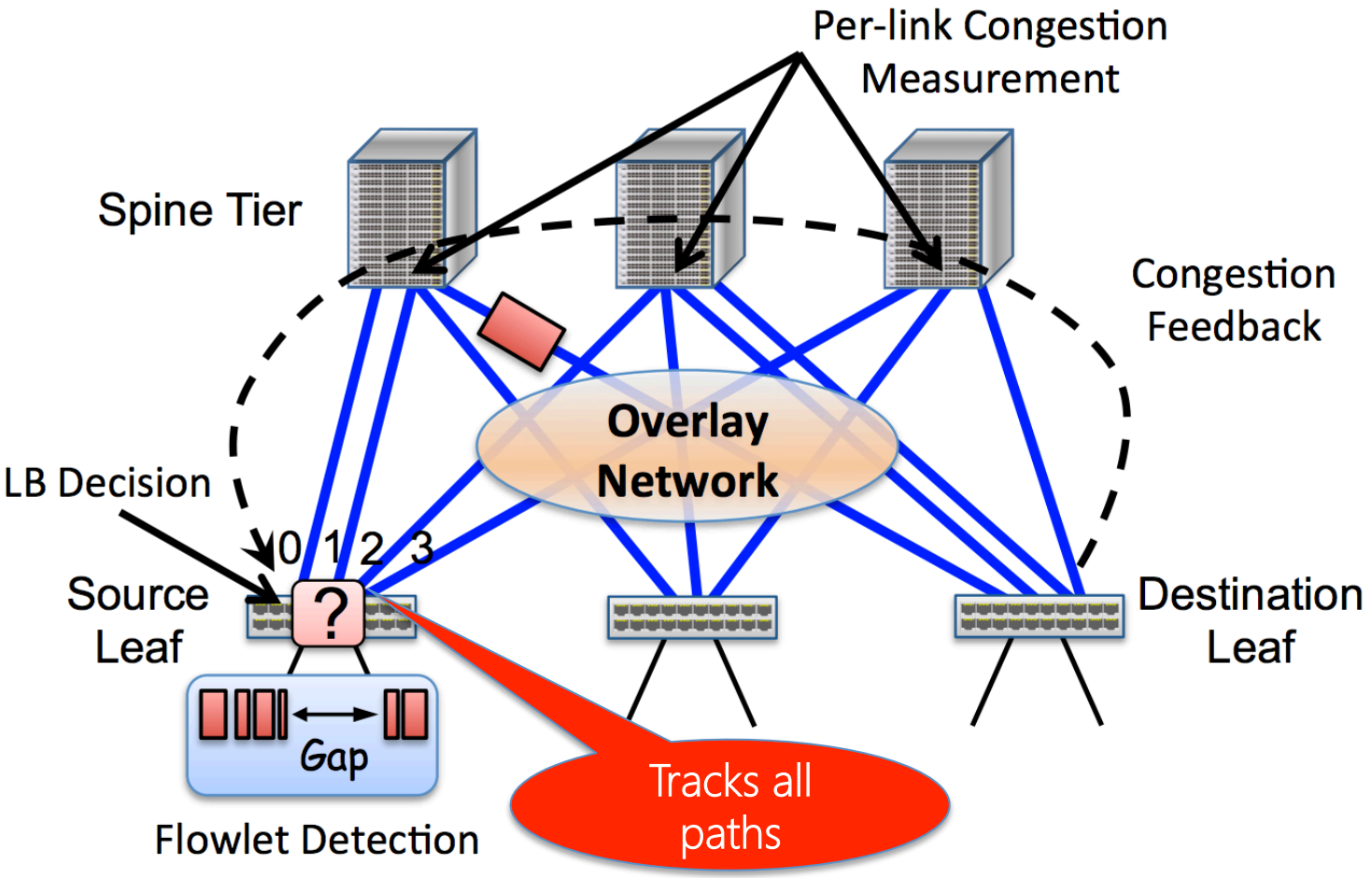
A Good Load Balancer

- Multiple network paths
  - Track path performance
  - Choose best path
- High bisection bandwidth
  - Fine grained load balancing
- Volatile traffic
  - In-dataplane
- Multiple tenants
  - In-network

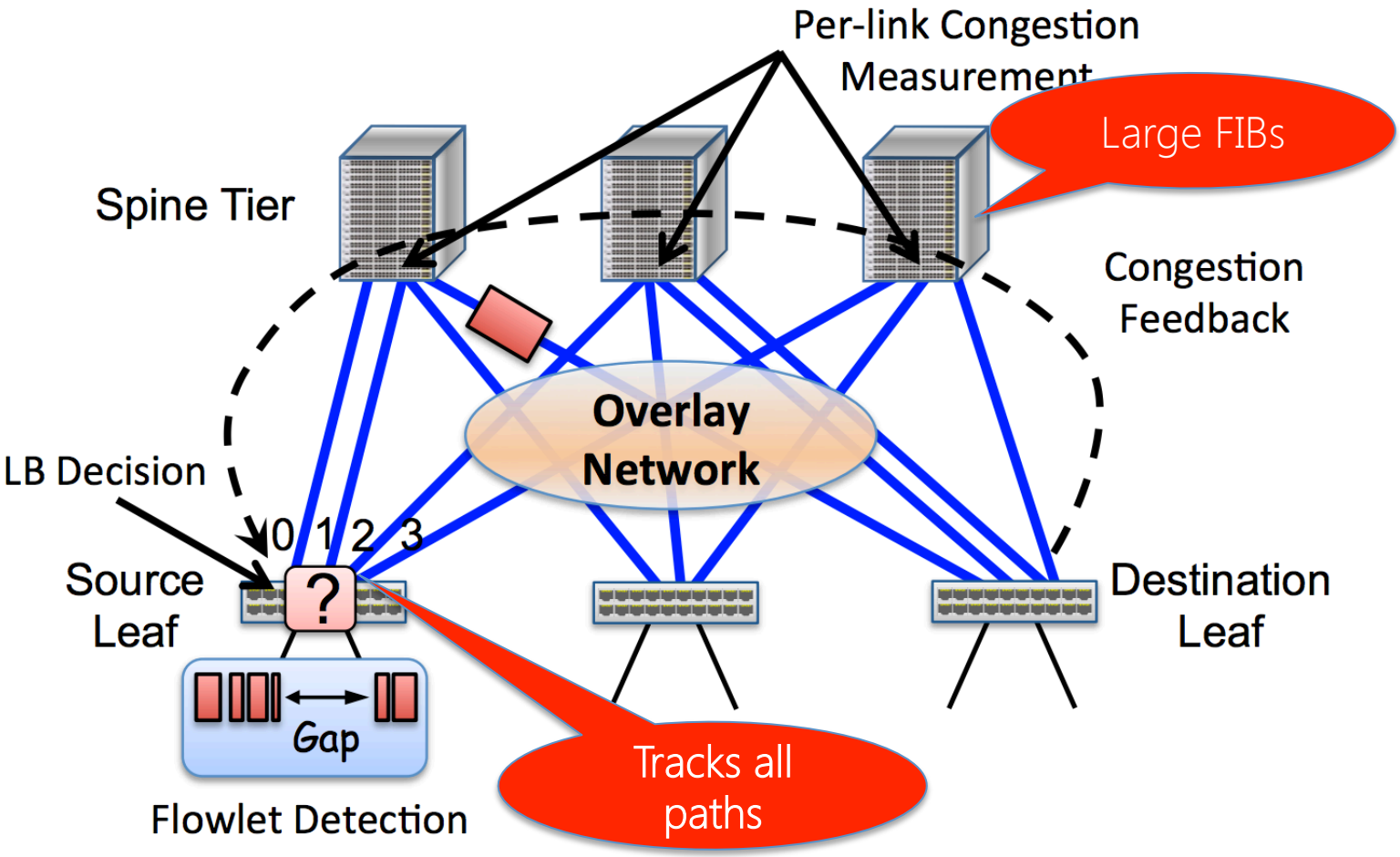
CONGA (SIGCOMM'14)



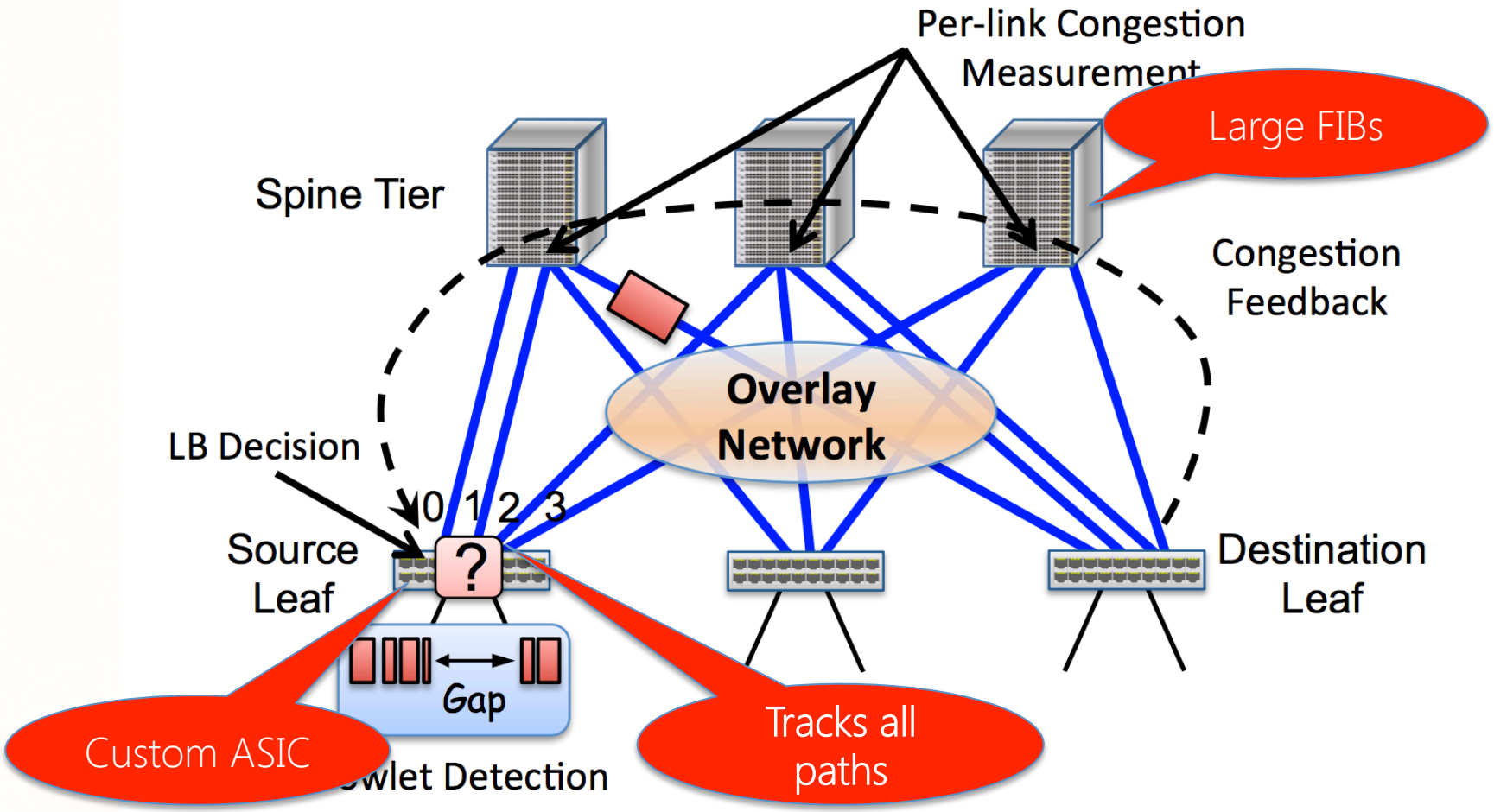
Datapath LB: Challenges



Datapath LB: Challenges



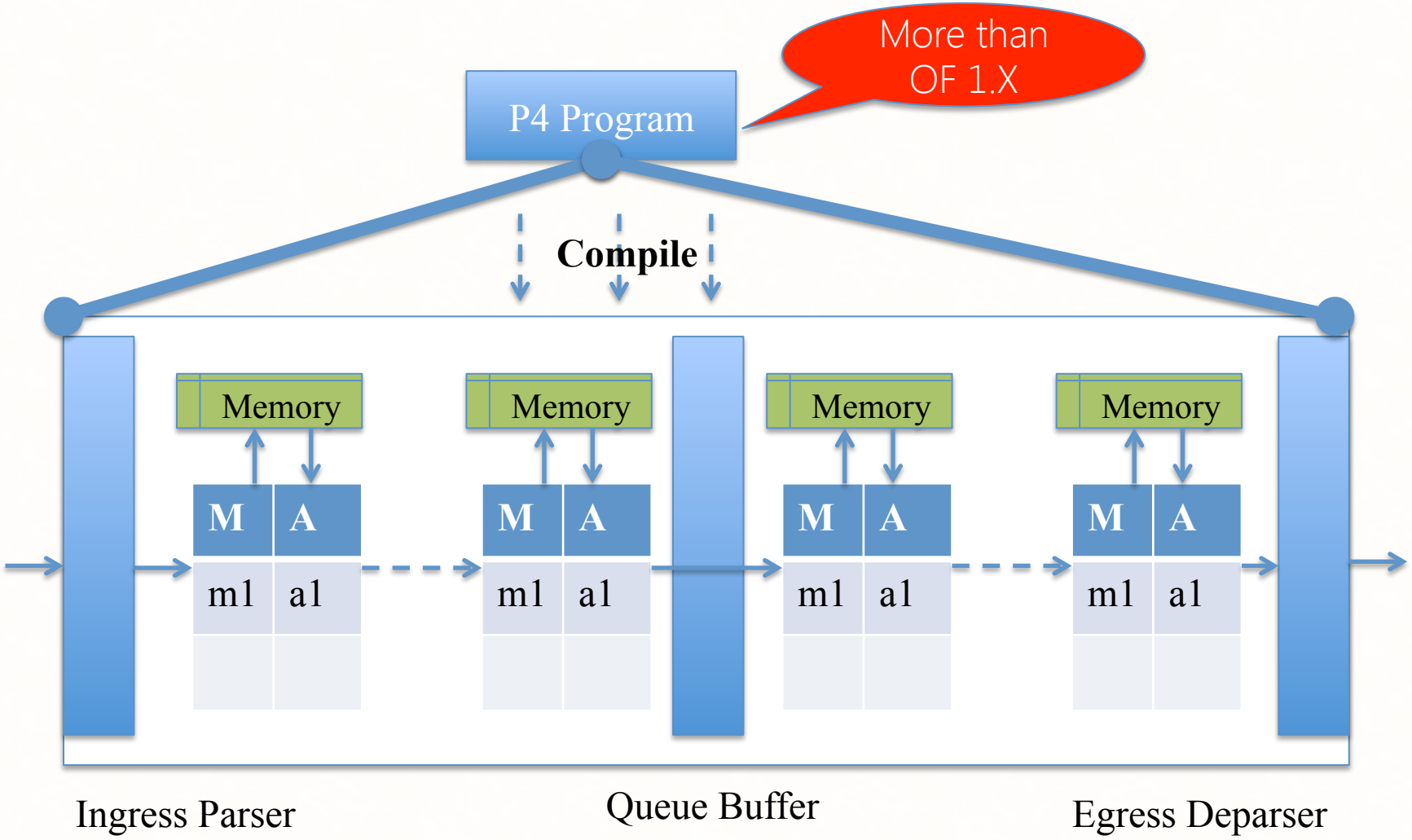
Datapath LB: Challenges



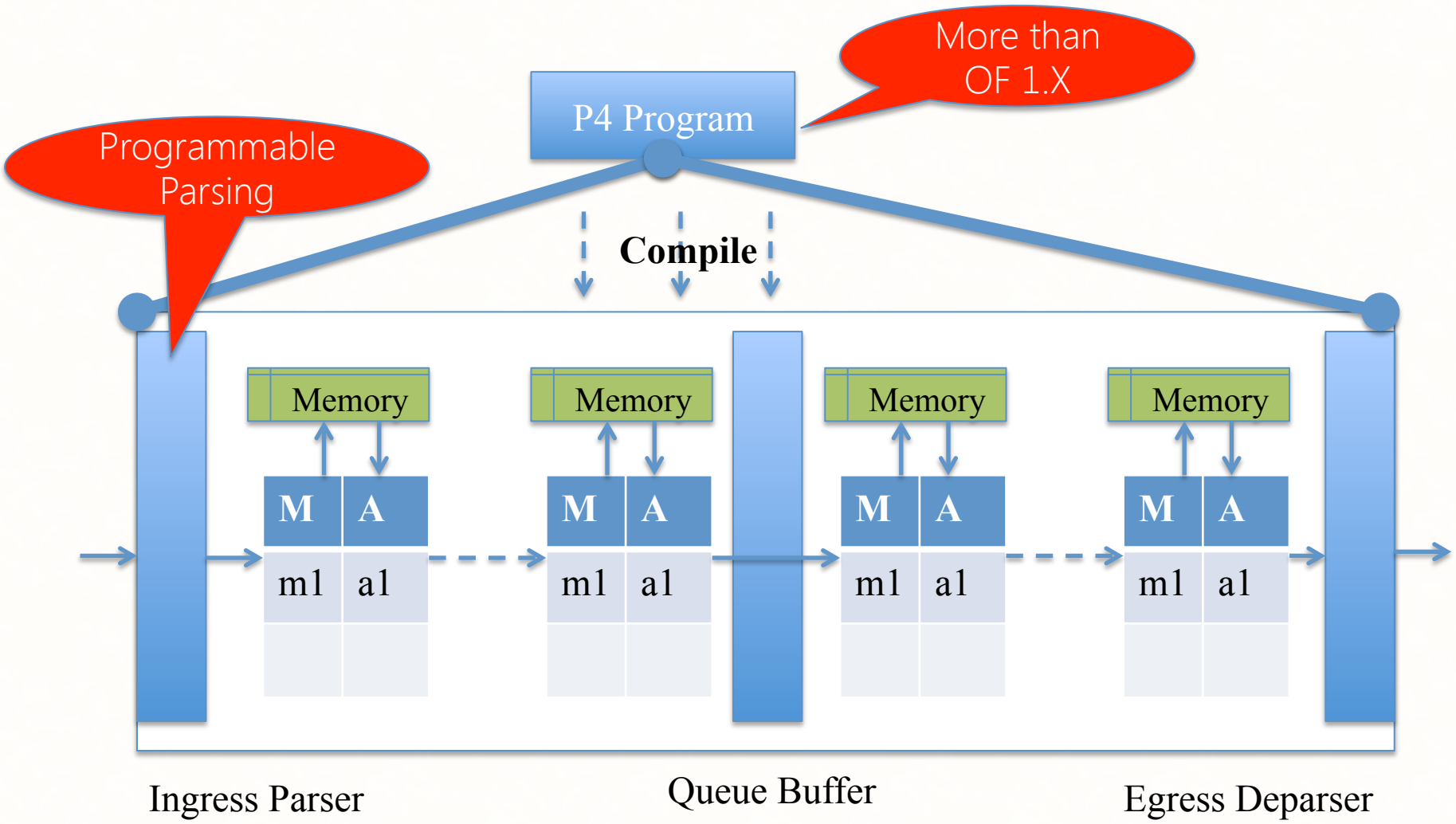
Programmable Commodity Switches

- Vendor agnostic
  - Uniform programming interface (P4)
  - Today's trend -> cheaper
- Reconfigurable in the field
  - Adapt or add dataplane functionality
- Examples
  - RMT, Intel Flexpipe, Cavium Xpliant, etc.

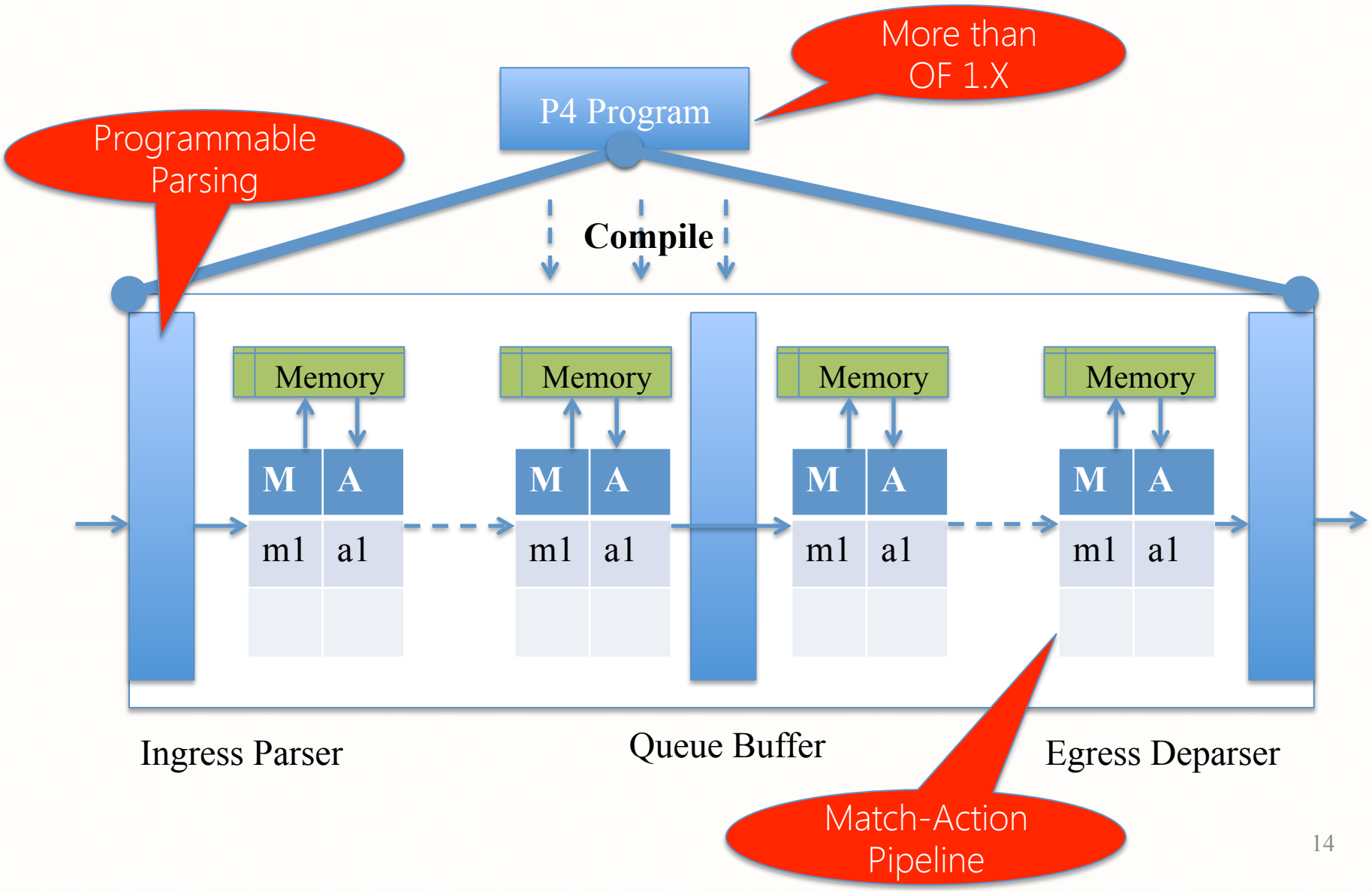
Programmable Switches - Capabilities



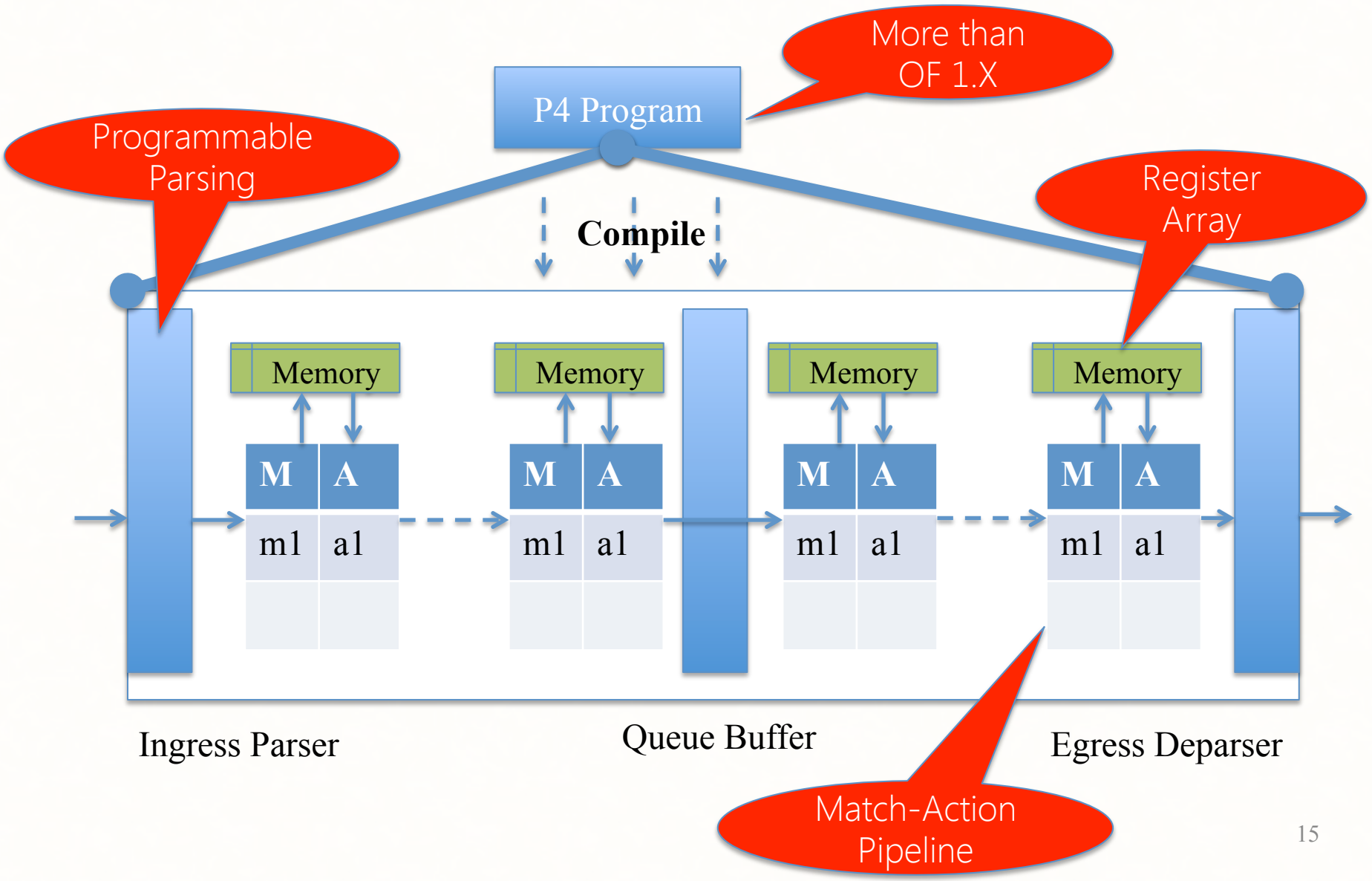
Programmable Switches - Capabilities



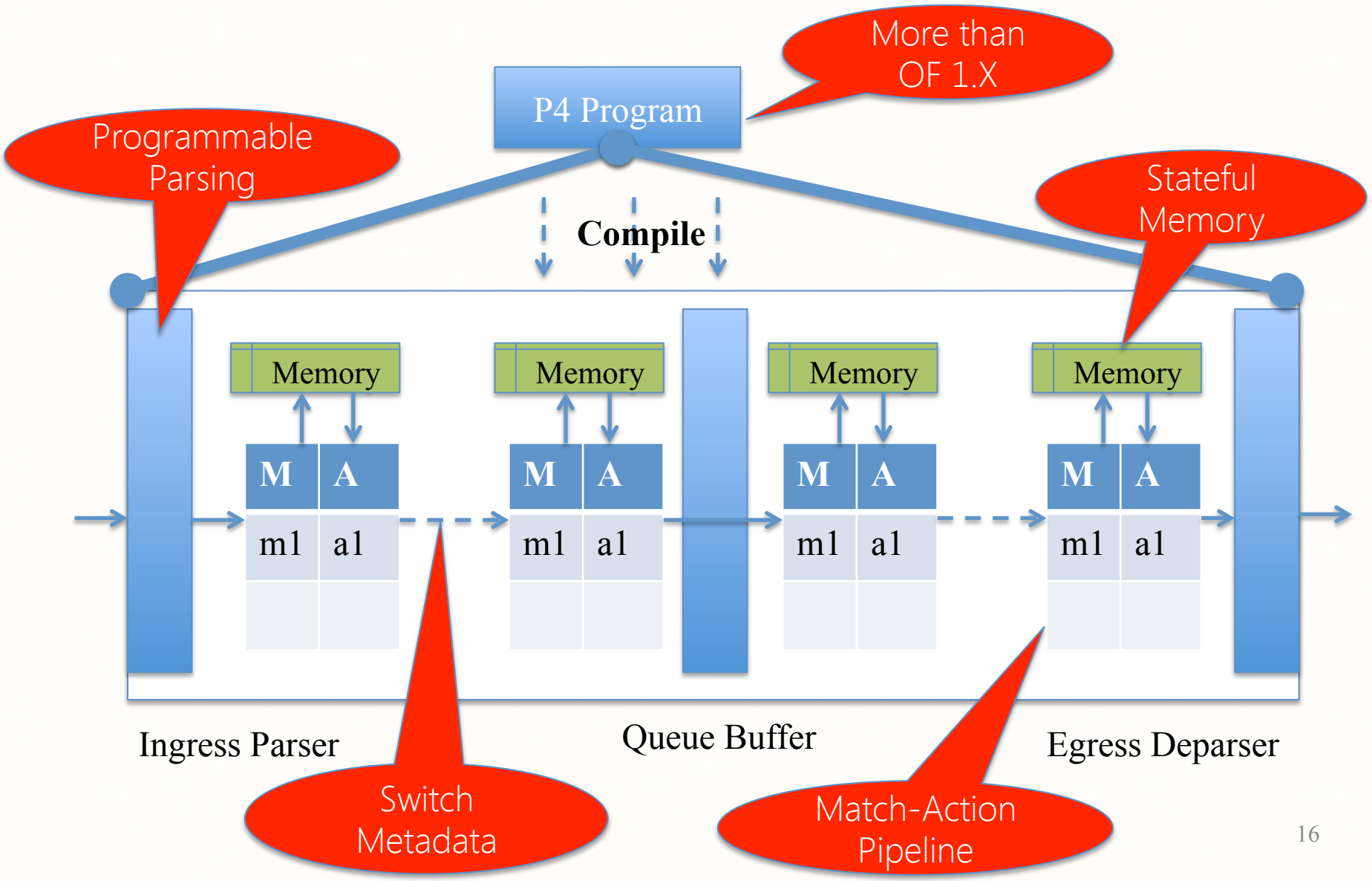
Programmable Switches - Capabilities



Programmable Switches - Capabilities



Programmable Switches - Capabilities



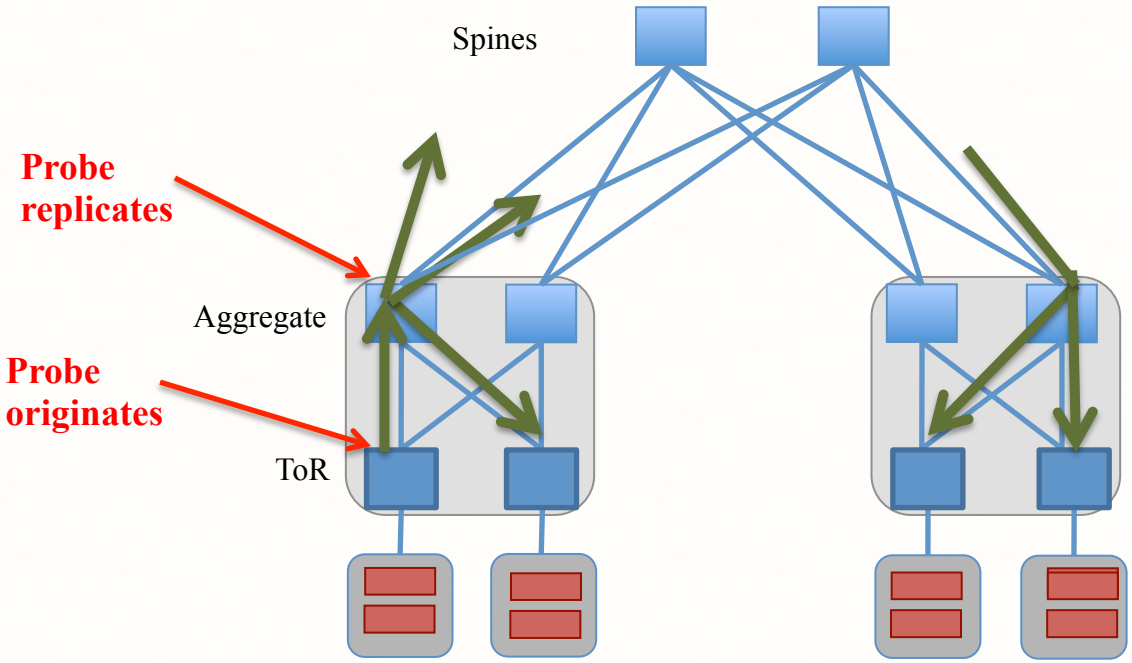
Hop-by-hop Utilization-aware Load-balancing Architecture

- Distance-vector like propagation
  - Periodic probes carry path utilization
- Each switch chooses best downstream path
  - Maintains only best next hop
  - Scales to large topologies
- Programmable at line rate
  - Written in P4.

HULA: Scalable and Programmable

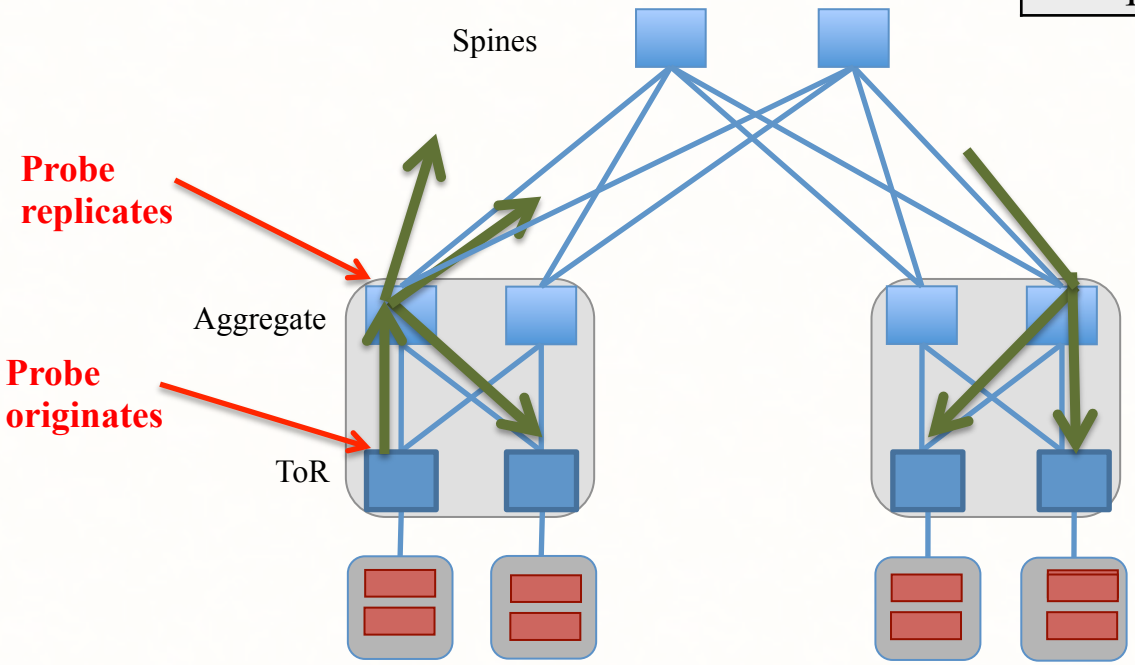
Objective	P4 feature
Probe propagation	**Programmable parsing**
Monitor path performance	**Link state metadata**
Choose best path, route flowlets	**Stateful memory and comparison operators**

Probes carry path utilization

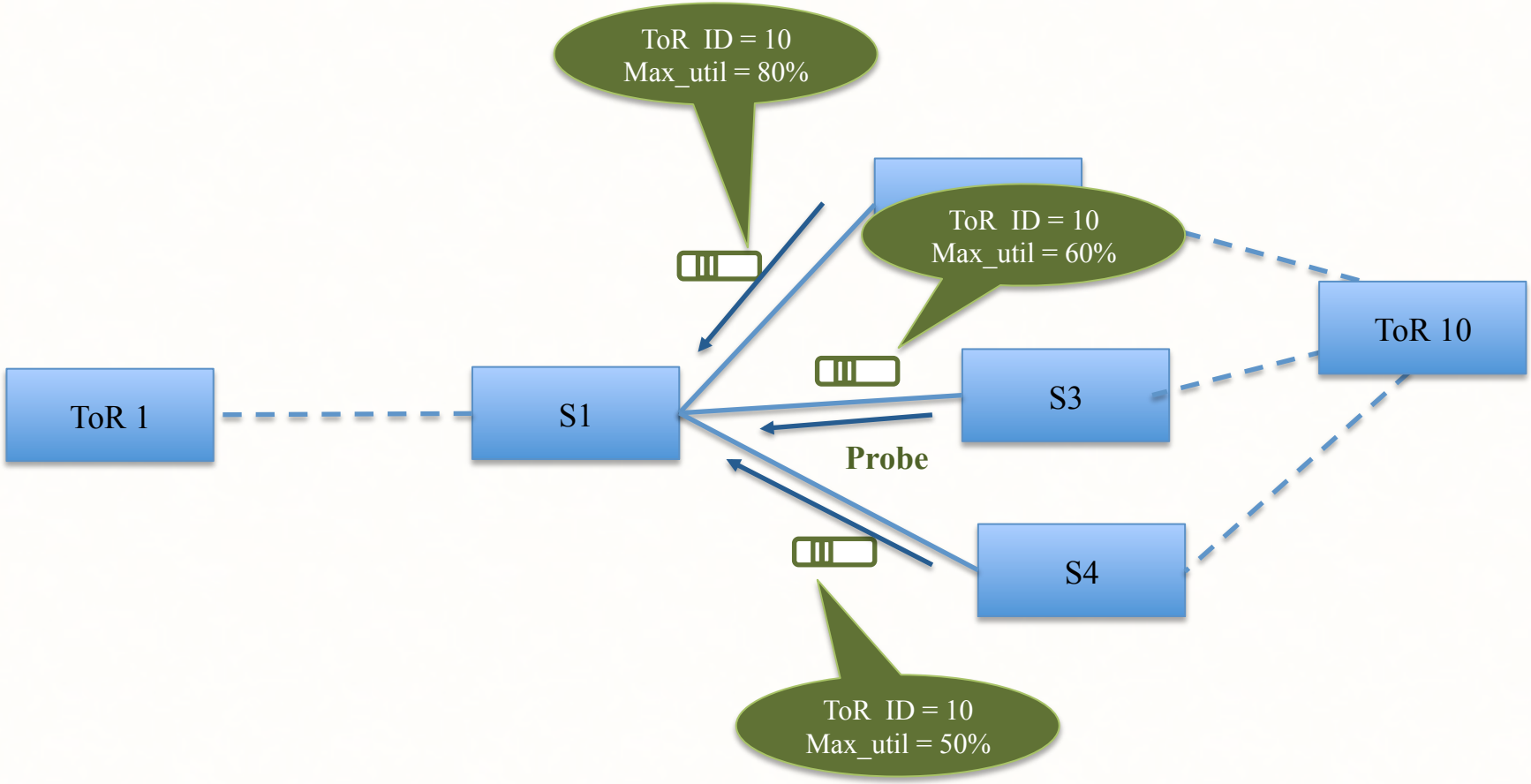


Probes carry path utilization

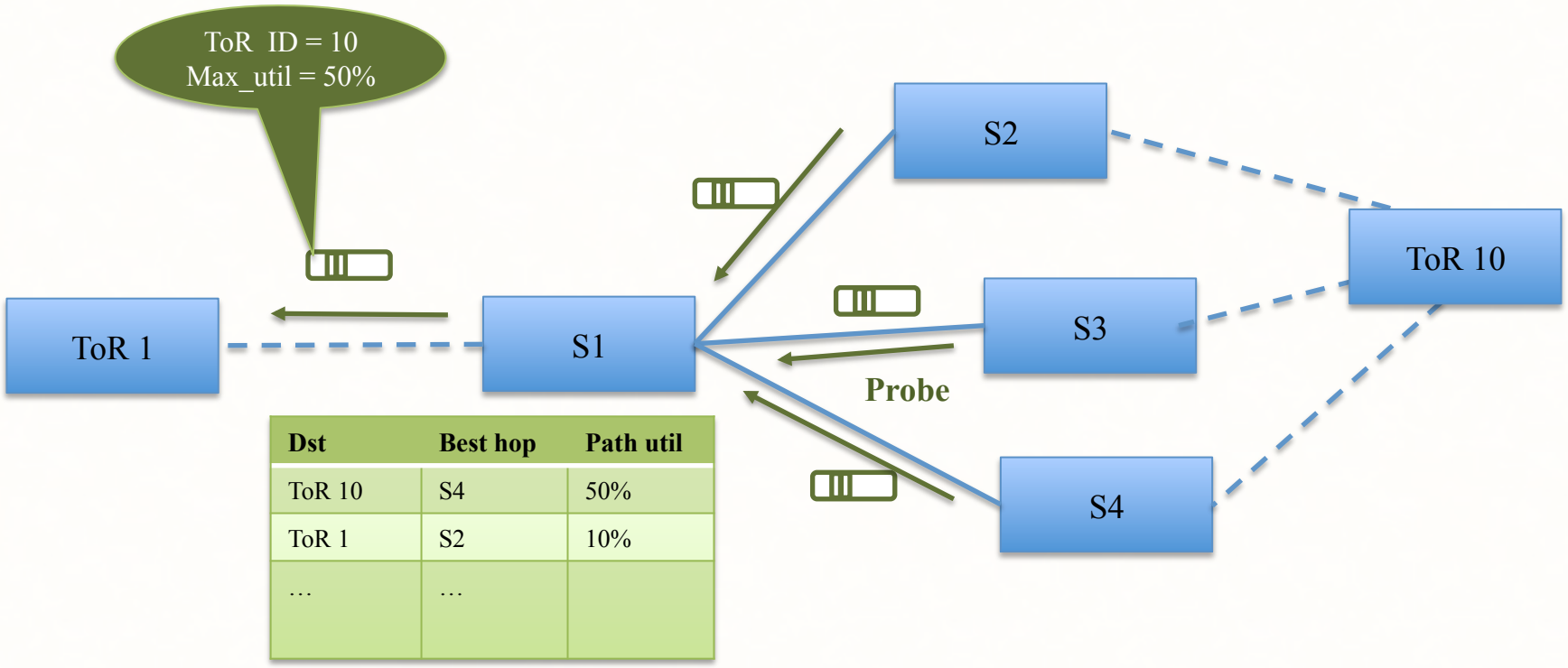
P4 primitives
New header format
Programmable Parsing
RW packet metadata



Probes carry path utilization



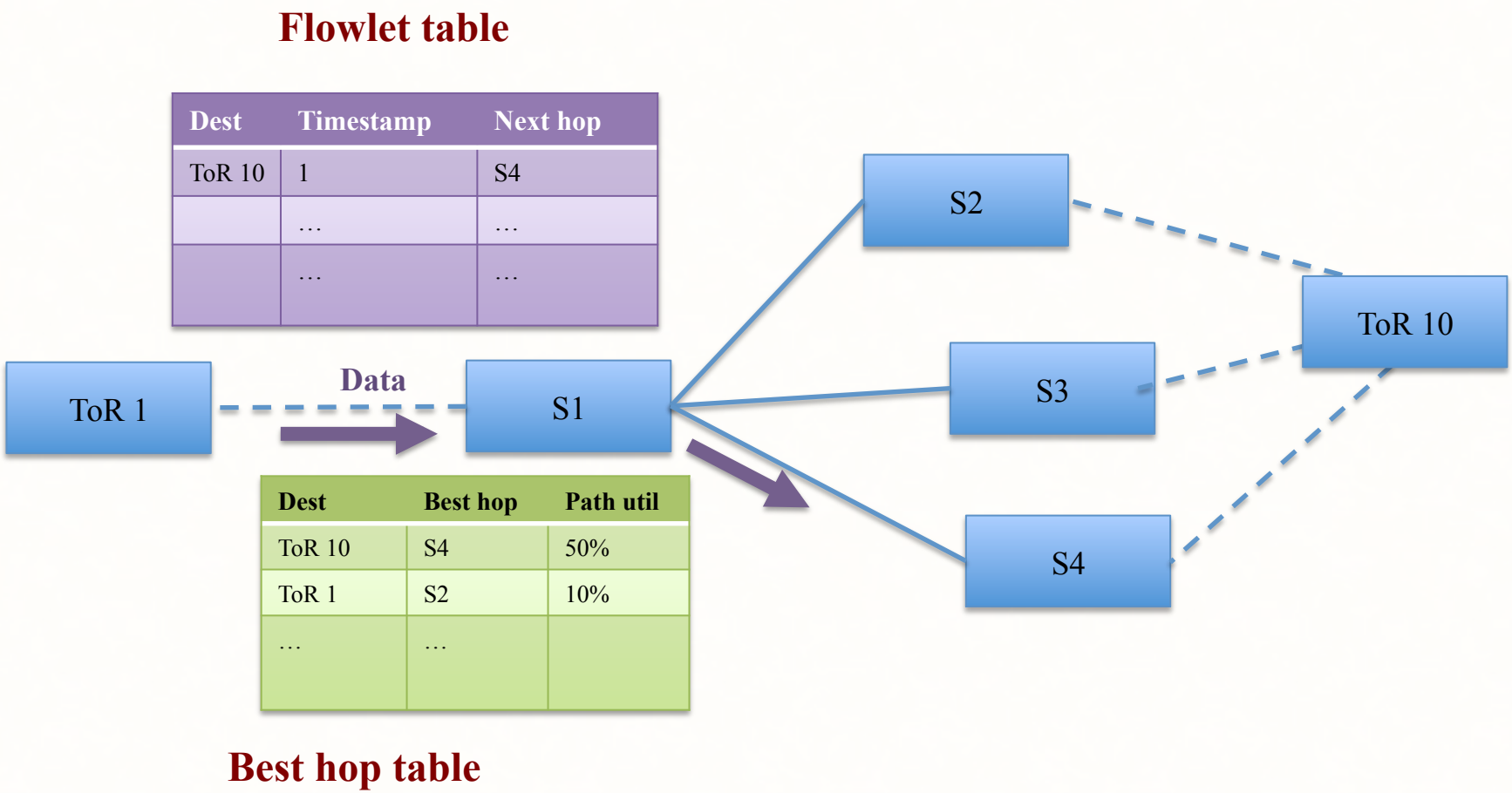
Each switch identifies best downstream path



Dst	Best hop	Path util
ToR 10	S4	50%
ToR 1	S2	10%
…	…	

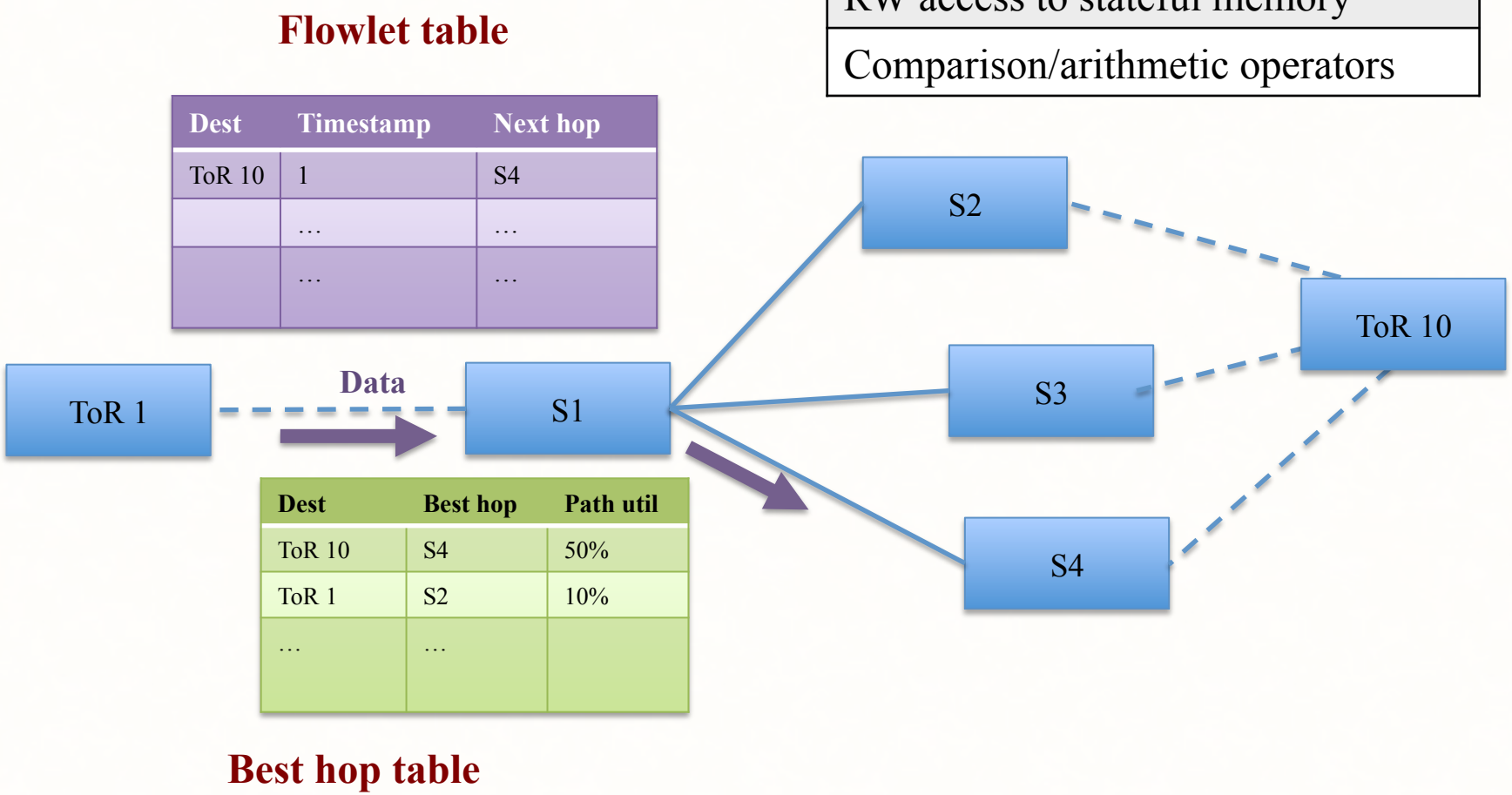
Best hop table

Switches load balance flowlets



Switches load balance flowlets

P4 primitives
RW access to stateful memory
Comparison/arithmetic operators



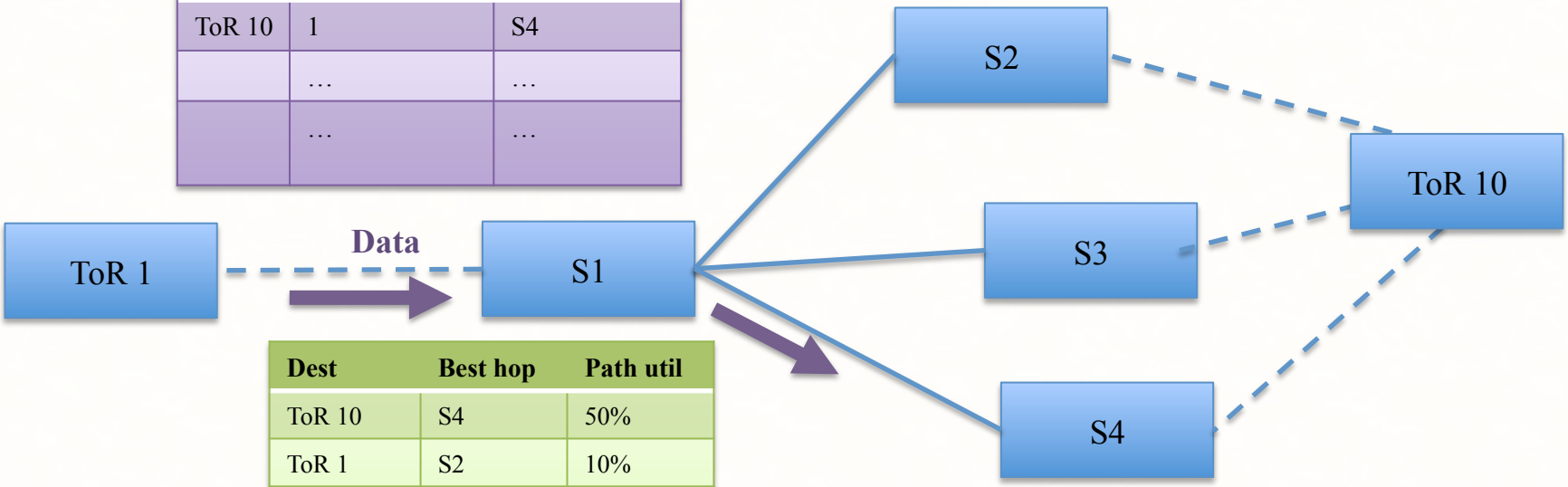
Switches load balance flowlets

P4 code snippet

```
if(curr_time - flowlet_time[flow_hash] > THRESH) {
        flowlet_hop[flow_hash] =best_hop[packet.dst_tor];
}
metadata.nxt_hop = flowlet_hop[flow_hash];
flowlet_time[flow_hash] = curr_time;
```

Flowlet table

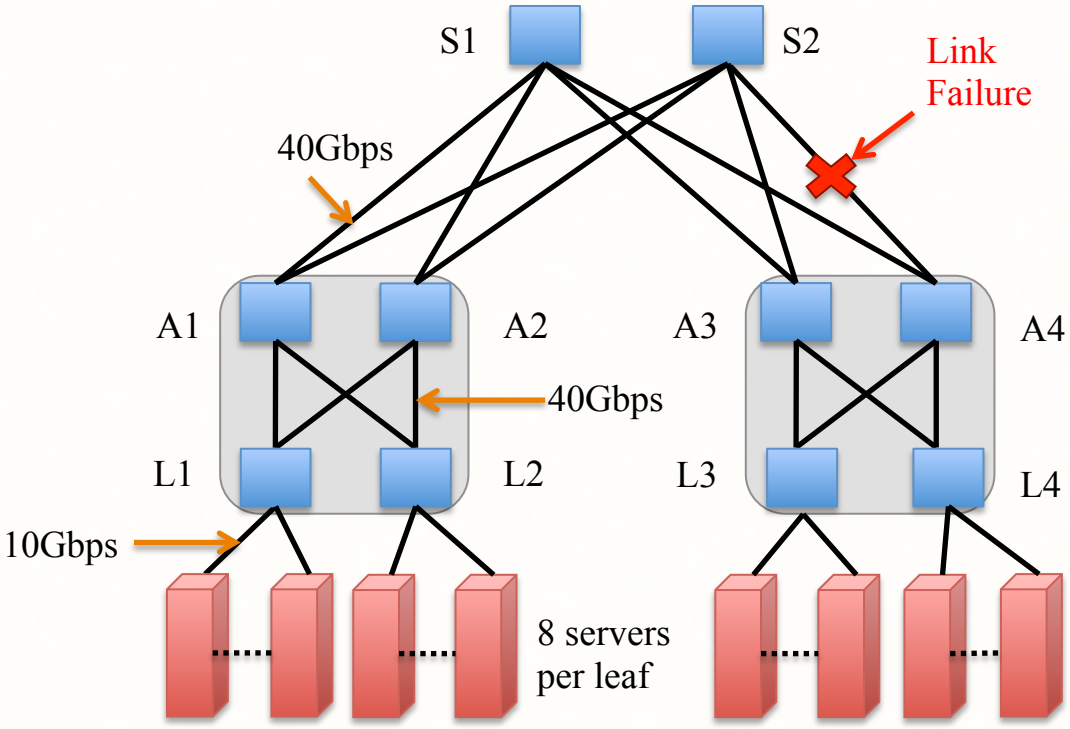
Dest	Timestamp	Next hop
ToR 10	1	S4
	…	…
	…	…



Dest	Best hop	Path util
ToR 10	S4	50%
ToR 1	S2	10%
…	…	

Best hop table

Evaluated Topology



Evaluation Setup

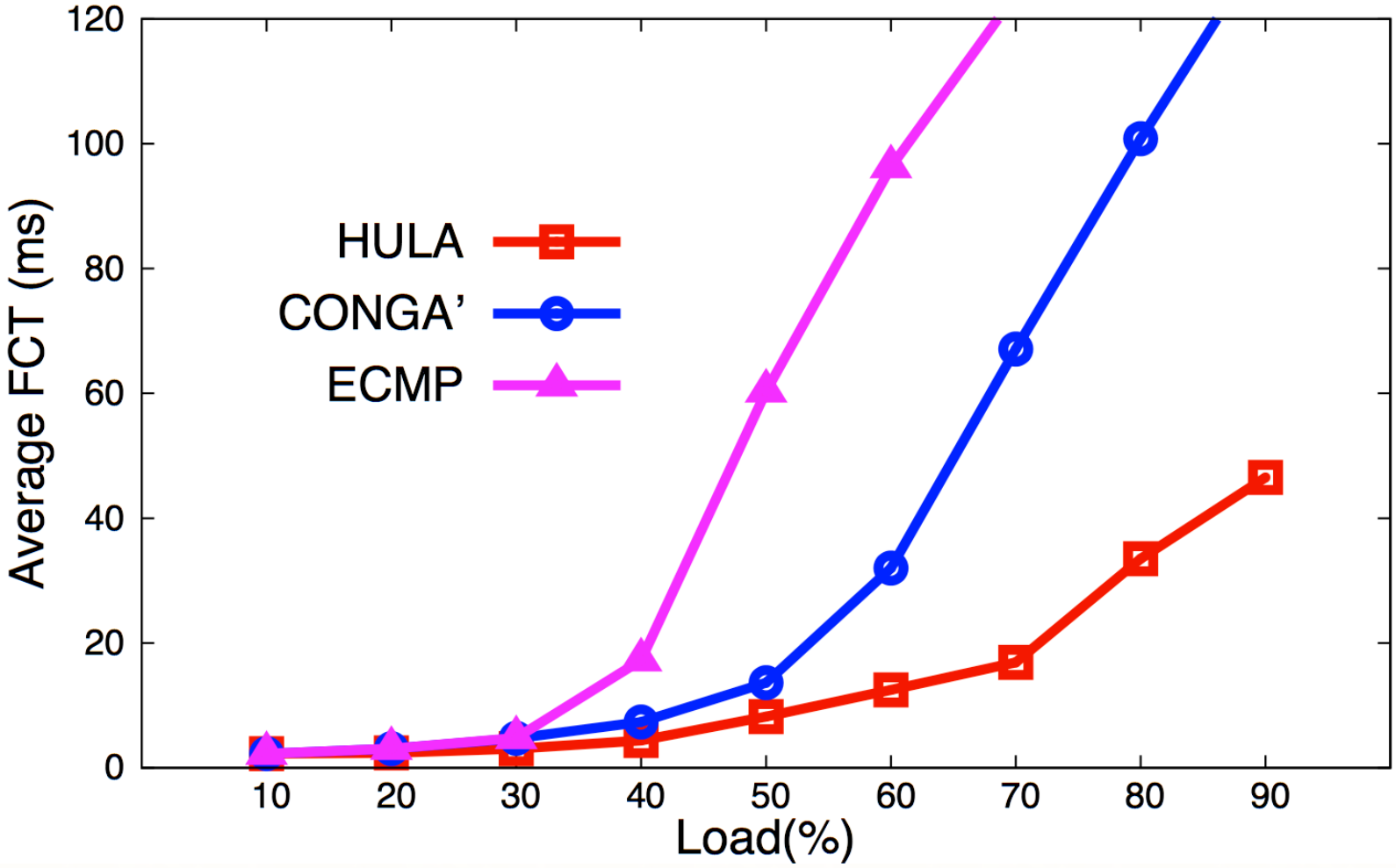
- NS2 packet-level simulator
- RPC-based workload generator
  - Empirical flow size distributions
  - Websearch and Datamining
- End-to-end metric
  - Average Flow Completion Time (FCT)

Compared with

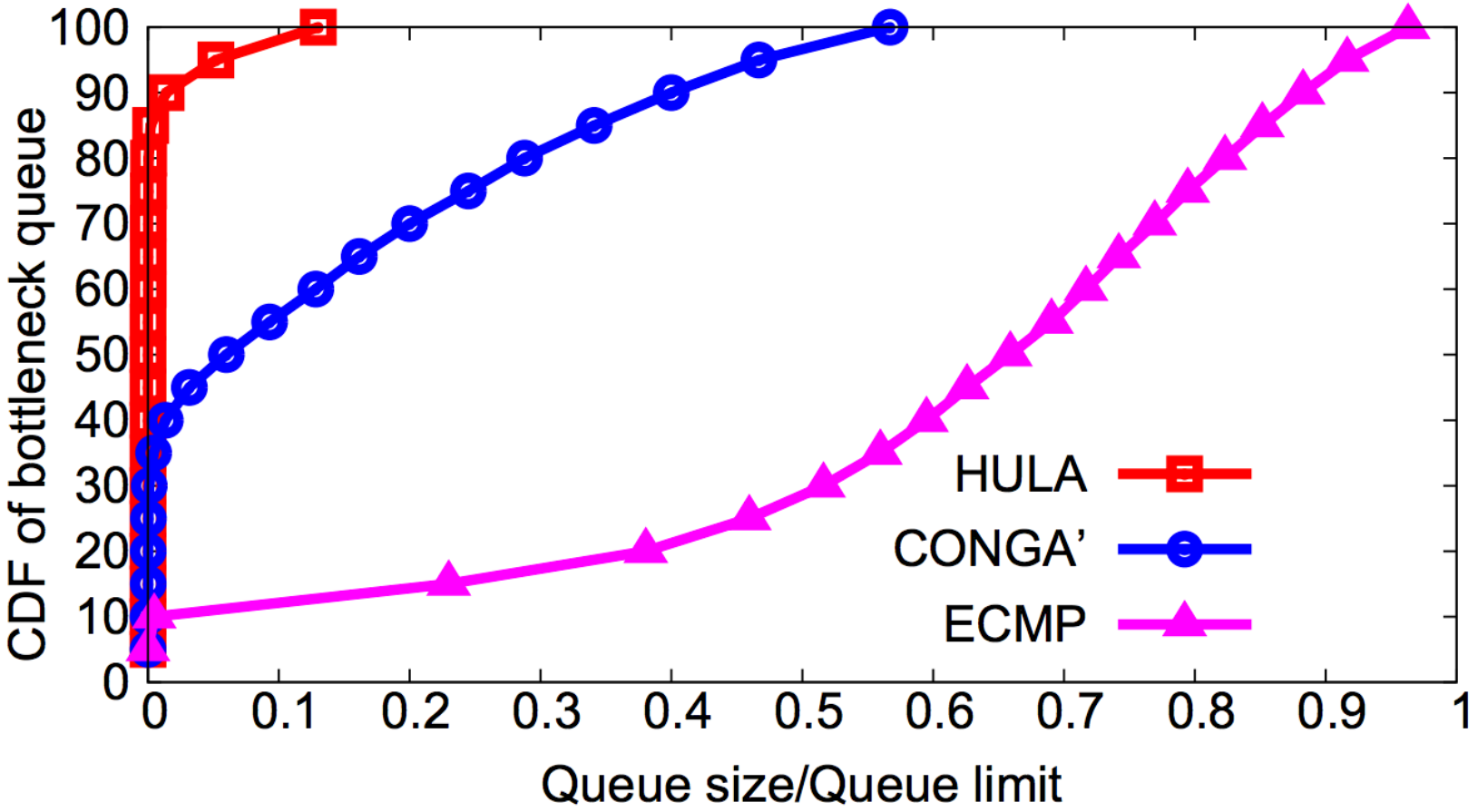
- ECMP
  - Flow level hashing at each switch
- CONGA'
  - CONGA within each leaf-spine pod
  - ECMP on flowlets for traffic across pods[1]

1. Based on communication with the authors

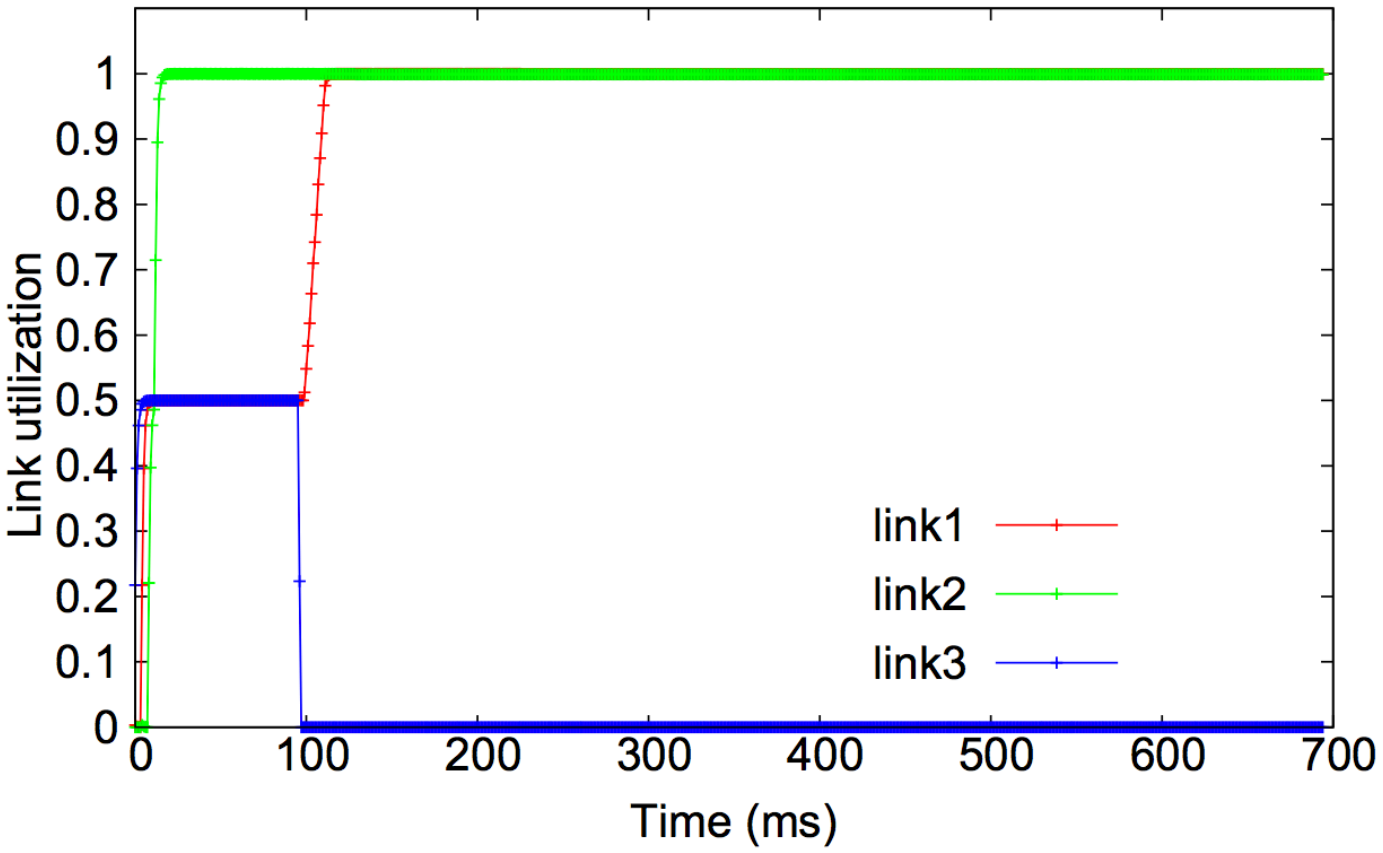
HULA handles high load much better



HULA keeps queue occupancy low



HULA is stable on link failure



HULA - Summary

- **Scalable** to large topologies
  - HULA distributes congestion state
- **Adaptive** to network congestion
- **Proactive** path probing
- Reliable when failures occur
- **Programmable** in P4!

Backup

HULA: Scalable, Adaptable, Programmable

LB Scheme	Congestion aware	Application agnostic	Dataplane timescale	Scalable	Programmable dataplanes
ECMP		✓	✓	✓	
SWAN, B4	✓	✓			
MPTCP	✓		✓	✓	
CONGA	✓	✓	✓		
HULA	✓	✓	✓	✓	✓