Comparative Evaluation of CEE-based Adaptive Routing

Daniel Crisan, Mitch Gusat and Cyriel Minkenberg

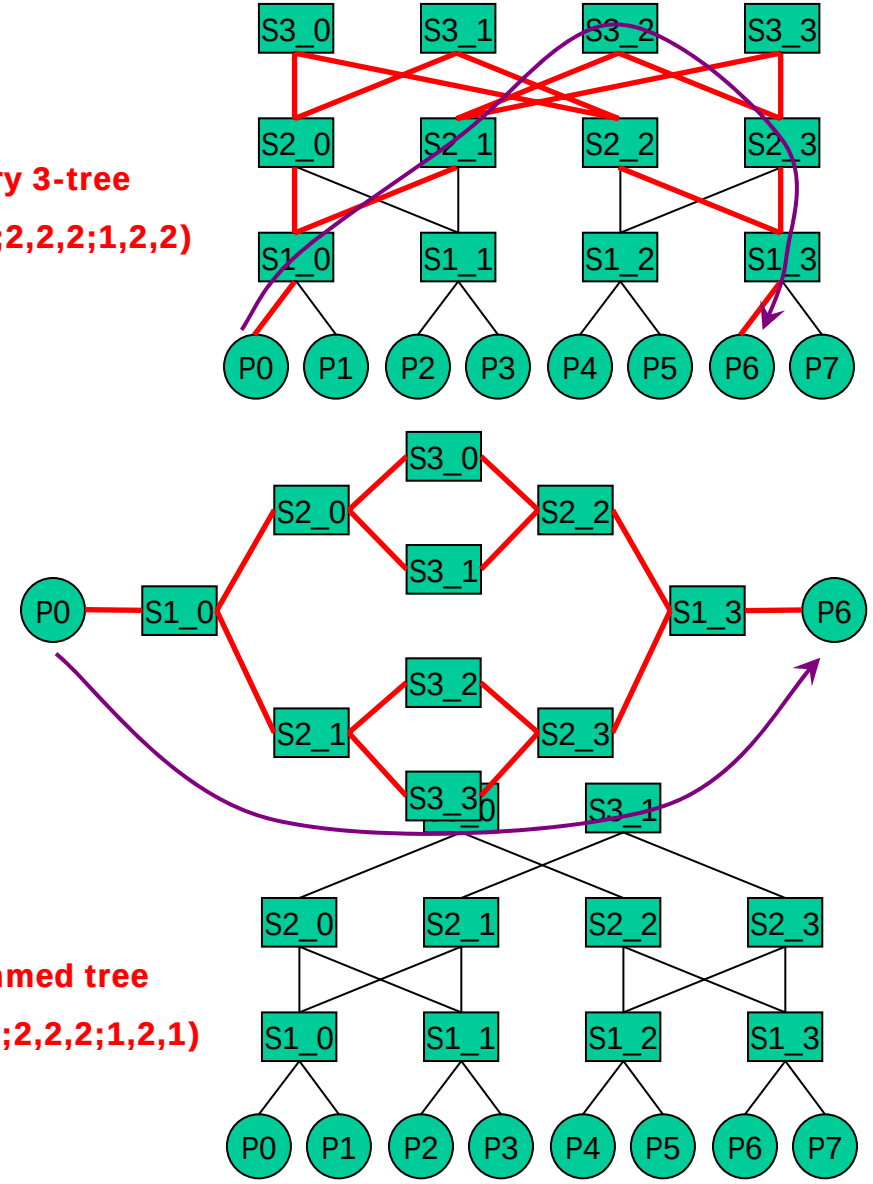
IBM Research GmbH, Zürich Research Laboratory

Outline

- Intro. Fat-tree DCNs
- Oblivious routing, best of
- CEE adaptive routing, switch & SRC
- Evaluation
- Conclusions

Fat-trees

- Most used in DCN
  - reliability
  - B_{bis}=ct.
- Fat tree *[Leiserson'85]*
- k-ary n-tree *[Petrini'97]*
- XGFT *[Ohring'95]*
- Enable multi-paths

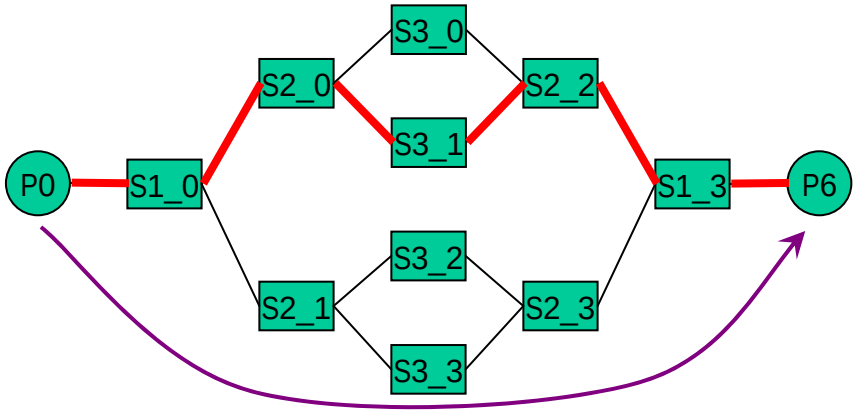


Oblivious Routing (1)

- Deterministic routing

  [Leiserson'92]...[Gomez'07]

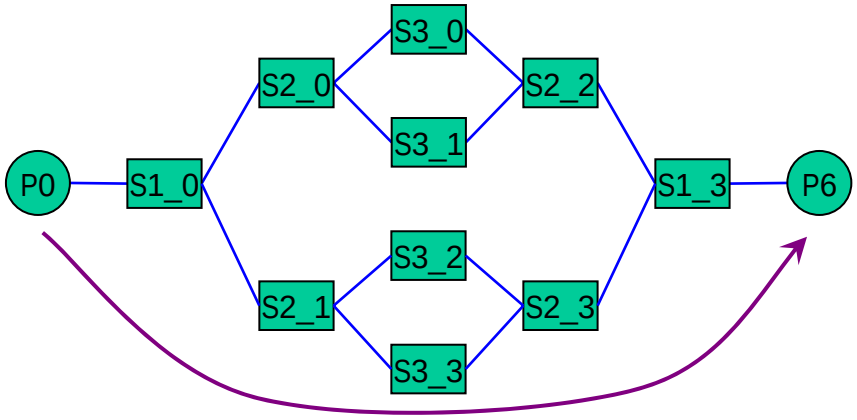
  - single fixed path



- Random routing

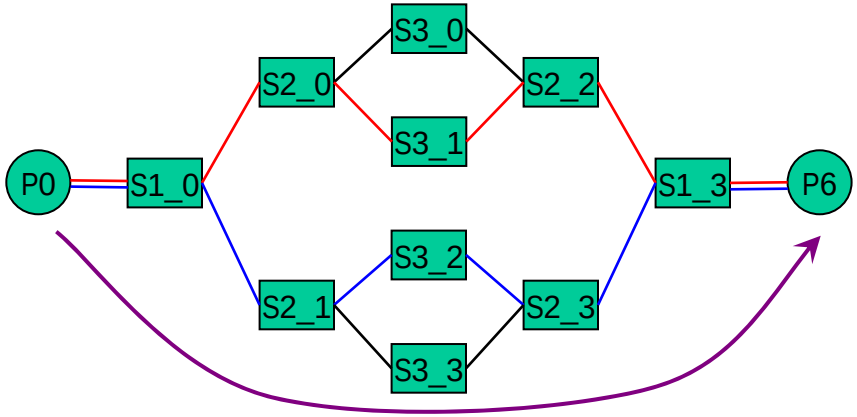
  [Valiant'81] [Greenberg'85]

  - use all paths with equal probability



Oblivous Routing (2)

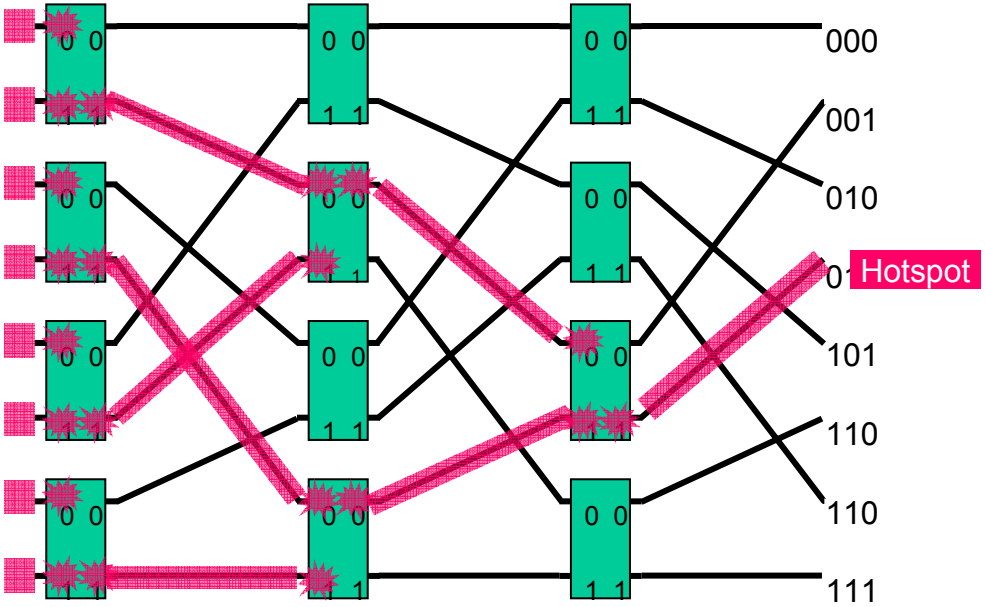
- Hashed routing
  - FlowID: 5-tuple
  - Hash function:
    {flow id} → {paths}

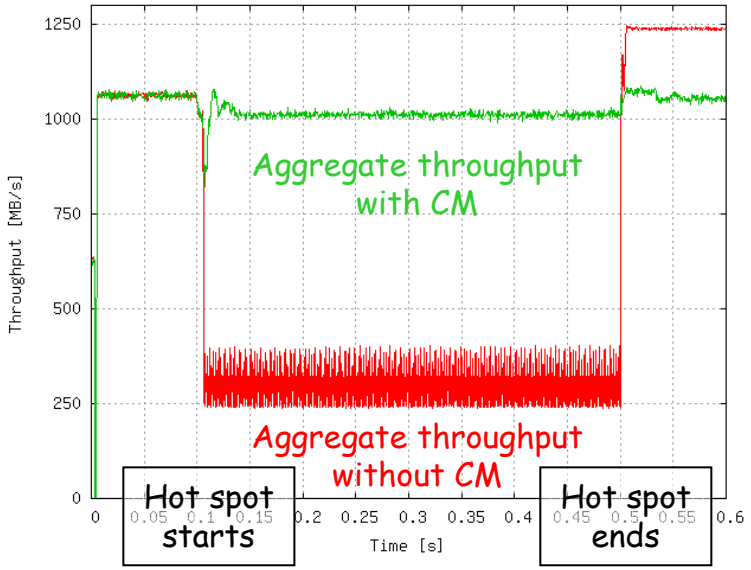


- Upper boundary
  - Few long flows → Deterministic
- Lower boundary
  - Many short flows → Random

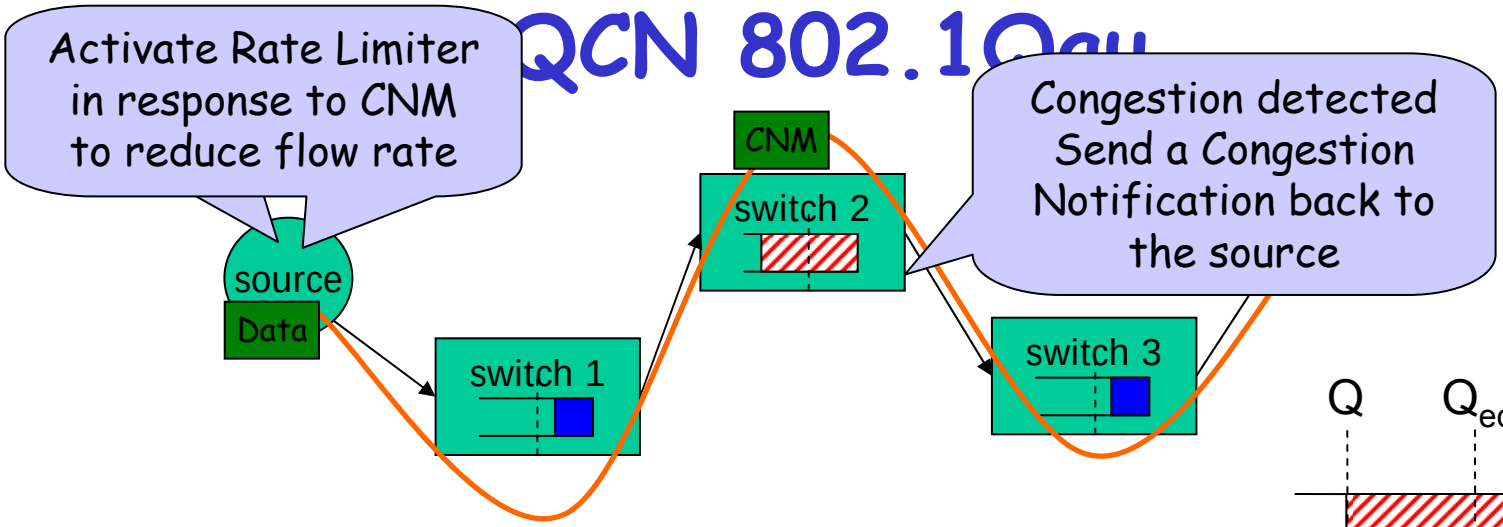
CEE Layer 2 Congestion Management

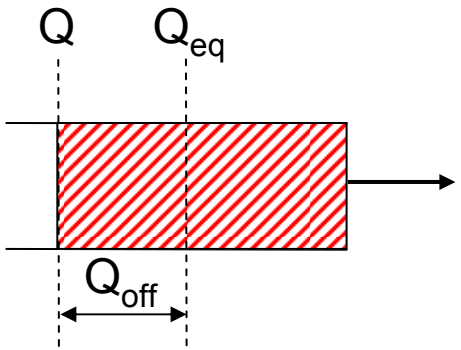
- Network congestion can lead to severe performance degradation
  - Lossy networks suffer from the “avalanche” effect: High load ➔ drops ➔ retransmissions ➔ increased load ➔ even more drops
  - Lossless CEE DC networks suffer from *saturation tree congestion:* Link-level flow control (PFC) can cause congestion to roll back from switch to switch
- Congestion management (CM) is needed to deal with long-term (sustained) congestion
  - PFC is ill suited to this task, dealing only with short-term (transient) congestion
  - Push congestion from the core towards the edge of the network





QCN 802.1Qau





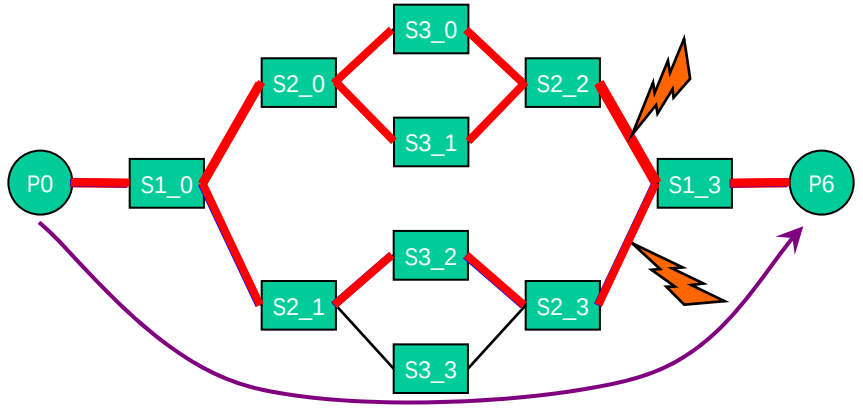
- Congestion detection
  - measures taken every ~100 packets
  - compute Q_{off} and Q_{delta} (position and velocity)
  - $F_b = -(Q_{off} + w \cdot Q_{delta})$ sent back only if is negative
- Rate Limiter
  - reduce rate proportional with the F_b
  - recover rate using a byte counter and a timer

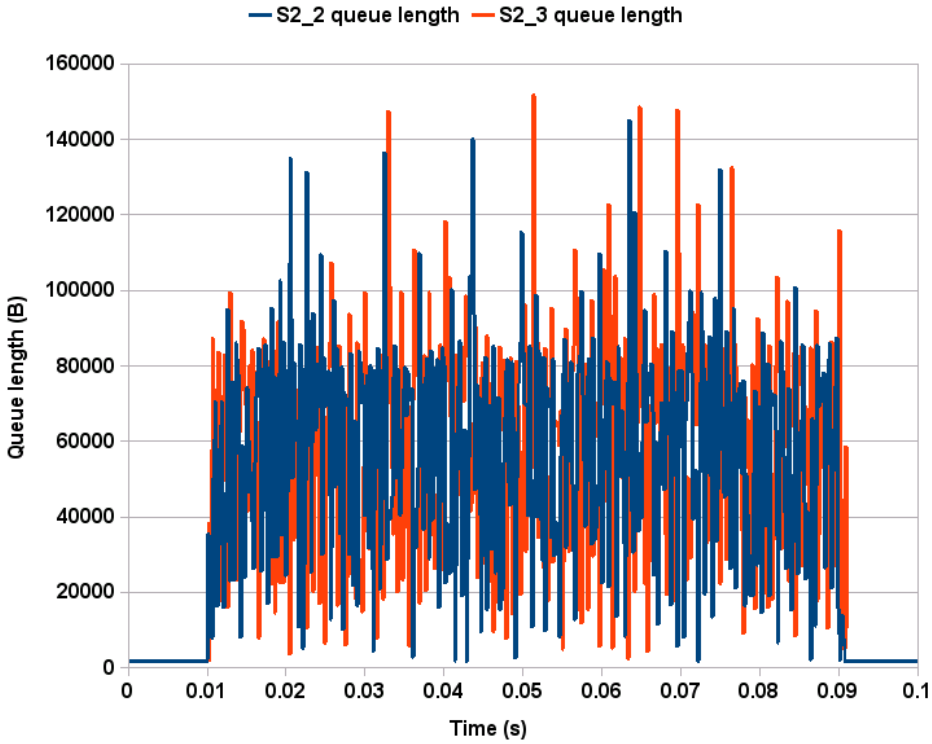
Rate or Route?
Congestion Management vs. Adaptive Routing

- CM solves congestion by reducing injection rate
  - Useful for saturation tree congestion, where many “innocent” flows suffer because of backlog of some hot flows
  - Does not exploit path diversity
  - Typical data center topologies offer high path diversity
    - Fat tree, mesh, torus
- Adaptive routing (switch AR) basic approach
  - Allow multi-path routing
  - By default route on shortest path (latency)
  - Detect downstream congestion by means of QCN
  - In case of congestion
    - First try to reroute hot flows on alternative paths
    - Only if no uncongested alternative exists, reduce send rate

Switch Adaptive Routing

- QCN feedback provide "congestion price"
- Algorithm
  [Minkenberg&Gusat'09]
  - switches snoop the CNs
  - based on feedback – steer the traffic
- Advantages
  - Congestion avoidance
  - Use of alternative paths
- Oscillations possible
- Routing controlled by switches





Source AR: R^3C^2 Concept

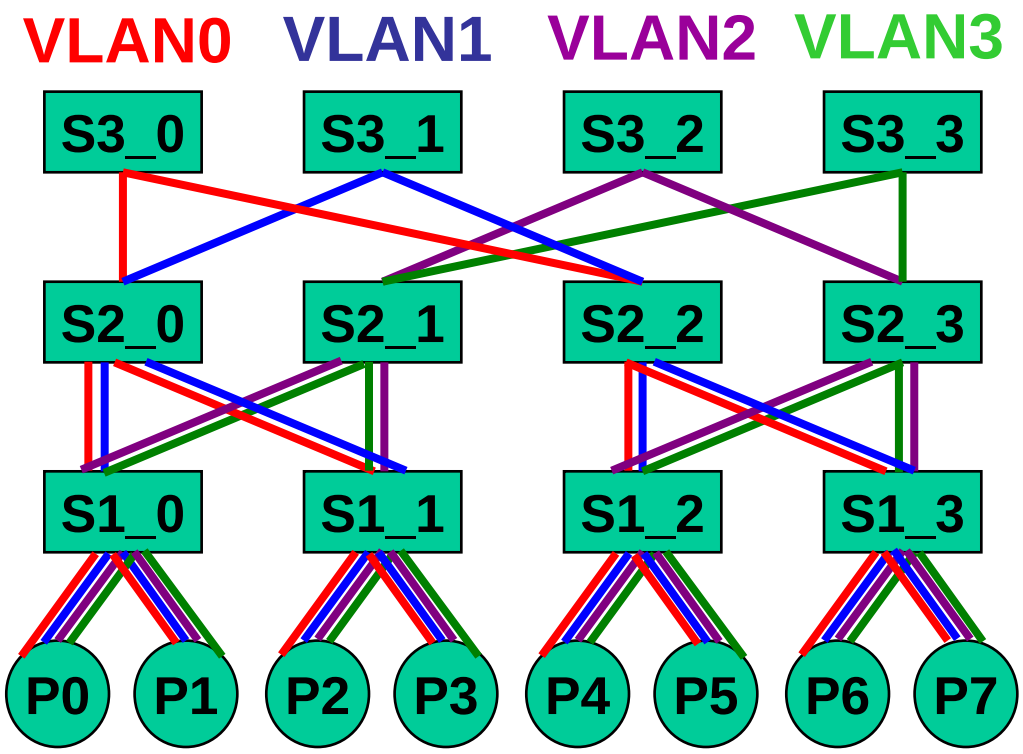
Take advantage of CNMs at the source for adaptive load-balancing

- **Congestion Point** issues CNMs
  - Where is the hotspot?
  - How severe is the hotspot?

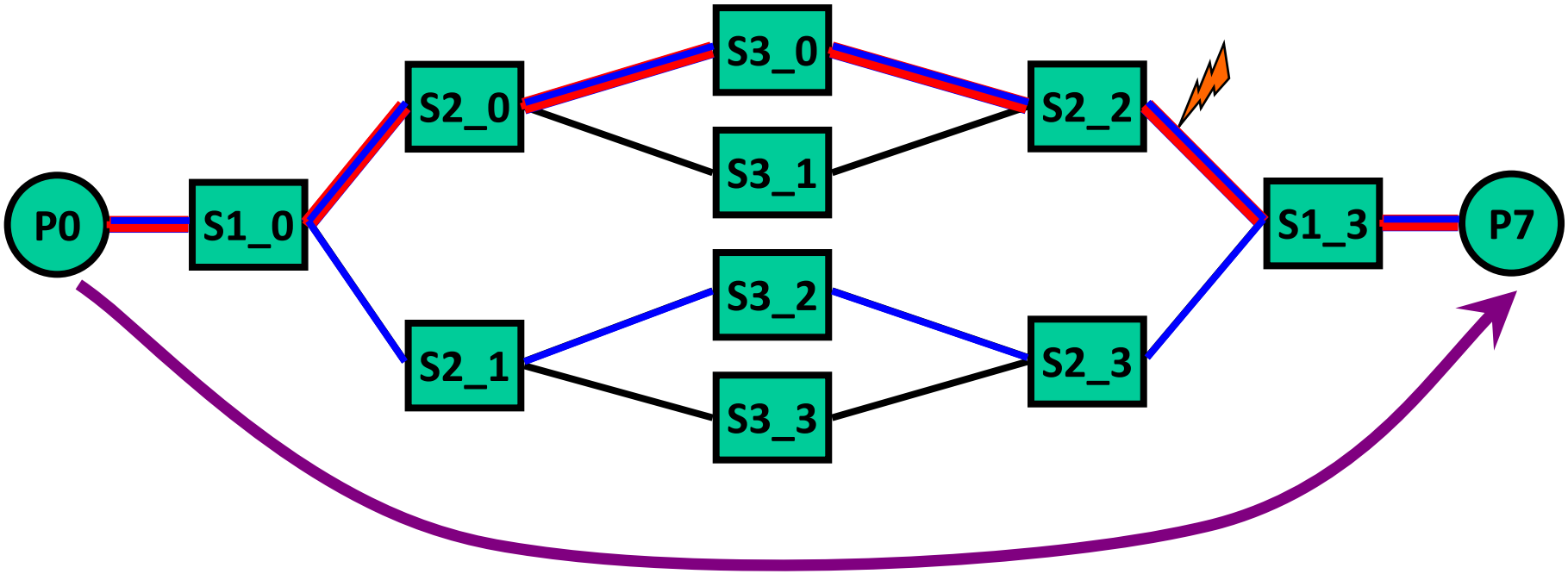
- **Source** receives the CNMs
  - Identifies the most severe hotspots
  - Reroutes traffic around the hotspots
  - Splits flows and rate-limits subflows

Source Routing in CEE: VLAN

- Ethernet is not source-routed
- Solution: VLAN
  - One tree per VLAN
- Source
  - Set VLAN# at injection → path selection

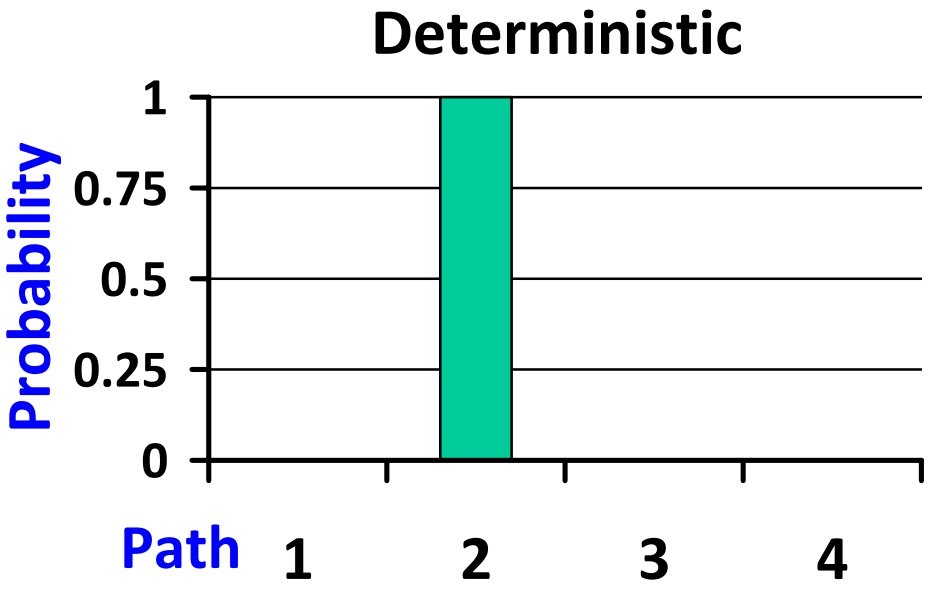


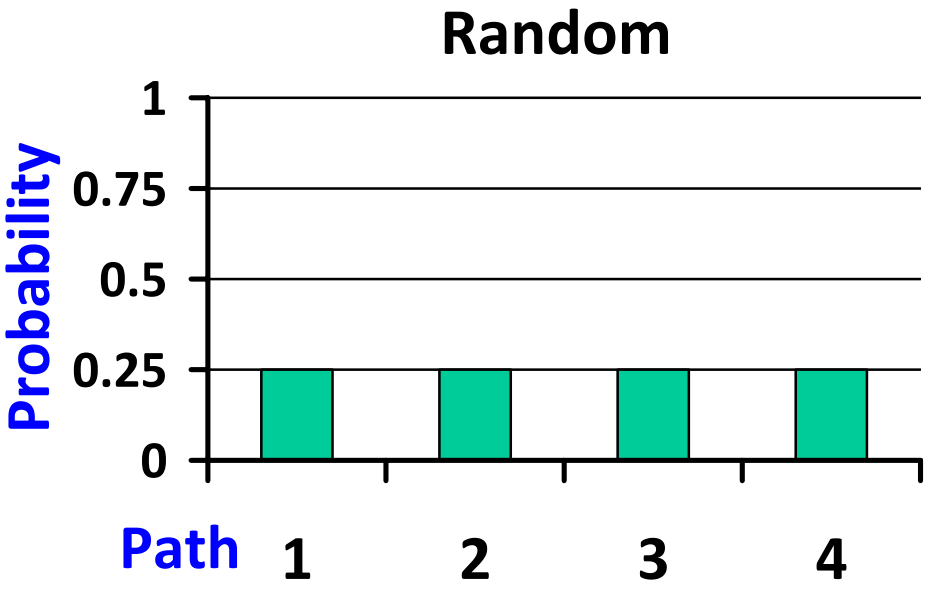
R^3C^2 Algorithm

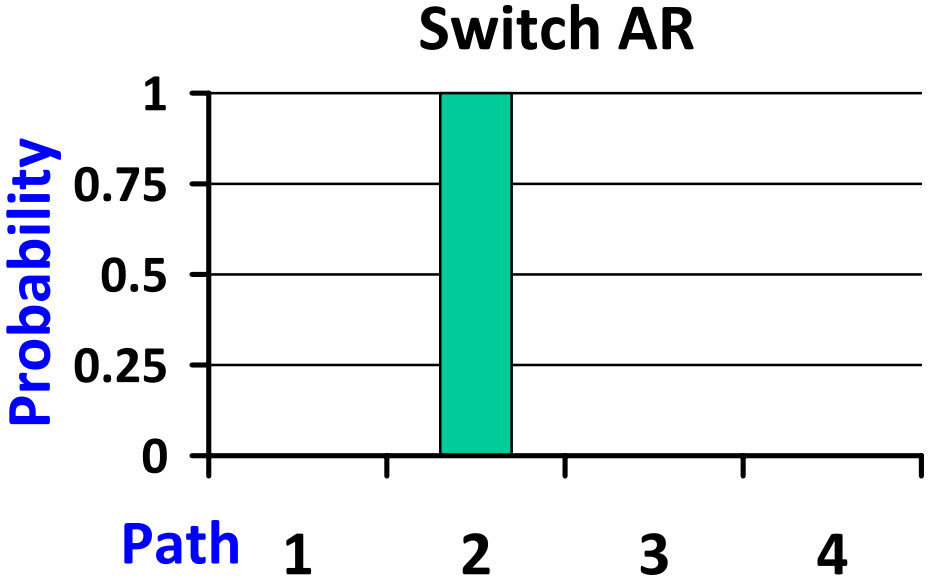


- No overload: Deterministic single path
- Congestion: Activate additional paths
- Path activation: avoid hotspots
- Use RL along each path

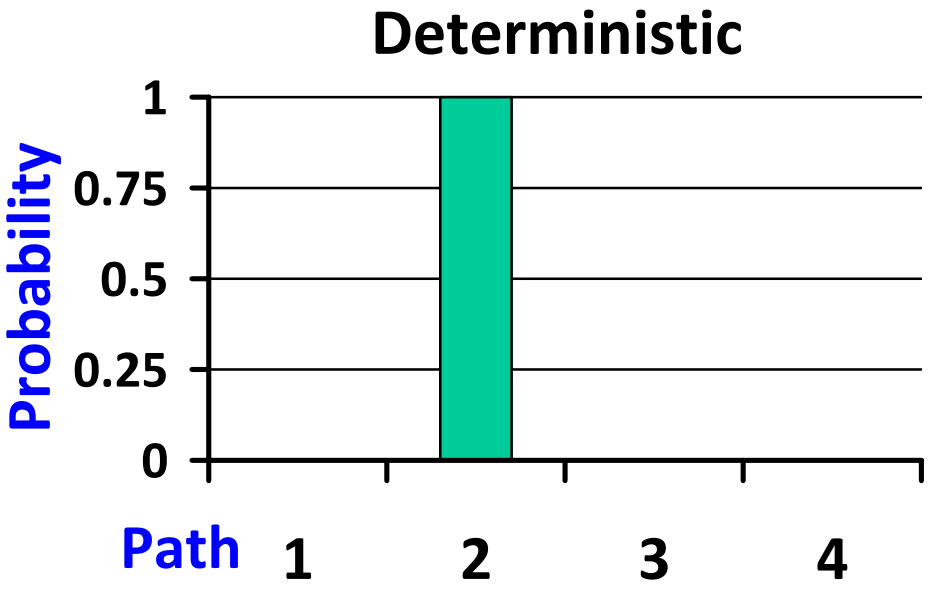
Routing Schemes

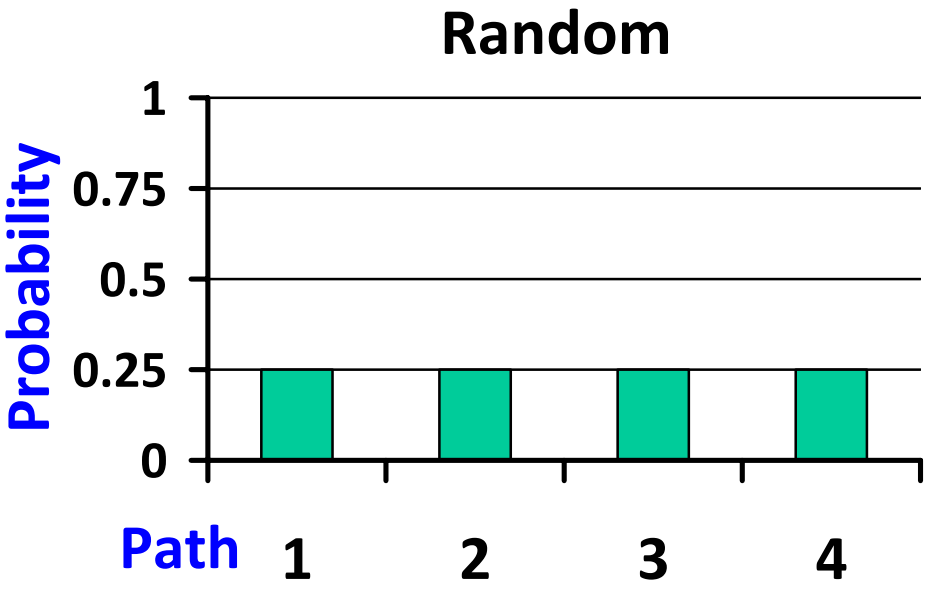


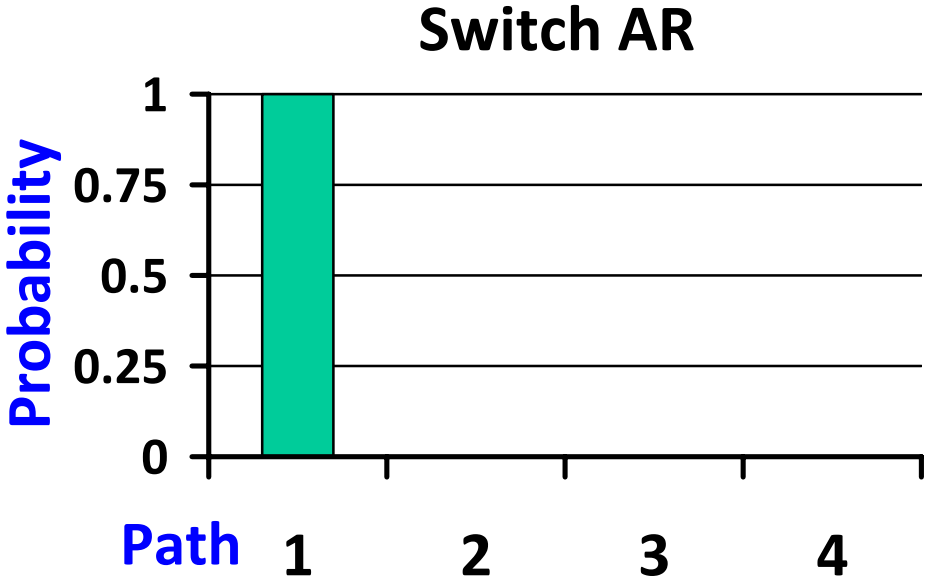




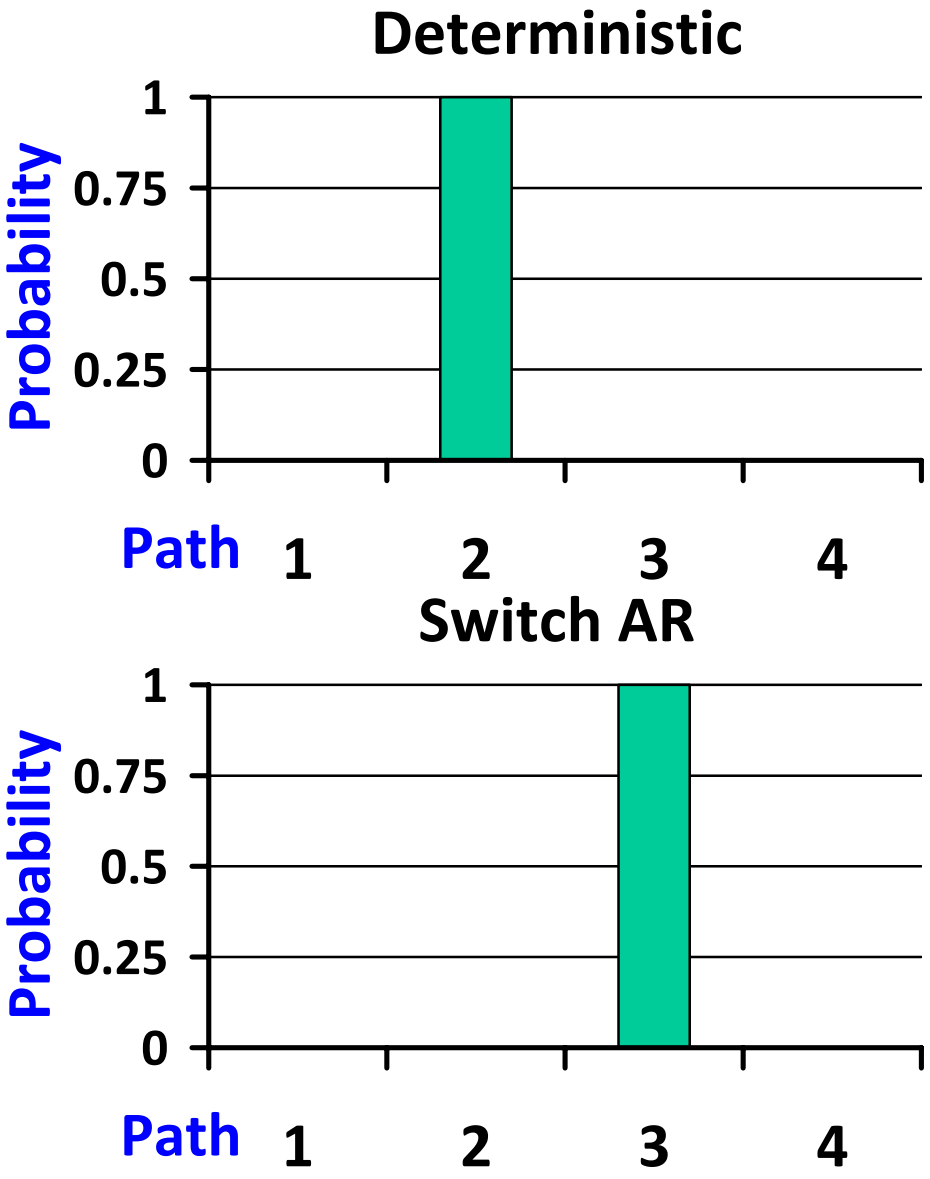
Routing Schemes

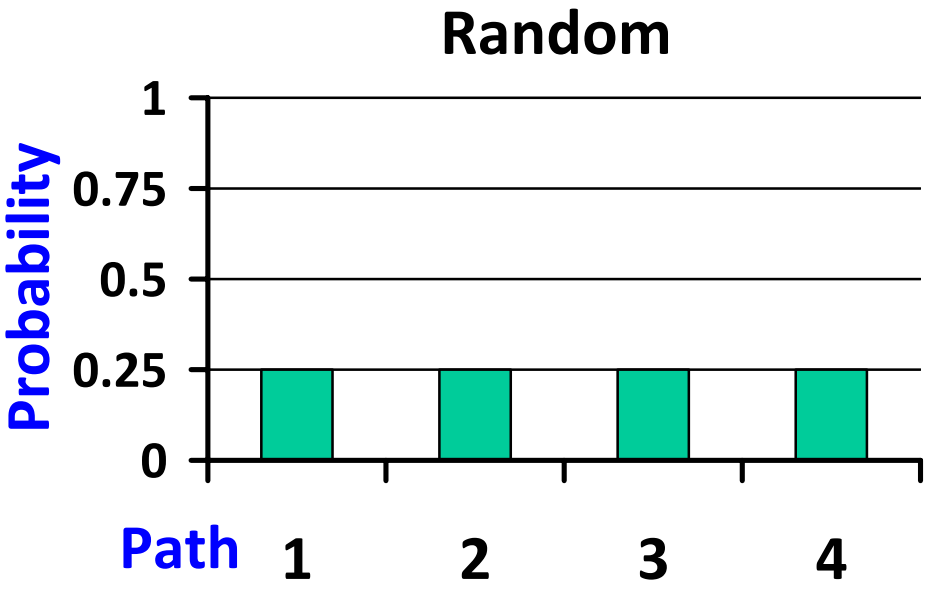




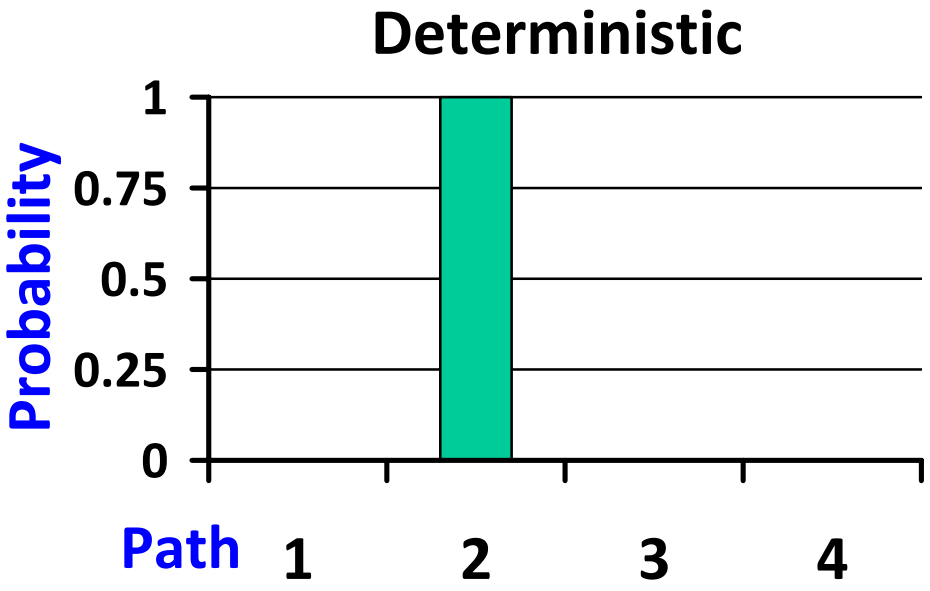


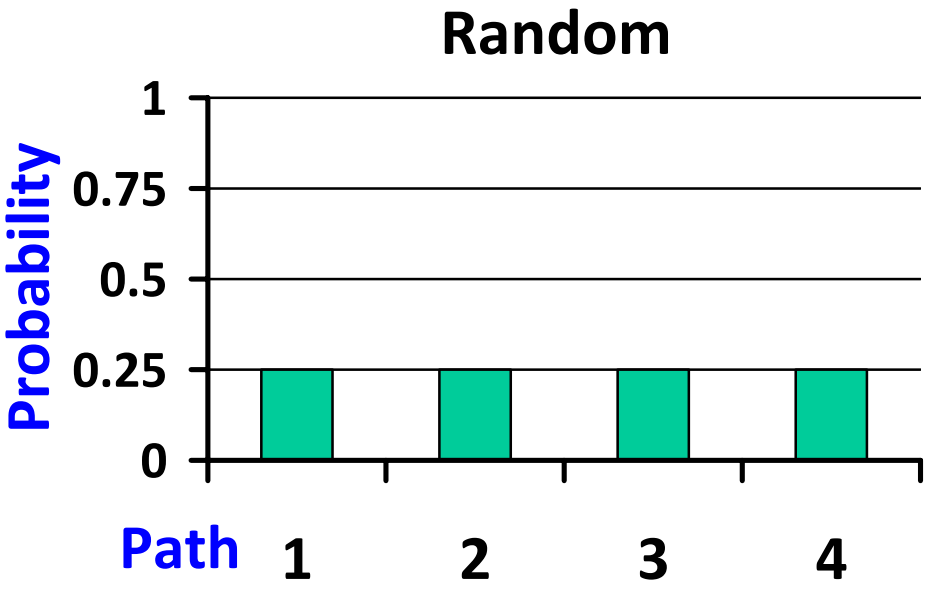
Routing Schemes

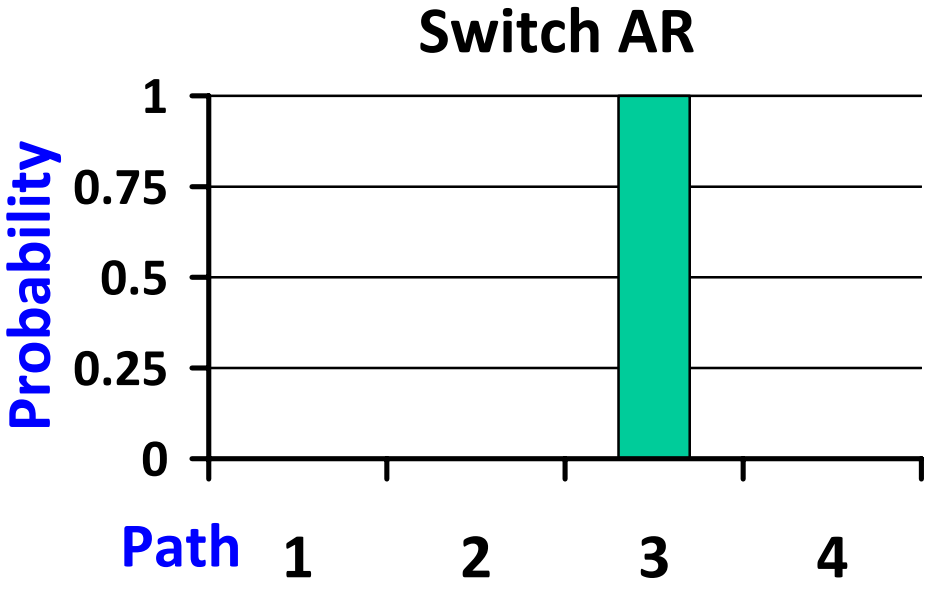


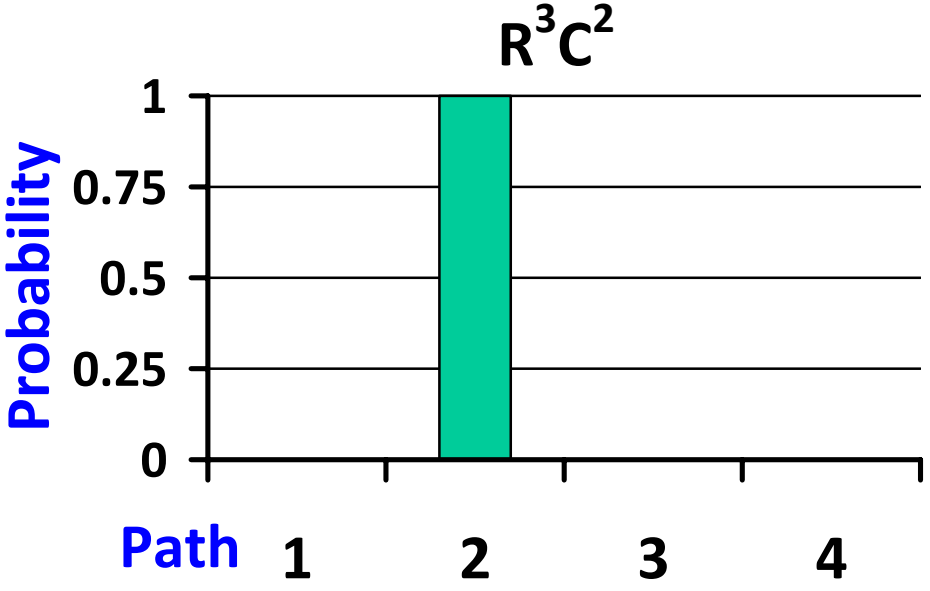


Routing Schemes

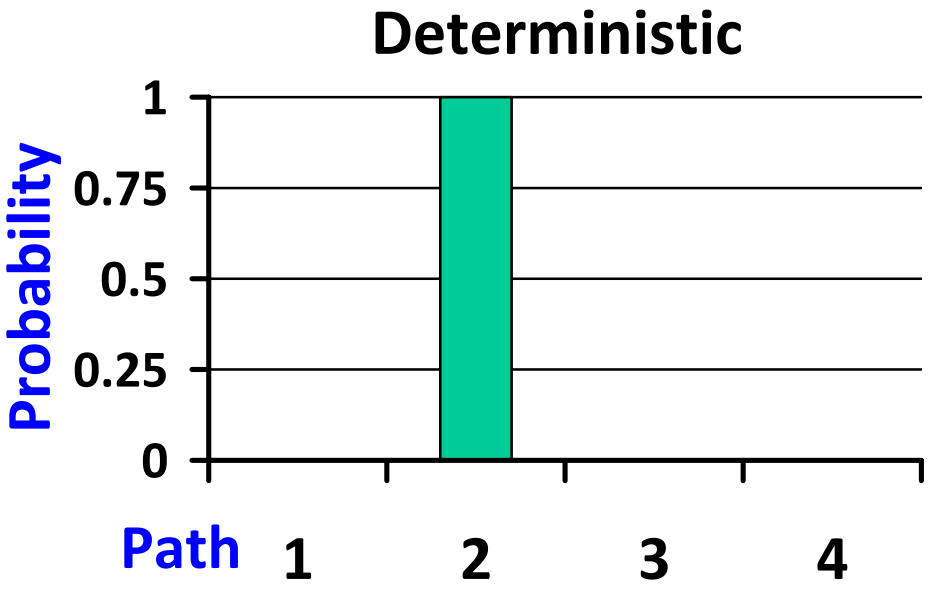


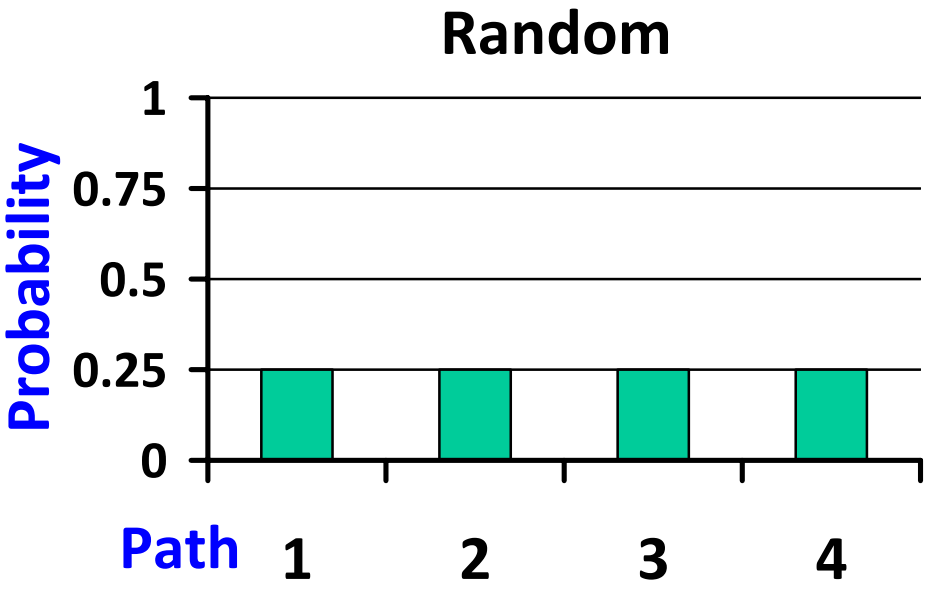


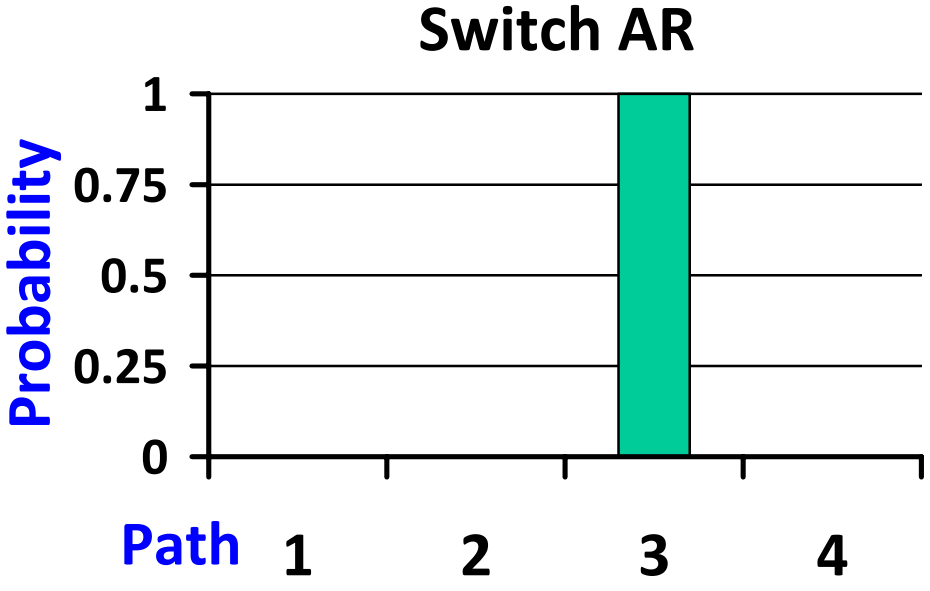


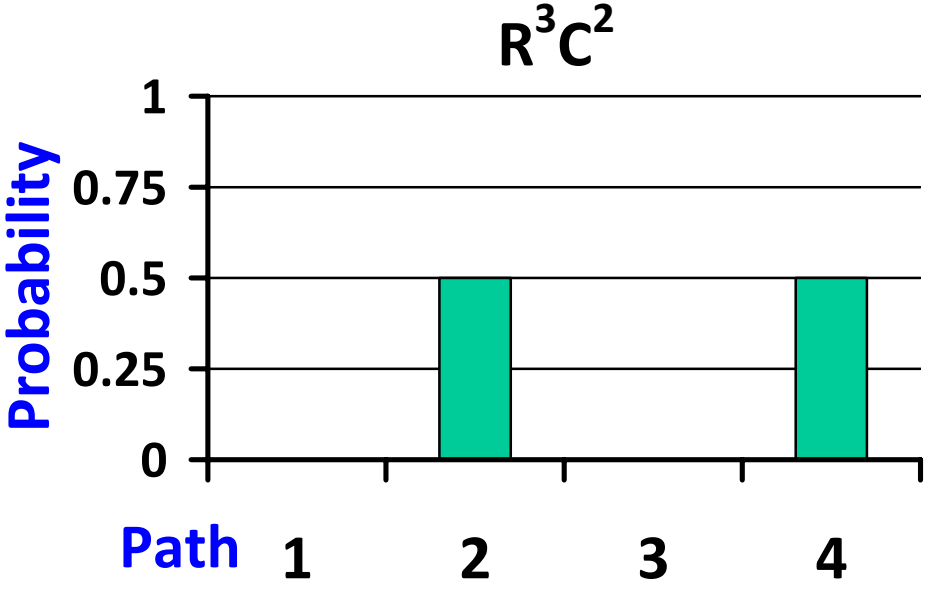


Routing Schemes

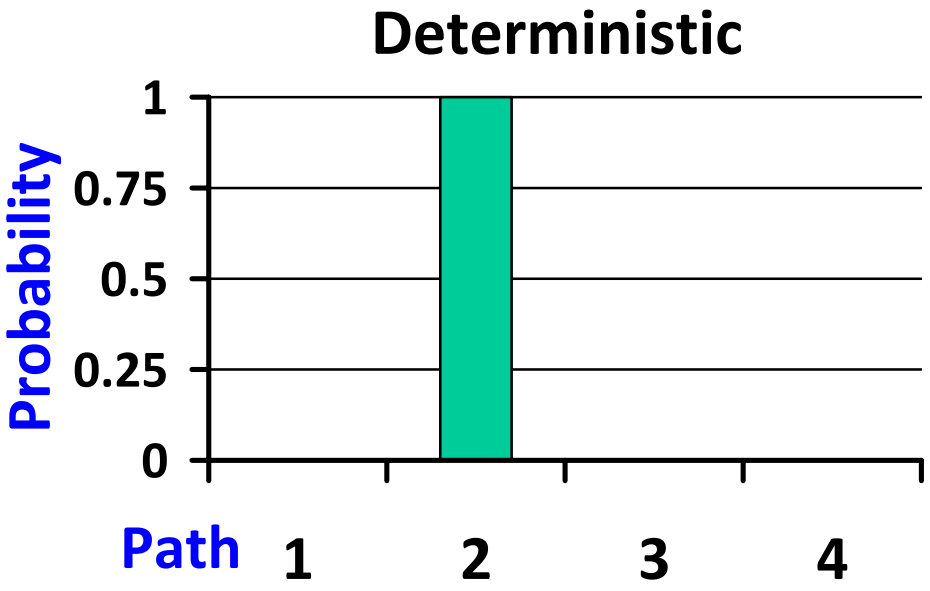


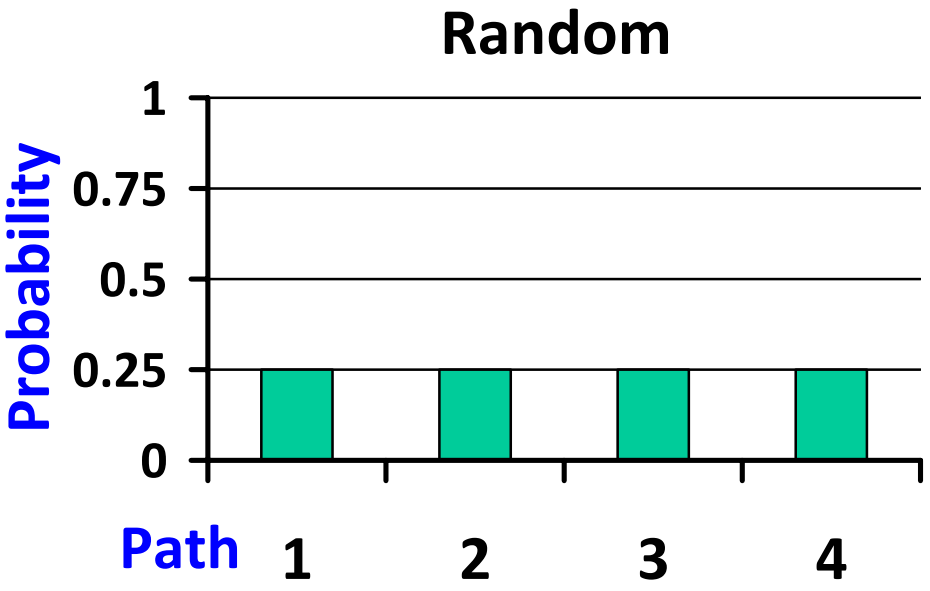


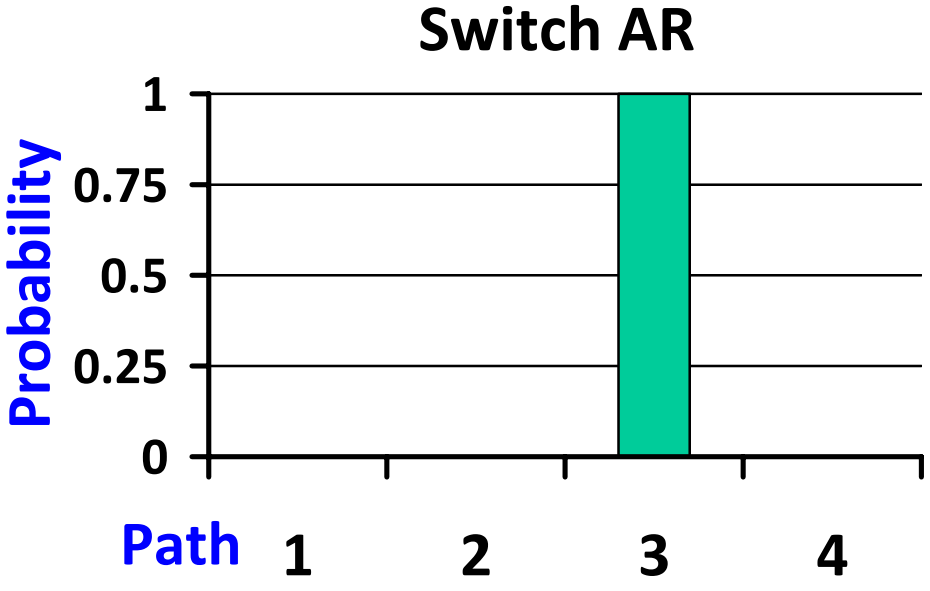


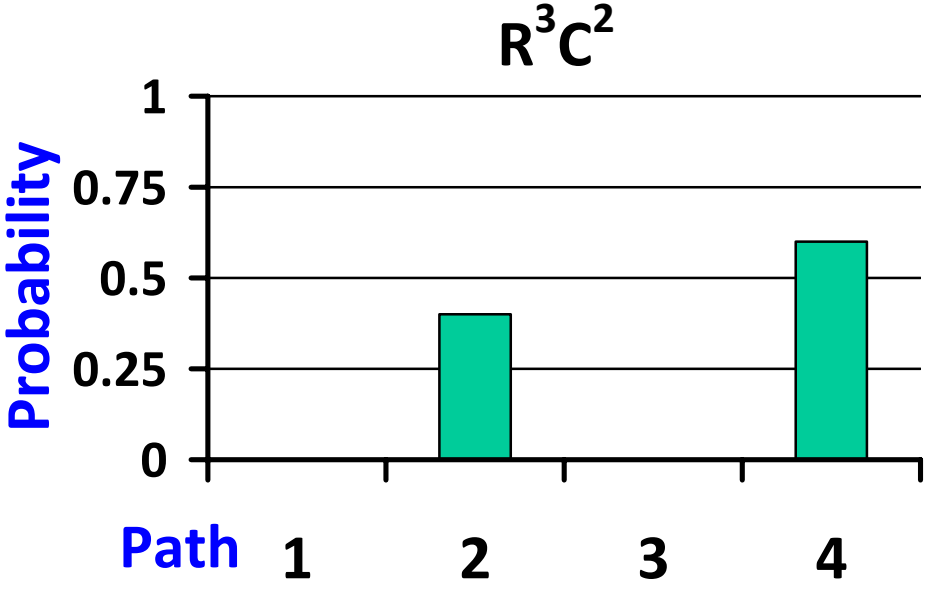


Routing Schemes

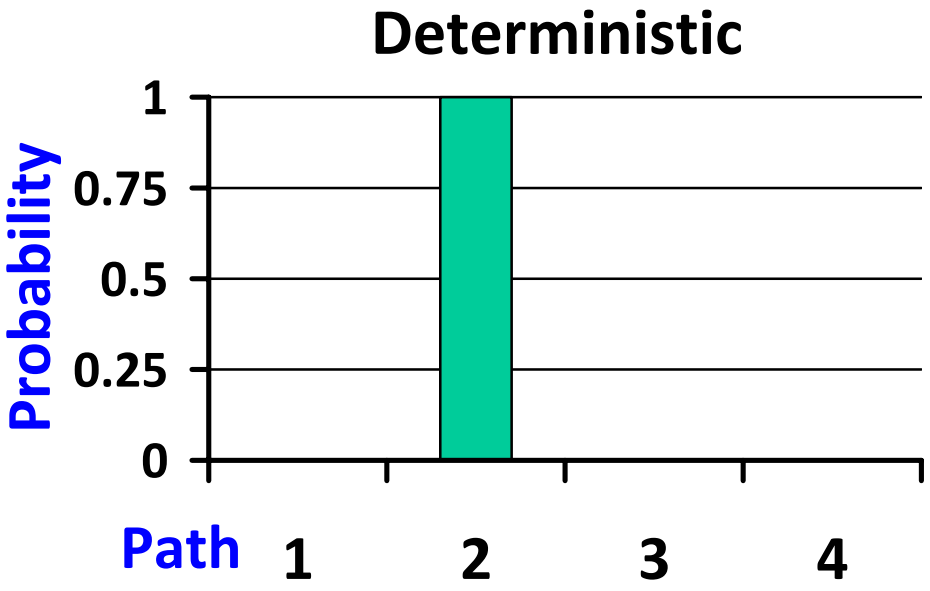


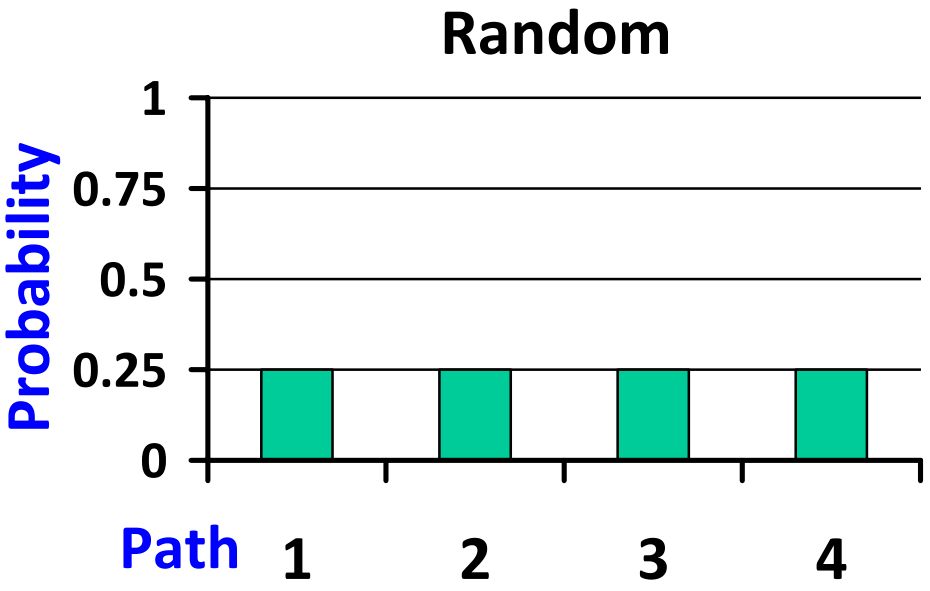


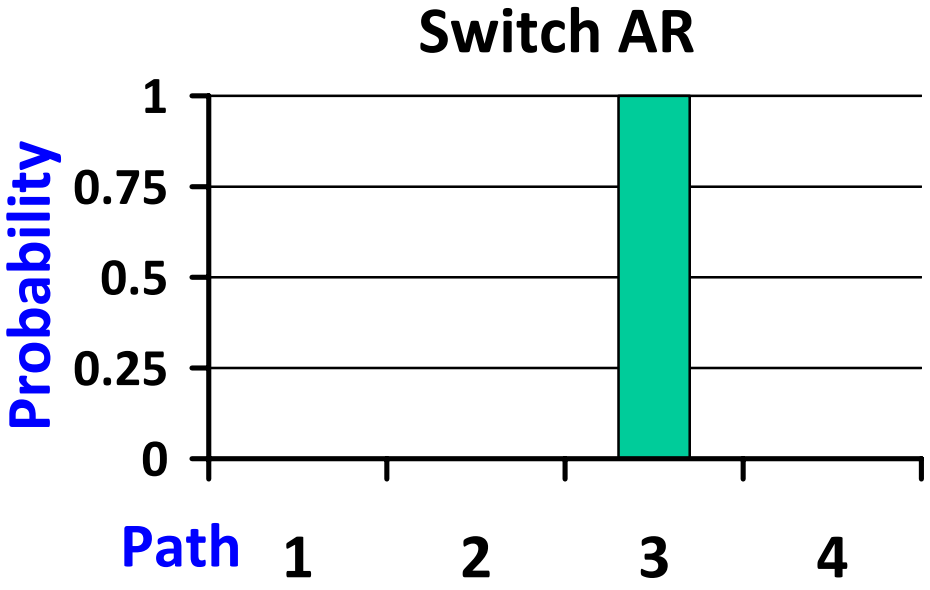


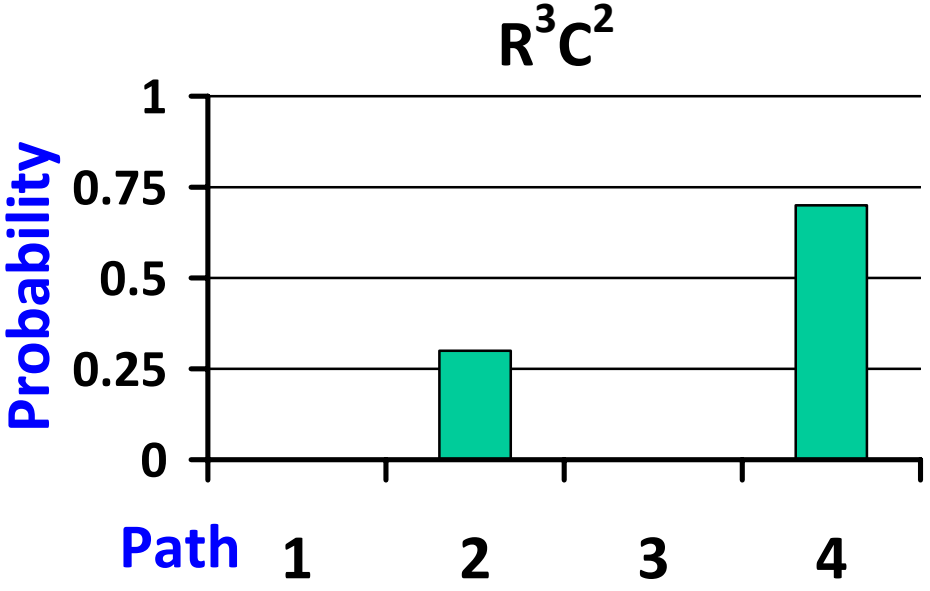


Routing Schemes









Evaluation Methodology

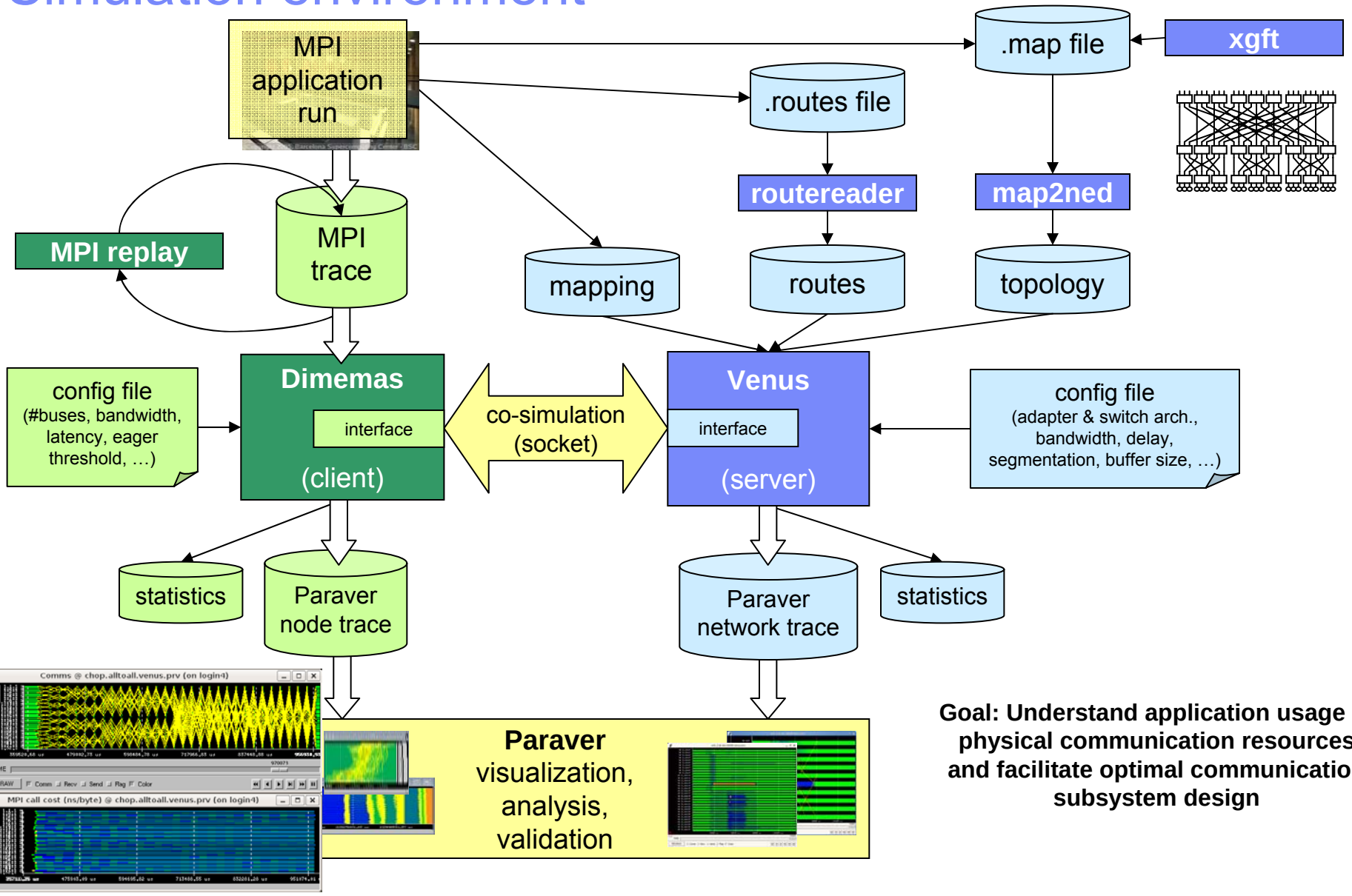
- Venus + Dimemas simulator

- Traffic
  - Synthetic: permutations + hotspot
  - HPC Traces:
    - NAS: BT, CG, FT, IS, MG
    - WRF, NAMD, Liso, Airbus

- Model parameters
  - 10Gbps CEE with MTU = 1500B
  - QCN and PFC: 802 DCB settings

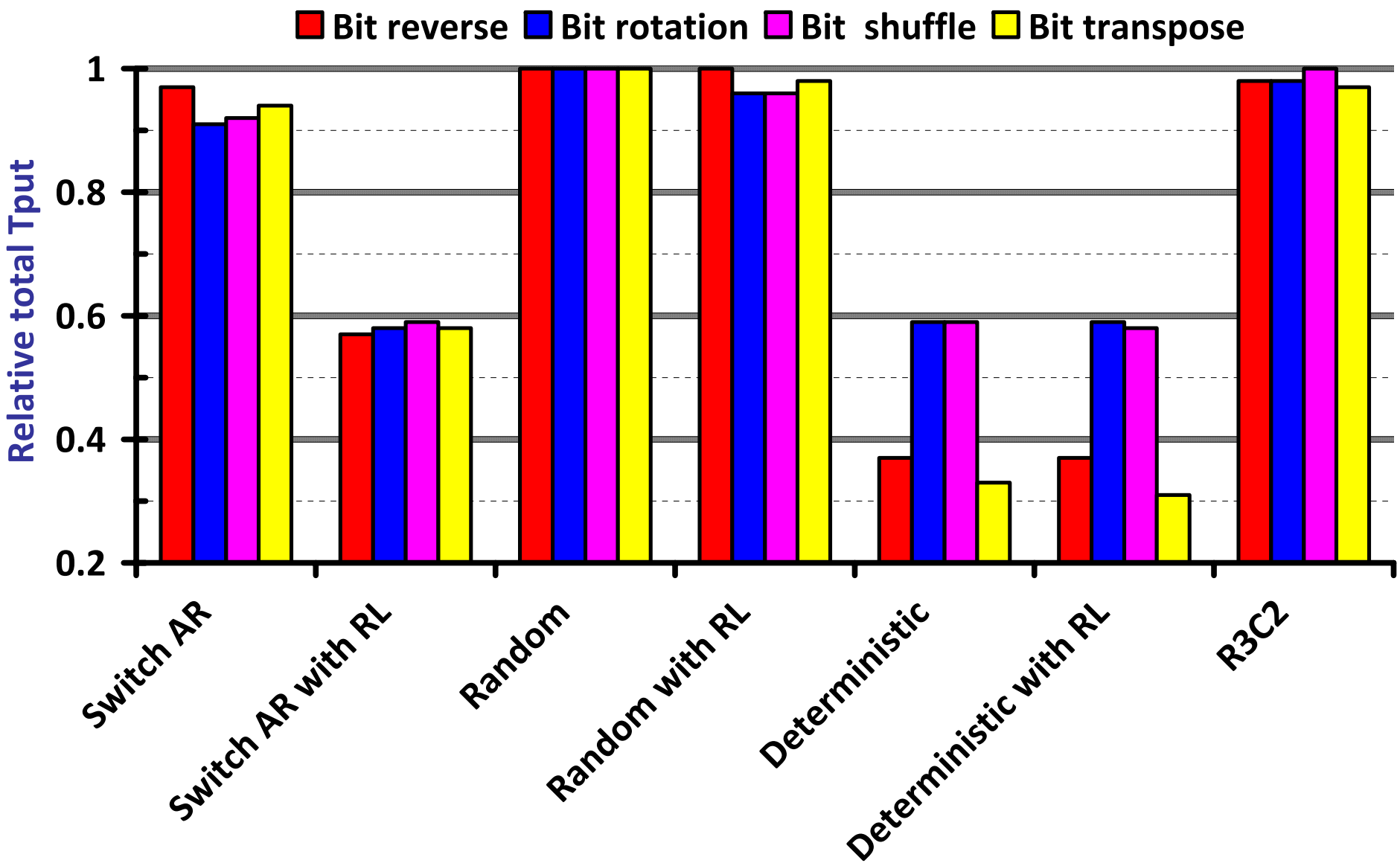
- Topology: 2-ary n-tree

Simulation environment

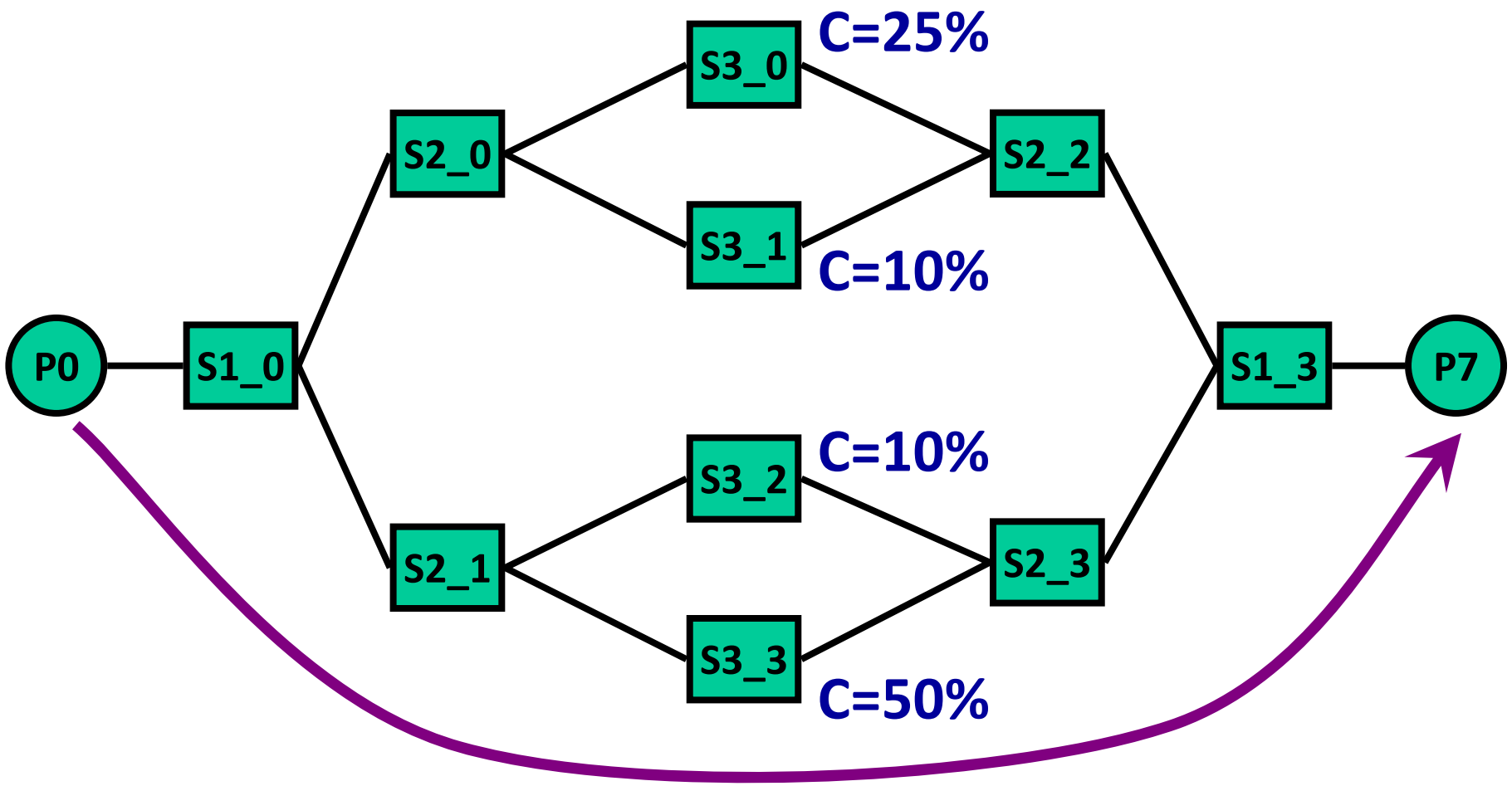


Goal: Understand application usage of physical communication resources and facilitate optimal communication subsystem design

Permutation Traffic

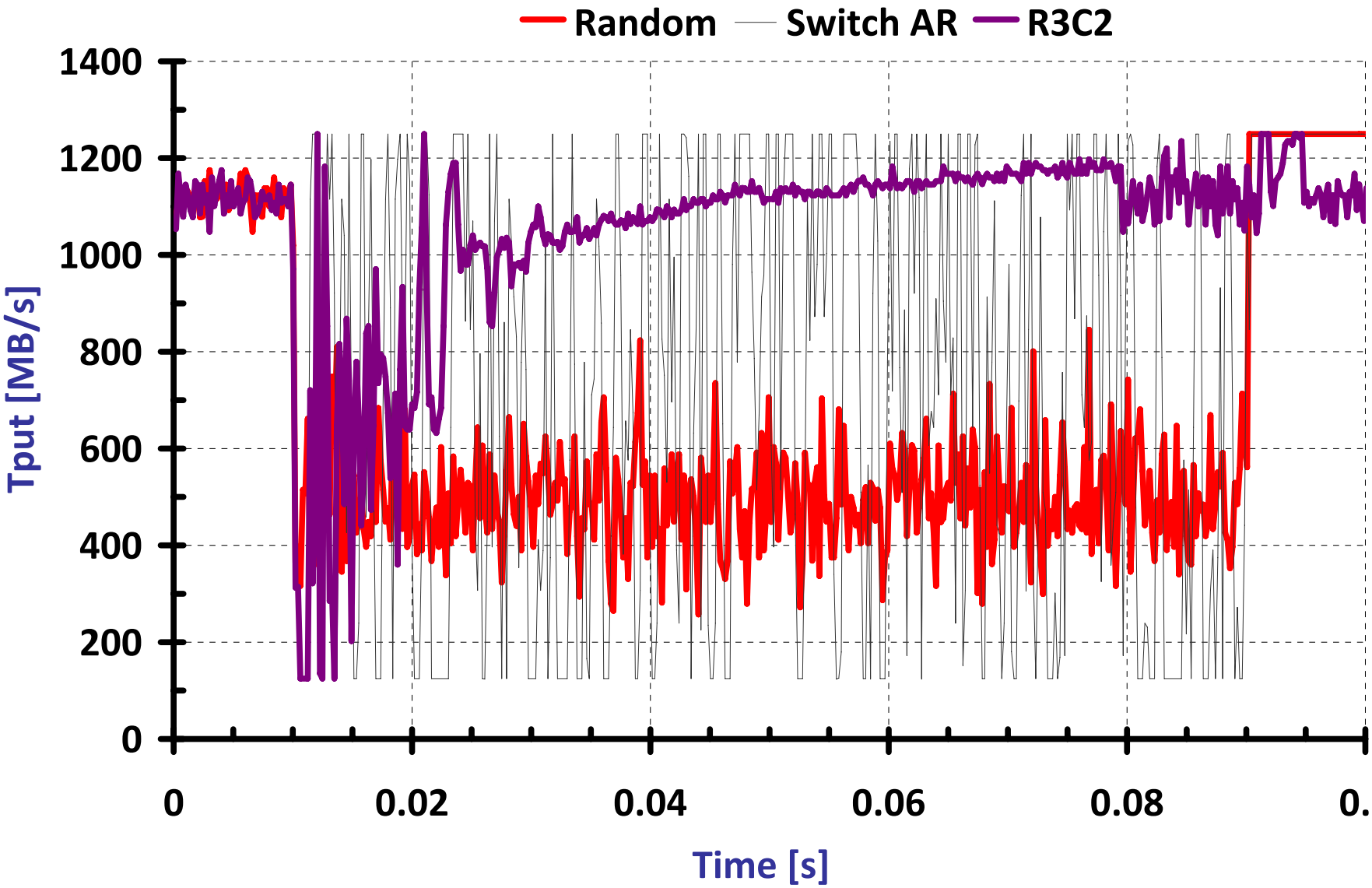


Hotspot Traffic Scenario

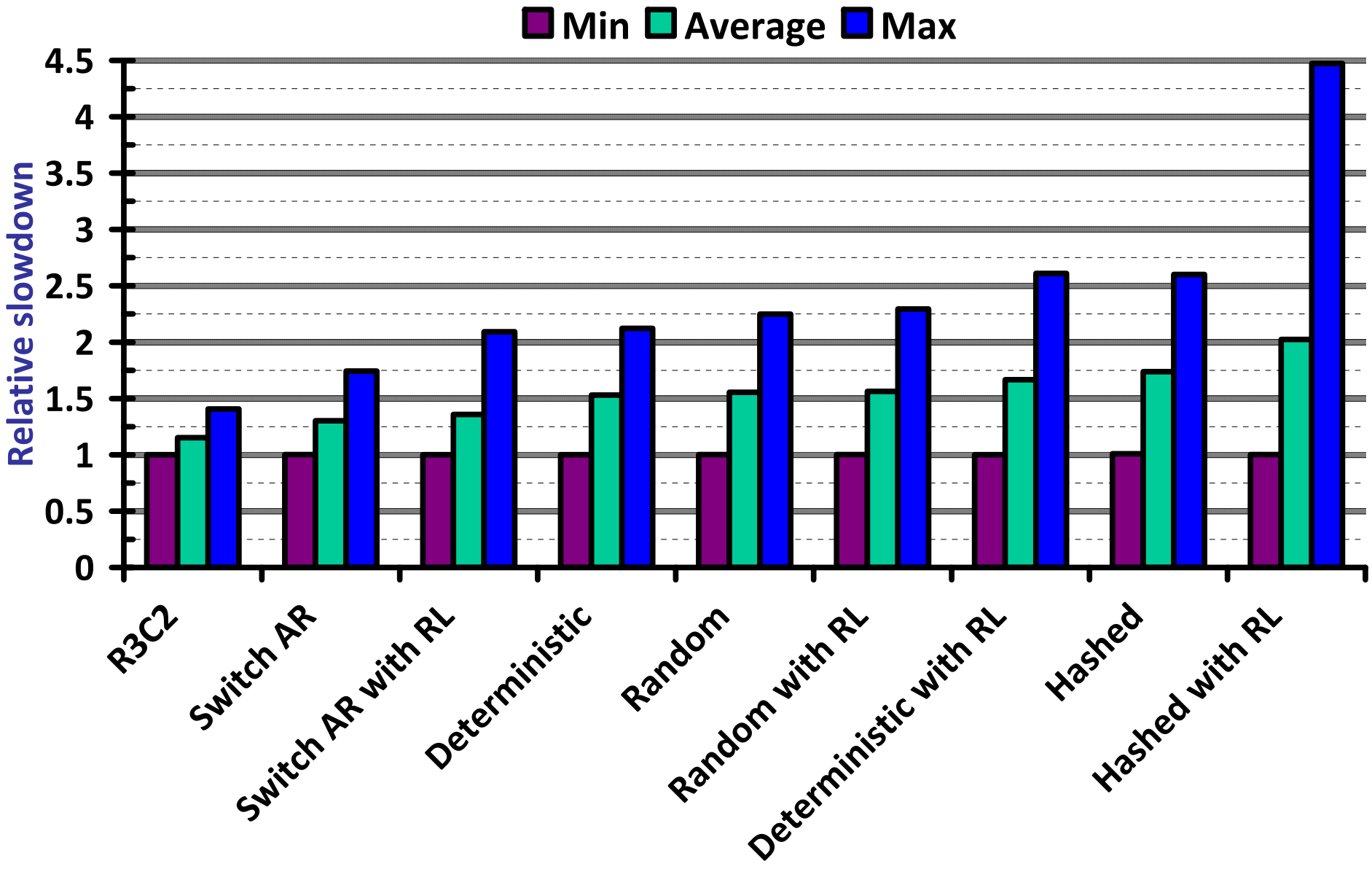


Single 95% flow

Hotspot Traffic



HPC Traces: Hotspot



Conclusions

- DCN offer abundant multipaths
  - Load balancing and reliability options
- Best routing: a qualified answer…
  1. D-mod-k deterministic: simple + no OOO delivery
  2. Random (-OOO) and hash: win under ideal DCN conditions, single prio, no failures or local overloads, w/ 'easy' traffic
  3. Adaptive (-OOO): best trade-off under realistic DCN scenarios… Performance benefits:

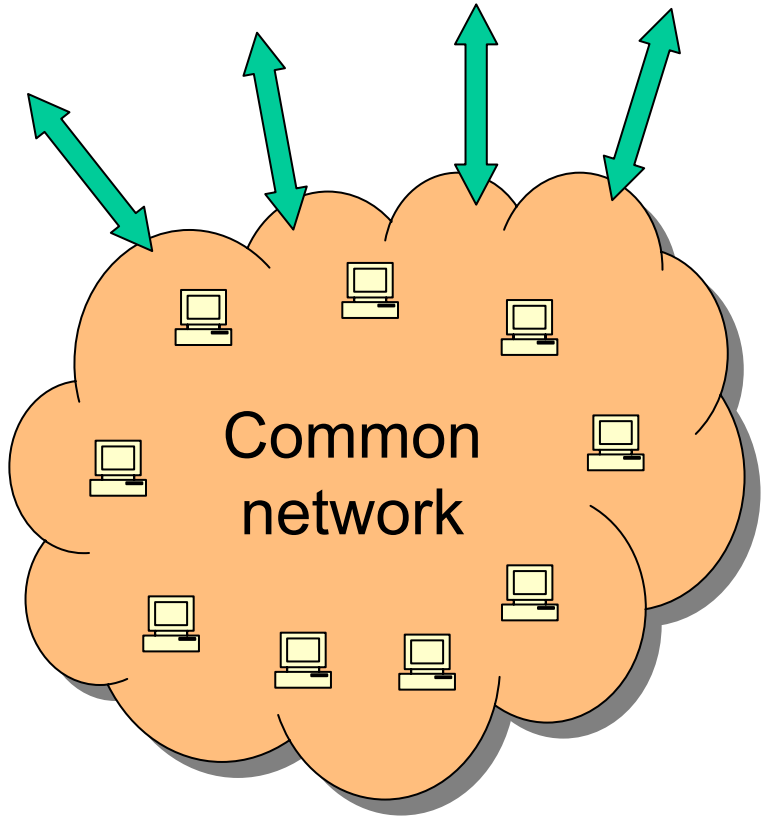
     80% over Deterministic

     40% over Random
- Rate or route ➔ Dual Route & Rate control
  - Improved stability and performance
- Open: ordering and additional cost vs. hashing

Backup

Datacenter Networks

- Google architecture
  - 10k node clouds - single network
- Node functions
  - storage – GFS chunkserver or BigTable tabletserver
  - processing – MapReduce worker
  - web server – GWS instance
  - controller – GFS master, BigTable master, MapReduce master
- Virtualization and convergence



[Barroso'03][Dean'04][Ghemawat'03]
[Chang'06][Barroso'09]

CEE-based Switch AR

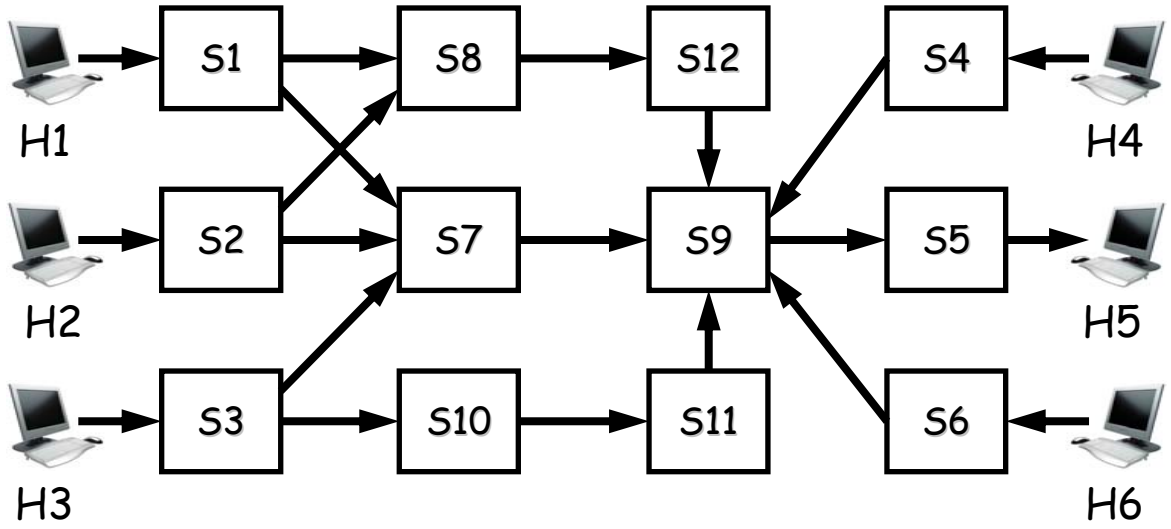
- Concept
  - Upstream switches **snoop** congestion notifications,
  - **annotate** routing tables with congestion information, and
  - **modify routing decisions** to route to the **least congested** port among those enabled for a given destination
- Routing table
  - Maps a destination MAC to one or more switch port numbers, listed in order of preference, e.g., shortest path first
- Congestion table
  - Maps a key **<**destination MAC, switch port number**>** to a congestion entry comprising the following information:
- Receiver checks frame order and performs resequencing if needed

field	type	meaning
congested	boolean	Flag indicating whether port is congested
local	boolean	Flag indicating whether congestion is local or remote
fbCount	integer	Number of notifications received
feedback	integer	Feedback severity value

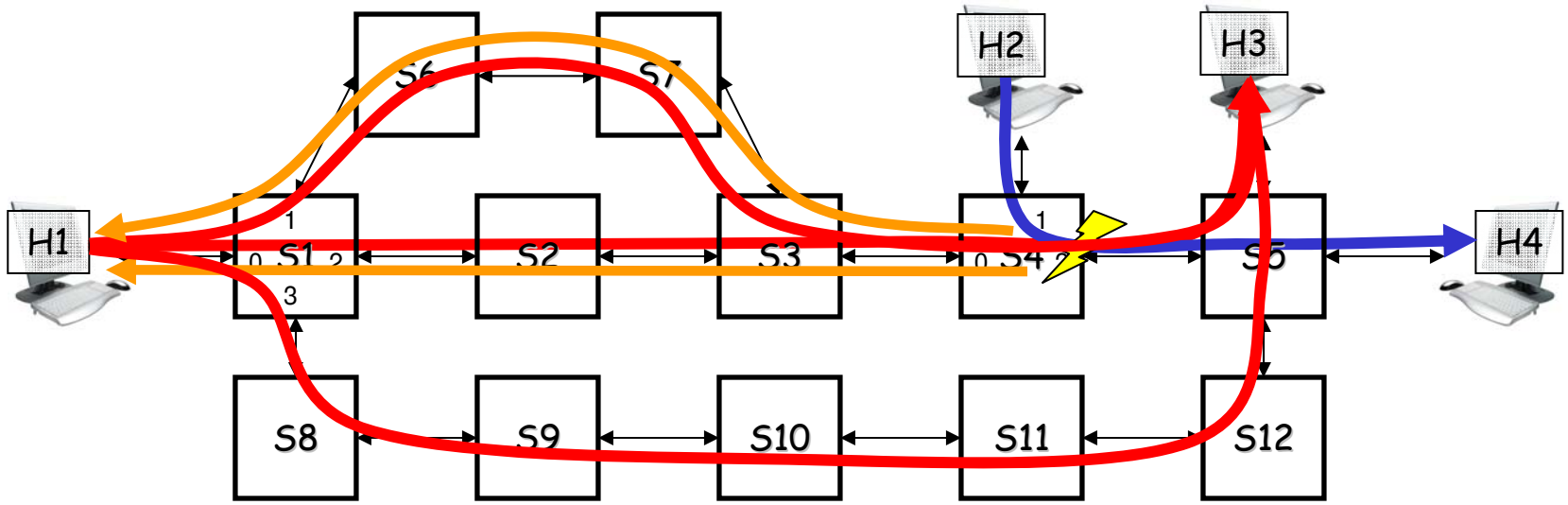
Routing decisions & configuration

- For a frame destined to MAC address *d*
  - try eligible ports in order of preference
  - select first port not flagged as congested
  - if all ports flagged as congested, route to port with minimum *feedback* value

- To ensure *productive* and *loop-free* routing without deadlocks
  - For each destination node *n*, construct a directed acyclic graph connecting all nodes ≠ *n* to node *n*

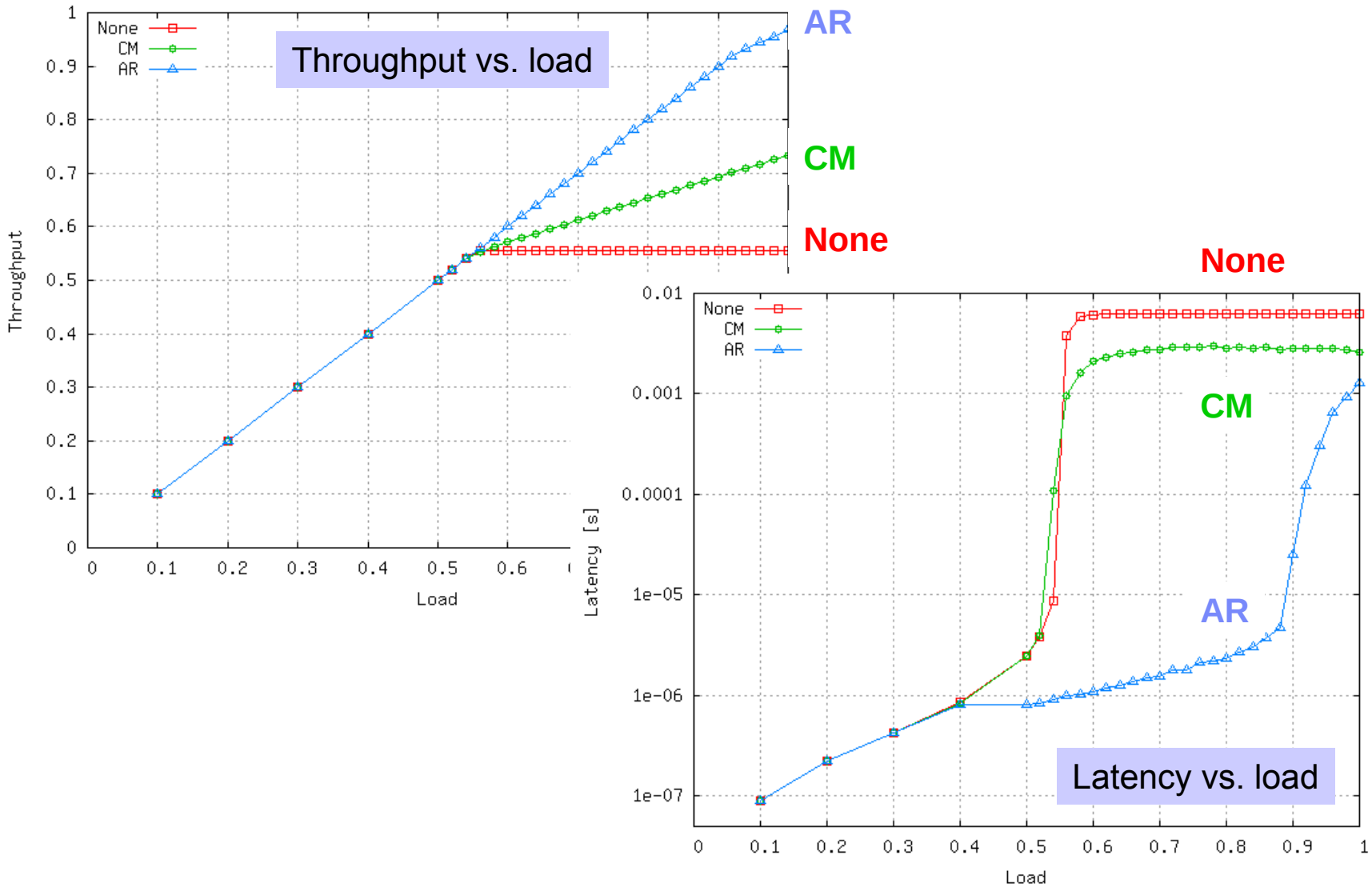


Routing of congestion notifications



- Congestion notifications need to be routed on all alternative paths leading to the CP
- When generated at CP, notification is routed to port on which sampled frame arrived
- In upstream switches, notification is routed on a random port leading to the sampled flow's source

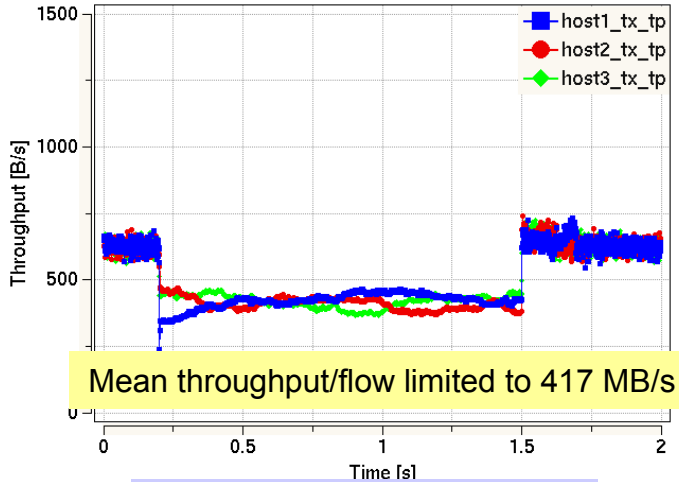
Rate/CM vs. Route/AR: Bernoulli Traffic Simulation



Results with 3-flow congestion scenarios

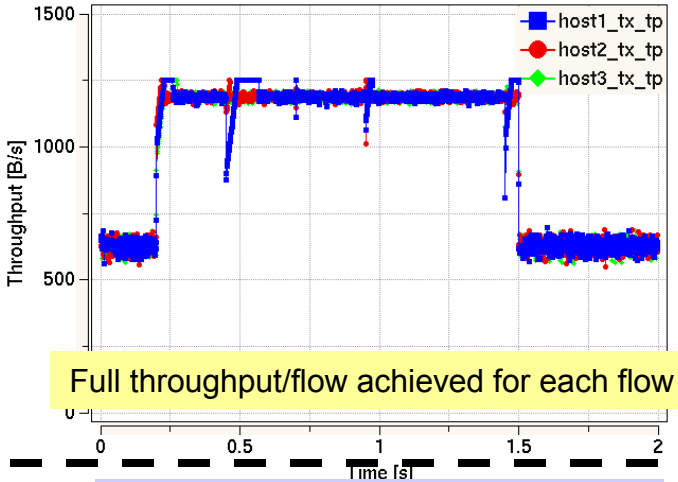
Without end-point contention

Flow throughput without adaptive routing



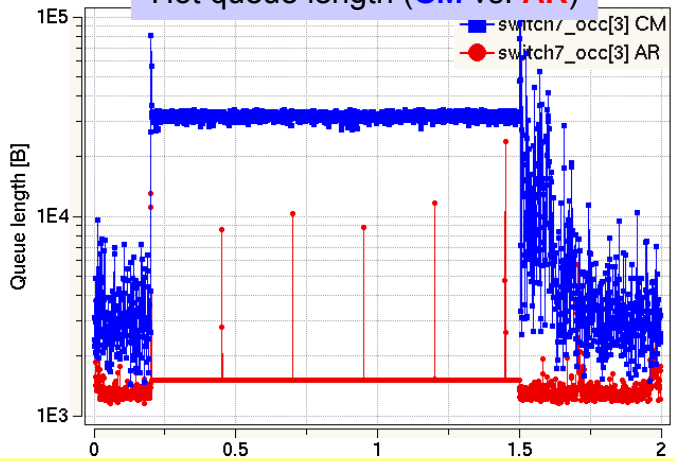
Mean throughput/flow limited to 417 MB/s

Flow throughput with adaptive routing



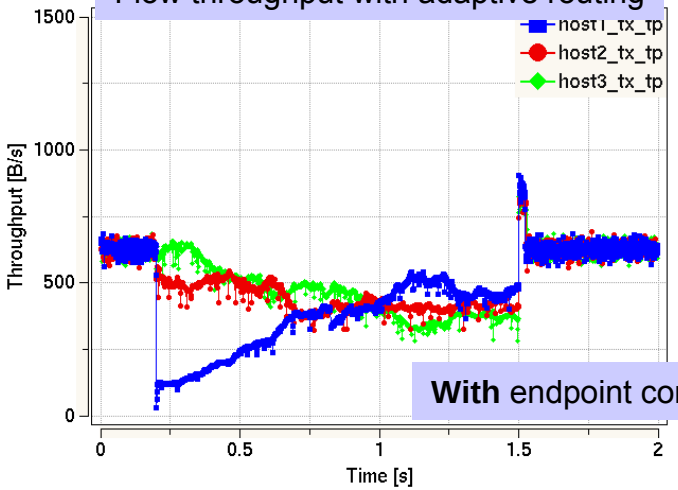
Full throughput/flow achieved for each flow

Hot queue length (**CM** vs. **AR**)



With CM, hot queue is controlled around equilibrium
With AR, hot spot disappears owing to re-routing
Reset spikes every 250 ms clearly visible

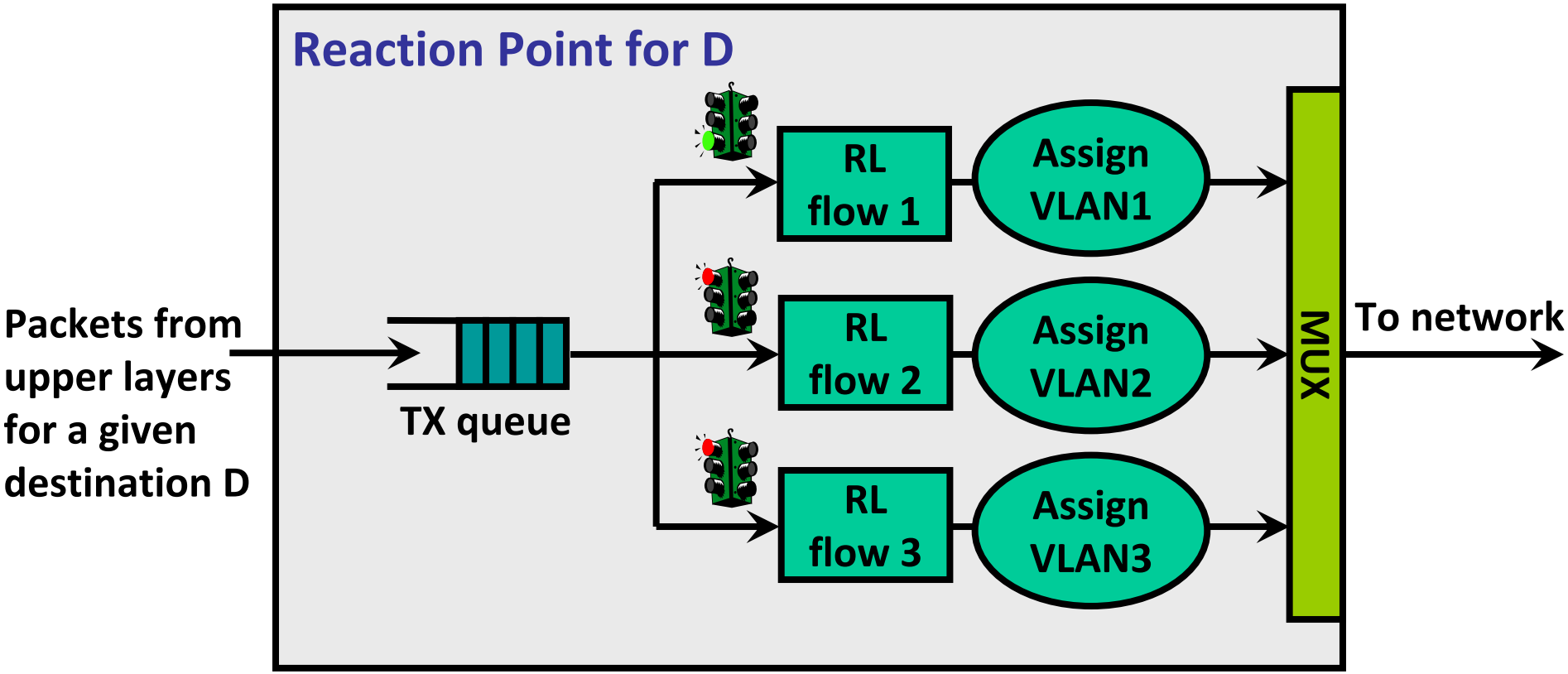
Flow throughput with adaptive routing



With endpoint contention

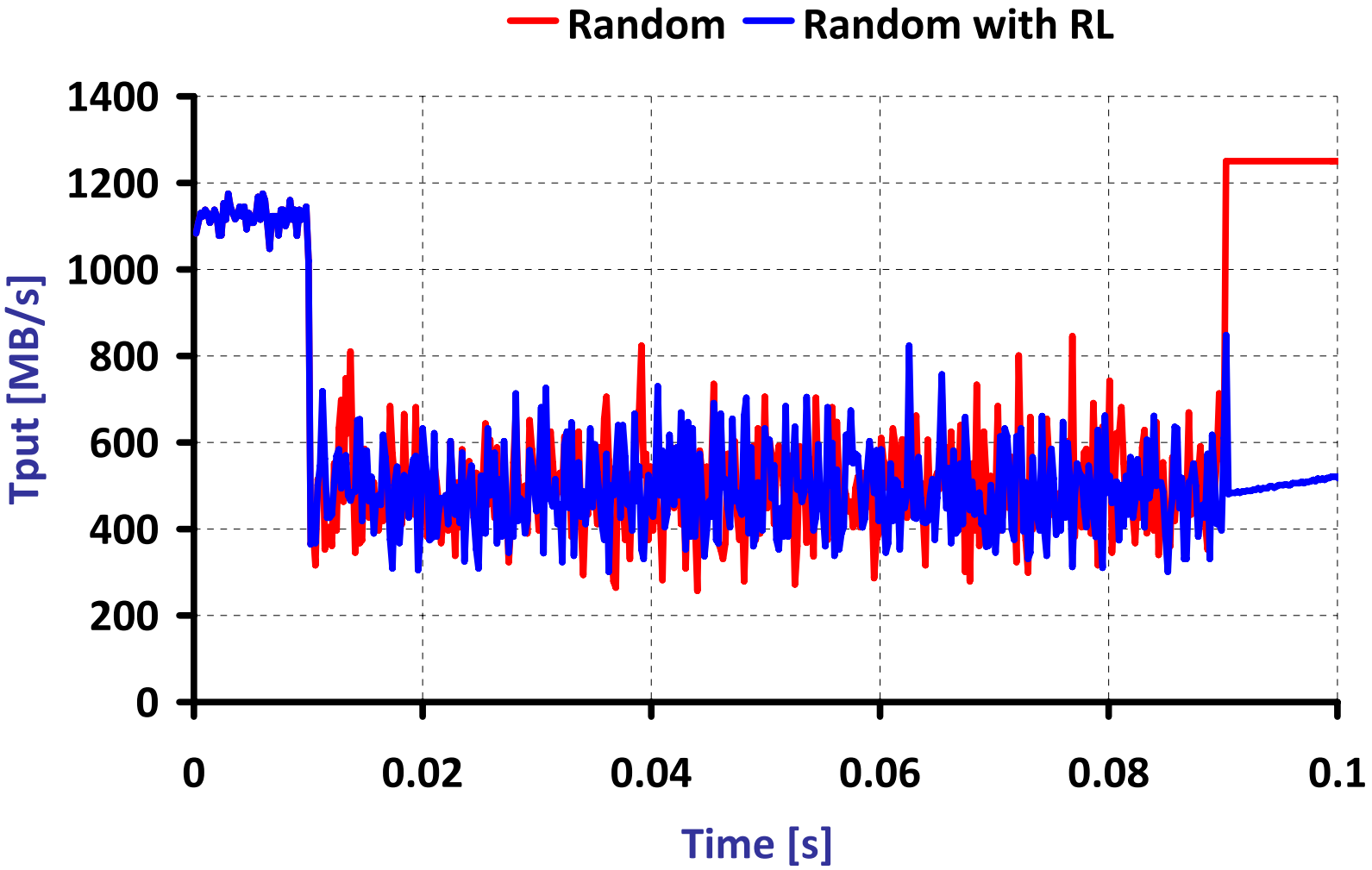
Hot queue lengths are controlled, indicating that CM still works well in conjunction with AR

R^3C^2 Reaction Point

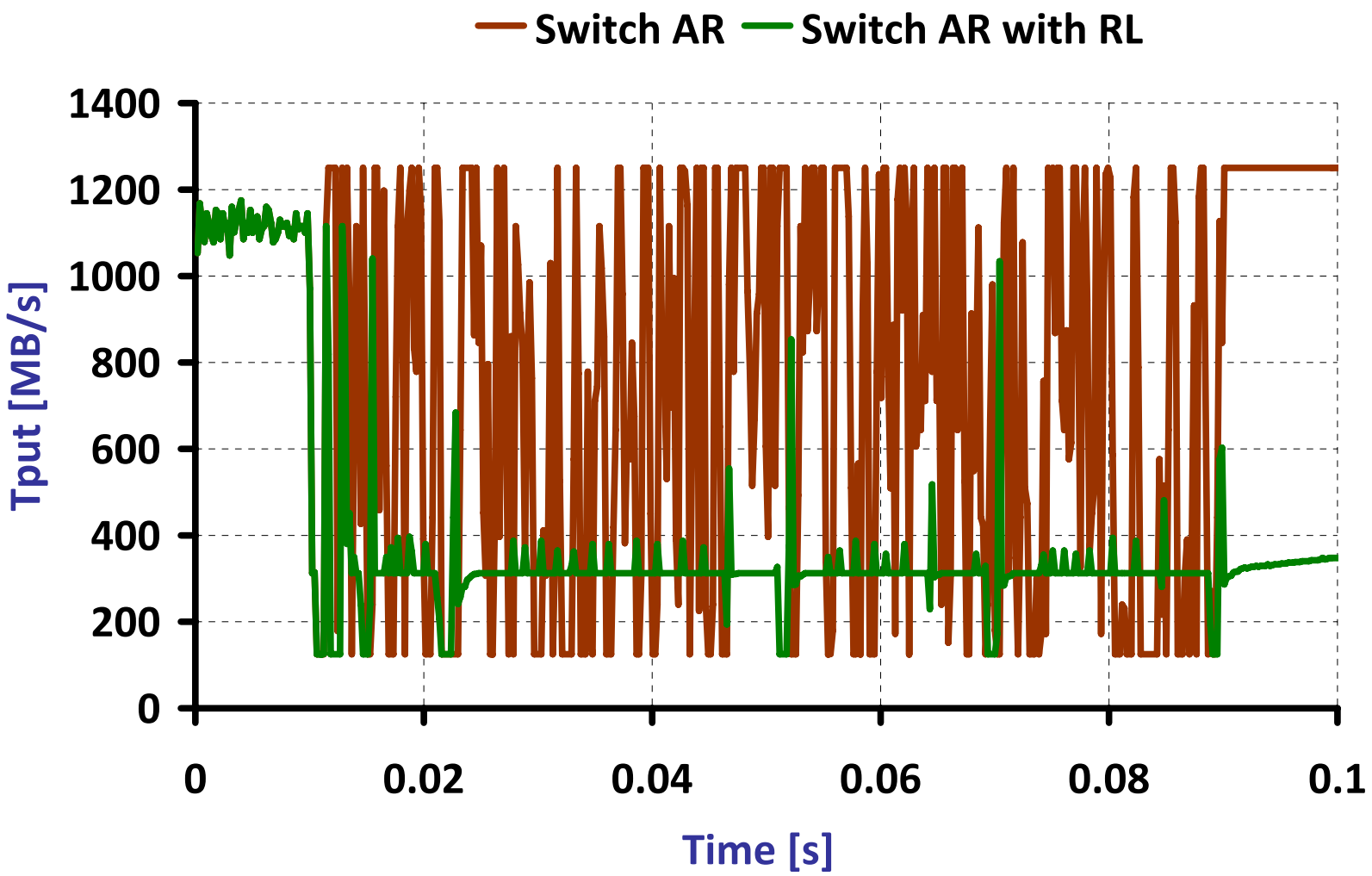


- Packet assigned the VLAN# of the 1st eligible Rate Limiter

Hotspot Traffic (1)



Hotspot Traffic (2)



Hotspot Traffic (3)