MAY 18-20 2021 P4 Workshop

Hosted by ONF

PL2: Towards Predictable Low Latency in Rack-Scale Networks

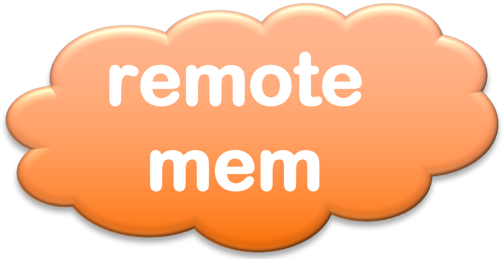
Yanfang Le, Software Engineer,
UW-Madison & Intel

UW-Madison: Aditya Akella, Michael Swift
VMware Research: Radhika Niranjan Mysore, Lalith Suresh, Gerd Zellweger, Sujata Banerjee

Rack-Scale Architecture

Rack-Scale Architecture

Rack-Scale Architecture



Rack-Scale Architecture

Intel® Rack Scale Design

NetCache

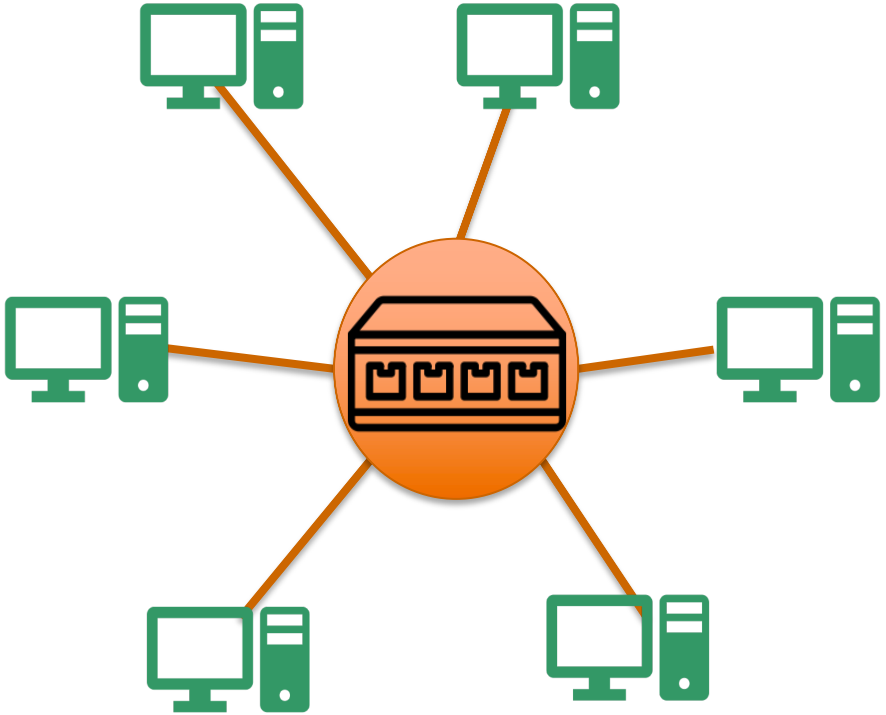
remote mem

Rack-Scale Architecture

Intel® Rack Scale Design

NetCache

remote mem



Motivation

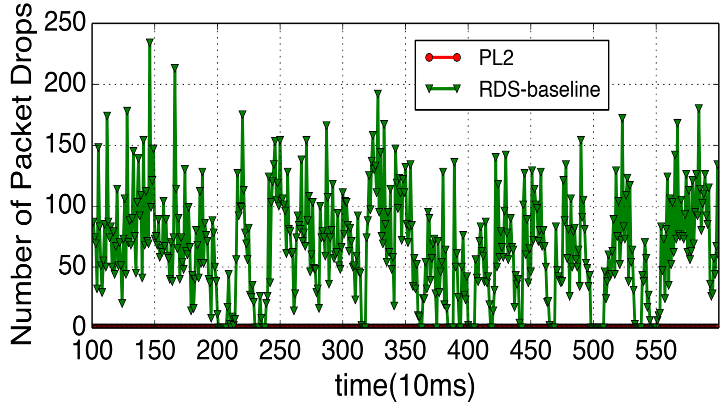
- Network densities increases, switch buffer does not increase as fast
- Incast ratio is up to 5000:1[**Swift SIGCOMM’20**]

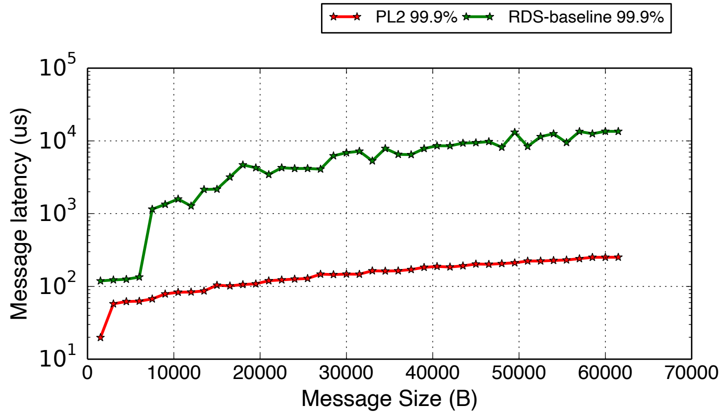
Motivation

- Network densities increases, switch buffer does not increase as fast
- Incast ratio is up to 5000:1[**Swift SIGCOMM’20**]
- Packets drop caused by incast
  - Line rate increases , RTT does not get smaller, RTT bytes increases
  - State-of-the-art CC starts at line rate
  - Drops cause the most noticeable tails

Motivation

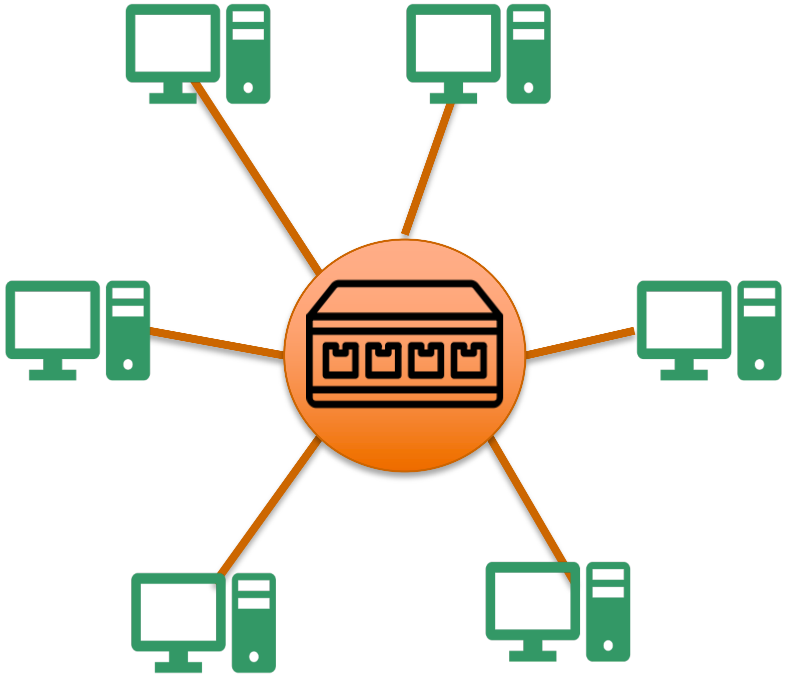
- Network densities increases, switch buffer does not increase as fast
- Incast ratio is up to 5000:1[**Swift SIGCOMM’20**]
- Packets drop caused by incast
  - Line rate increases , RTT does not get smaller, RTT bytes increases
  - State-of-the-art CC starts at line rate
  - Drops cause the most noticeable tails





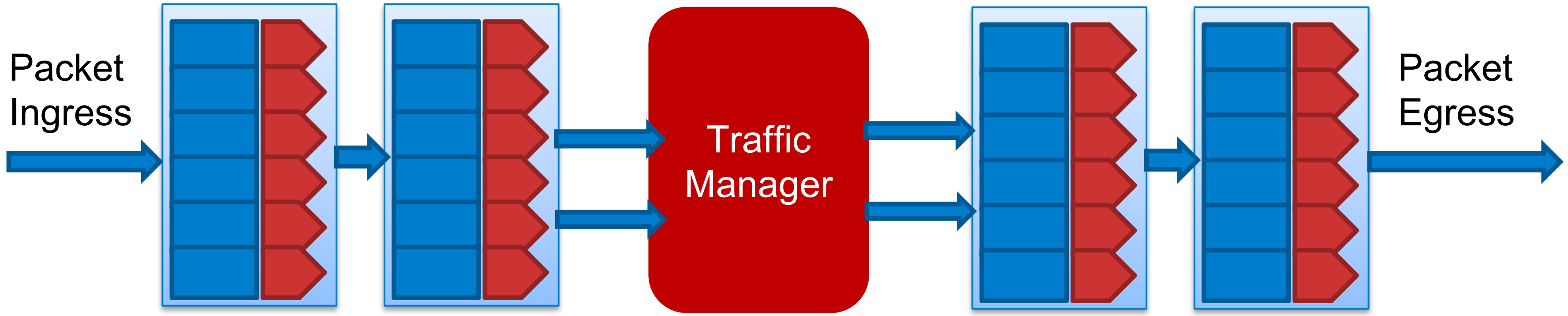
Architecture for Low Latency Network

- A single (ToR) switch to which all hosts are connected
  - Global knowledge of the demand
  - Global coordination across end-points



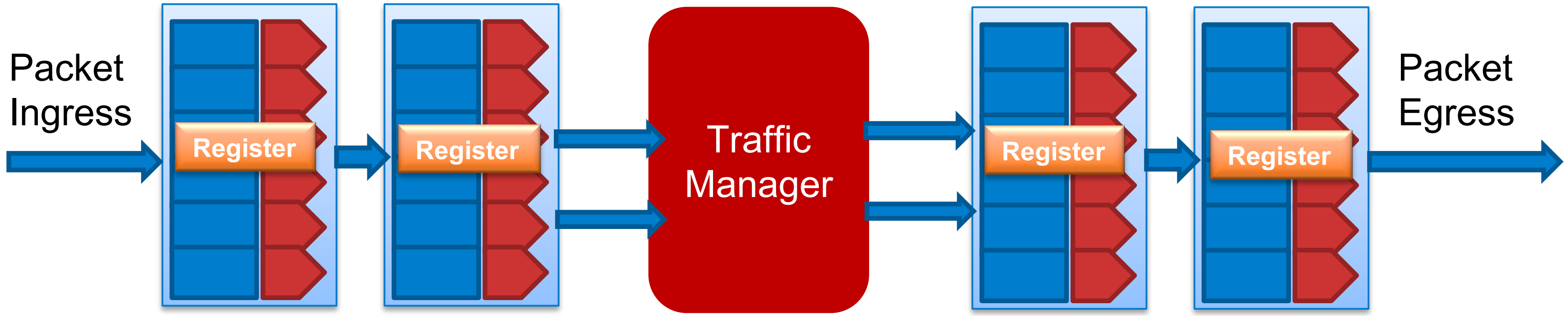
Programmable Switch

- Programmable switch offers in-transit packet processing and in-network state



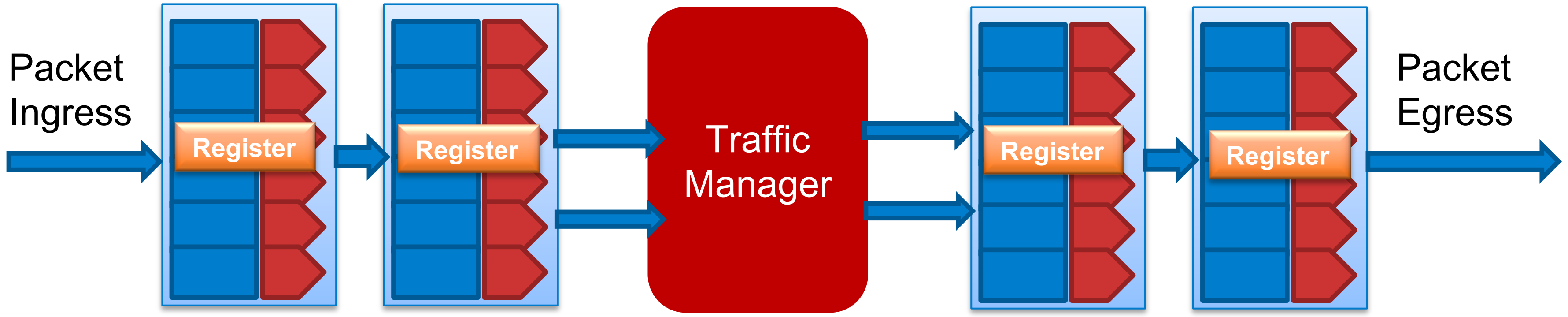
Programmable Switch

- Programmable switch offers in-transit packet processing and in-network state



Programmable Switch

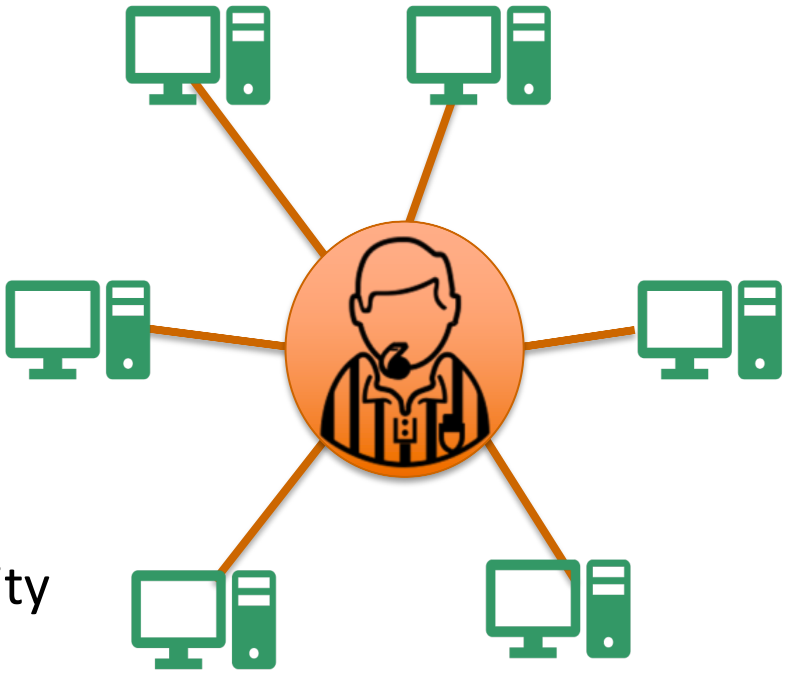
- Programmable switch offers in-transit packet processing and in-network state



- Programmable switch as the centralized packet scheduler

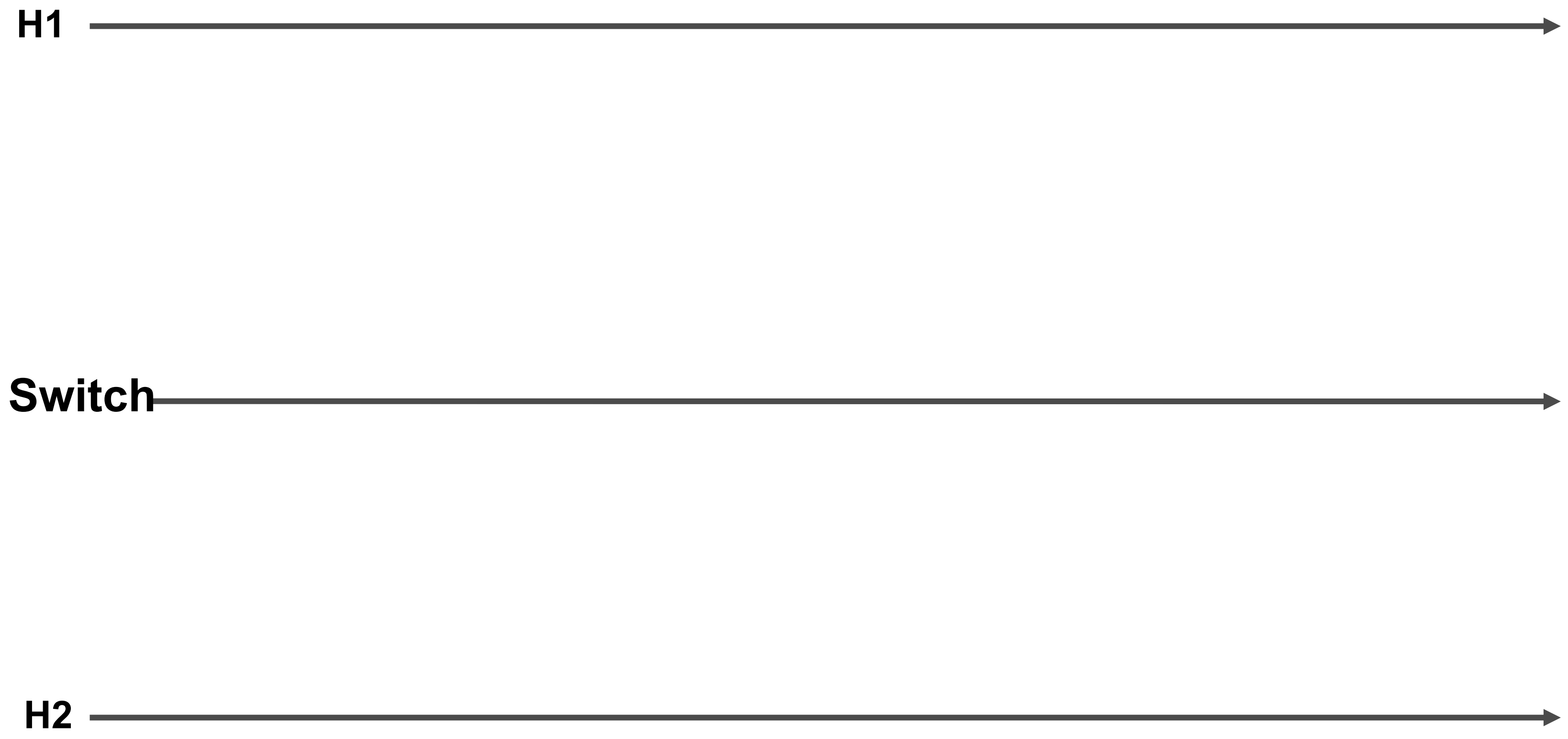
PL2: Centralized Scheduler

- Switch acts as an arbiter
  - Global visibility into packet reservation requests from all the endpoints
- Switch algorithm should fit
  - microseconds-level scheduling overhead
  - switch architecture, memory capacity
  - restricted programming model and memory access model



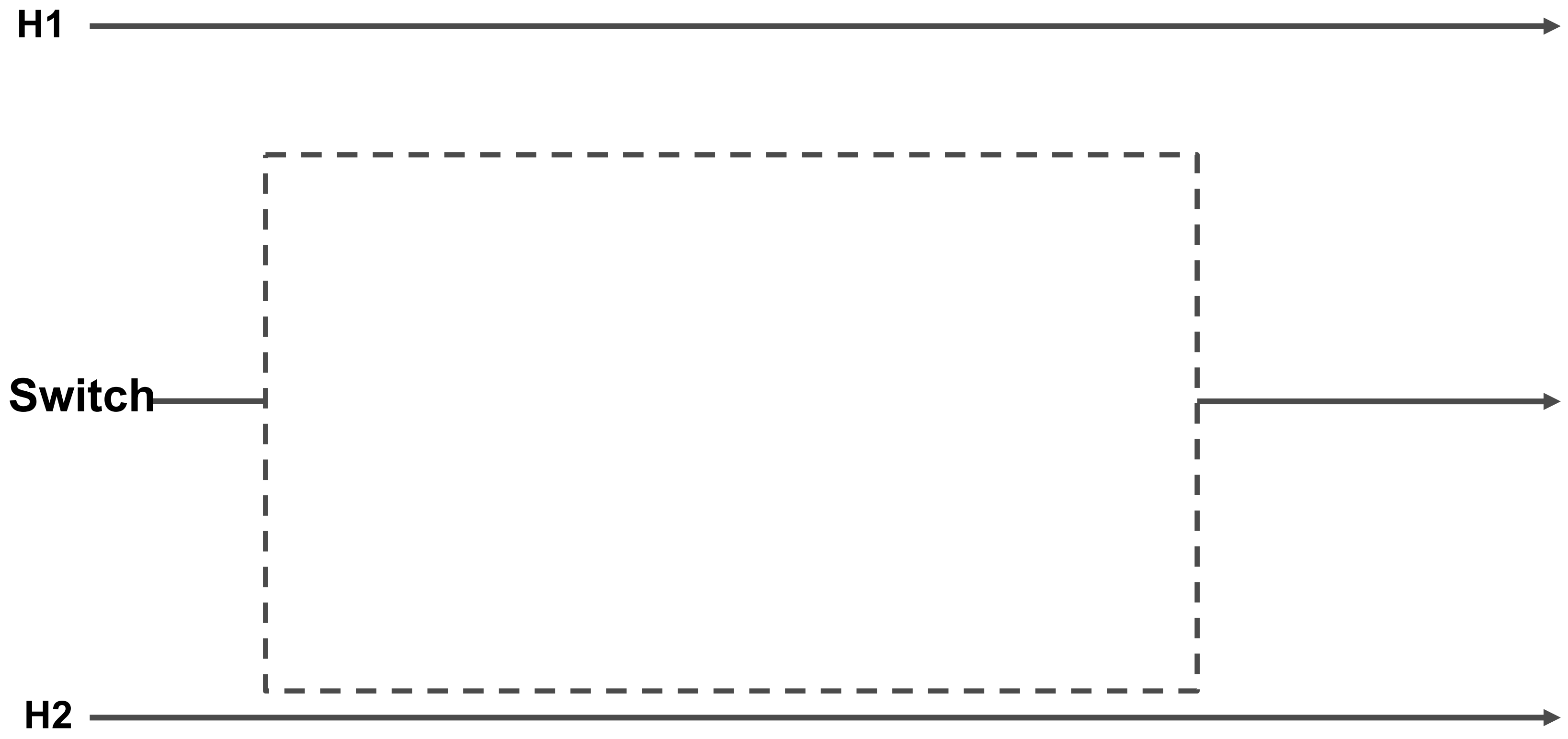
PL2 Overview

🟥 **Reserved timeslot** ☐ **Available timeslot**

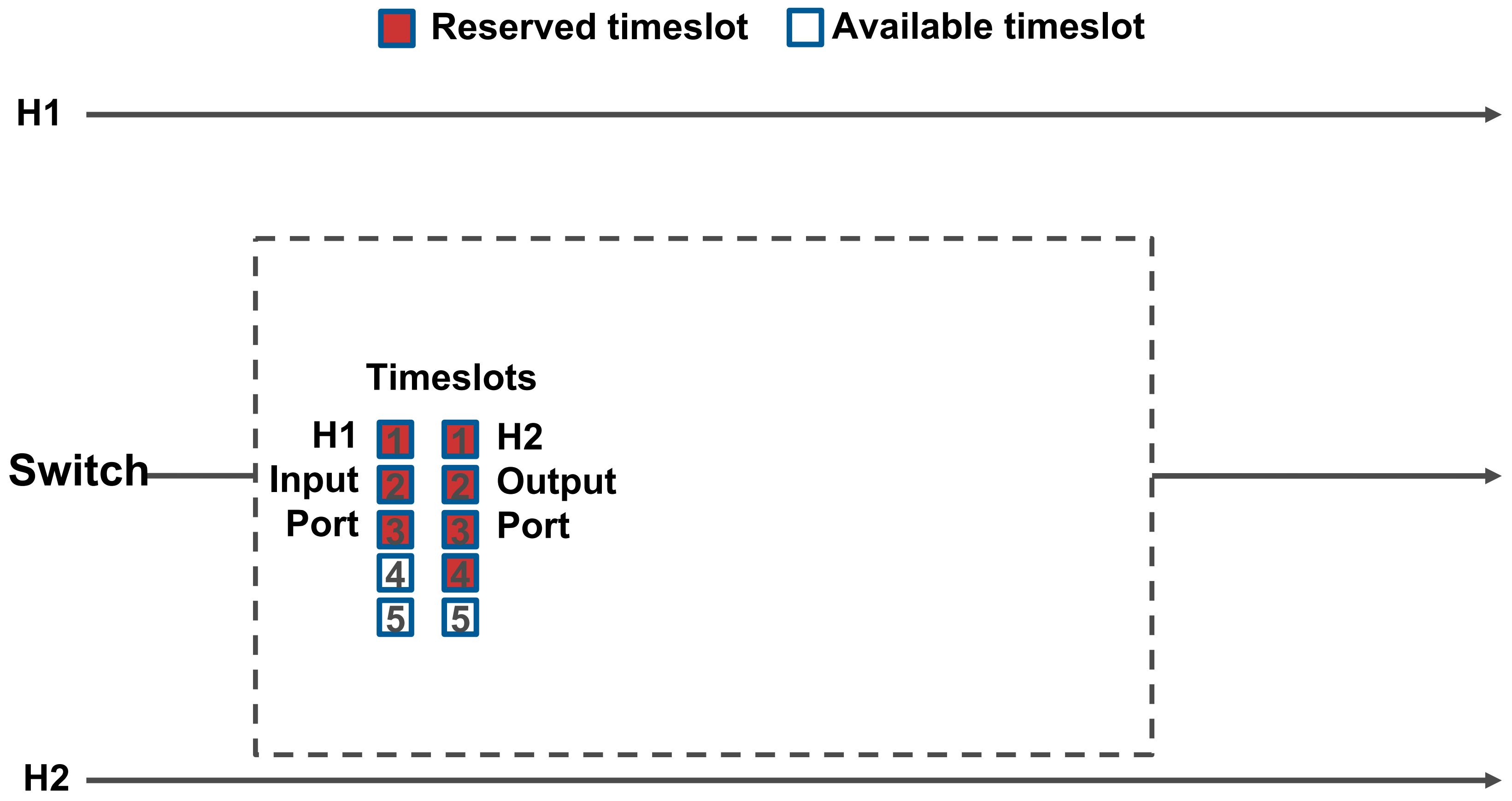


PL2 Overview

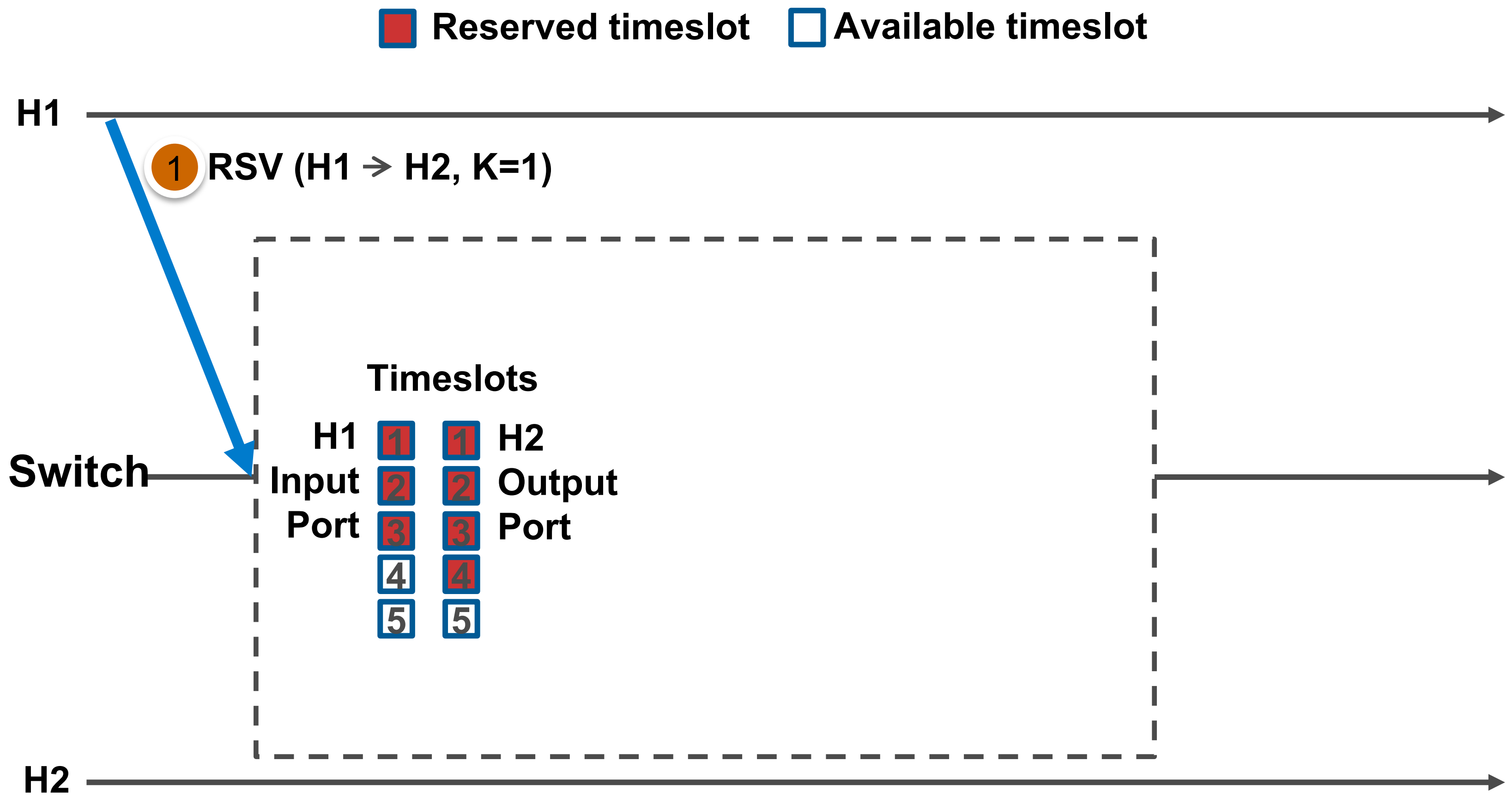
🟥 Reserved timeslot ☐ Available timeslot



PL2 Overview

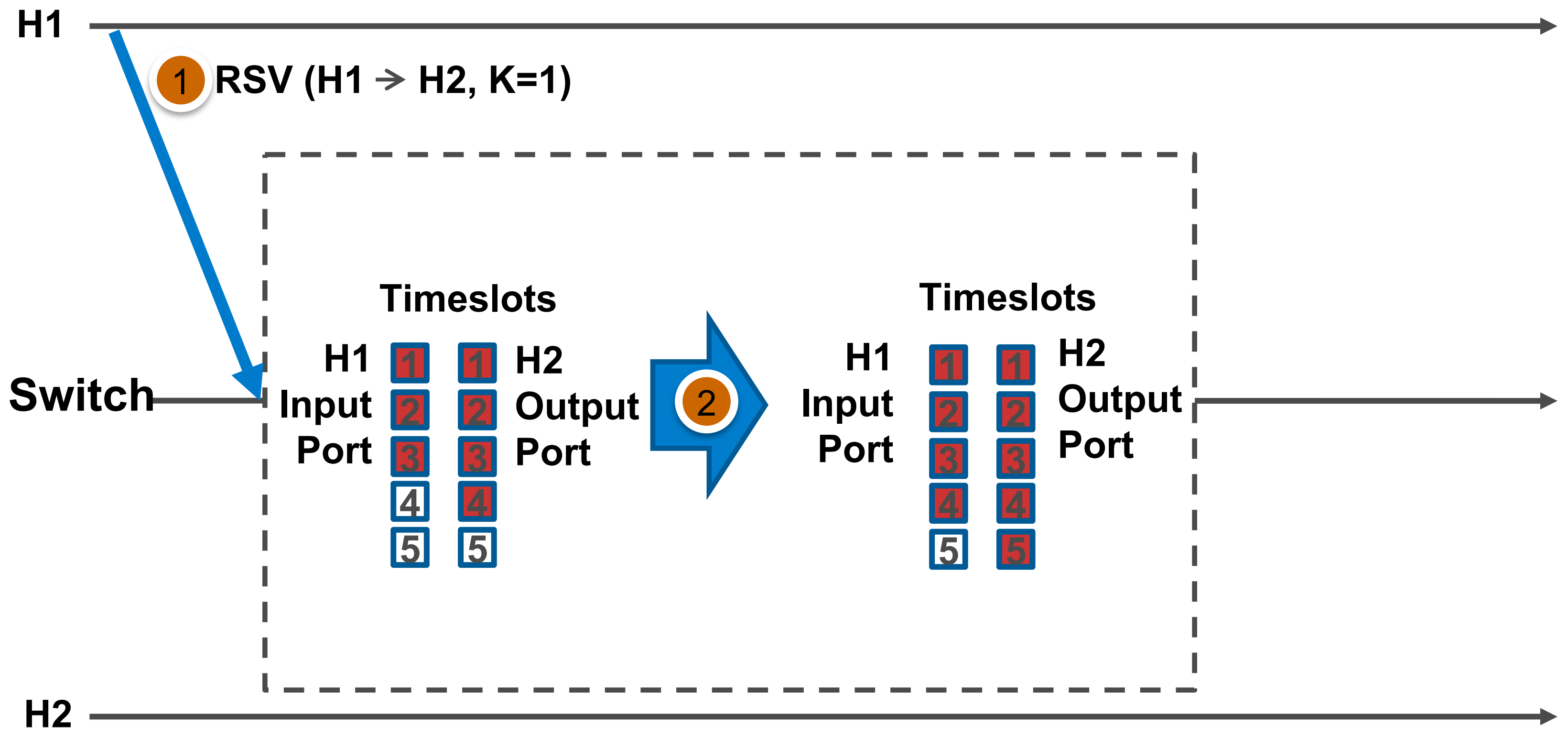


PL2 Overview

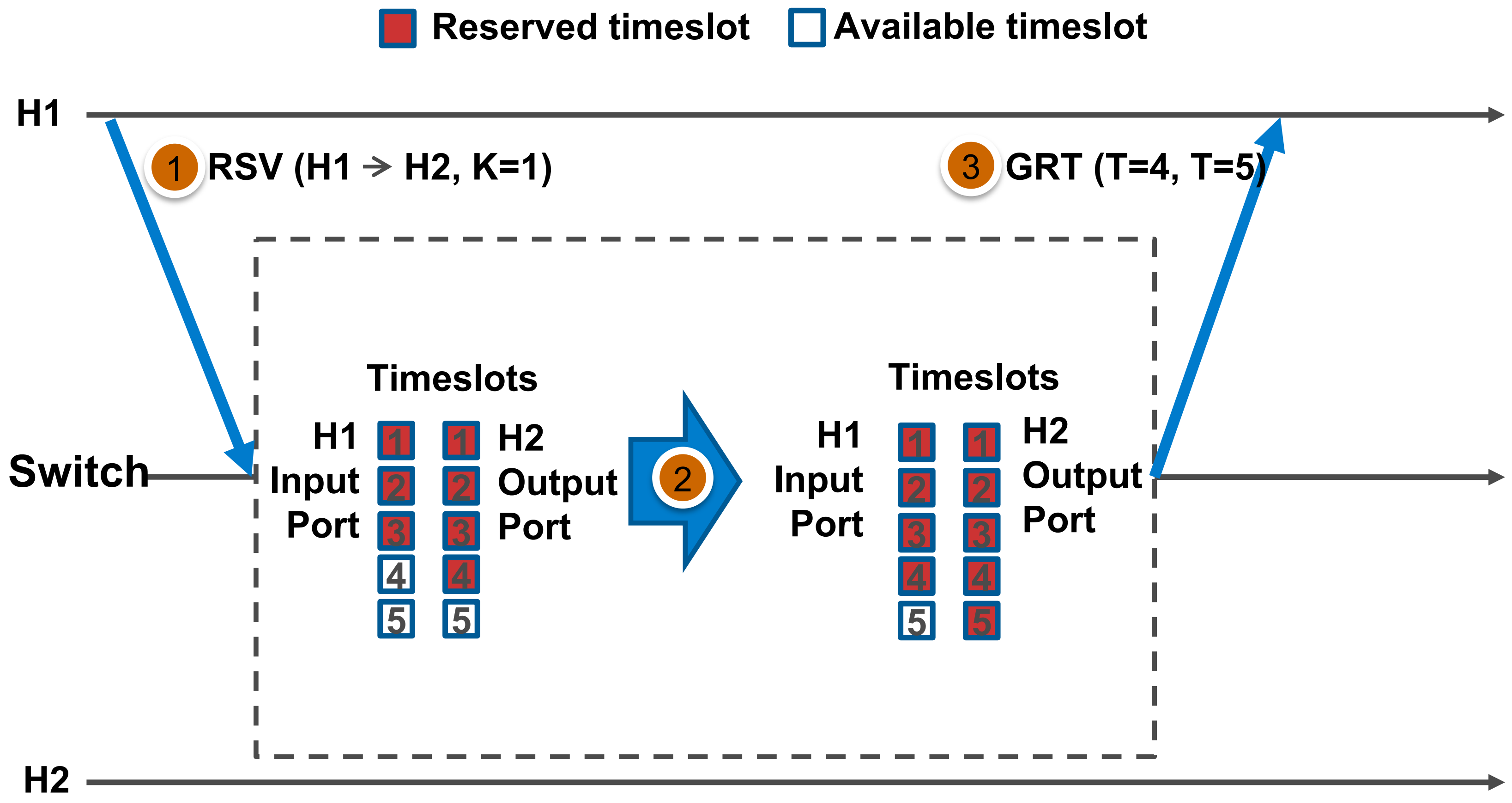


PL2 Overview

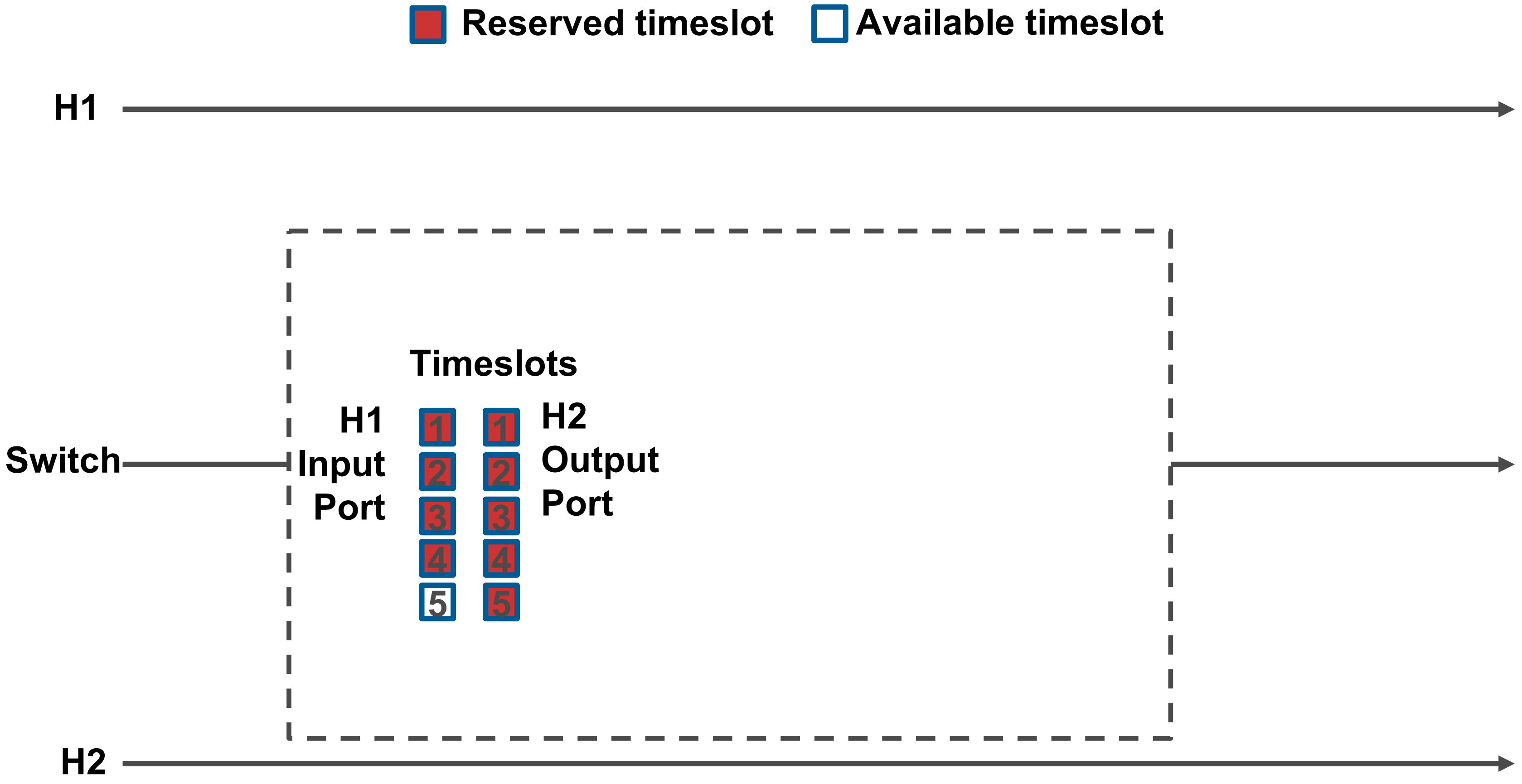
🟥 Reserved timeslot ☐ Available timeslot



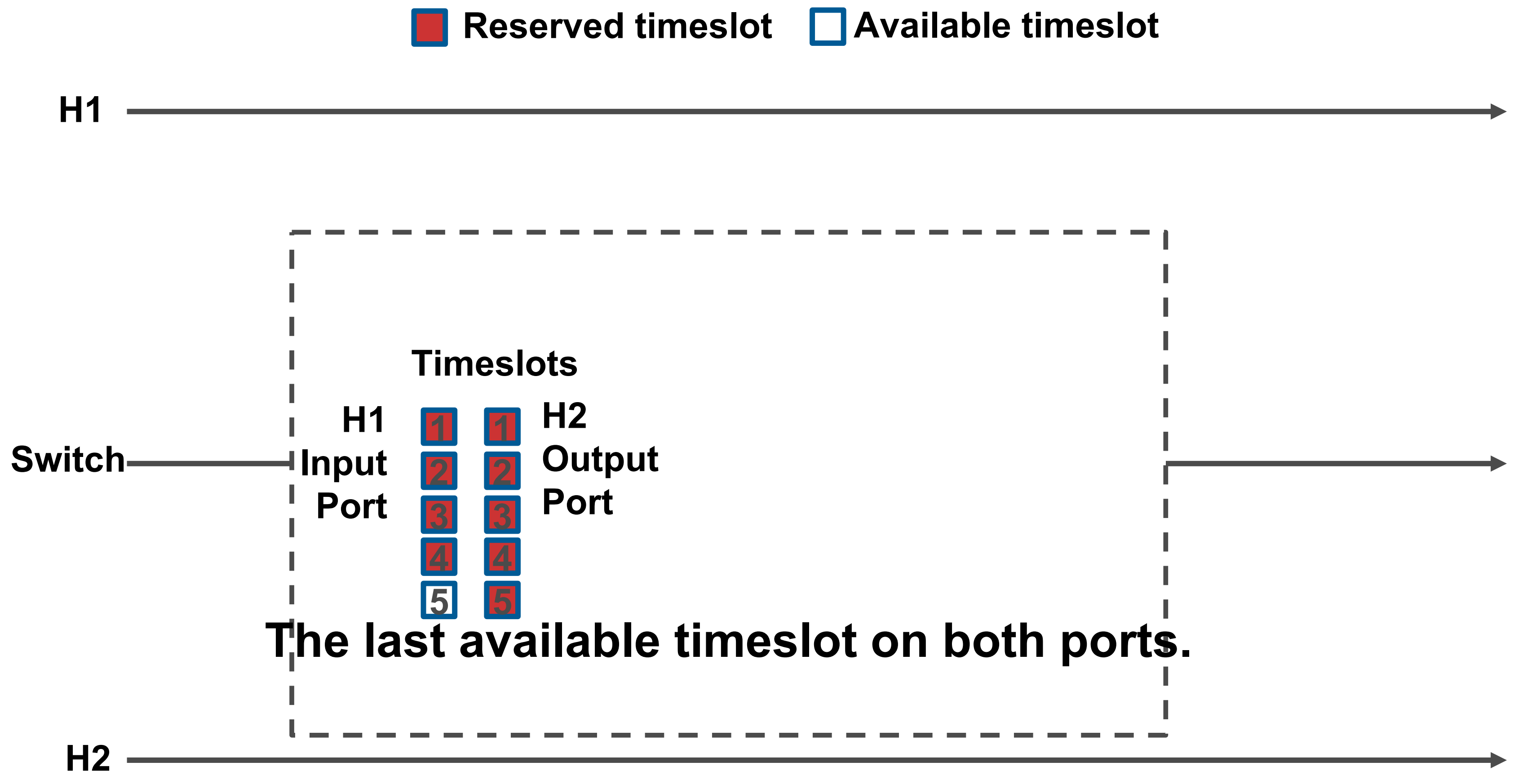
PL2 Overview



PL2 Overview



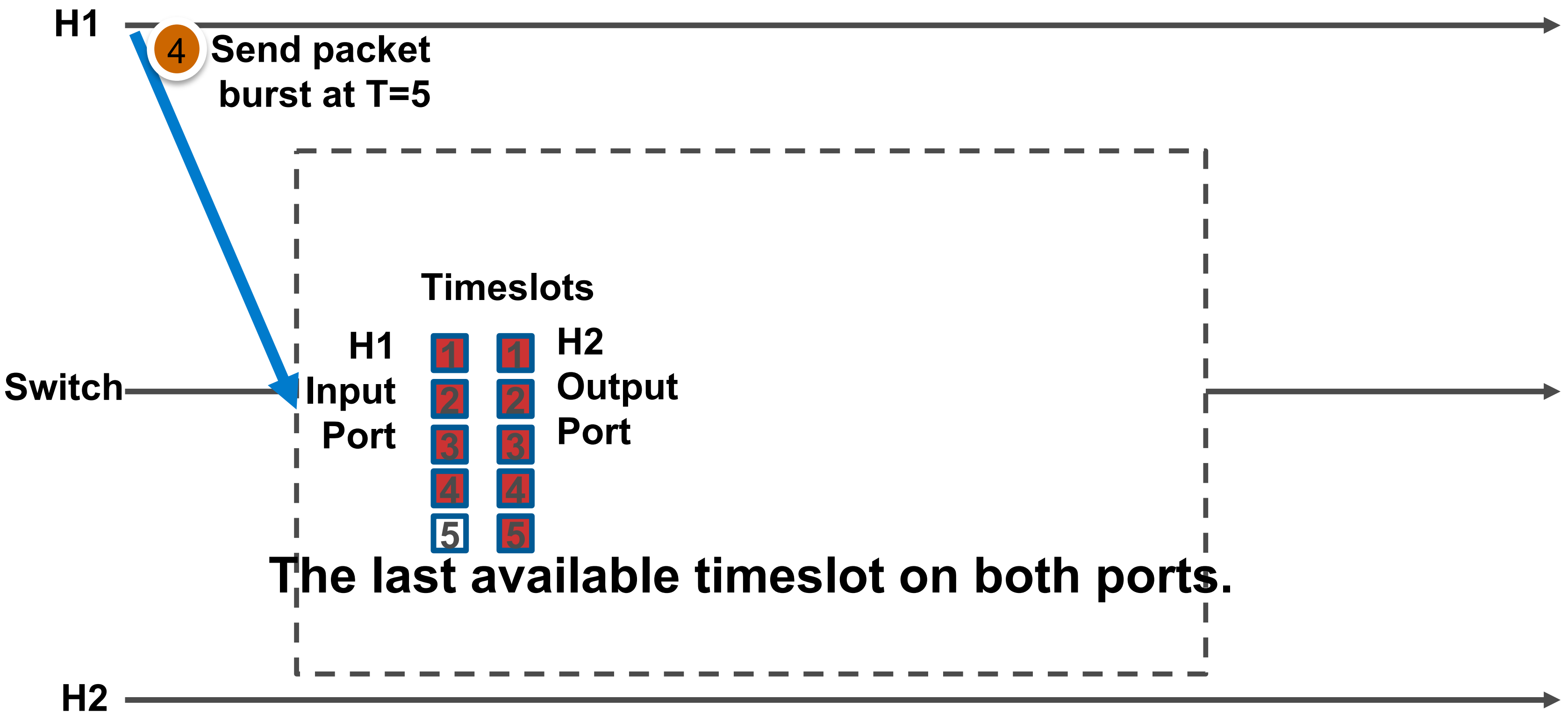
PL2 Overview



The last available timeslot on both ports.

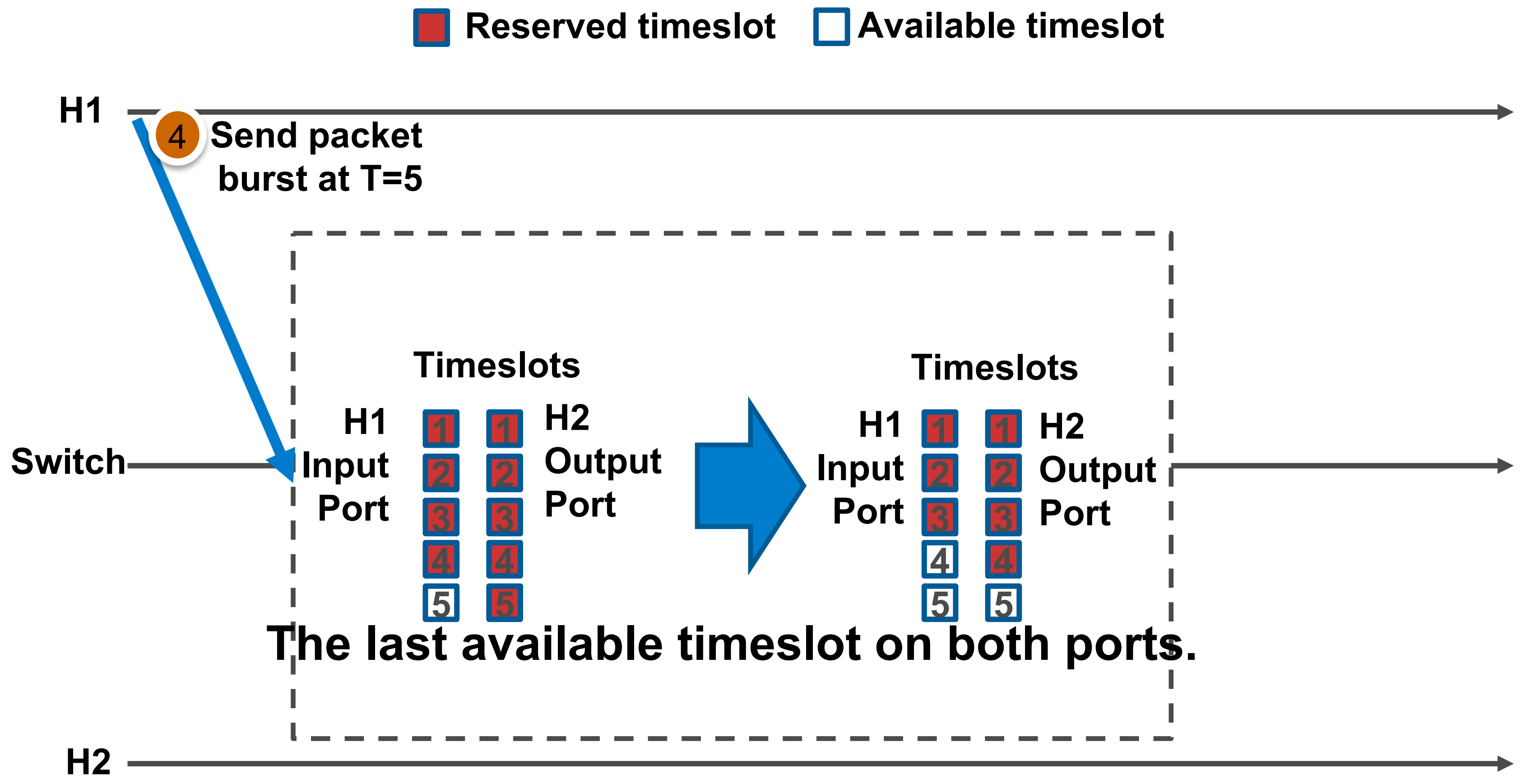
PL2 Overview

🟥 Reserved timeslot ☐ Available timeslot



The last available timeslot on both ports.

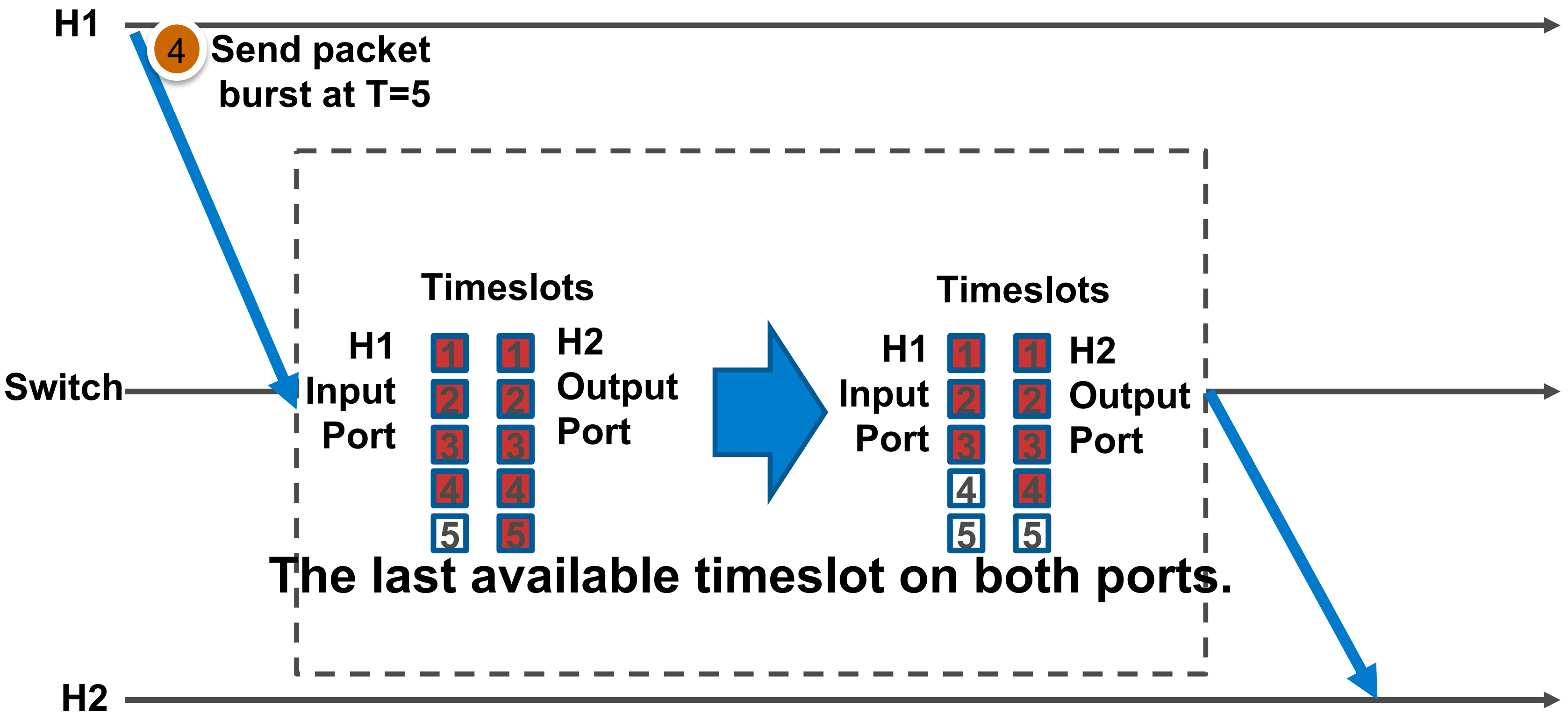
PL2 Overview



The last available timeslot on both ports.

PL2 Overview

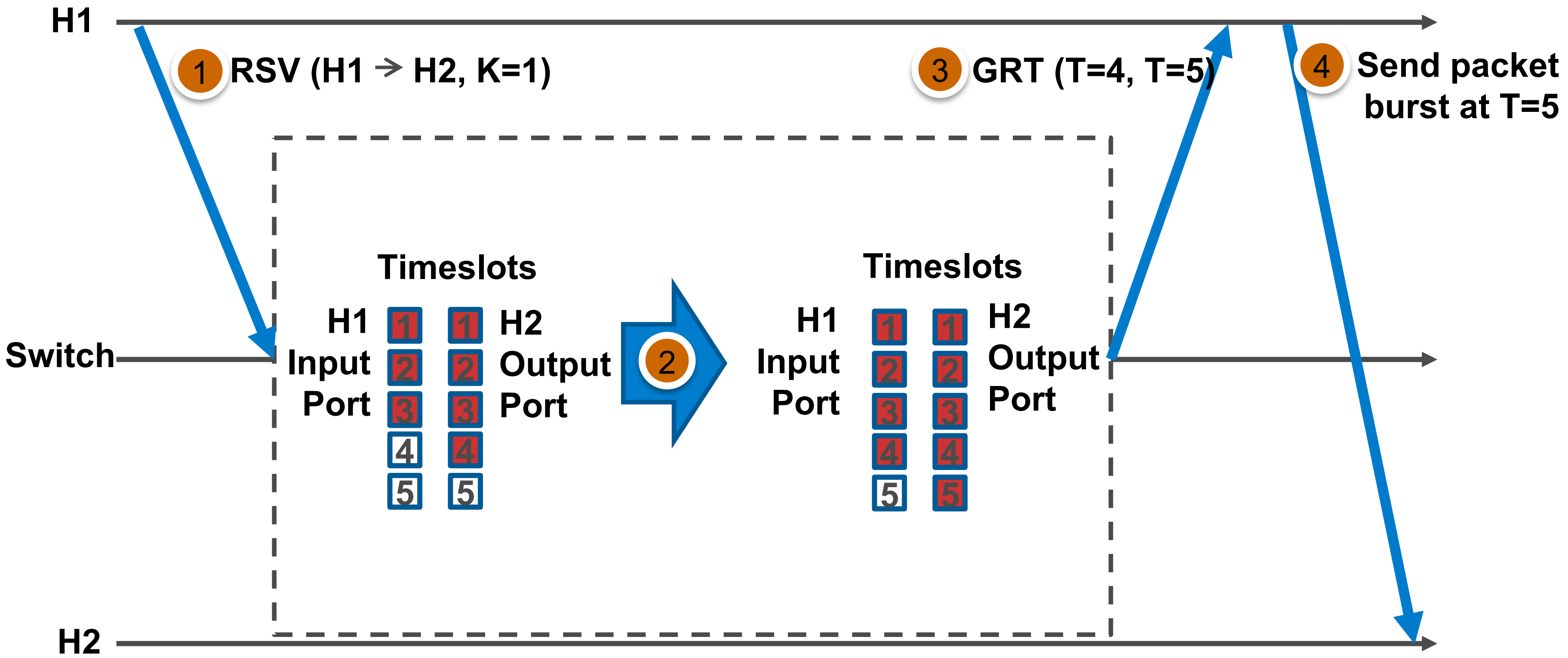
🟥 Reserved timeslot ⬜ Available timeslot



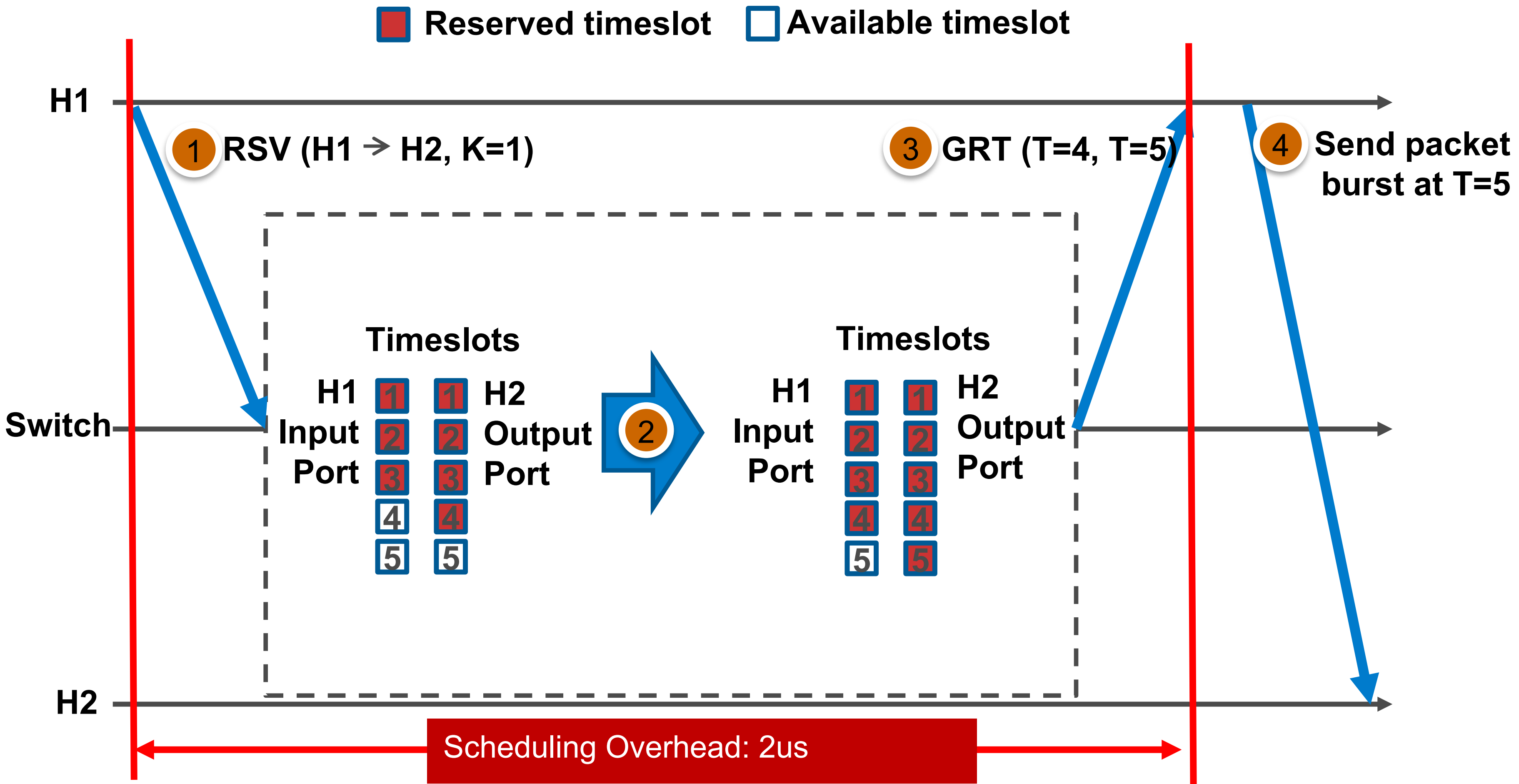
The last available timeslot on both ports.

Scheduling Overhead

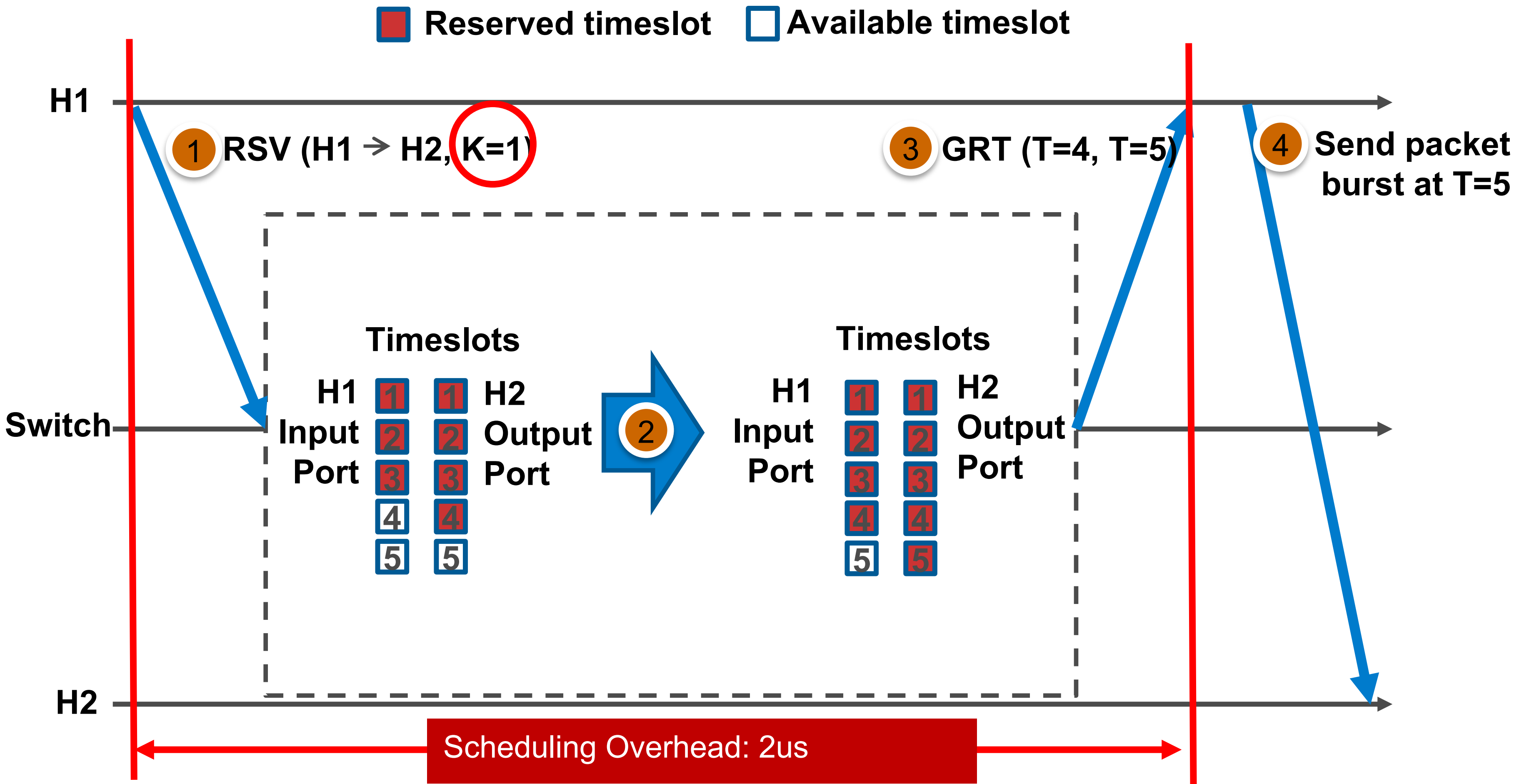
🟥 Reserved timeslot ☐ Available timeslot



Scheduling Overhead



Scheduling Overhead



Scheduling Overhead

- Reduce the scheduling frequency, K=4
- Minimize on the RSV-GRT delay
    - Burst RSV packet with last burst data packets

Scheduling Overhead

- Reduce the scheduling frequency, K=4
- Minimize on the RSV-GRT delay
    - Burst RSV packet with last burst data packets
- **Predicable latency:**
    - For each burst (K=4), the message delivery time = waiting time + data transmission
    - The waiting time increases as the reservation slots increase

Scheduling Overhead

- Reduce the scheduling frequency, K=4
- Minimize on the RSV-GRT delay
    - Burst RSV packet with last burst data packets
- **Predicable latency:**
    - For each burst (K=4), the message delivery time = waiting time + data transmission
    - The waiting time increases as the reservation slots increase

End-to-end message latency is unbounded!

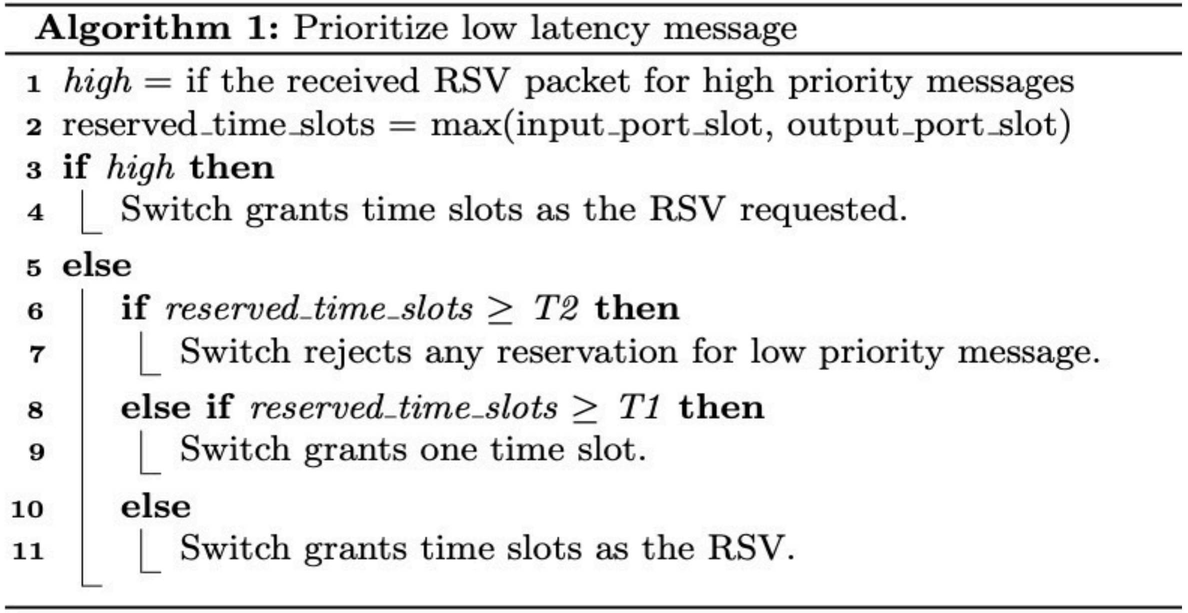
Prioritize Low Latency Message

- Switch grants a limited time slots (e.g., T2), for packets transmission

Prioritize Low Latency Message

- Switch grants a limited time slots (e.g., T2), for packets transmission
- Two properties
  - The maximum queue length is bounded by T2.
  - The `message` latency can be bounded by T2 + RSV-GRT-delay
    - When the network does not have more than T2 packets reservation from the high priority message
    - Set a bit for high priority message’s RSV packet

Prioritize Low Latency Message

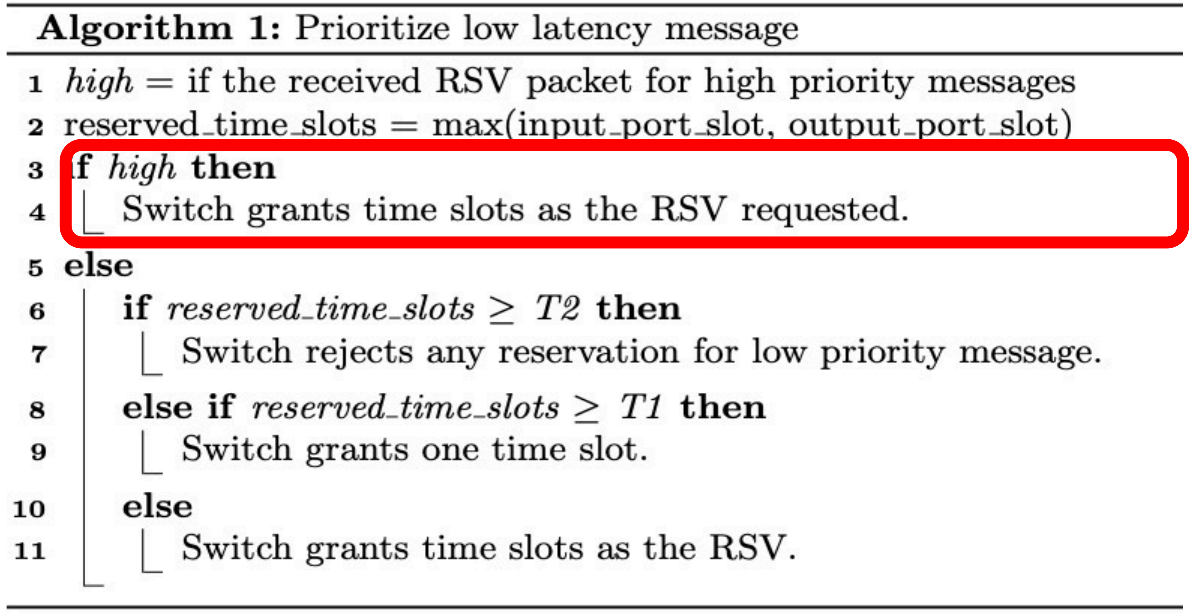


Algorithm 1: Prioritize low latency message

```
high = if the received RSV packet for high priority messages
reserved_time_slots = max(input_port_slot, output_port_slot)
if high then
    Switch grants time slots as the RSV requested.
else
    if reserved_time_slots ≥ T2 then
        Switch rejects any reservation for low priority message.
    else if reserved_time_slots ≥ T1 then
        Switch grants one time slot.
    else
        Switch grants time slots as the RSV.
```

T1=64, T2 = 128

Prioritize Low Latency Message

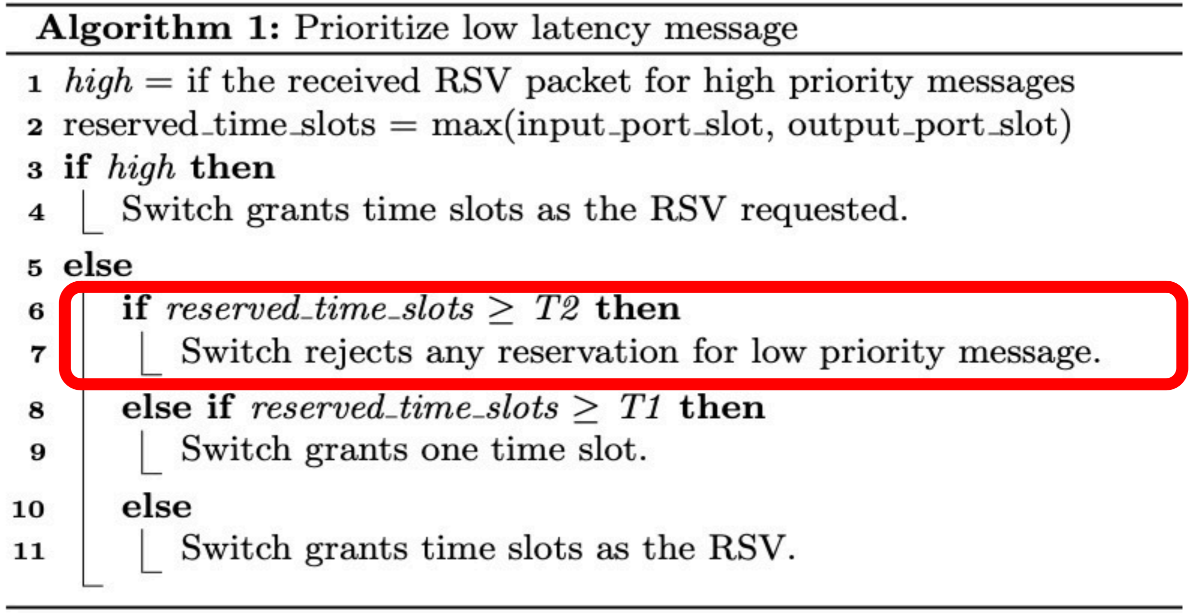


Algorithm 1: Prioritize low latency message

```
high = if the received RSV packet for high priority messages
reserved_time_slots = max(input_port_slot, output_port_slot)
if high then
    Switch grants time slots as the RSV requested.
else
    if reserved_time_slots ≥ T2 then
        Switch rejects any reservation for low priority message.
    else if reserved_time_slots ≥ T1 then
        Switch grants one time slot.
    else
        Switch grants time slots as the RSV.
```

T1=64, T2 = 128

Prioritize Low Latency Message

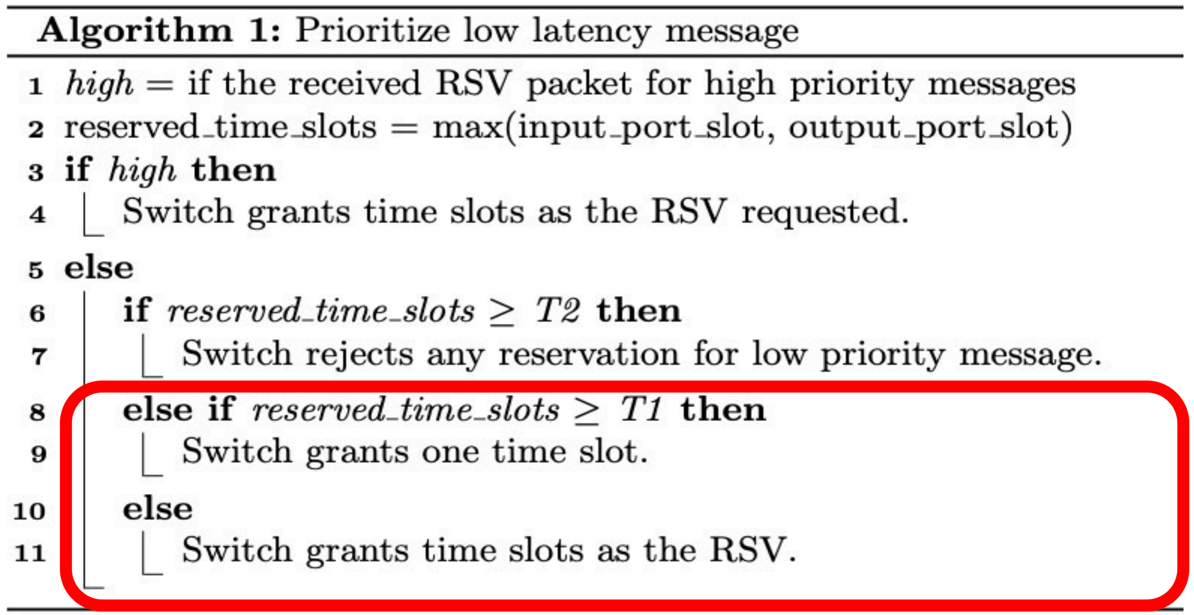


Algorithm 1: Prioritize low latency message

```
1  $high$ = if the received RSV packet for high priority messages
2  reserved_time_slots = max(input_port_slot, output_port_slot)
3  if $high$ then
4      Switch grants time slots as the RSV requested.
5  else
6      if $reserved\_time\_slots \geq T2$ then
7          Switch rejects any reservation for low priority message.
8      else if $reserved\_time\_slots \geq T1$ then
9          Switch grants one time slot.
10     else
11         Switch grants time slots as the RSV.
```

T1=64, T2 = 128

Prioritize Low Latency Message



Algorithm 1: Prioritize low latency message

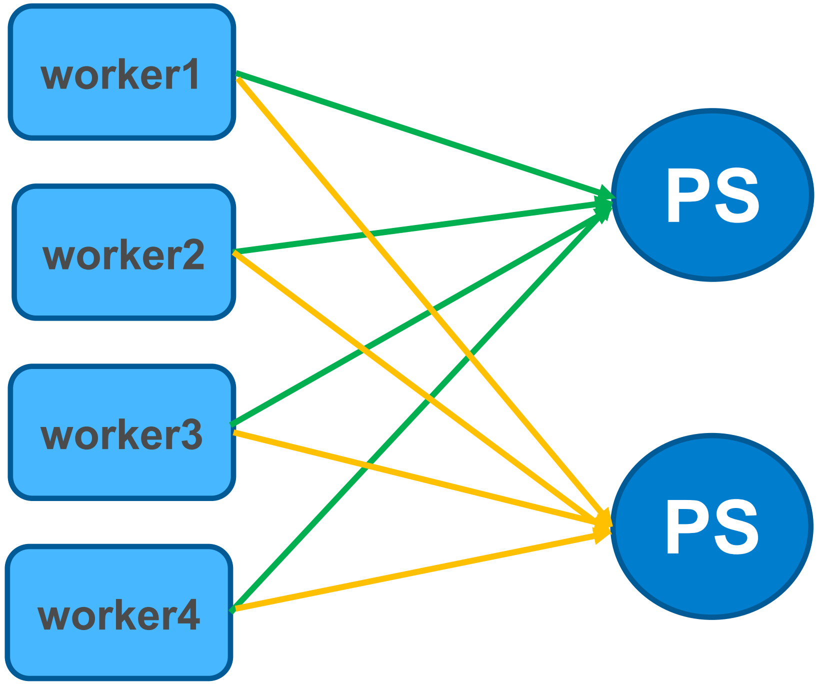
```
$high$ = if the received RSV packet for high priority messages
reserved_time_slots = max(input_port_slot, output_port_slot)
if $high$ then
    Switch grants time slots as the RSV requested.
else
    if $reserved\_time\_slots \geq T2$ then
        Switch rejects any reservation for low priority message.
    else if $reserved\_time\_slots \geq T1$ then
        Switch grants one time slot.
    else
        Switch grants time slots as the RSV.
```

T1=64, T2 = 128

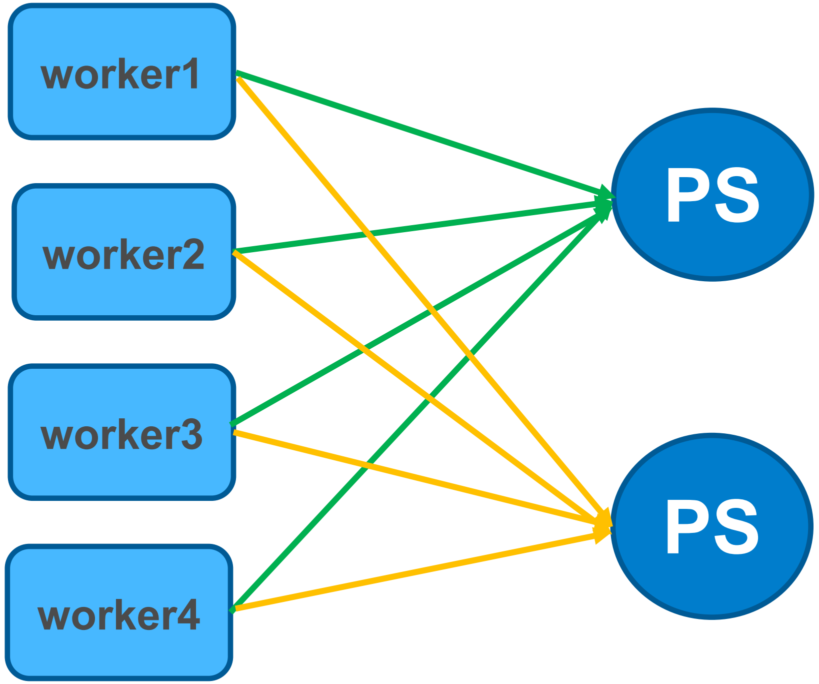
Implementation and Evaluation

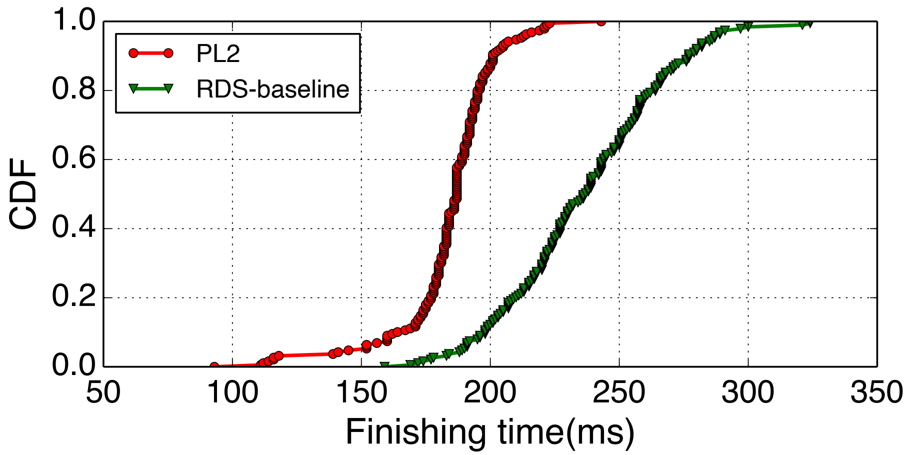
- Implementation
  - A customized networking stack and Mellanox VMA, Userspace TCP stack
  - Use P4 to implement the centralized scheduling algorithm at Tofino switch
- Evaluation Setup
  - Testbed setup: 6 hosts connect to one Tofino switch
  - Baseline: Receiver-Driven Scheme (RDS); TCP+Cubic and UDP
  - Workloads: Memcached, VGG16, Workload trace (W1-W5) from Homa

PL2 v.s. RDS on VGG16

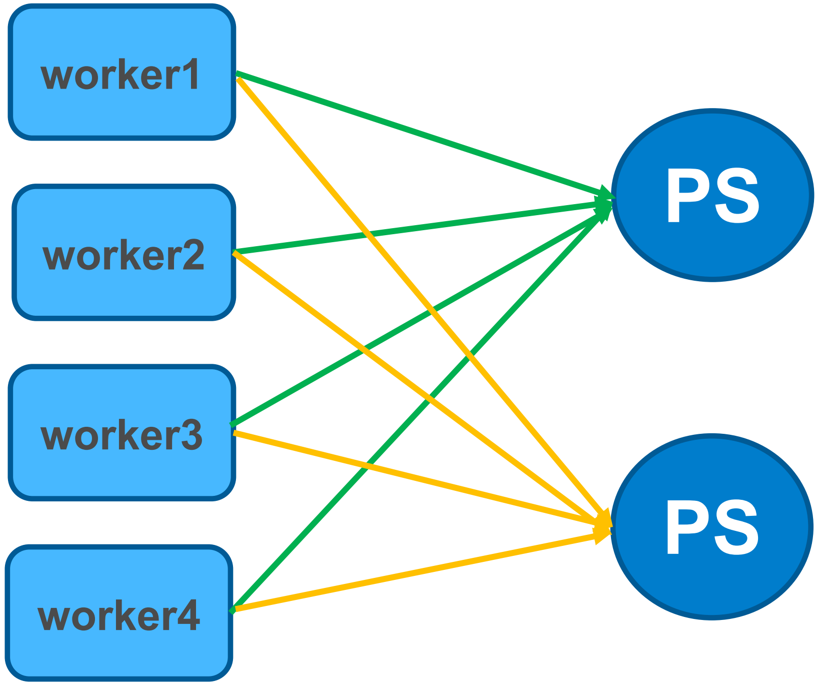


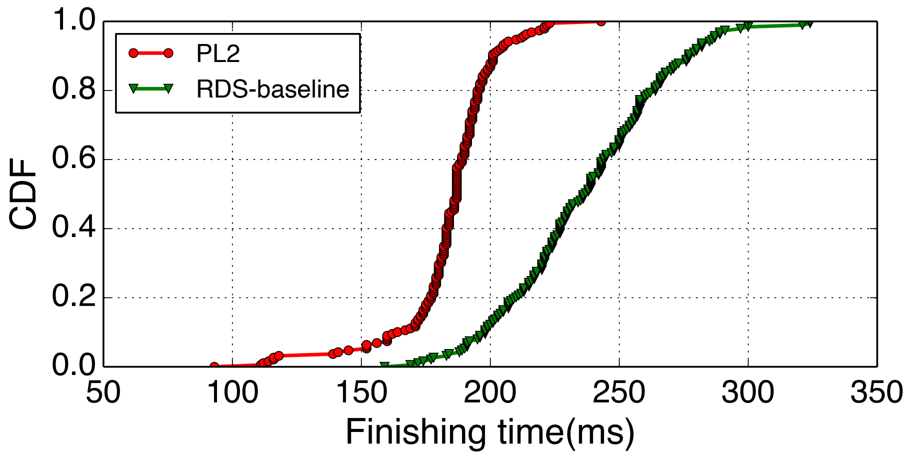
PL2 v.s. RDS on VGG16





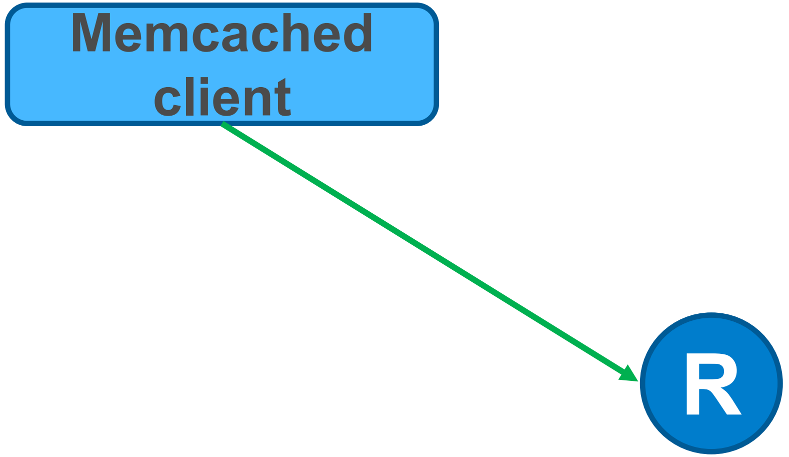
PL2 v.s. RDS on VGG16

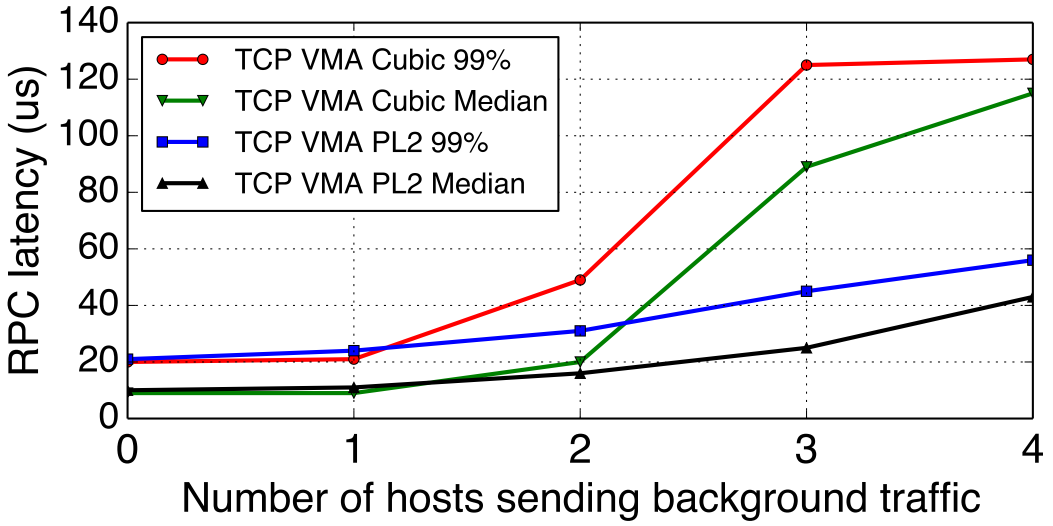




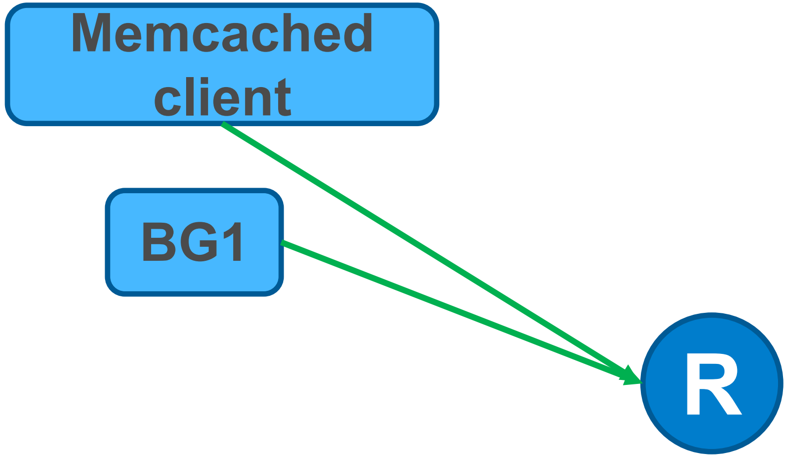
PL2 improves training latencies for VGG16 by 30%

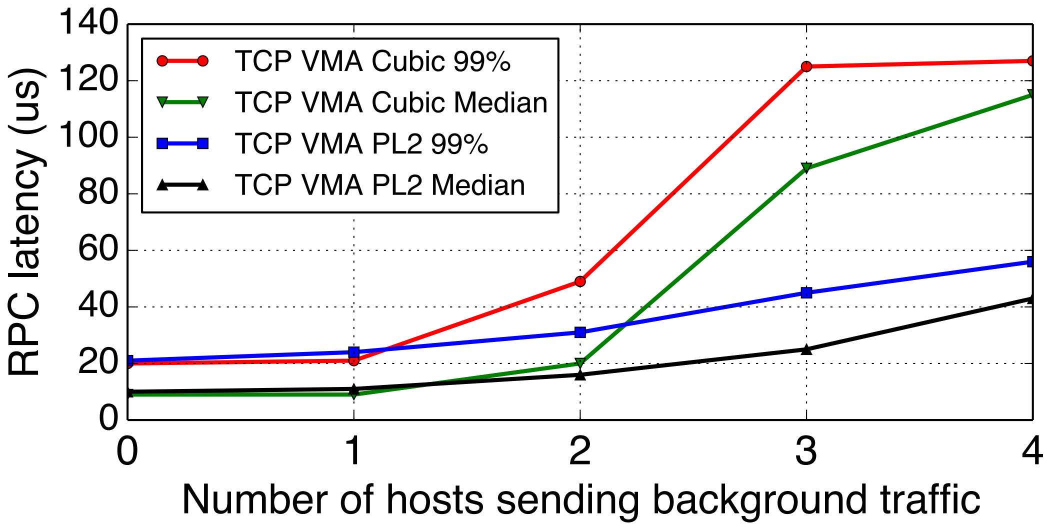
Memcached competing Background traffic



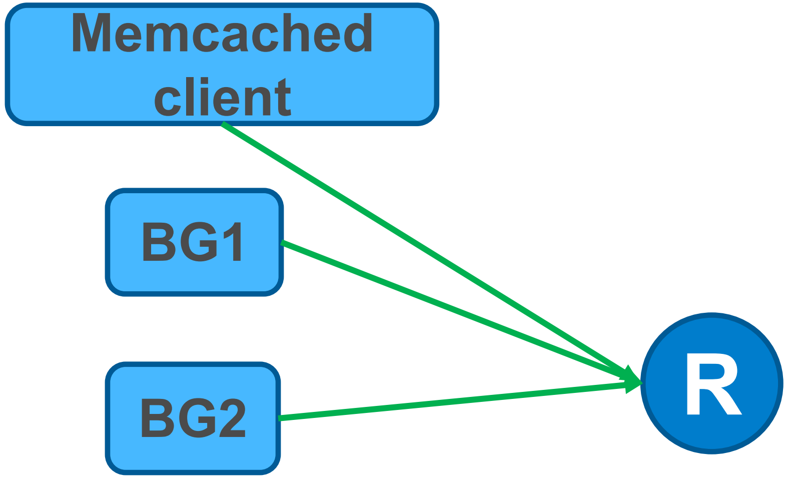


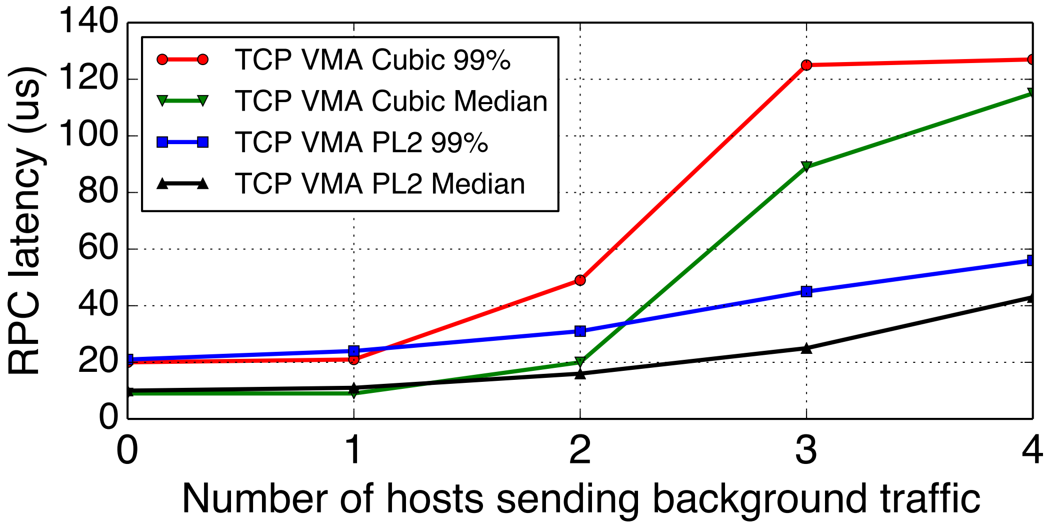
Memcached competing Background traffic



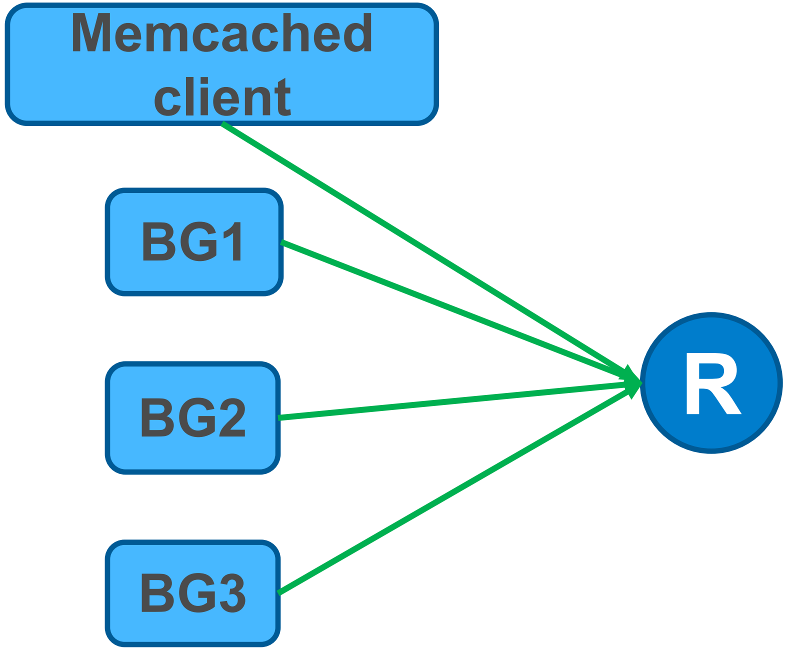


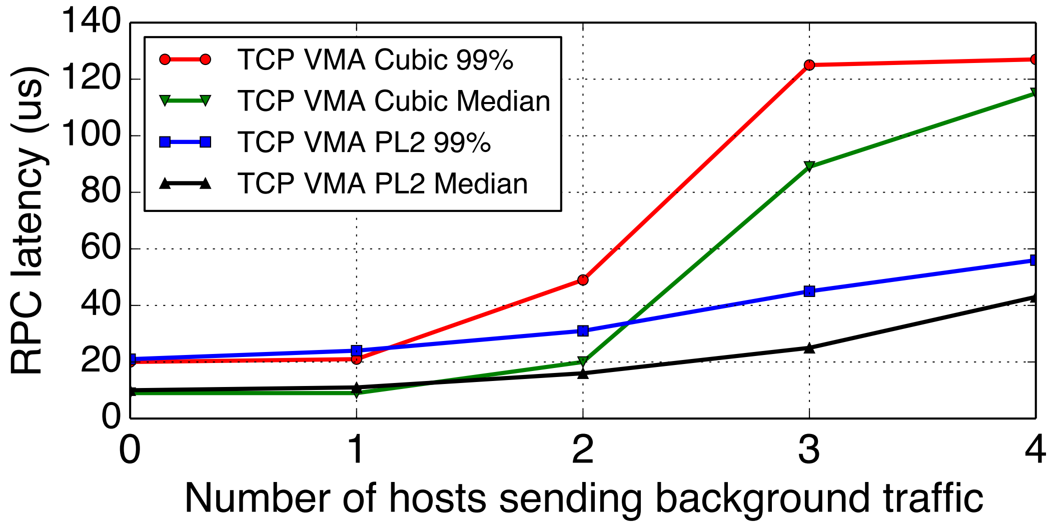
Memcached competing Background traffic



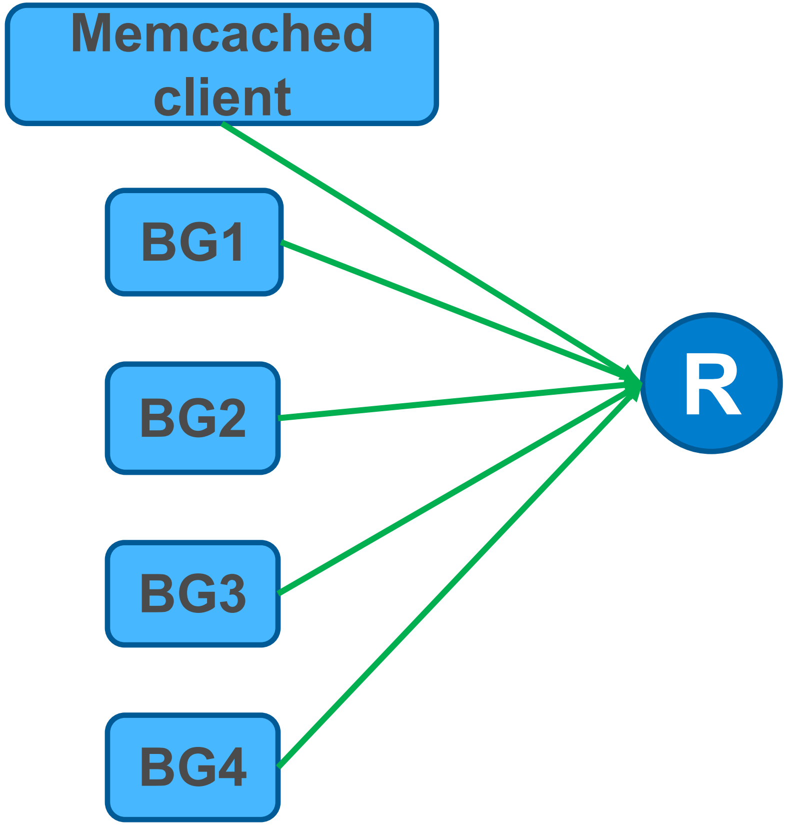


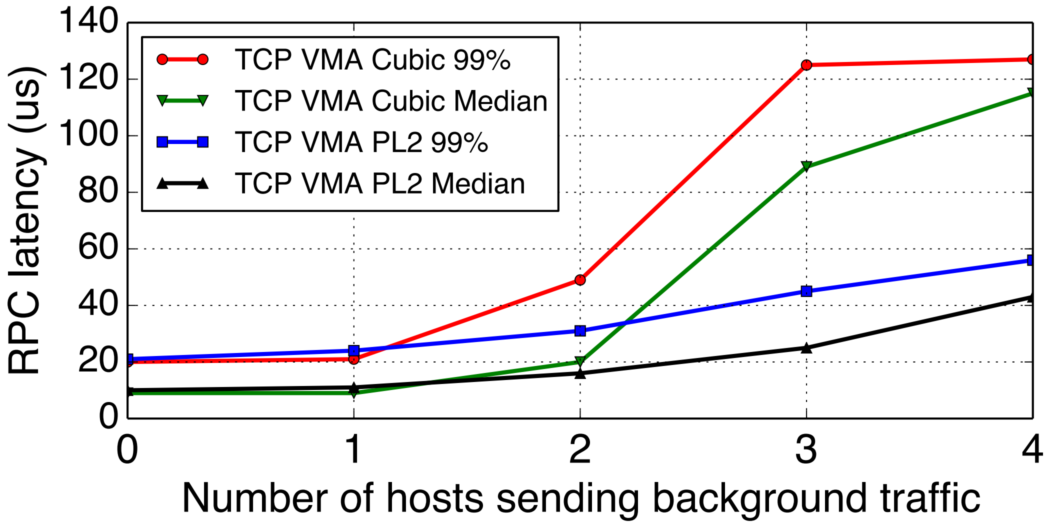
Memcached competing Background traffic



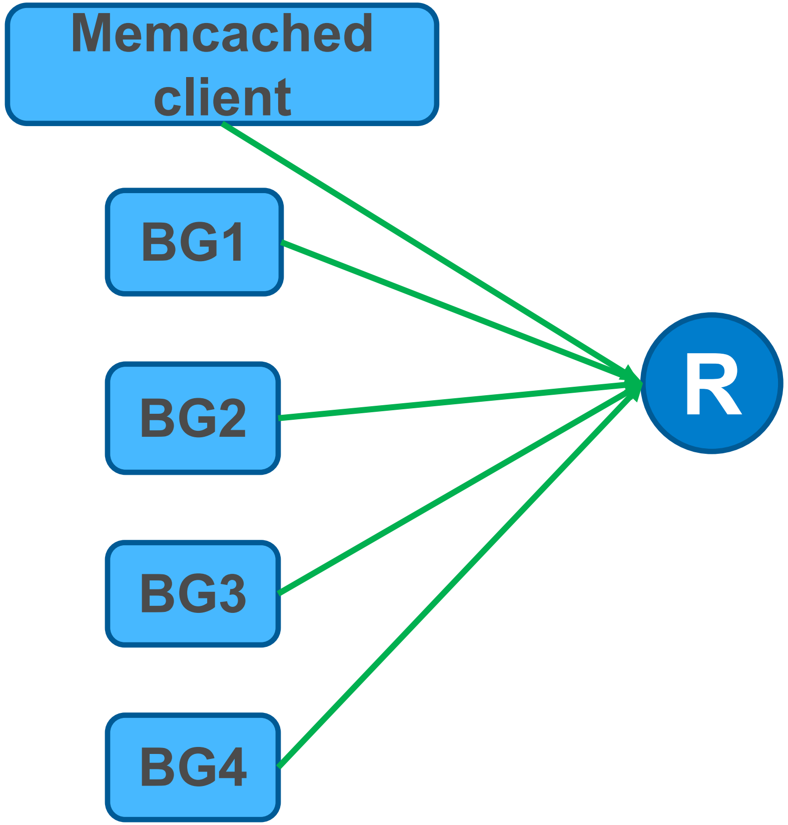


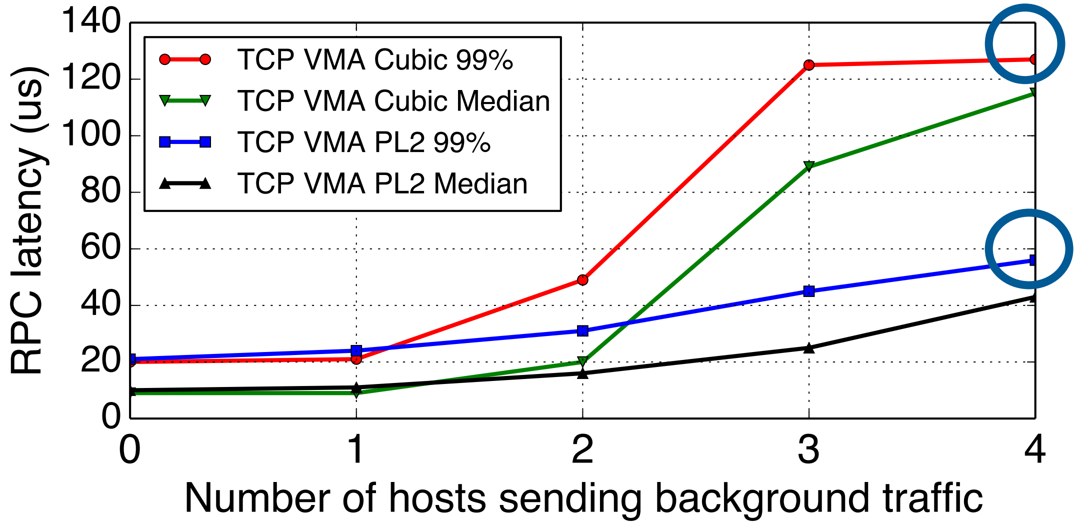
Memcached competing Background traffic



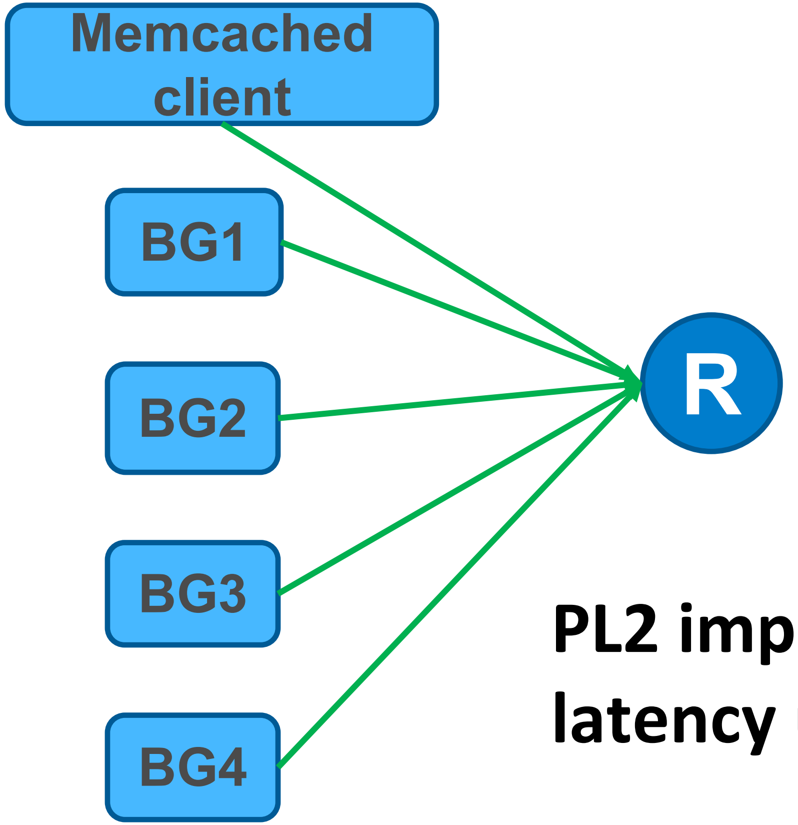


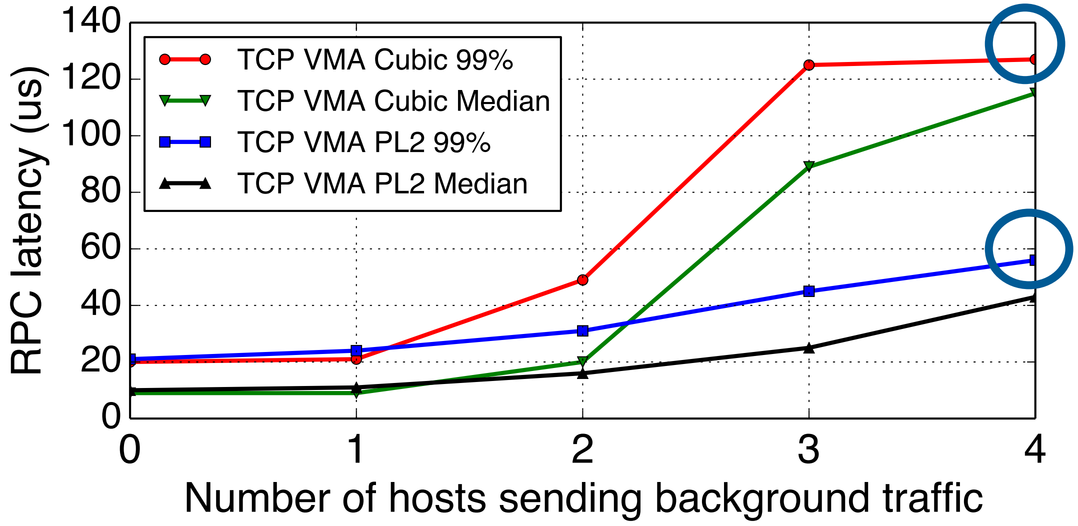
Memcached competing Background traffic





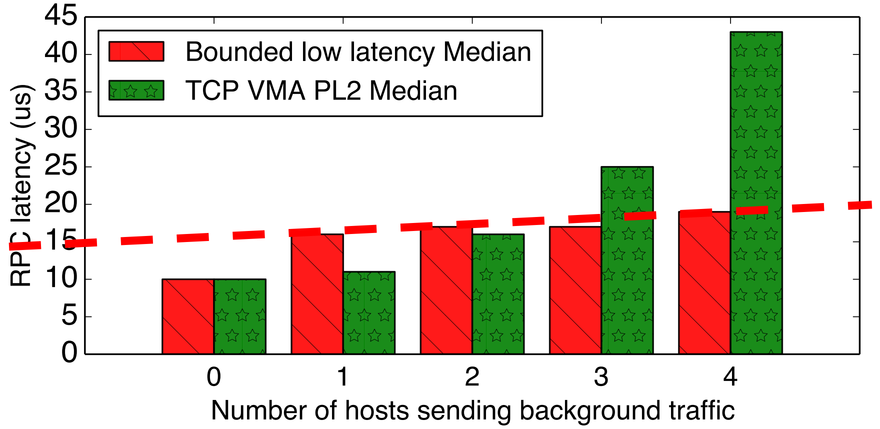
Memcached competing Background traffic

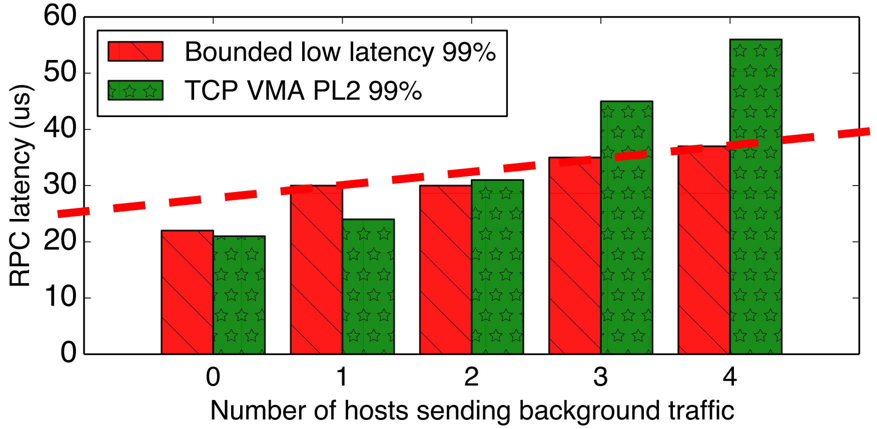




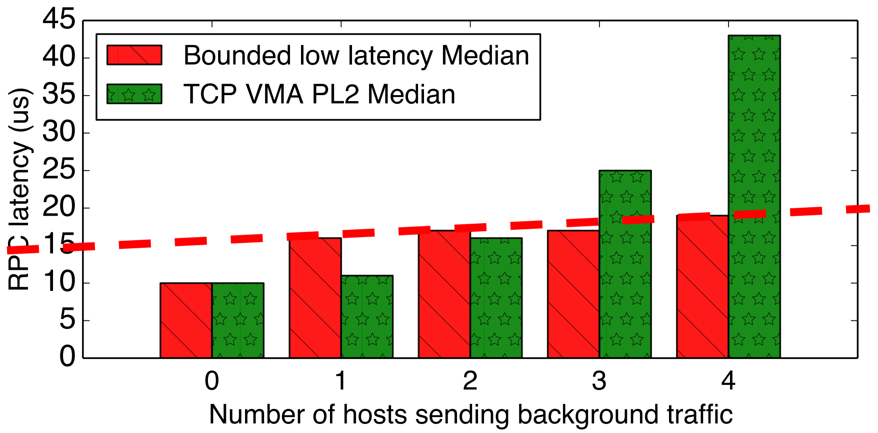
PL2 improves the 99th-percentile RPC latency up to 3X compared with Cubic.

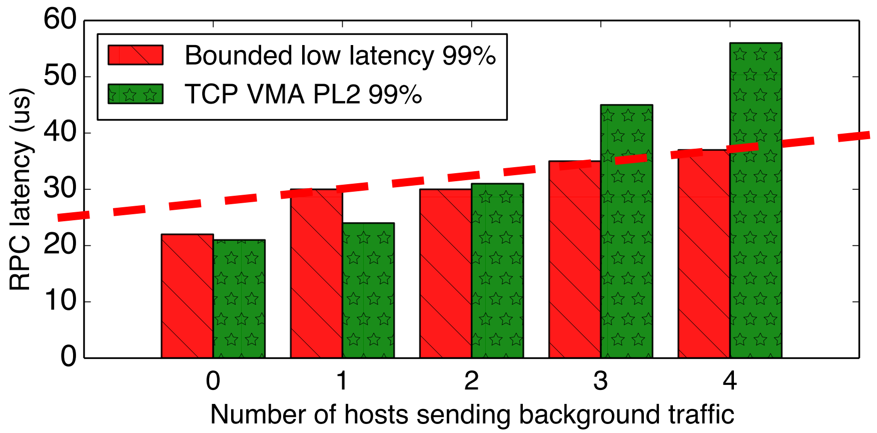
Memcached (Bounded Low Latency)





Memcached (Bounded Low Latency)





PL2 bounds message latency regardless of heavy background traffic.

Summary

- PL2 is a centralized packet scheduler towards the predictable low latency network in rack-scale network
- Co-design the switch logic and end host logic
    - Reduce the packet scheduling overhead
    - Bounded low latency for high priority message

MAY 18-20
2021 P4
Workshop
Hosted by ONF

Thank You

Yanfang Le, Yanfang@cs.wisc.edu