Congestion Detection in Lossless Networks

Authors : Yiran Zhang, Yifan Liu, Qingkai Meng, Fengyuan Ren

SIGCOMM '21: Proceedings of the 2021 ACM SIGCOMM 2021 Conference

Presented by Aishwarya Shrestha

Agenda

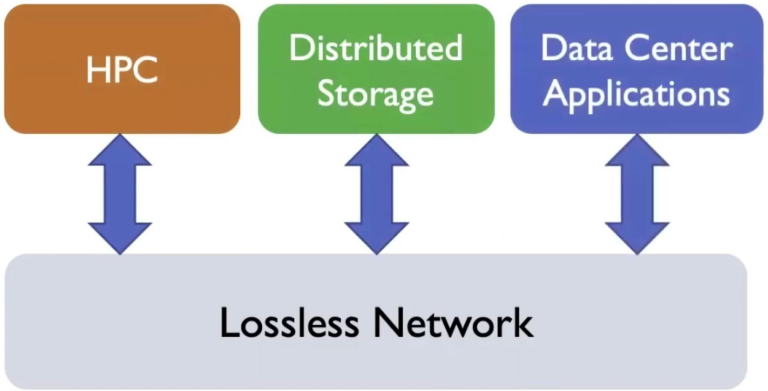
- ❏ Background
- ❏ Design
- ❏ Implementation
- ❏ Testbed
- ❏ Results
- ❏ Conclusion
- ❏ Critique

Background

Reliable | Efficient | Min delay | No packet loss

Lossless Network

- A lossless network is a type of computer network in which **no data packets are lost** or dropped during transmission.
- Employs a variety of techniques such as error correction codes, flow control mechanisms, and congestion avoidance algorithms
- Achieved by use of congestion control mechanisms
  - ensure that the network is not overloaded with traffic beyond its capacity.



Background

It is mechanism for regulating the flow of data

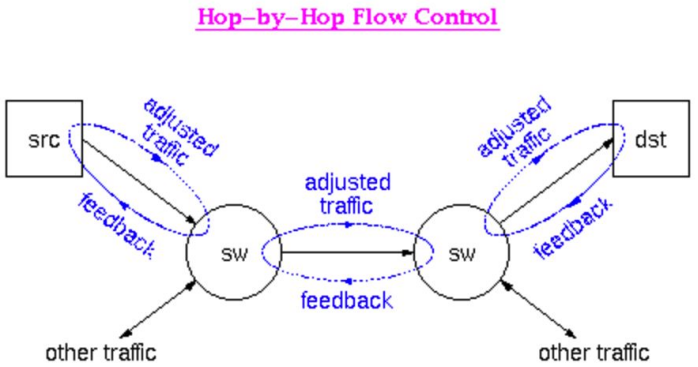
Hop-by-Hop flow control

- Lossless networks **rely on Hop-by-Hop flow** control to guarantee zero packet loss and normal operation.
- Hop-by-hop flow control regulates **data flow** by controlling the transmission rate between adjacent network nodes.
  - Node individually manages the flow of data based on feedback received from the adjacent nodes
  - Only forwards packets when it is confident that the receiver has enough buffer space to store the packets.

- Hop-by-hop flow control is used in

  1. Converged Enhanced Ethernet (CEE)

  2. InfiniBand networks (IB)



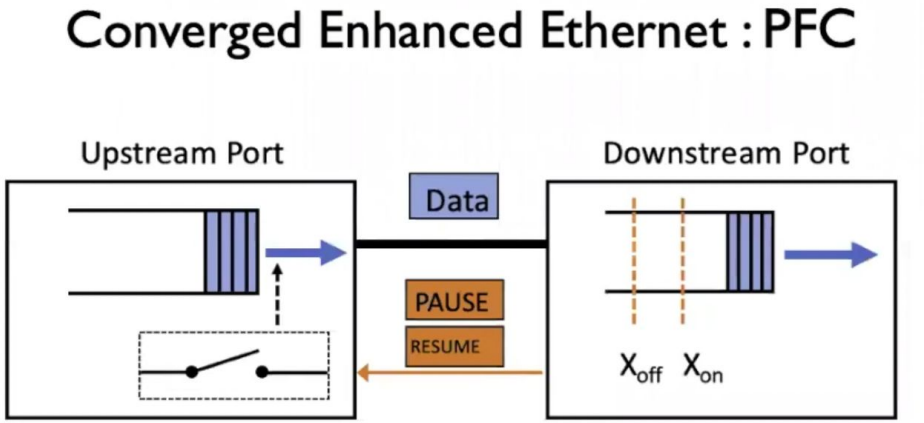
Existing Congestion Detection Mechanism

Converged Enhanced Ethernet (CEE)

- In CEE, **Priority Flow Control (PFC)** is used to provide **hop-by-hop flow control.**
  - Network switches pauses the flow of traffic on certain high-priority links

- Each upstream port communicates with its downstream port to negotiate the priority classes that will be used for traffic flow control.

- Allows network switches to **pause** the flow of traffic on certain high-priority links.

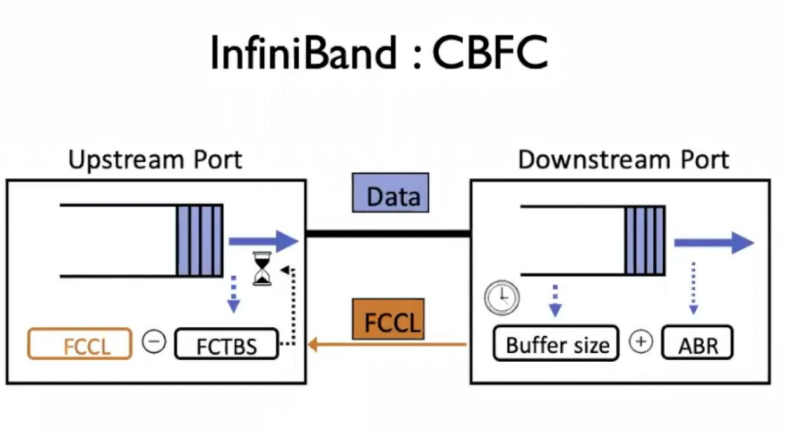


Existing Congestion Detection Mechanism

InfiniBand networks

Credit-Based Flow Control (CBFC)

- In InfiniBand networks, hop-by-hop flow control is achieved using Virtual Lanes (VLs)
  - InfiniBand is considered a Credit-Based Flow Control (CBFC) network - it allows network nodes to regulate the flow of traffic on each link independently.
  - Network switches pauses the flow of traffic on certain high-priority links



- The downstream switch maintains an Adjust Block Register (ABR) to record total received blocks

- The downstream switch sends a Flow Control Credit Limit (FCCL) message to the upstream switch periodically, which contains the sum of allocated buffer size and ABR

- After receiving the FCCL message, the upstream switch can only send a packet when there are available credits

Existing Congestion Detection Mechanisms

Two lossless networks 1) Converged Enhanced Ethernet (CEE) and 2) InfiniBand

CEE (Converged Enhanced Ethernet)

Congestion management developed by Data Center Bridging Task Group:

1. Congestion Point (CP) for Congestion detection
2. Reaction Point (RP) for rate control

InfiniBand

InfiniBand specification specifies the framework of congestion control:

1. The switch detects congestion
2. Channel Adapter (CA) conducts injection throttling

Congestion detection is conducted by switches in existing lossless networks.

Existing Congestion Detection Mechanisms

CEE

DCQCN (Data Center Quantized Congestion Notification) is adopted in CEE, detects congestion based on queue size.

- Congestion Point (CP) marks packets Explicit Congestion Notification (ECN) flags according to with the Random Early Detection (RED) algorithm.
- The ECN marking is a one-bit indicator

InfiniBand (IB)

It combines credits information and queue size to detect congestion:

- If the queue of an output port exceeds a threshold and there are available credits to send packets, marks packets with Forward ECN (FECN).
- A single bit FECN marking indicates the presence of congestion.

In both, the switch ports alternate between sending (ON) and pausing (OFF) to regulate the flow of traffic. They work in a similar way to guarantee a drop-free property.

Summary : Existing Congestion Detection Mechanisms

CEE: Priority Flow Control

InfiniBand (IB) : Class-Based Flow Control

- Both work in the same way to guarantee the drop rate. The switch ports alternate between sending(ON) and pausing(OFF) to guarantee the drop-free property.
  - With PFC, the switch uses priorities to determine which packets to pause first
  - With CBFC, the switch uses classes to determine which packets to pause first

Summary : Current congestion detection methods essentially detect incipient congestion by queue size, similar to traditional lossy networks

1. Existing congestion detection mechanisms fail to cognize the impact of ON-OFF sending patterns, leading to inaccurate congestion detection results in lossless network
2. A port may have the same queue length evolution and sending pattern, but the real congestion state is diverse

Design

Understanding port states in lossless network

- This paper presents

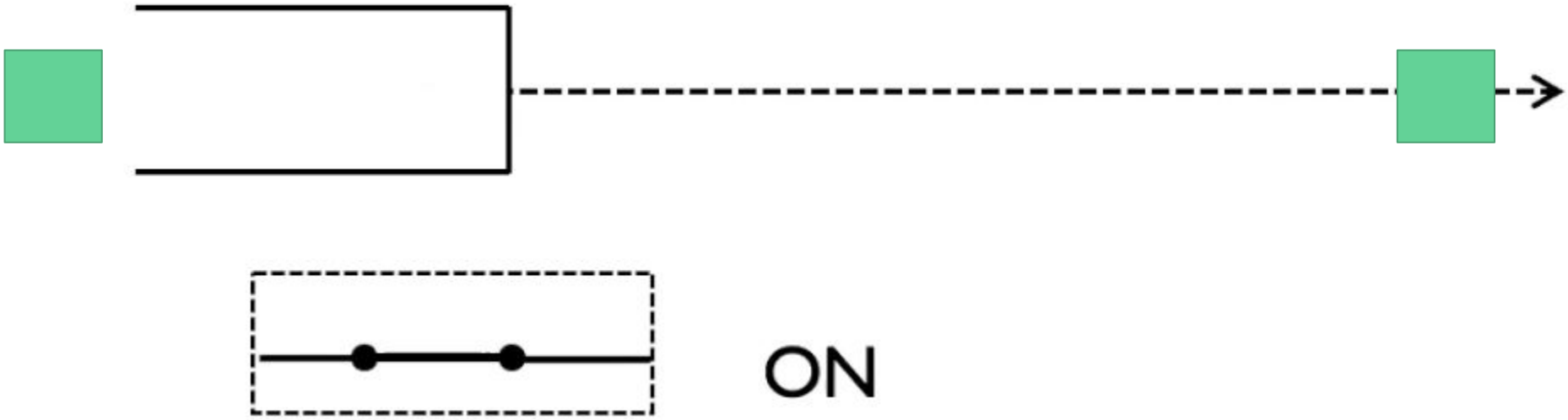
  1. **Ternary states** of switch ports and
  2. Proposes **Ternary Congestion Detection (TCD)** for lossless networks
  3. Provides Ternary congestion notification to enhance Congestion Control decisions

Ternary states

(1) congestion
(2) non-congestion and
(3) undetermined

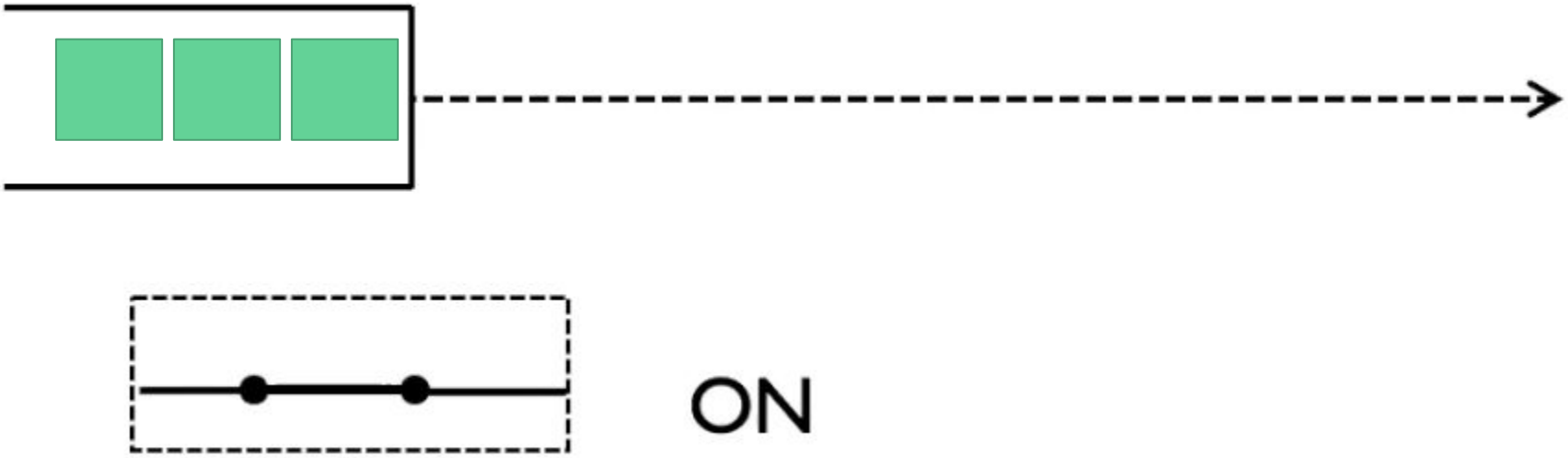
Understanding port states in lossless network

1. Non-congestion : The port is persistently ON and without queue buildup



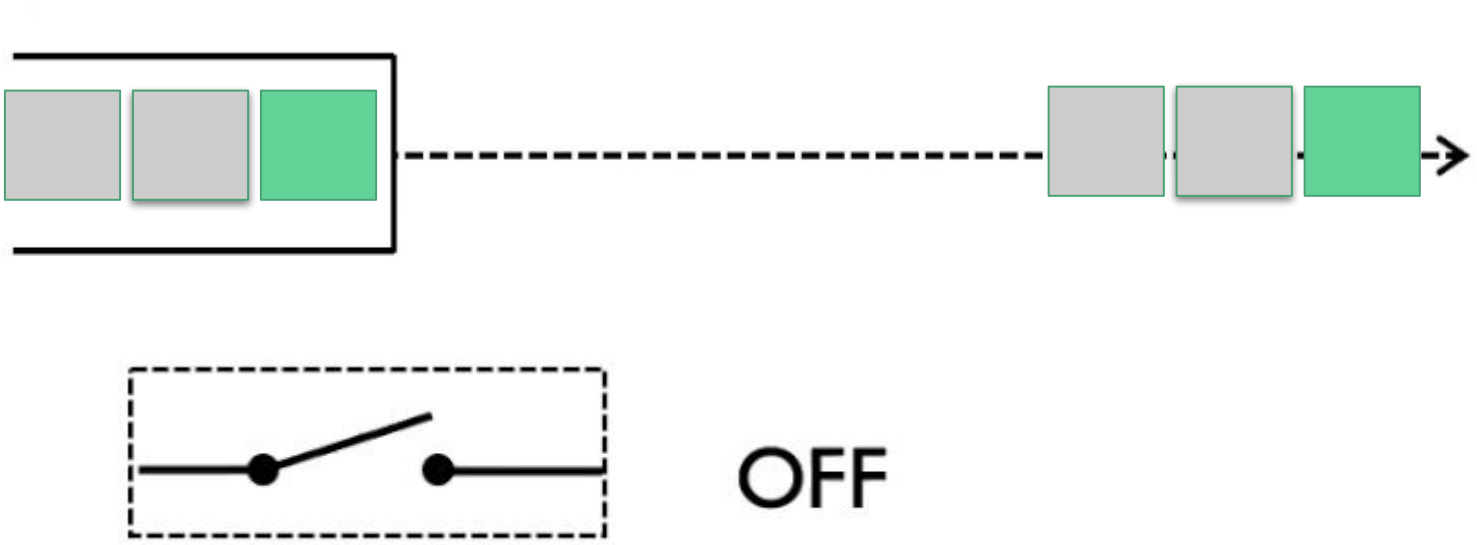
Understanding port states in lossless network

2. Congestion : The port is persistently ON. The output rate is at full rate with queue buildup not caused by OFF.



Understanding port states in lossless network

1. Undetermined : The output rate is in an ON-OFF style.



Ternary states of switch ports

1. Non-congestion (0): The port is persistently ON and without queue buildup.

2. Congestion (1): The port is persistently ON. The output rate is at full rate with queue buildup not caused by OFF.

3. Undetermined (/): The output rate is in an ON-OFF style.

Port may experience queue buildup

The cause of queue build up may be ambiguous.

- Receiving PAUSEs/no credits
- Excessive input traffic

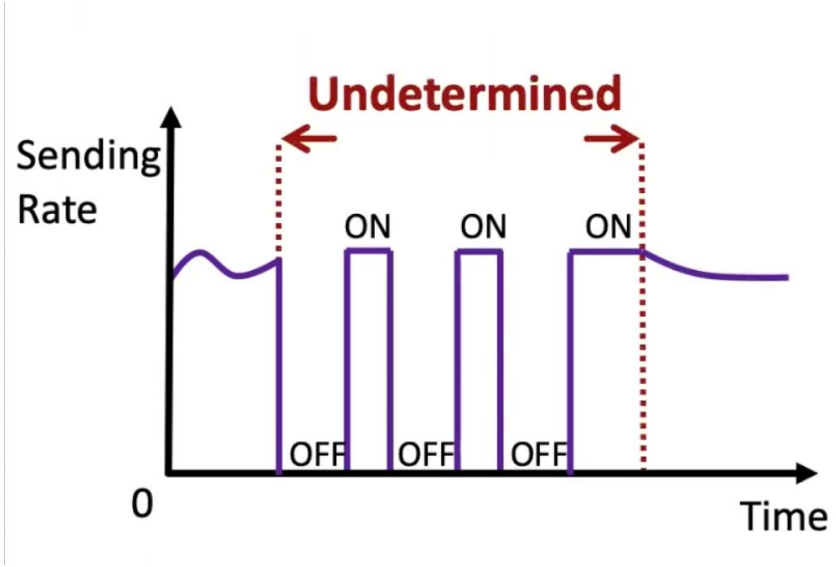
Understand the port states in lossless networks

Undetermined state: The output rate is in an ON-OFF style

The sending rate is regulated to an ON-OFF pattern.

Each OFF period, the port is pausing, and all arriving packets are queued.

Each ON period, the port sends at the full rate, dequeuing the accumulated packets.



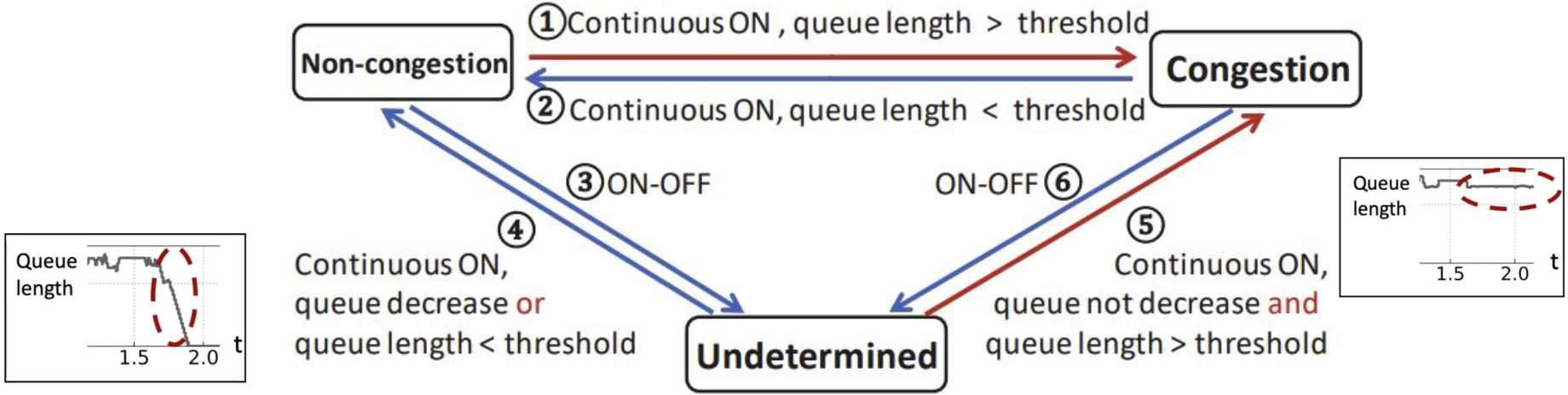
- ***Enter:***
  - Once the port is paused due to receiving PAUSEs/no credits

- ***Exit:***
  - Recovering to send packets in a continuous ON pattern.
  - The real state emerges after leaving the undetermined state : *non-congestion/congestion*

Ternary congestion notification to enhance Congestion Control decisions

Transitions among Ternary States

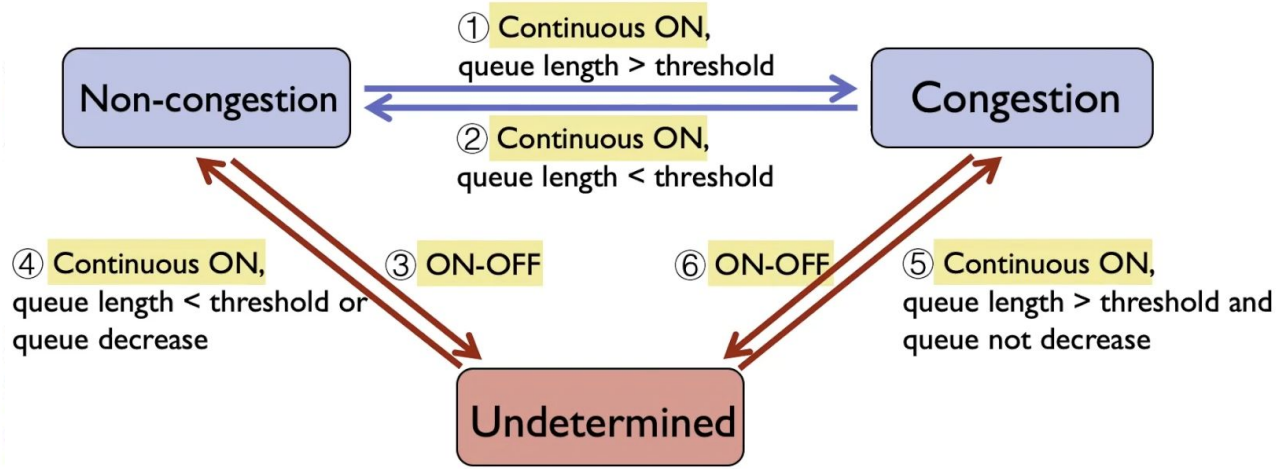
1. Transition between the non-congested (NC) state and the congested (C) state is same process as in lossy networks
2. In NC / C state, as soon as port enters ON-OFF sending pattern, the port switch to undetermined (U) state.
3. U to NC or C state involves evolution of queue length



Ternary congestion notification to enhance Congestion Control decisions

Transitions among Ternary States

1. To obtain the trend of queue length evolution, switches checks the queue size every period
2. The problem is to distuningh between **‘Continuous ON sending pattern**’ and ‘**ON-OFF sending pattern’**
3. Foundation of TCD is the parameter $max\ (Ton)$ (queue length)

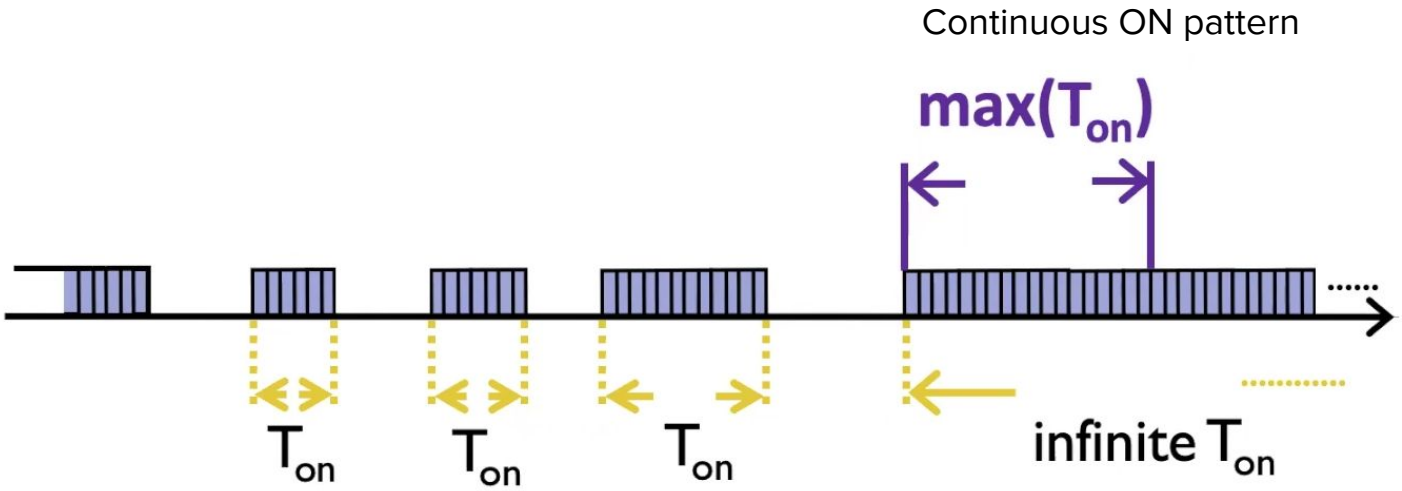


State transitions among ternary states

Key insight on capturing state transitions :

- A port in a continuous ON sending pattern has infinite T on

- A port in an ON-OFF sending pattern usually has a limited T on since the port is paused intermittently



Ternary Congestion Notification

Switches should mark packets to indicate when they detect congestion and notify endpoints about it.

1. They advocate that switches should also inform the transitions to the undetermined state to endpoints.
2. Switches enable end-to-end congestion controls to conduct different rate adjustments for undetermined flows and congested flows

3. To handle this, **TCD supports ternary congestion notification**.

Ternary congestion notification

- If a packet only goes through an undetermined port, the corresponding flow is undetermined.

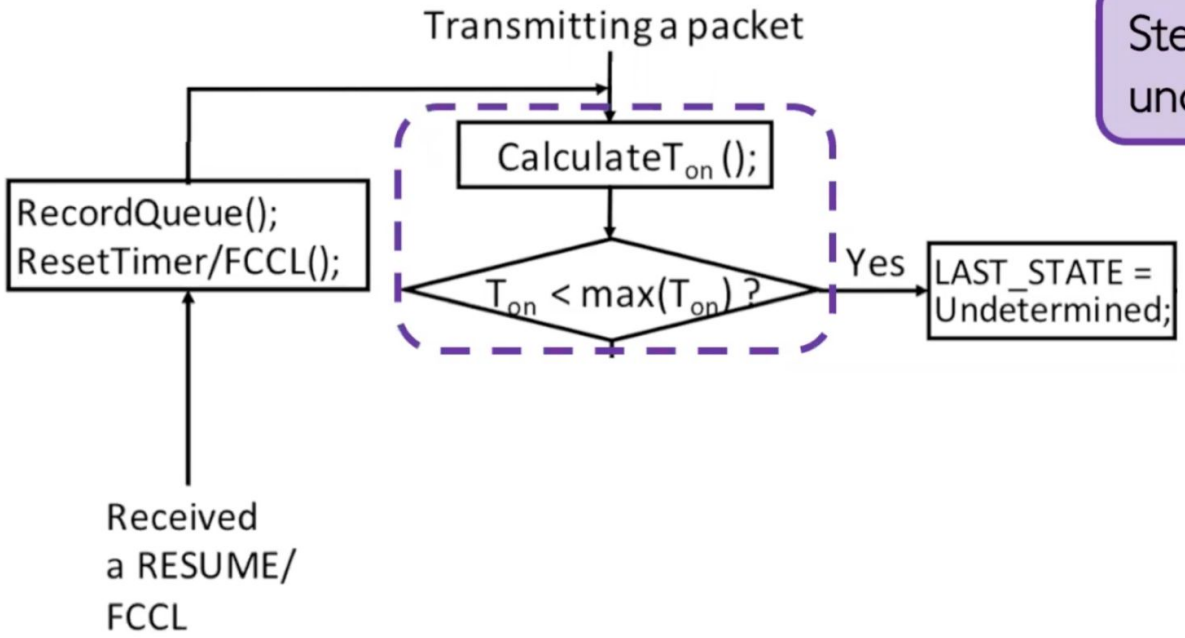
- Code point 10 is used to indicate the undetermined state encountered (UE).

- UE can only be marked when the current code point is not CE. Switches mark CE whenever the port is in a congestion state.

Table 1: TCD marking scheme.

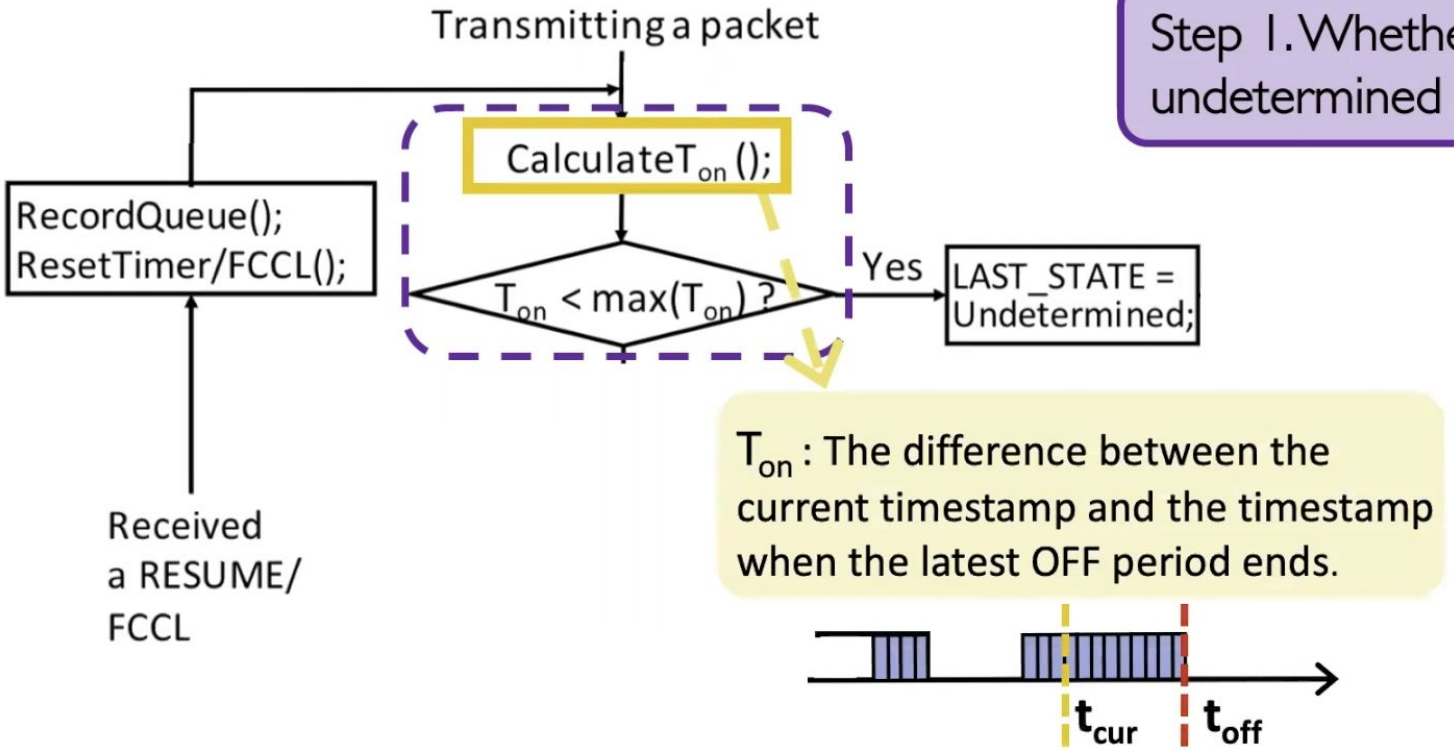
Code points	Meaning
00	Non TCD-Capable Transport
01	TCD-Capable Transport
10	Undetermined Encountered (UE)
11	Congestion Encountered (CE)

TCD Workflow

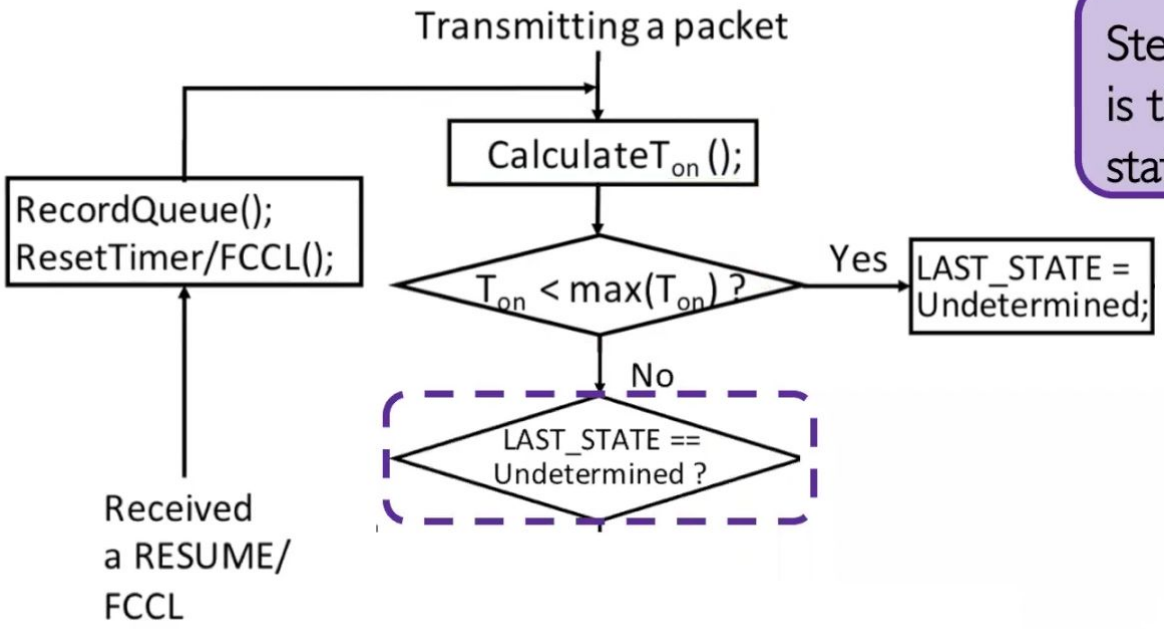


Step 1. Whether in the undetermined state ?

TCD Workflow

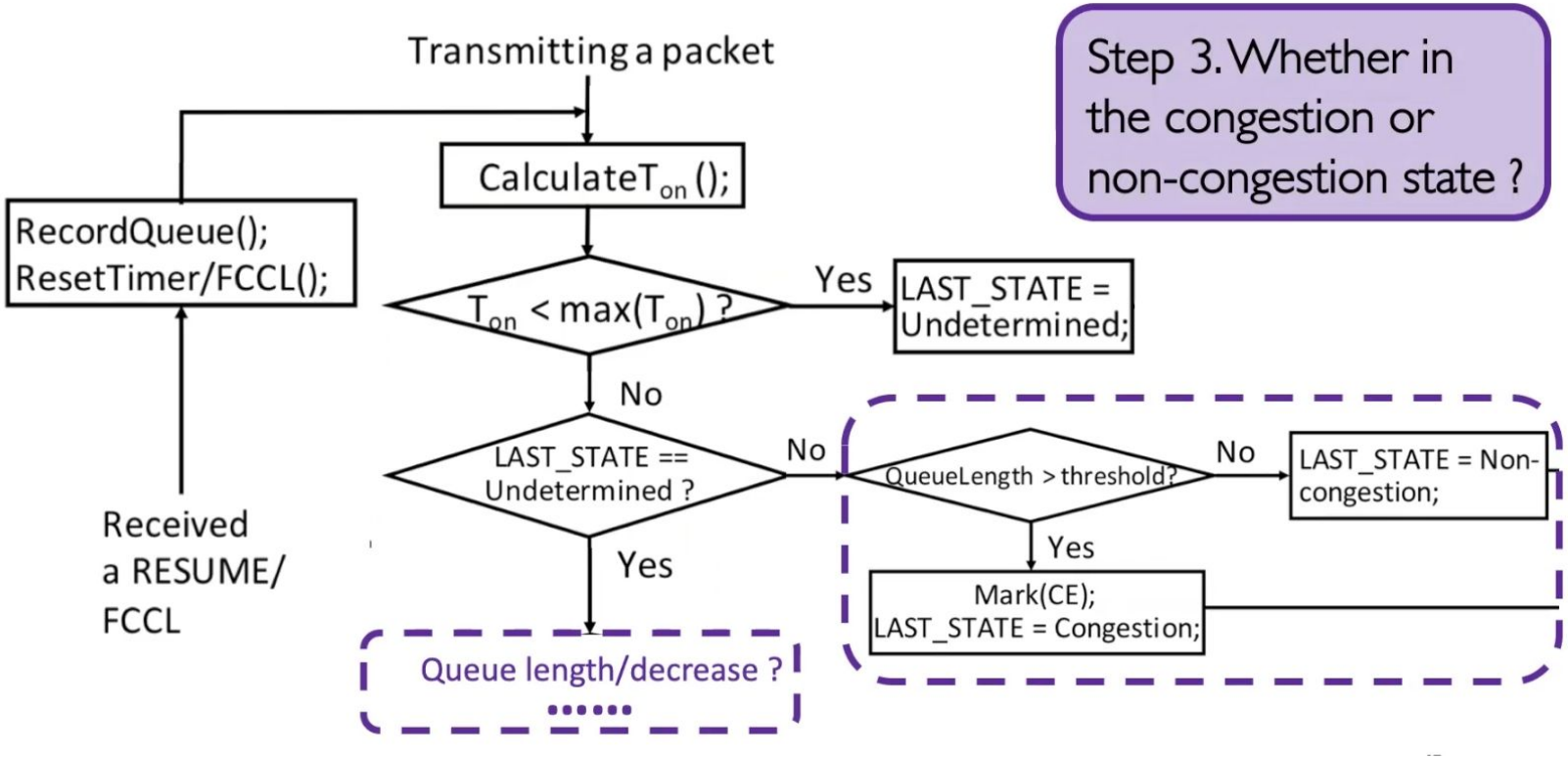


TCD Workflow

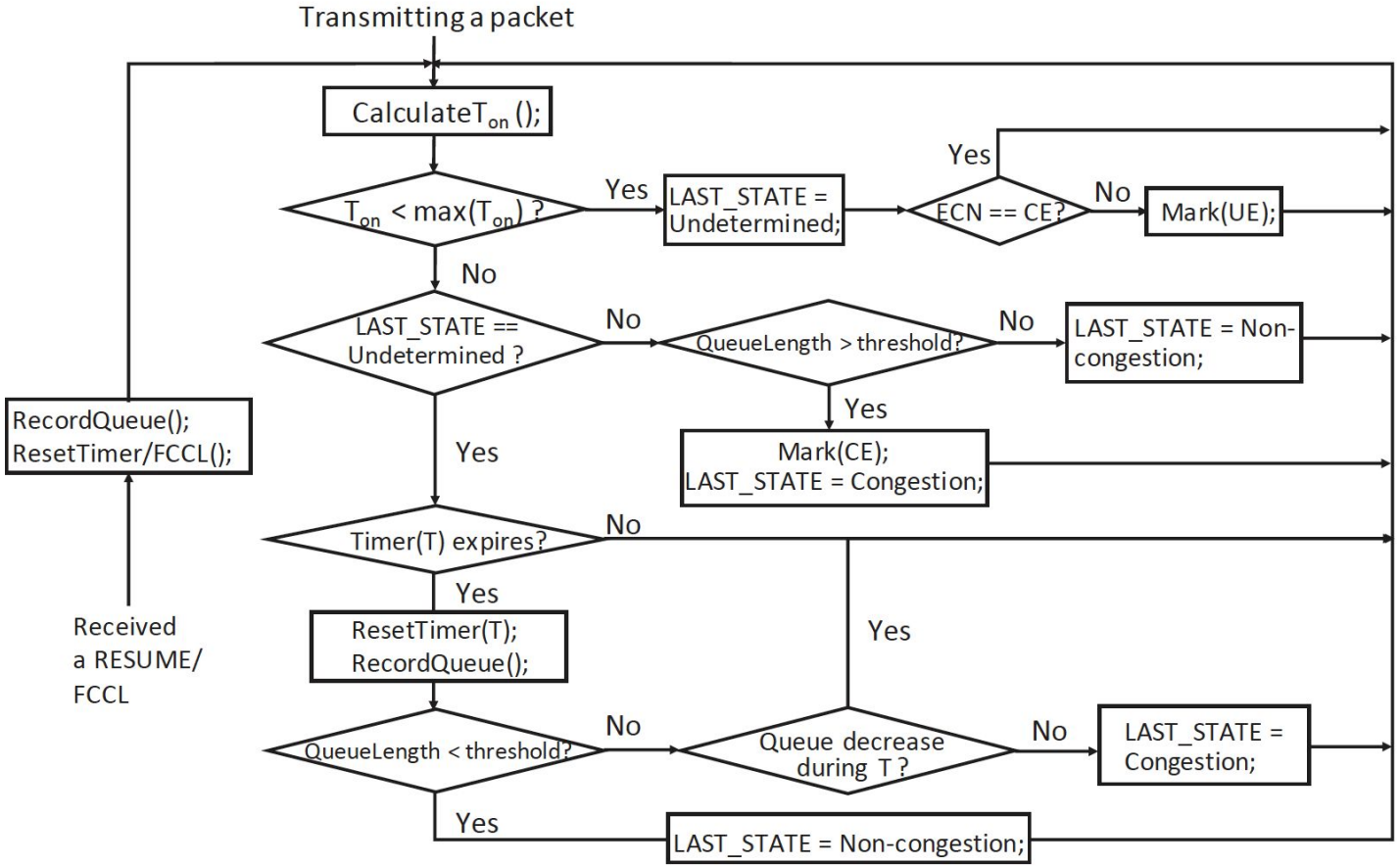


Step 2. LAST_STATE is the undetermined state ?

TCD Workflow



TCD Workflow



Testbed

- Compact topology, with switch T0 directly connecting to switch T2

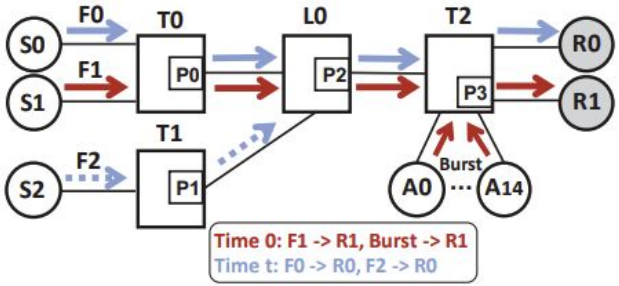
- Implementation: DPDK (Data Plane Development Kit) Intel 82599 NICs

- Two servers, each equipped with dual Intel Xeon E5-2620 v3 CPUs (6 cores, 2.4GHz), are used to build software switches.

- Each server is plugged with 2 dual-port Intel 82599 10G NICs, working as a four-port switch.

- To achieve line-rate sending/receiving operation at 10Gbps, each RX/TX module is implemented on an individual core.

- Implemented PFC according to IEEE 802.1Qbb and CBFC according to InfiniBand Specification

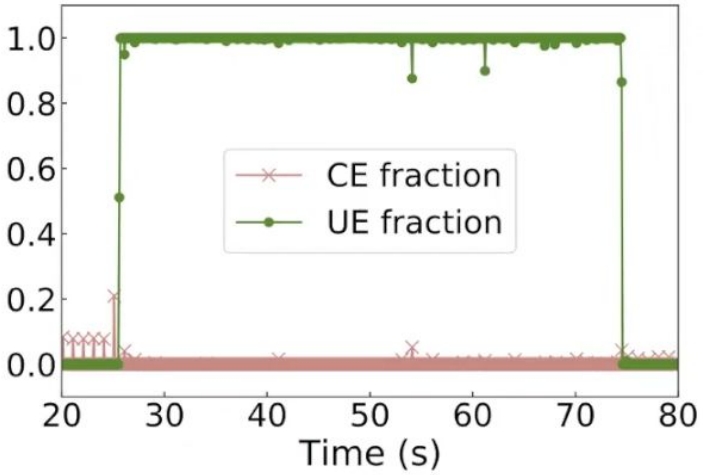


Results

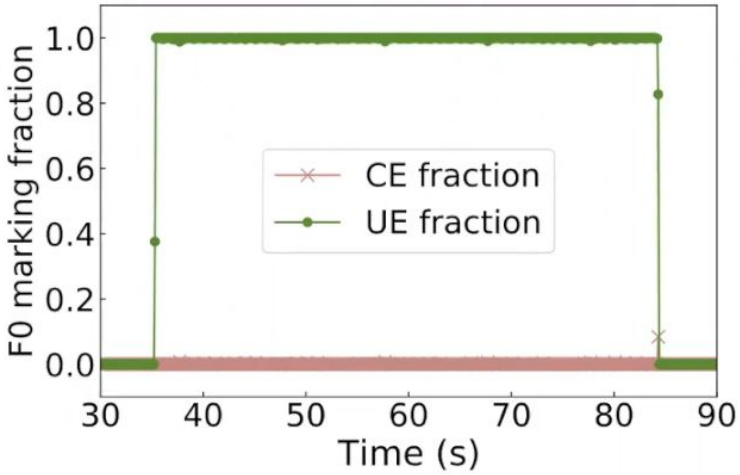
0: 1Gbps F1: 8 Gbps

Victim flow F0 is detected as undetermined other than congested

F0 marking fraction



CEE

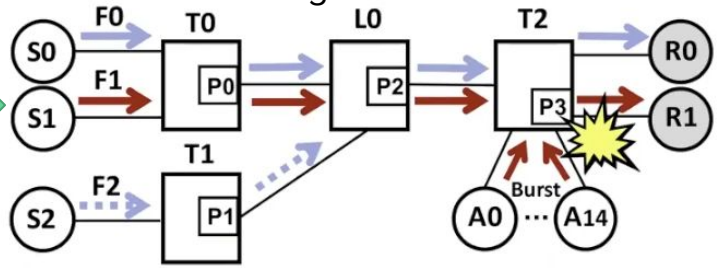


InfiniBand

Results

Typical Scenario Validation : Single congestion point scenario

Figure-1



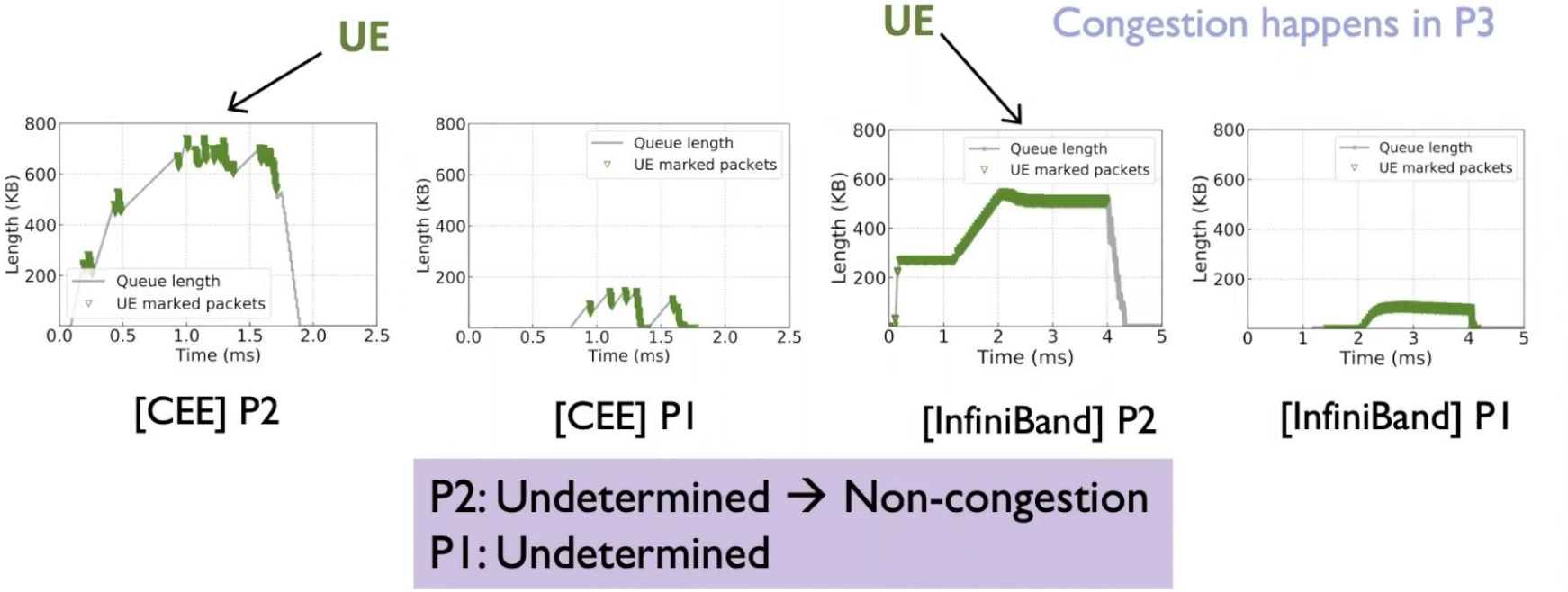
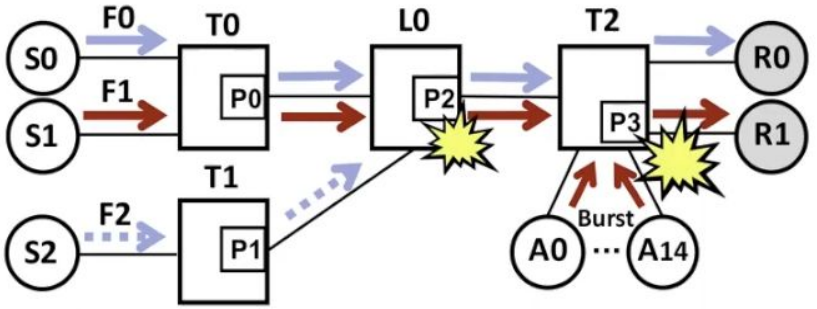
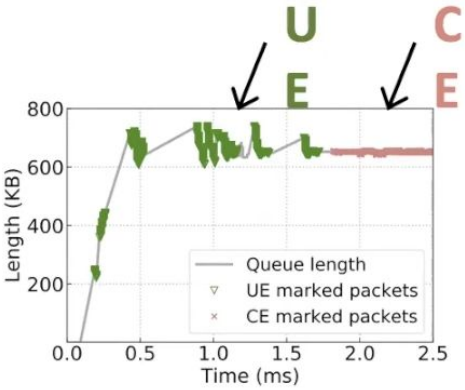


Figure-2

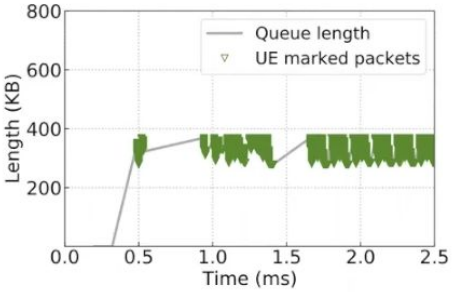
Typical Scenario Validation: Multiple congestion point scenario



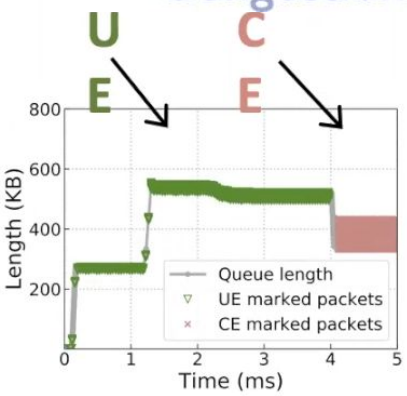
Congestion happens in P3 then P2



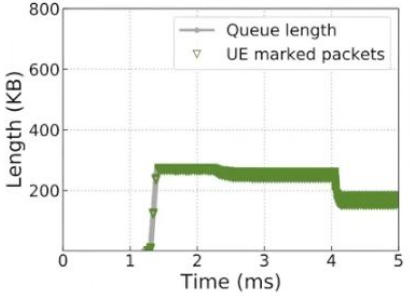
[CEE] P2



[CEE] P1



[InfiniBand] P2



[InfiniBand] P1

P2 state : Undetermined → Congestion
P1 state : Undetermined

Results

Victim Scenario

S0 flows marked with CE:

Scheme	Fraction
ECN (CEE)	26.6%
TCD (CEE)	0%
FECN (IB)	13.5%
TCD (IB)	0%

Victim flows marked with CE

Results

Victim Scenario

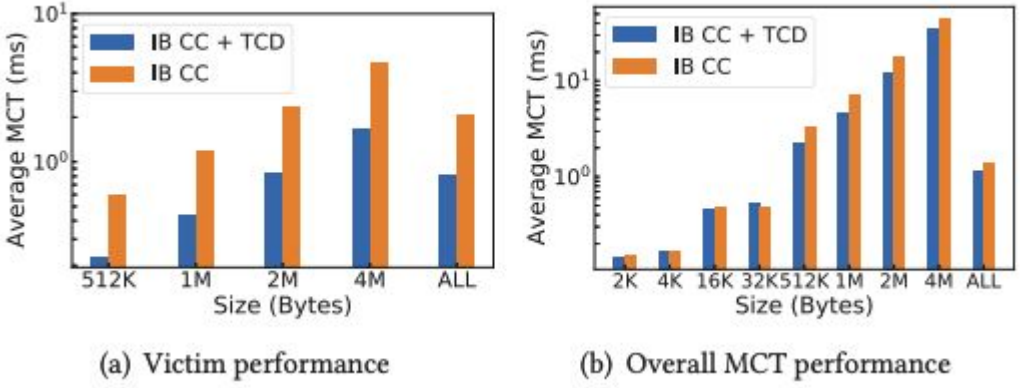
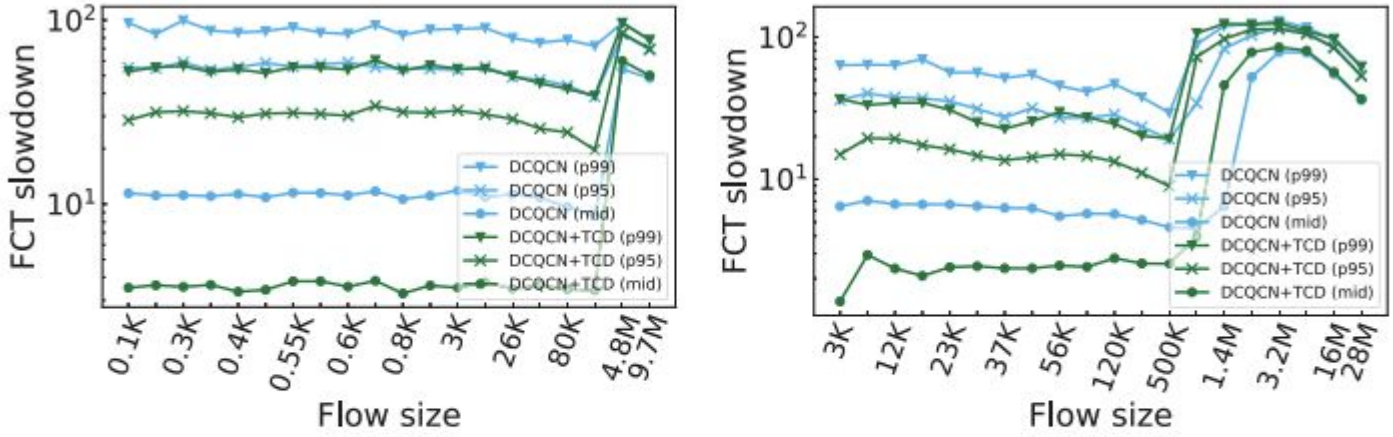


Figure 17: MCT performance (IB CC+TCD)

Results



(a) *Hadoop* workload

(b) *WebSearch* workload

Figure 16: Overall FCT slowdown (DCQCN+TCD)

Conclusion

- Proposed a new approach called t**ernary congestion detection (TCD)** for detecting congestion in lossless networks using CEE and InfiniBand

- Reveal a new port state called the **undetermined state** and **define ternary states**.

- Testbed and extensive simulations demonstrate that TCD can accurately detect congestion ports and identify congested flows as well as undetermined flows

- Experiments shows that existing congestion control algorithms can achieve better performance by combining with TCD, confirming that accurate congestion detection is significant for congestion controls

Critique

1. Novel contribution as it offers an incremental improvement by identifying an Undetermined state.

   Third state in addition to the typical binary congestion/non-congestion states

2. Well-designed research approach

3. They noted that the researchers also conducted experiments, but due to limitations in space, they did not present these findings in the paper.

Critique

3. The authors evaluate TCD through simulation experiments and demonstrate its effectiveness in improving network performance compared to existing congestion control mechanisms.

4. However, the paper could benefit from providing more details on the experimental setup and results.

5. To increase the ability to reproduce the study, it would have been advantageous to include additional information regarding the experimental setup and results.

Critique

6. Their proposed solution has the potential to enhance the effectiveness, impartiality, and flexibility of congestion control mechanisms

7. The authors have presented a compelling argument for their proposed solution.

8. Broader range of existing congestion control mechanisms could have been evaluated and compared them with their proposed solution.

9. Finally, the authors could have discussed the limitations of their proposed solution and how future research could address these limitations

10. TCD is simple and inexpensive to implement in switches

Thankyou!