PortFC: Designing High-performance Deadlock-free BCube Networks

Peirui Cao[1] **Rui Ning**[1] Hongwei Yang[2] Zhaochen Zhang[1]
Chang Liu[1] Rui Li[1] Yongqi Yang[1] Yunzhuo Liu[1] Chengyuan Huang[1]
Tao Sun[2] Xiaodong Duan[2] Guihai Chen[1] Chen Tian[1]

[1]Nanjing University [2]China Mobile

Server-centric topology is widely used

Training

OpenAI

Modular DC

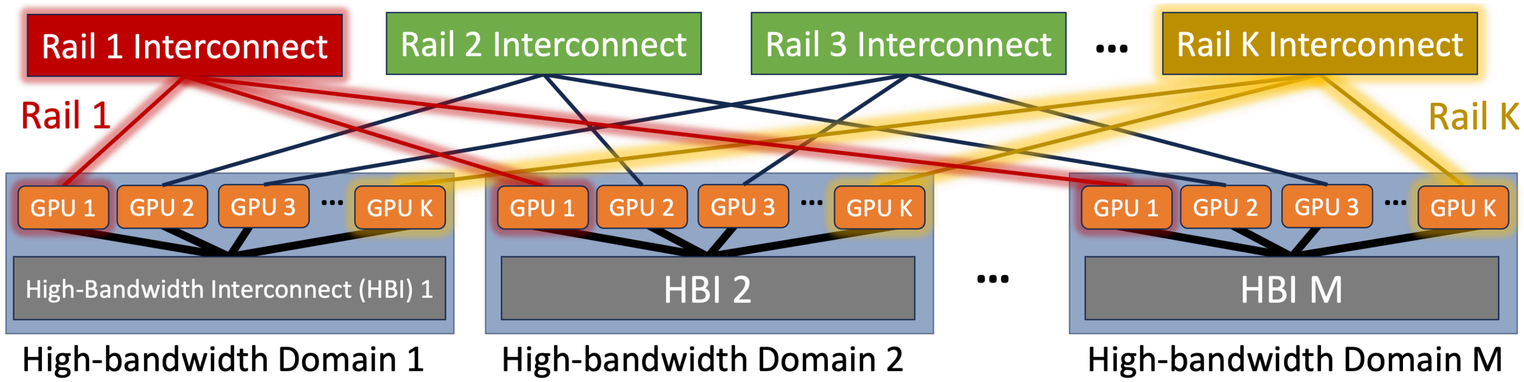
DELL EMC

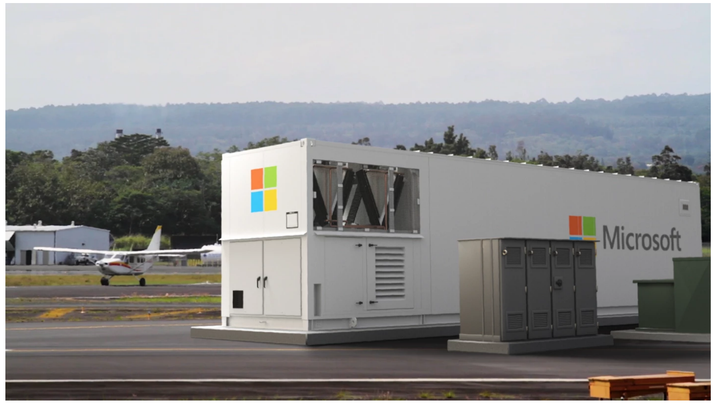
Portable DC

IBM

Cloud

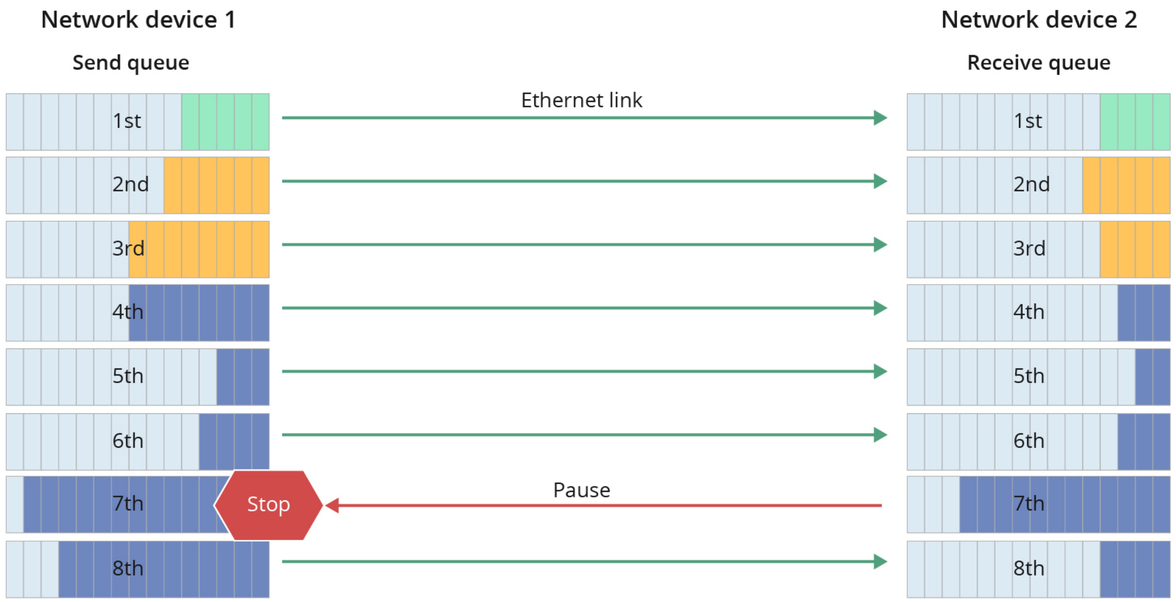
Azure





State-of-the-art Flow Control

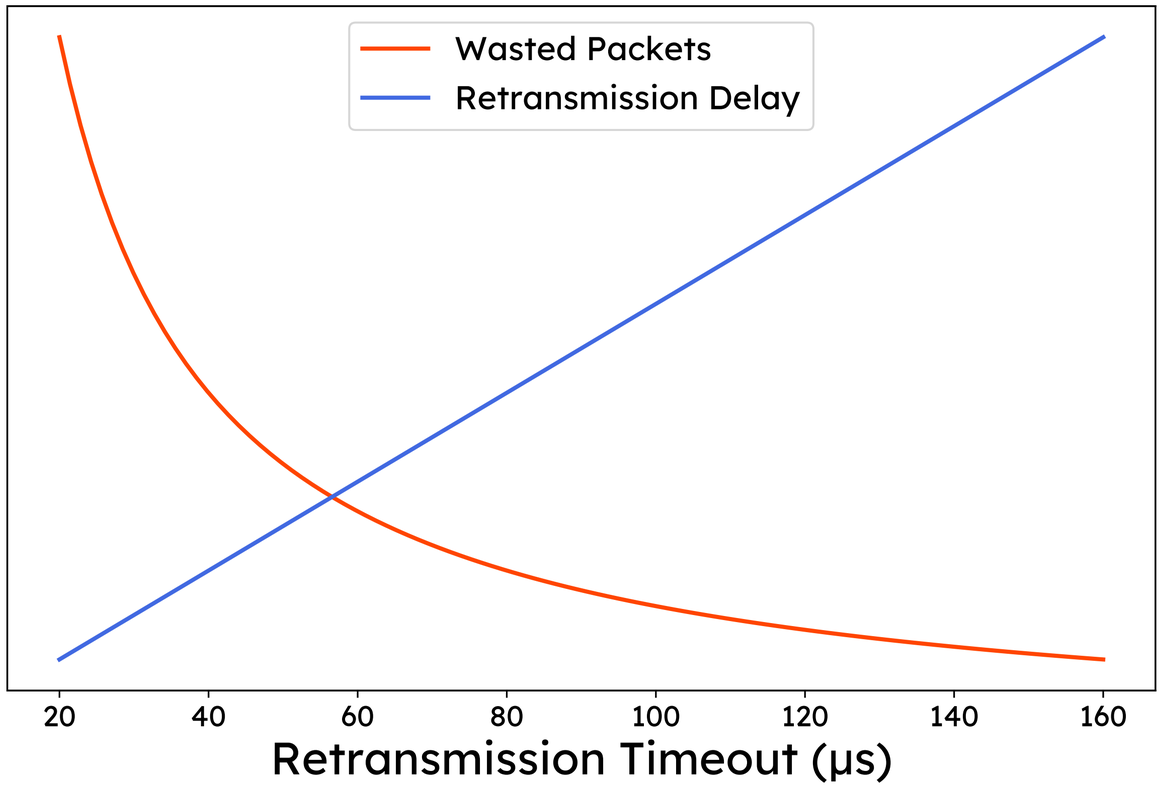
- Flow control mechanism tries to achieve lossless Ethernet in RoCE via controlling whether packets are allowed to be sent [PFC IEEE Std], [IRN SIGCOMM'18].



- Remove backpressure signal
- Relax the lossless requirement
- More efficient loss recovery

PFC ⟷ IRN

Problem of IRN: High ReTx Overhead



Lower RTO: ✅

More wasted packets 😭

Lower ReTx delay 😎

Higher RTO: ✅

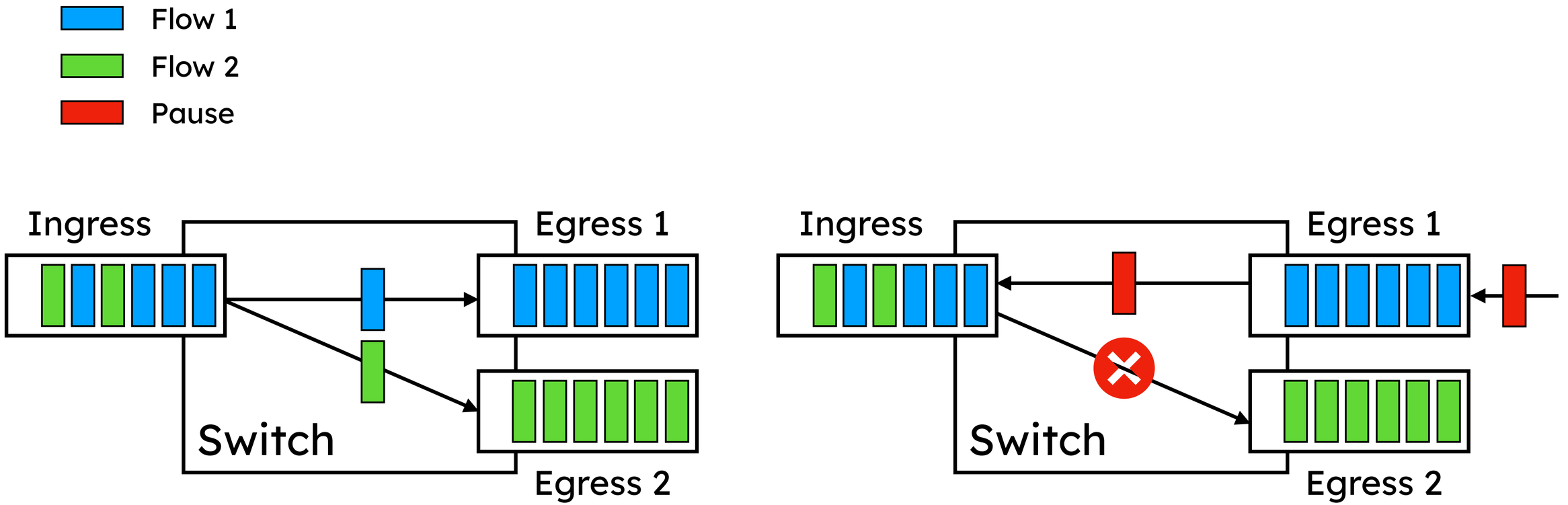
Fewer wasted packets 😎

Higher ReTx delay 😭

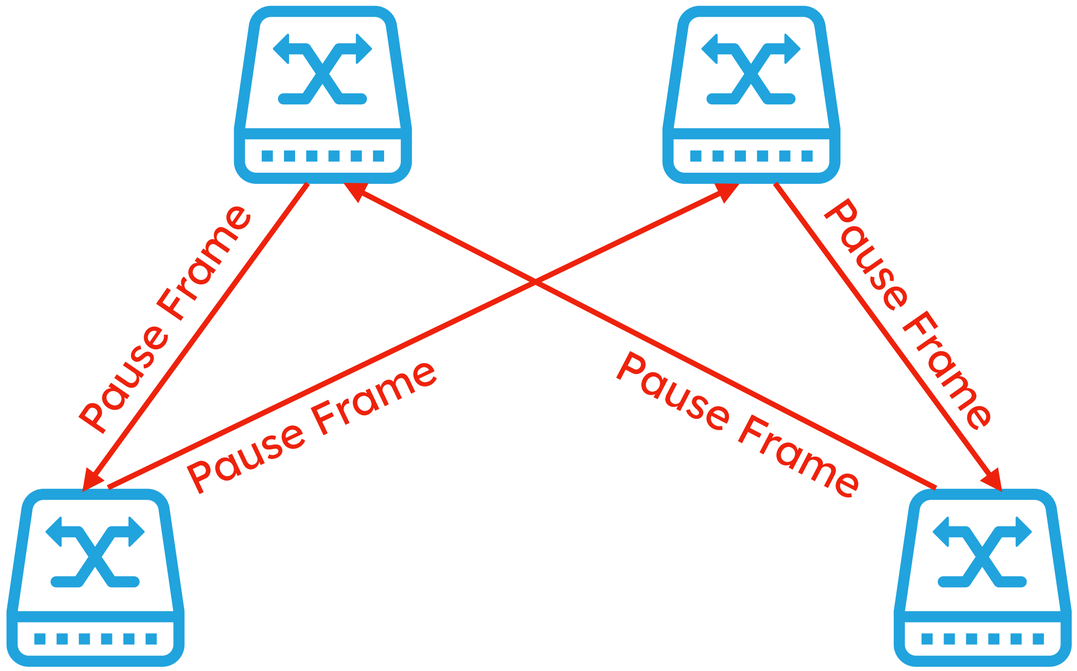
Optimal RTO: ❌

Hard to find in BCube Topo

Problem of PFC: Head-of-Line Blocking



Problem of PFC: Deadlock



Deadlock Detection: ✅

Simple to implement 😎

Deadlock again easily 😭

Deadlock Prevention: ✅

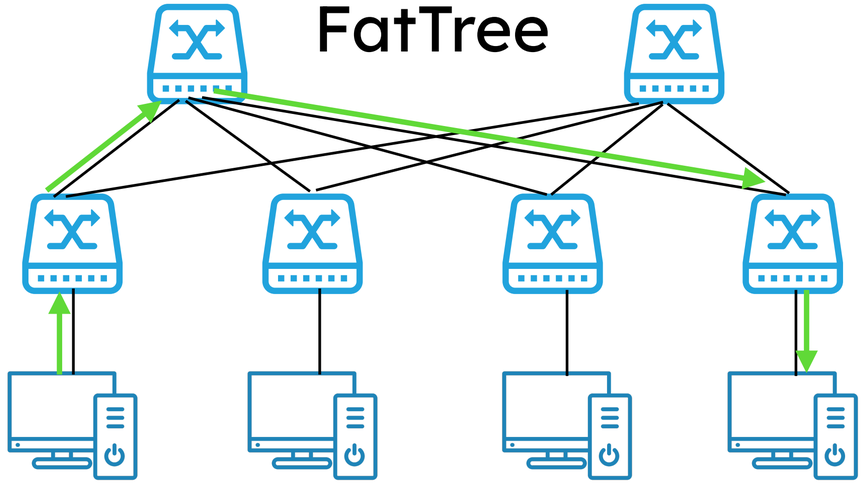
Hard to implement 😭

Ineffective in BCube 😭

No Deadlock: ❌

Unrealistic under PFC

Challenge: Unresolved in BCube Topo

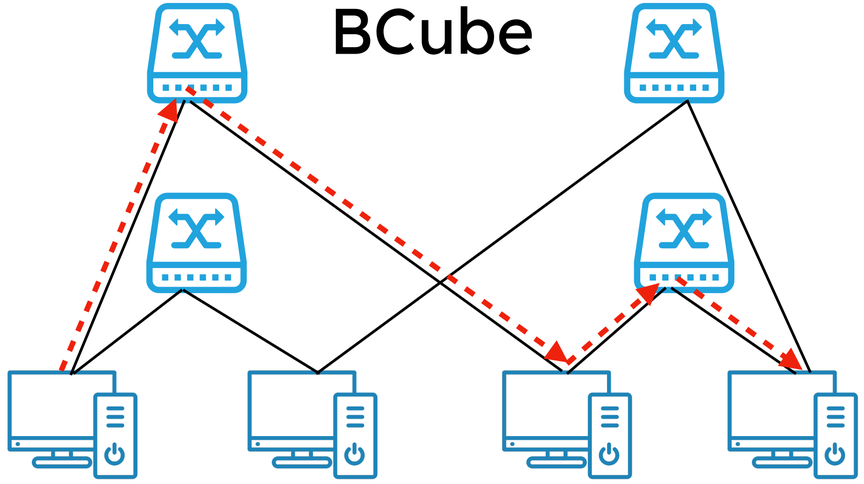


✔

Up-Down Routing:

Solves CBD in Fat-Tree / Clos 😎

Why? Server can forward!

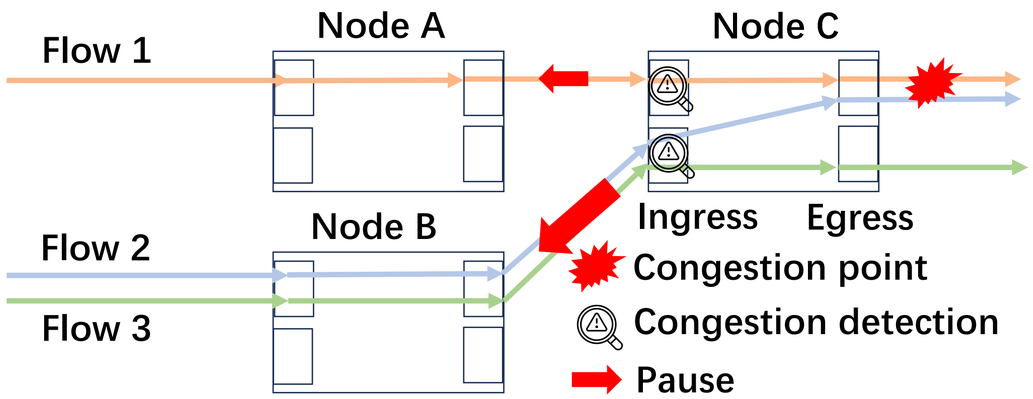


✘

Up-Down Routing:

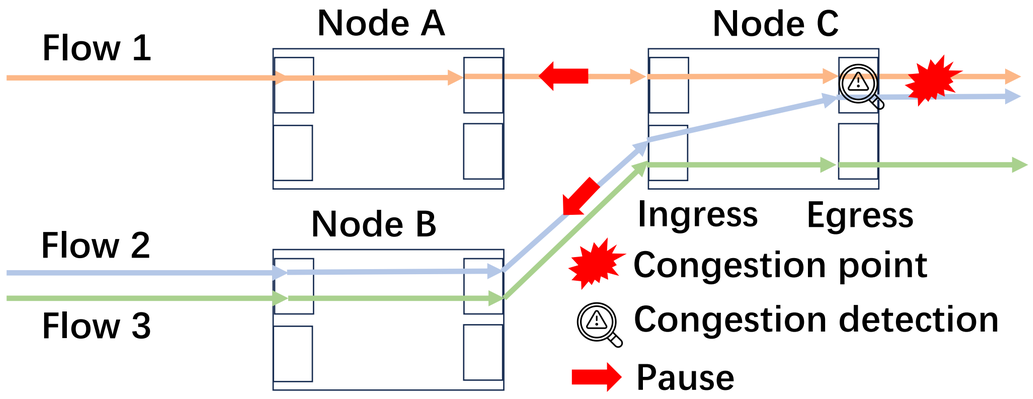
Not applicable to BCube 😭

Opportunity: Egress Congestion Detection



Ingress Detection:

HoLB Issue Happened 😭



Egress Detection:

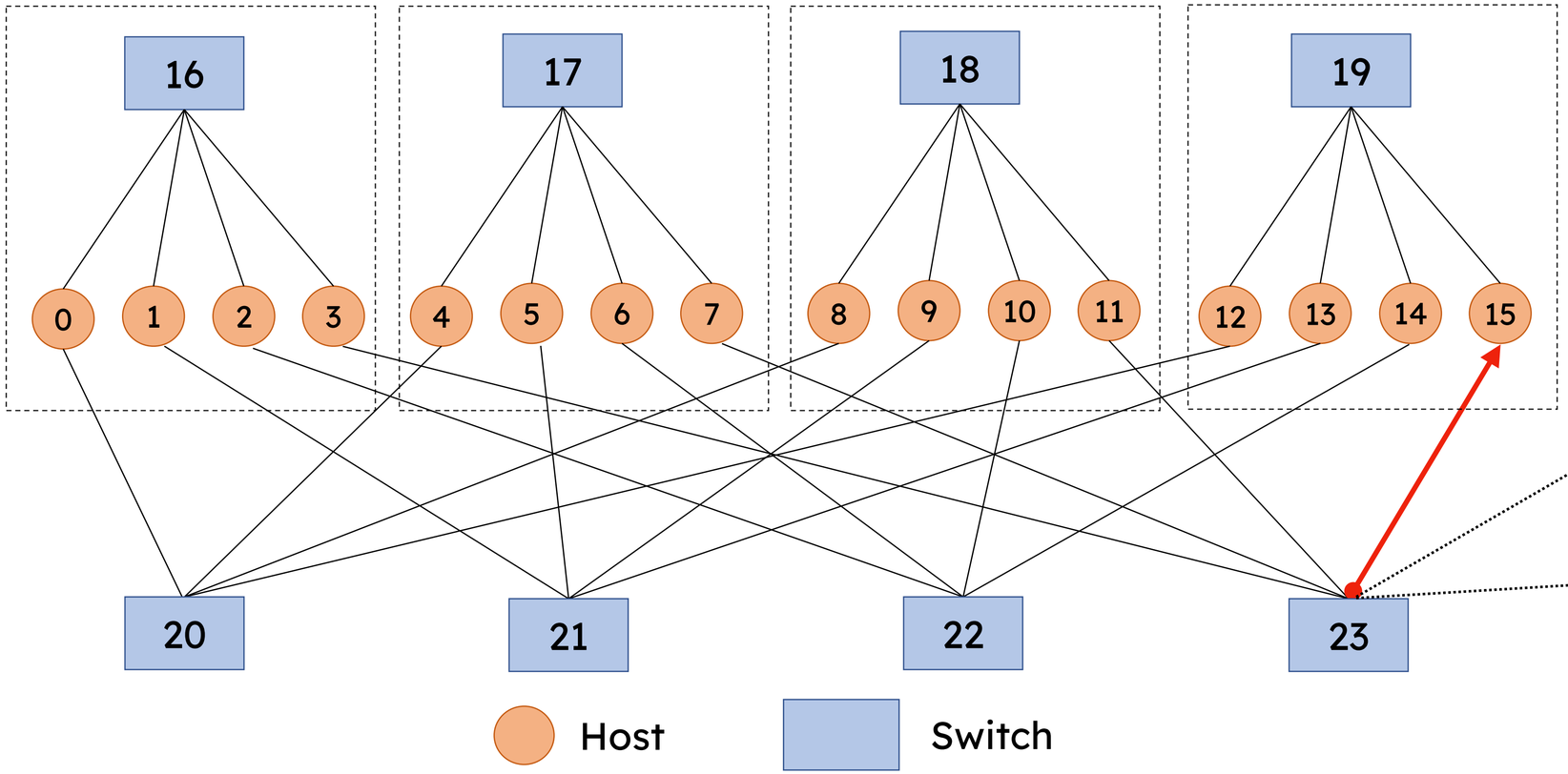
HoLB Issue Solved 😎

PortFC Overview

- Queue Allocation
  - Queue classifications
  - Differences between host and switch queue allocation
- Control Frame Reaction
  - Host / Switch side reaction
- Deadlock-free Proof Sketch

PortFC: Switch Queue Allocation

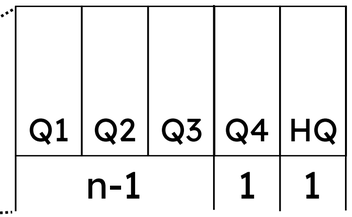
Select different queues based on the egress port of next-next-hop switch



Q1~3: Transit Flow

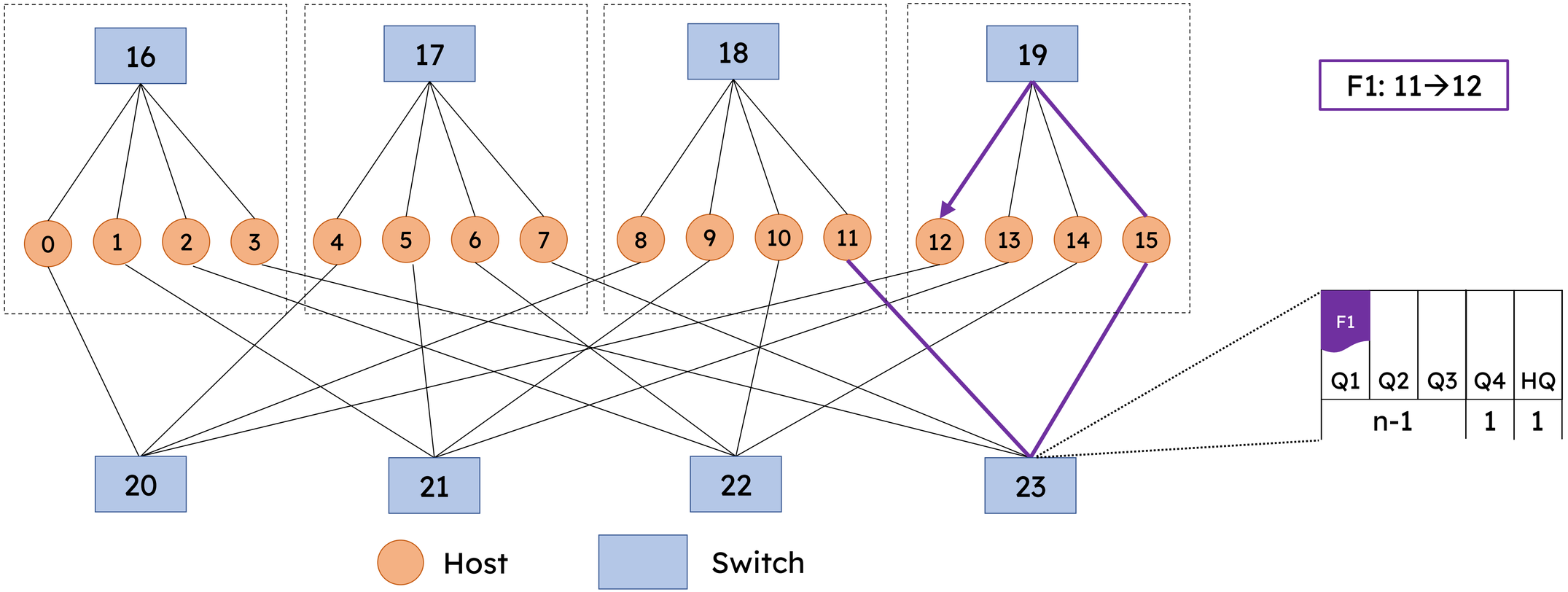
Q4: Last-hop Flow

HQ: High-Priority



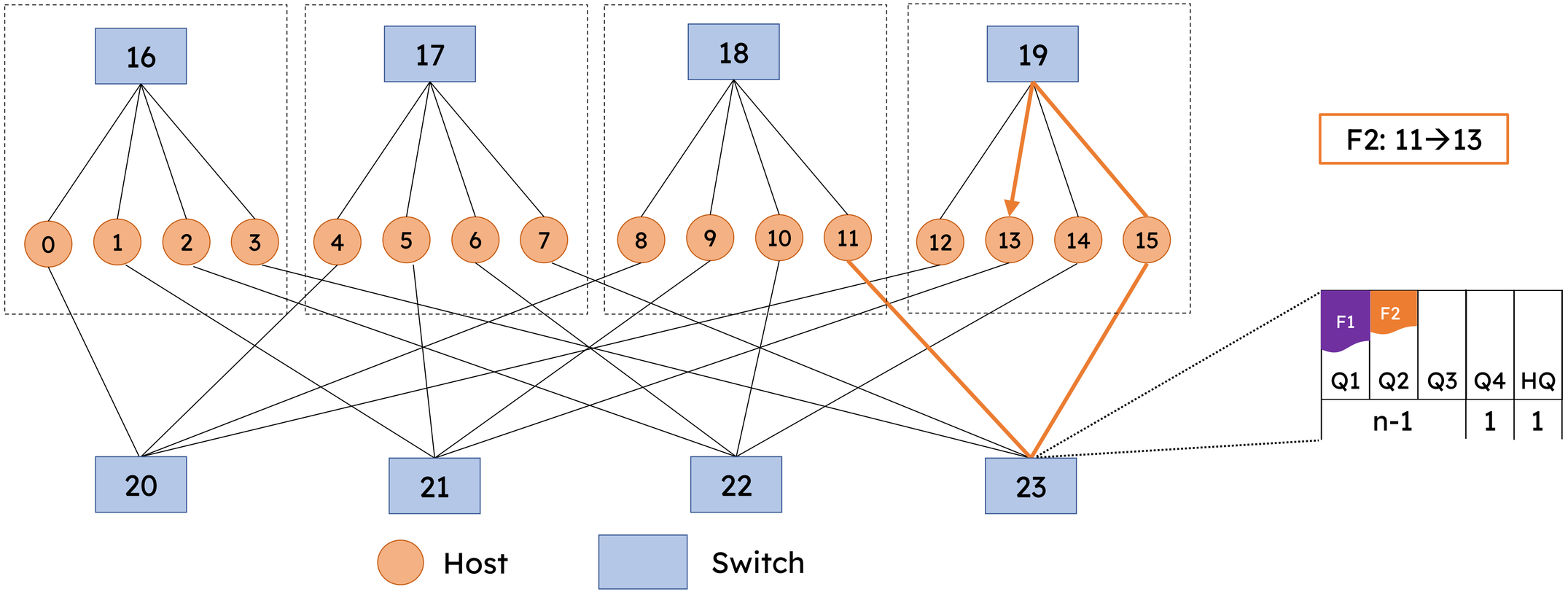
PortFC: Switch Queue Allocation

Consider a flow traversing the link from switch #23 to host #15



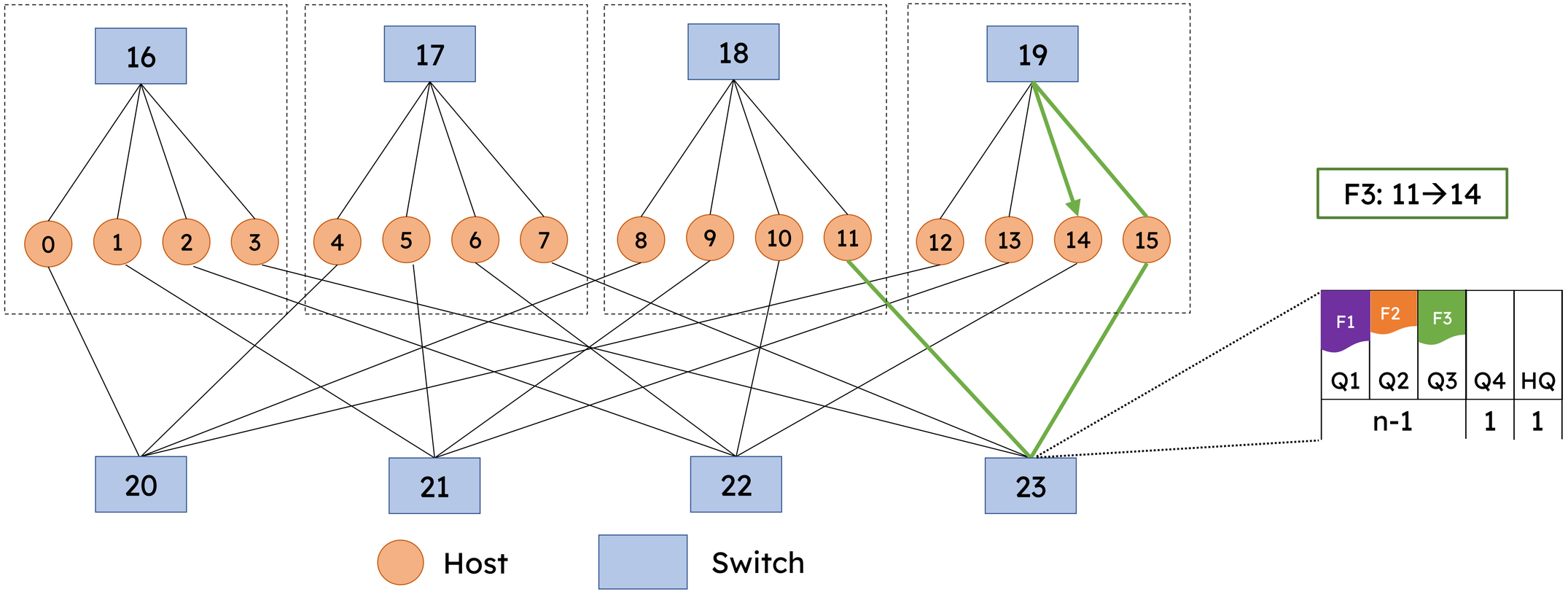
PortFC: Switch Queue Allocation

Consider a flow traversing the link from switch #23 to host #15



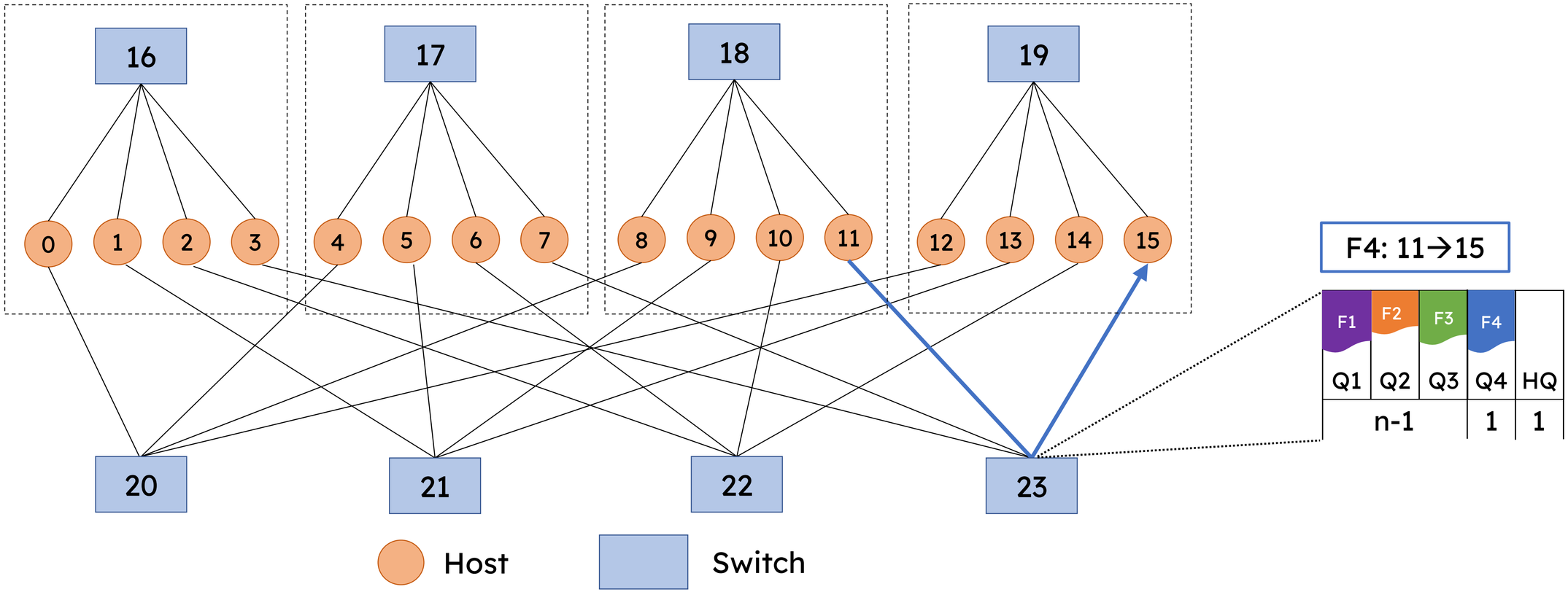
PortFC: Switch Queue Allocation

Consider a flow traversing the link from switch #23 to host #15



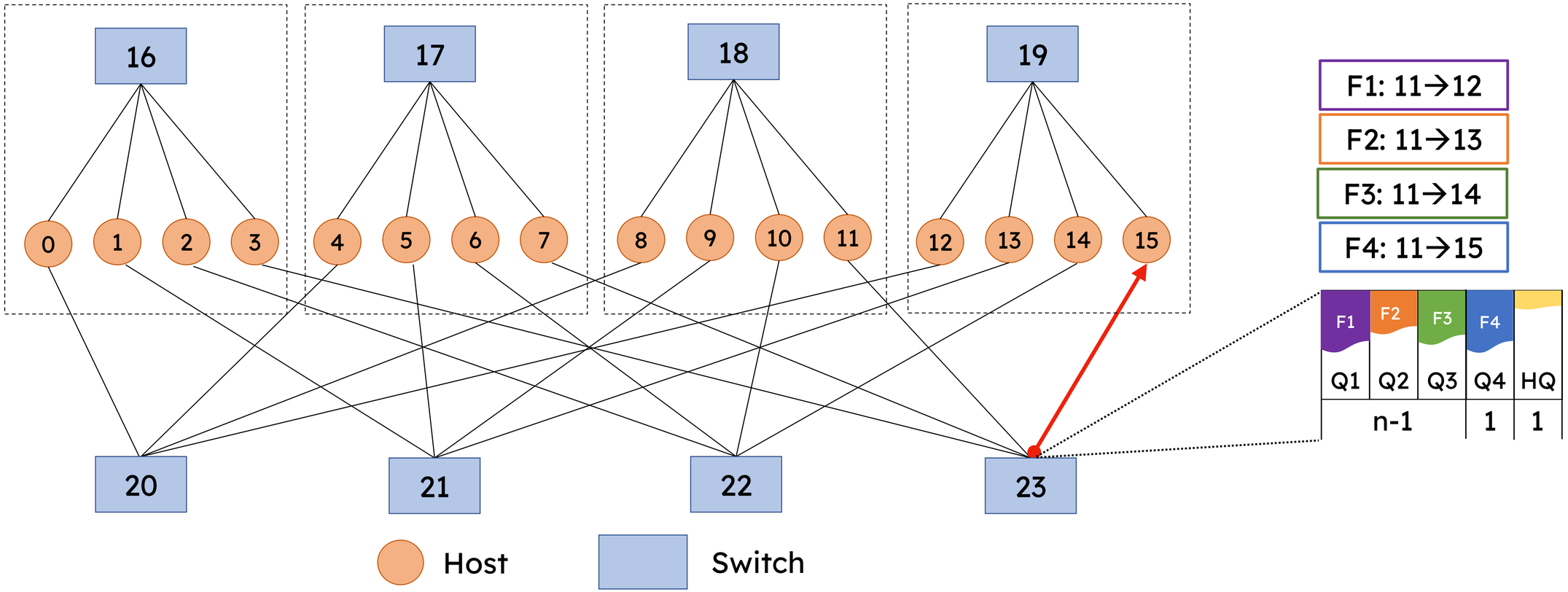
PortFC: Switch Queue Allocation

Consider a flow traversing the link from switch #23 to host #15



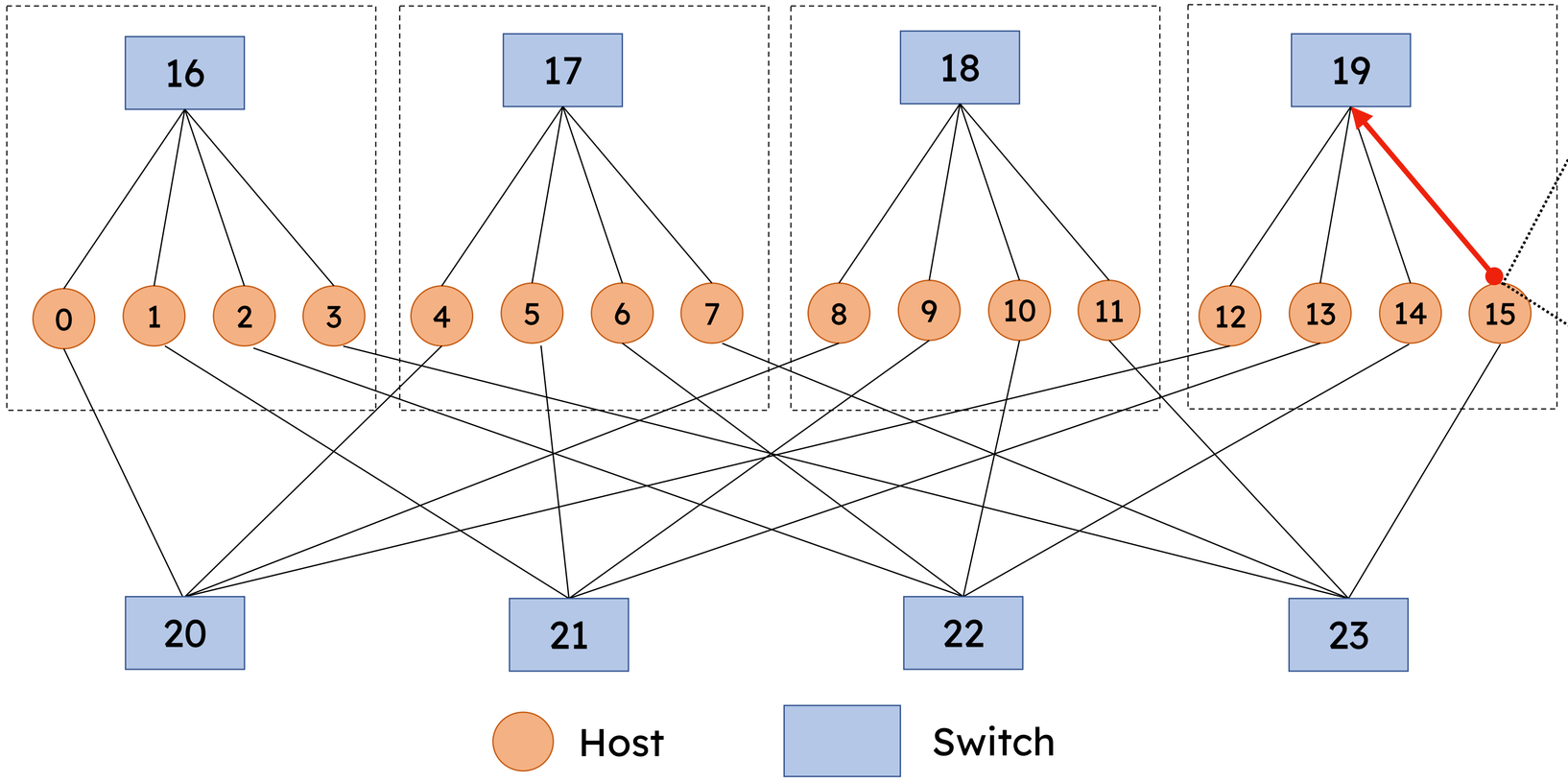
PortFC: Switch Queue Allocation

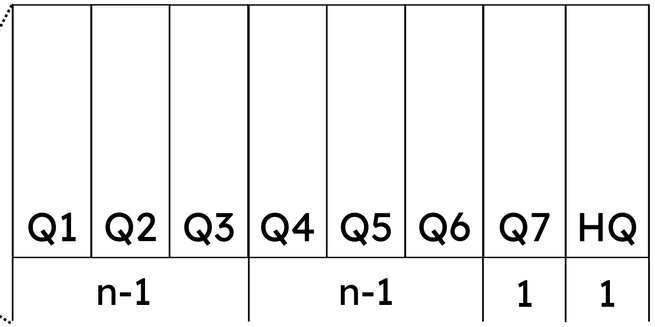
Consider a flow traversing the link from switch #23 to host #15



PortFC: Host Queue Allocation

Select different queues based on the egress port of next-hop switch





Q1~3: Local 4-hop Flow

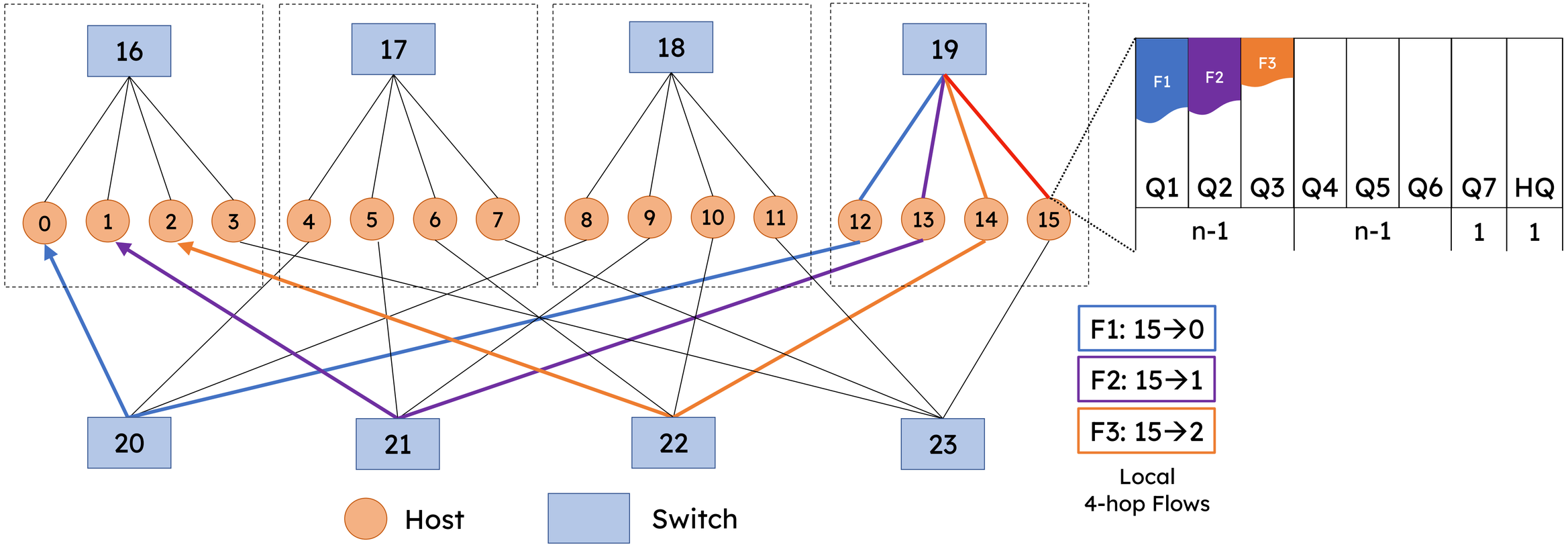
Q4~6: Local 2-hop Flow

Q7: Remote Flow

HQ: High-Priority

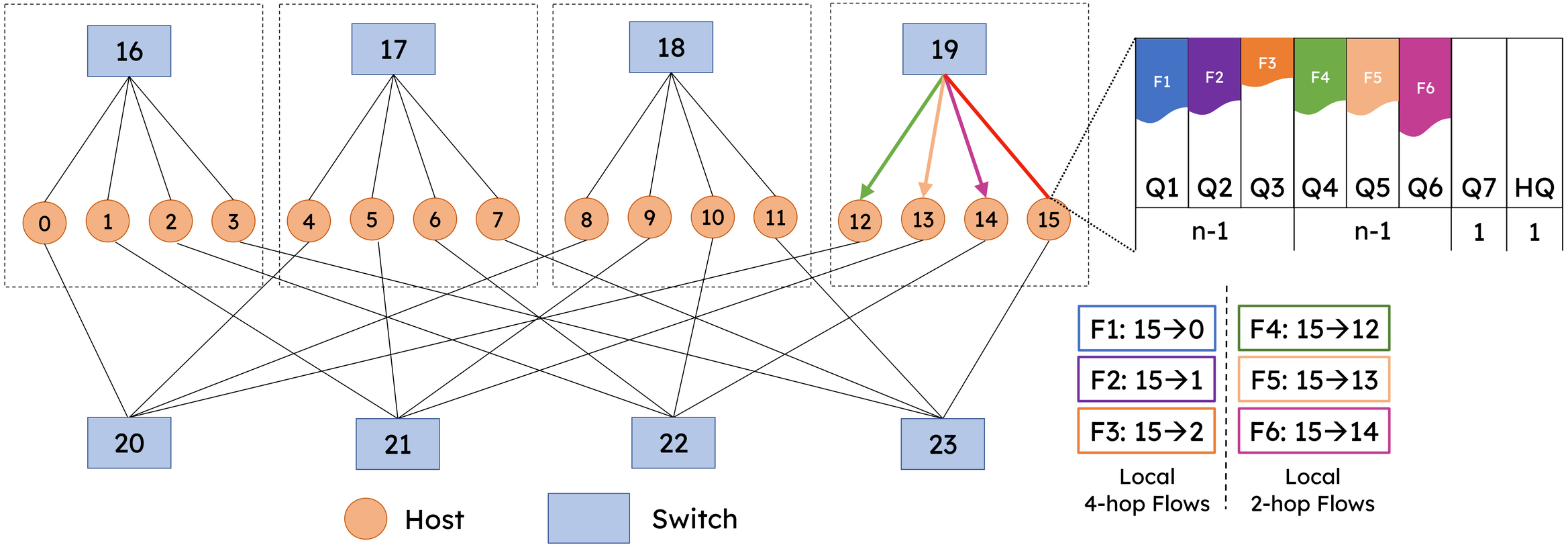
PortFC: Host Queue Allocation

Consider flow through host #15



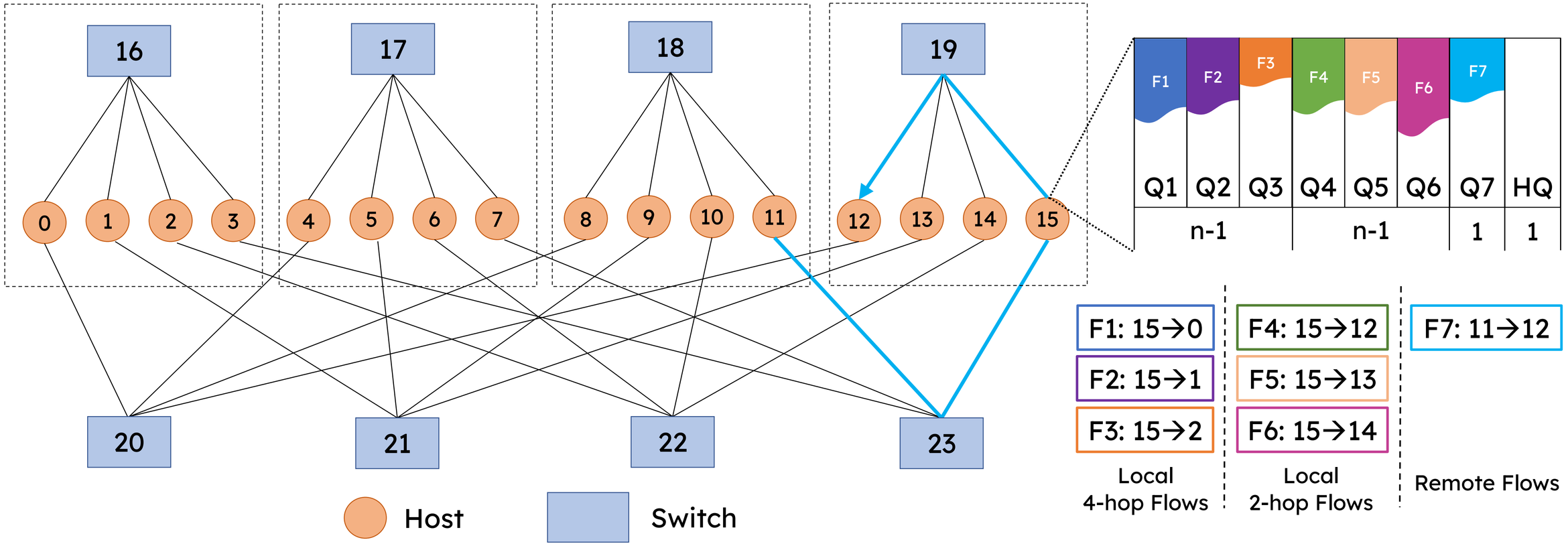
PortFC: Host Queue Allocation

Consider flow through host #15



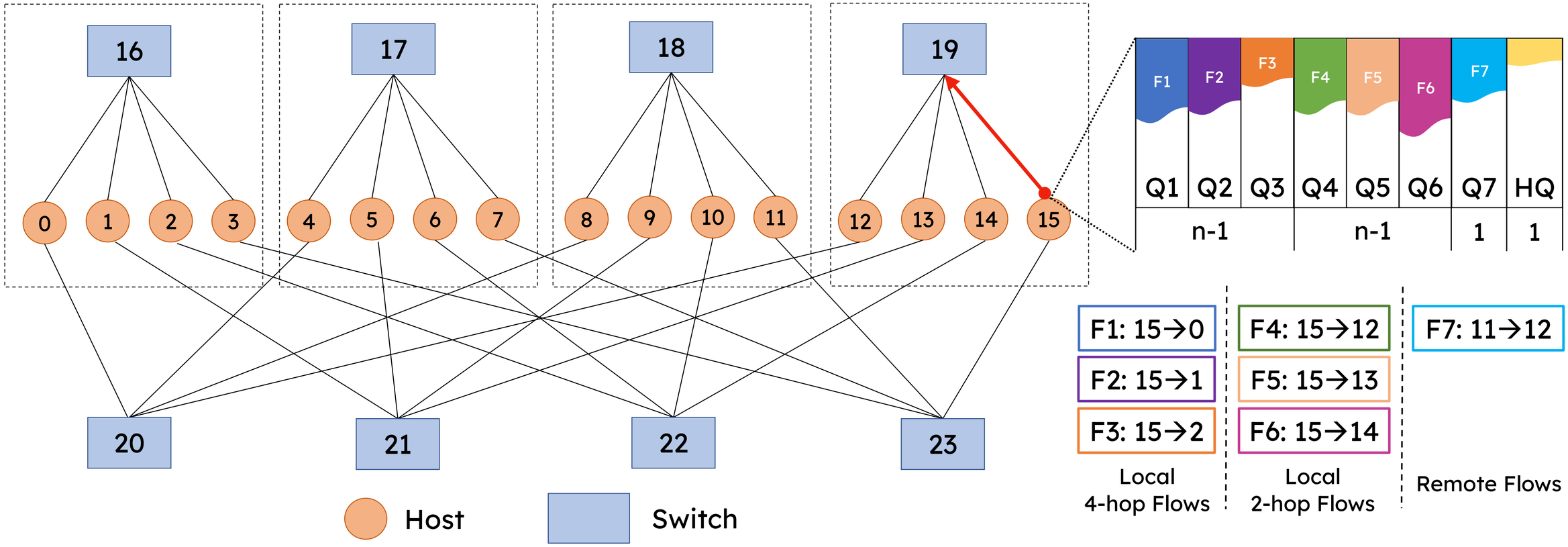
PortFC: Host Queue Allocation

Consider flow through host #15

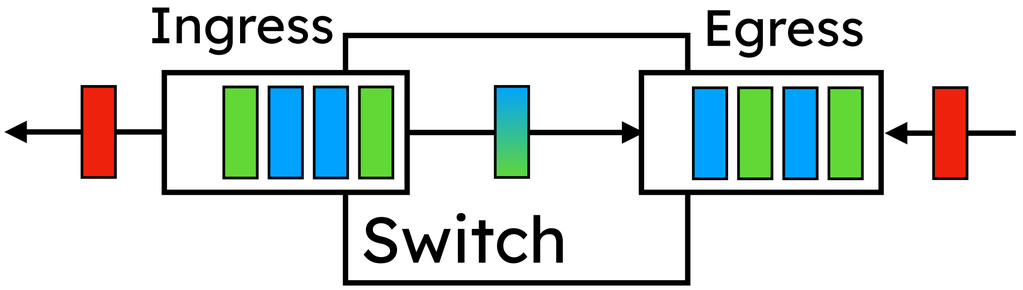


PortFC: Host Queue Allocation

Consider flow through host #15

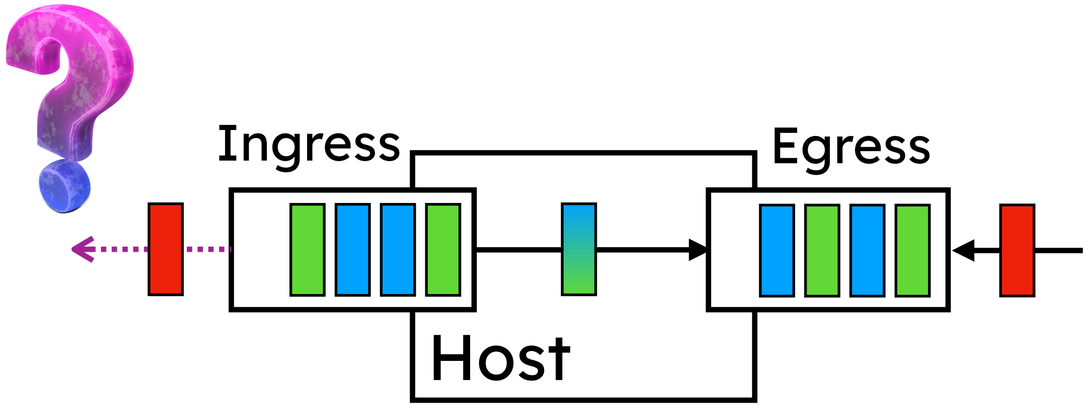


PortFC: Control Frame Reaction



- Pause / Resume Queue

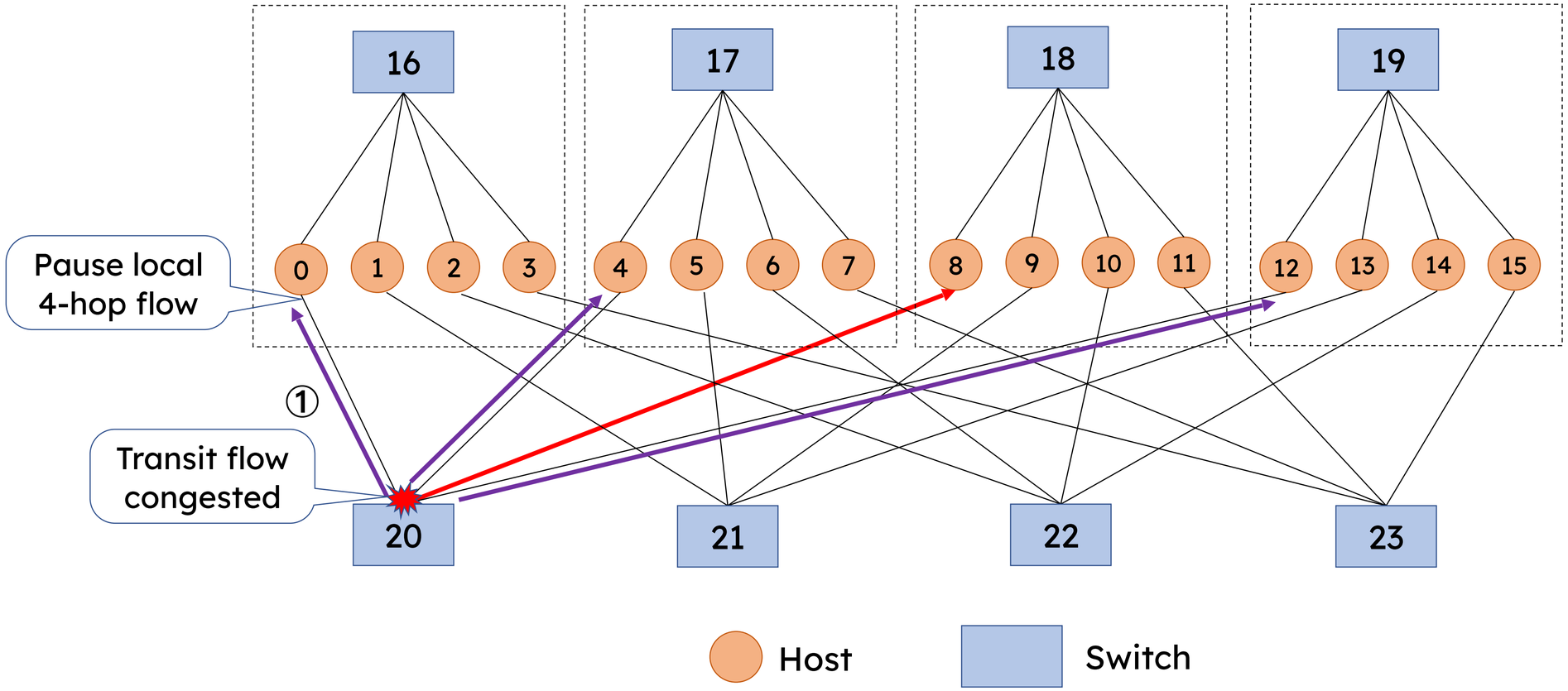
Key Design to Avoid Deadlock!



- Pause / Resume Queue
- **Forward to upstream***

PortFC: Deadlock Free Proof

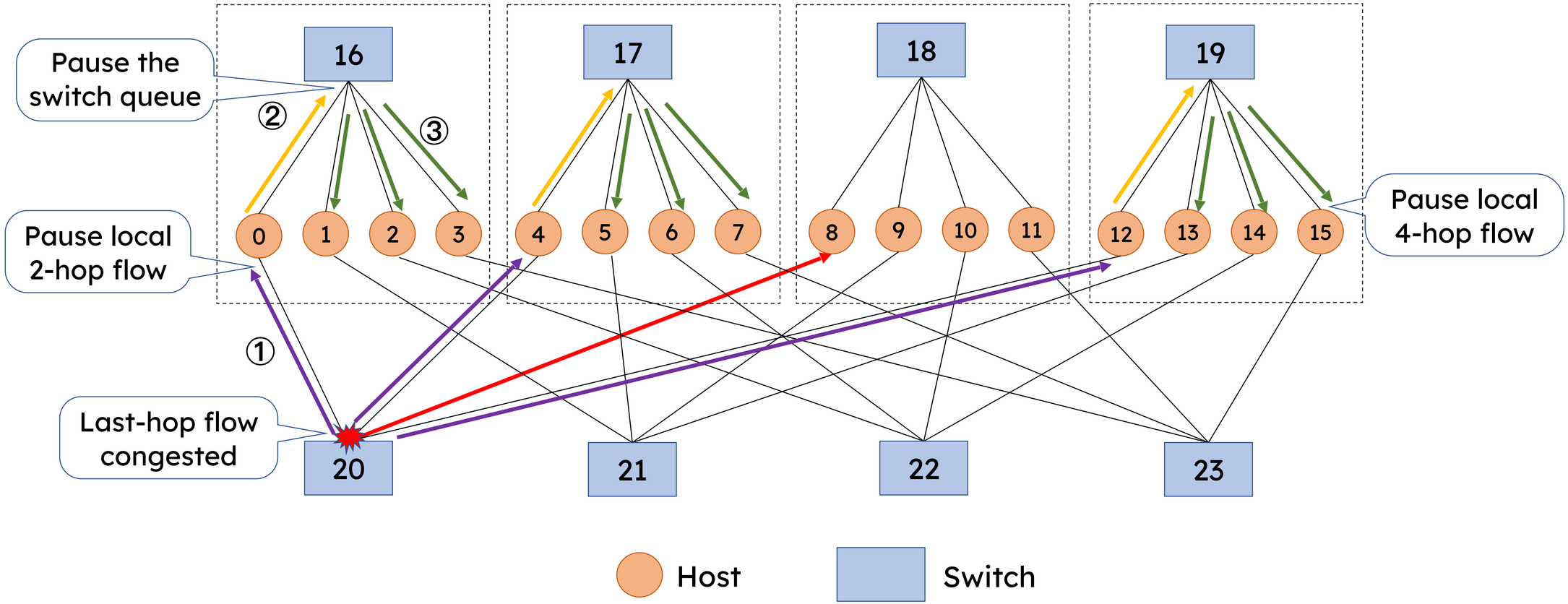
1. Congestion of transit flow stopped at last-hop host



Congestion propagation path of transit flow

PortFC: Deadlock Free Proof

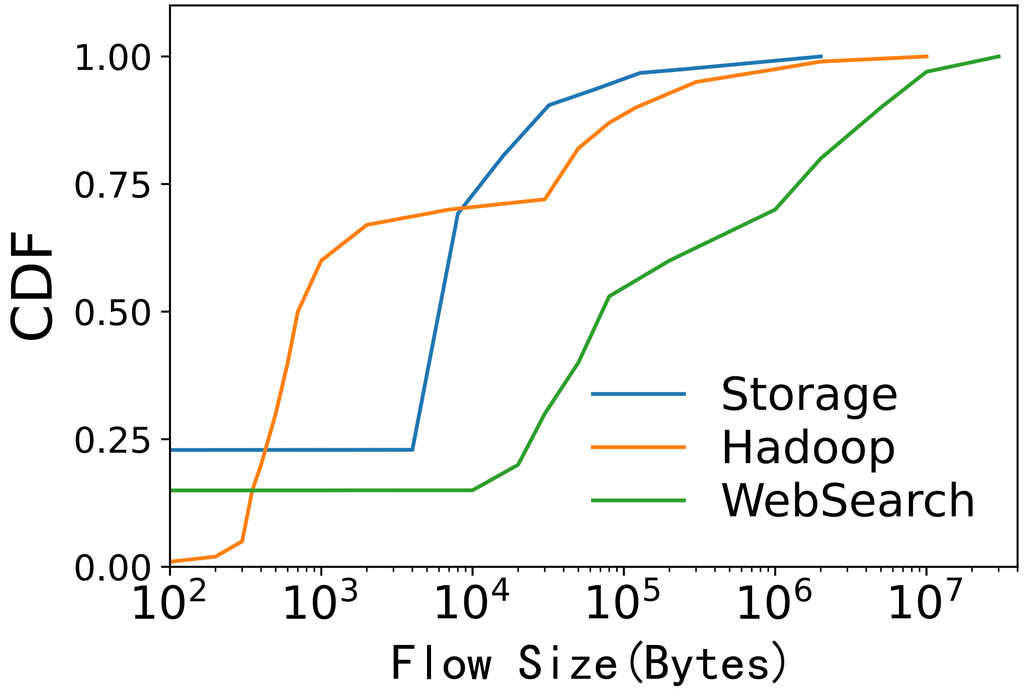
1. Congestion of transit flow stopped at last-hop host
2. Congestion of last-hop flow stopped at last-last-hop host



Congestion propagation path of last-hop flow

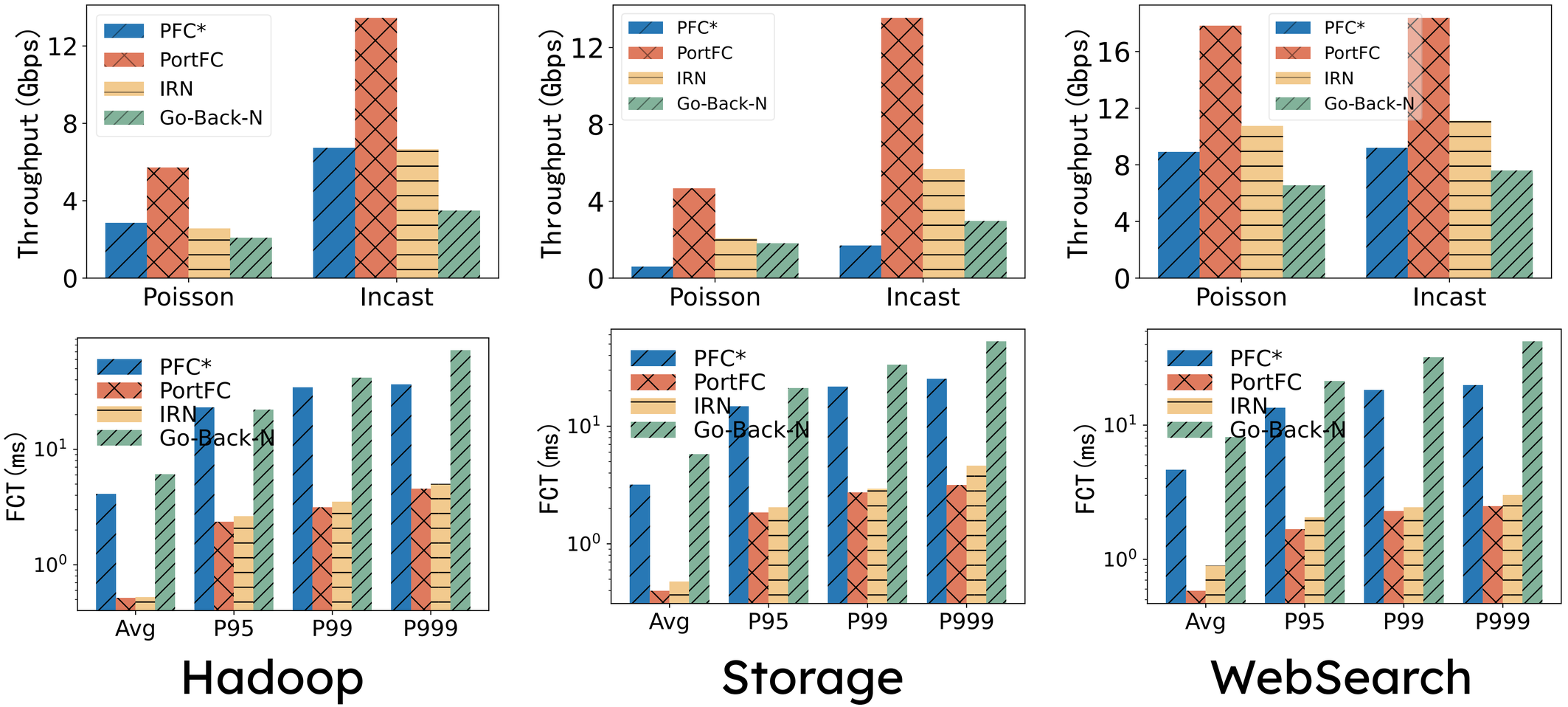
Evaluations

- Setup:
  - BCube(4, 1) & BCube(8, 1) Topology
  - 100 Gbps Link
- Workloads: Hadoop, Storage, WebSearch
- Metric:
  - Flow Completion Time
  - Throughput
  - Queue Length



Experiments

PortFC achieves 1.7×–8.0× higher end-to-end throughput.



Takeaway

- PortFC eliminates HoLB and deadlocks while ensuring losslessness.
  - Detect congestion based on egress instead of ingress queue
  - Split flows based on the egress port of next-hop switch
- PortFC achieves up to 8.0x higher throughput and reduces latency by up to 87.7% compared to SOTA systems.

✉️ caopeirui@nju.edu.cn

✉️ rning@smail.nju.edu.cn