NDSS SYMPOSIUM

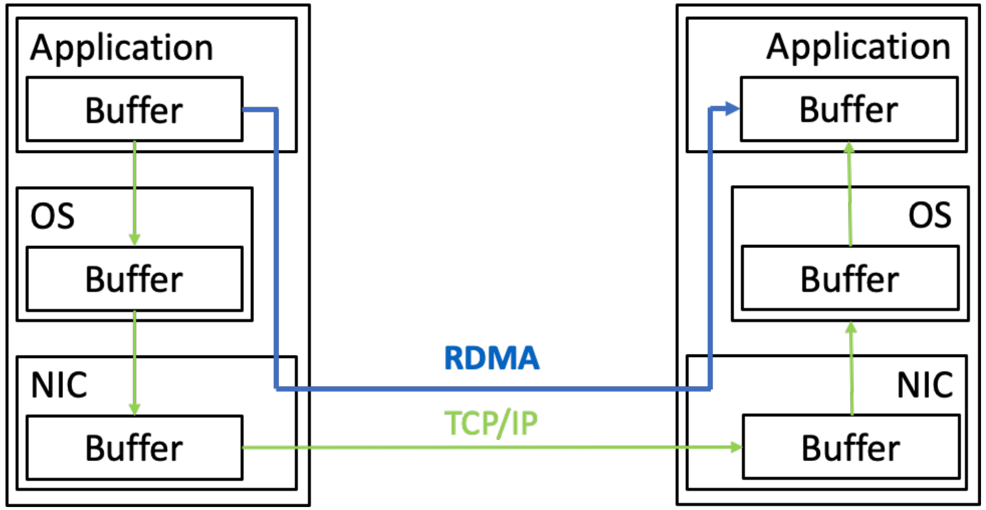
LoRDMA: A New Low-Rate DoS Attack in RDMA Networks

Shicheng Wang[1], Menghao Zhang[2], Yuying Du[3], Ziteng Chen[4],
Zhiliang Wang[1], Mingwei Xu[1], Renjie Xie[1], Jiahai Yang[1]

1. Tsinghua University **2. Beihang University** **3. Information Engineering University** **4. Southeast University**

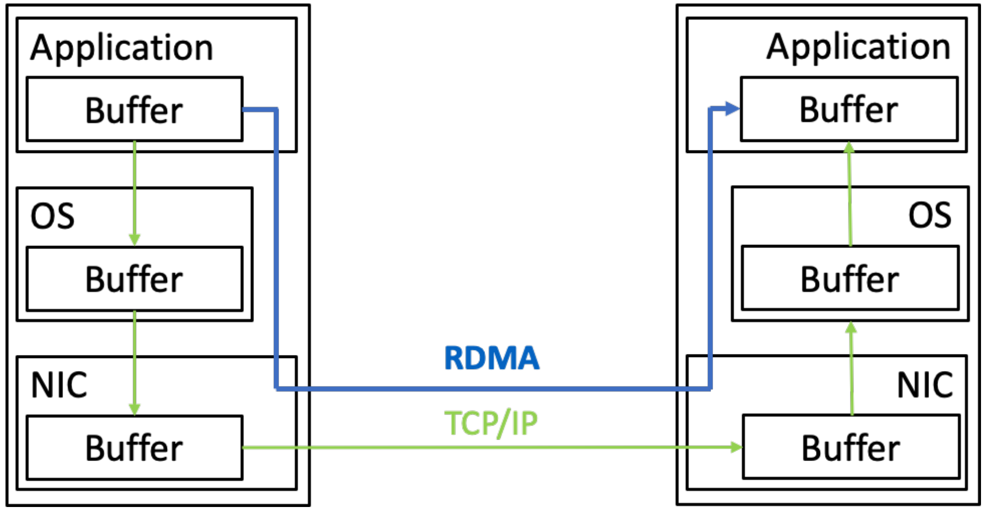
RDMA network

- RDMA (Remote Direct Memory Access) is becoming an attractive trend
  - Access remote host memory directly without CPU intervention
    - High bandwidth:10/40~100/400Gbps
    - Low delay: <100us



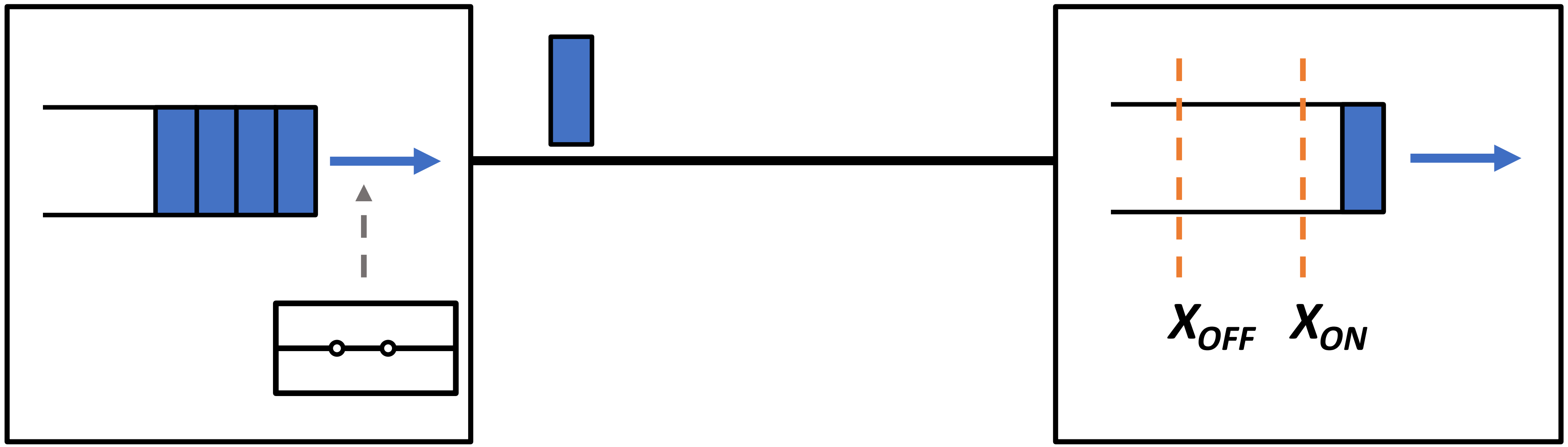
RDMA network

- RDMA (Remote Direct Memory Access) is becoming an attractive trend
  - Access remote host memory directly without CPU intervention
    - High bandwidth:10/40~100/400Gbps
    - Low delay: <100us
  - Application scenarios
    - Distributed machine learning
    - Distributed cloud storage
    - Search queries



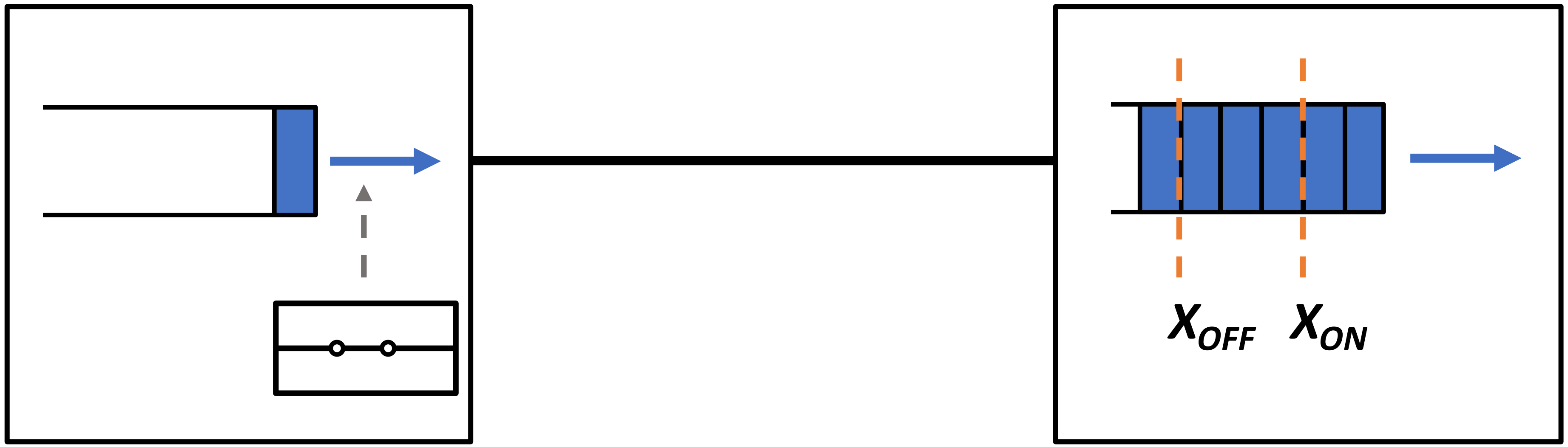
RDMA traffic control mechanism

- Priority Flow Control (PFC)



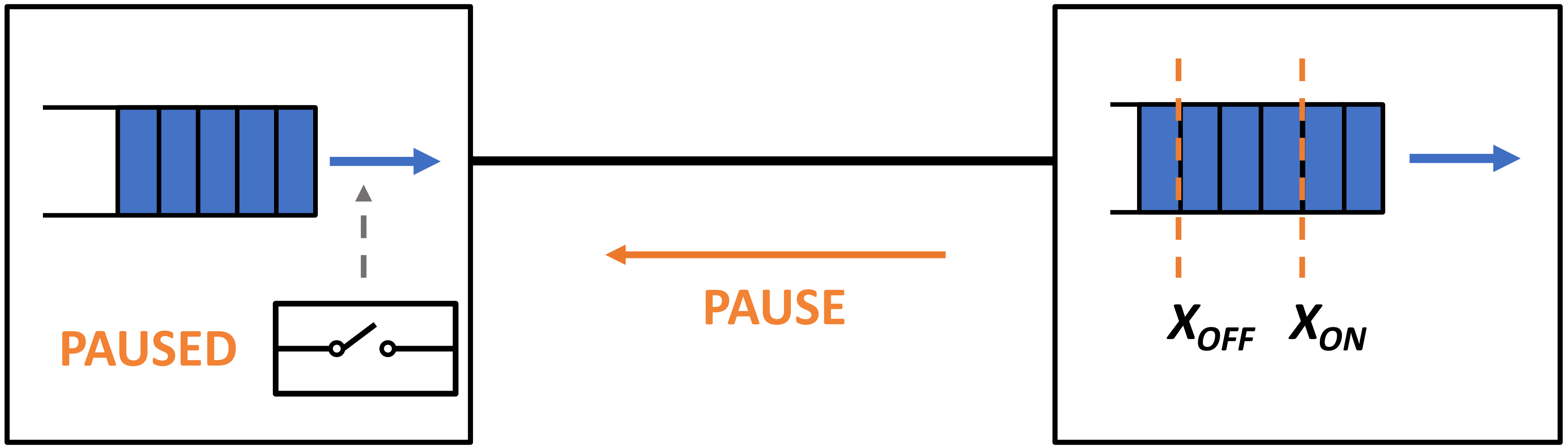
RDMA traffic control mechanism

- Priority Flow Control (PFC)
  - When **busy**, send **PAUSE** frame
    - Ingress queue length > X_{OFF} → PAUSE



RDMA traffic control mechanism

- Priority Flow Control (PFC)
  - When **busy**, send **PAUSE** frame
    - Ingress queue length > X_{OFF} → PAUSE

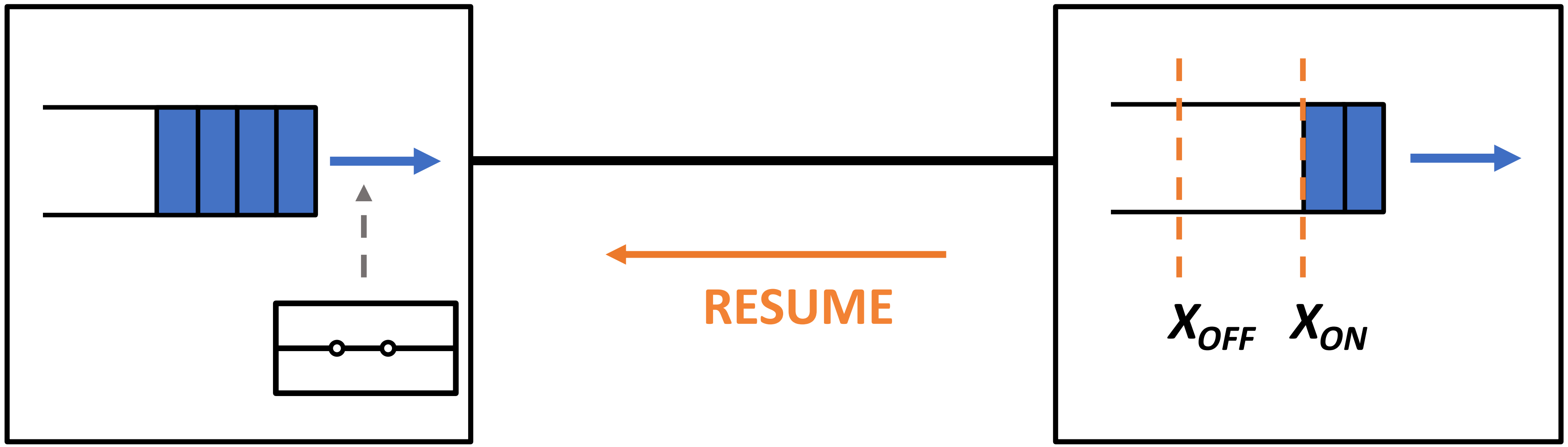


RDMA traffic control mechanism

- Priority Flow Control (PFC)
  - When **busy**, send **PAUSE** frame
    - Ingress queue length > X_{OFF} → PAUSE
  - When not **busy**, send **RESUME** frame
    - Ingress queue length < X_{ON} → RESUME

RDMA traffic control mechanism

- Priority Flow Control (PFC)
  - When **busy**, send **PAUSE** frame
    - Ingress queue length > X_{OFF} → PAUSE
  - When not **busy**, send **RESUME** frame
    - Ingress queue length < X_{ON} → RESUME

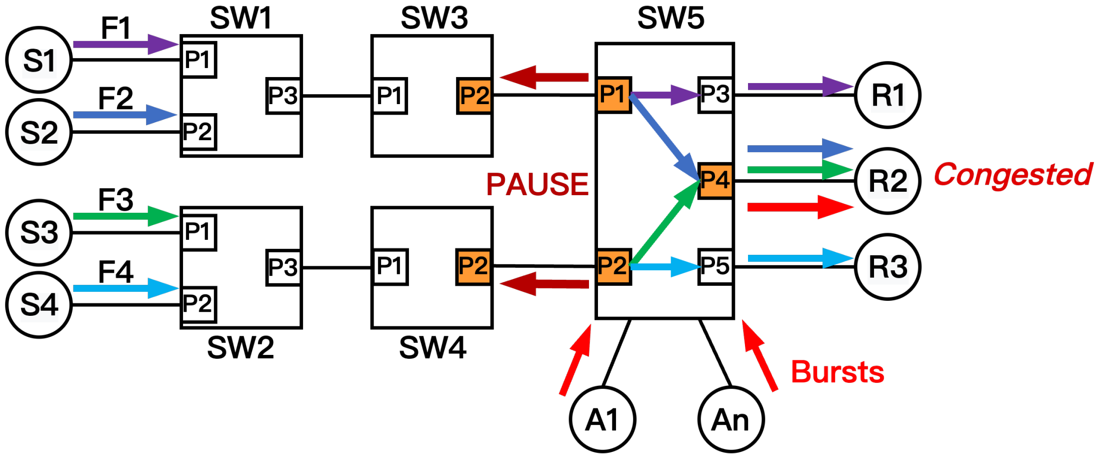


PFC issues

- Hop-by-hop congestion spread
  - Head-of-line blocking

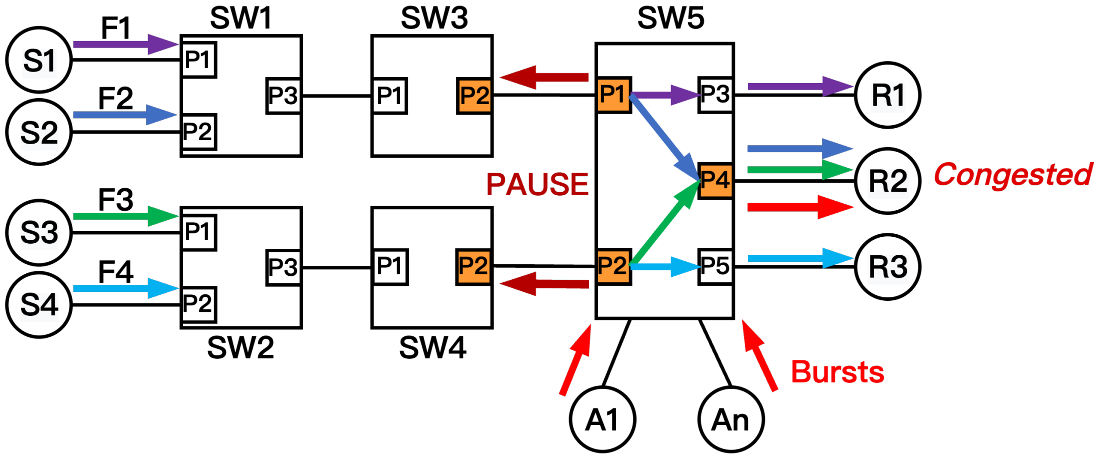
PFC issues

- Hop-by-hop congestion spread
  - Head-of-line blocking



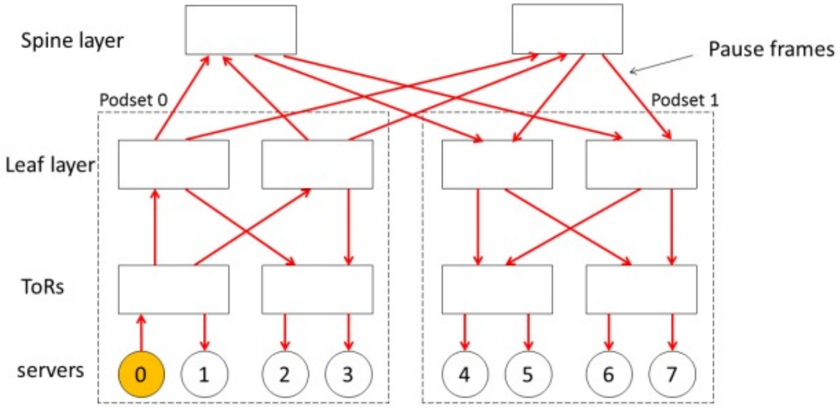
PFC issues

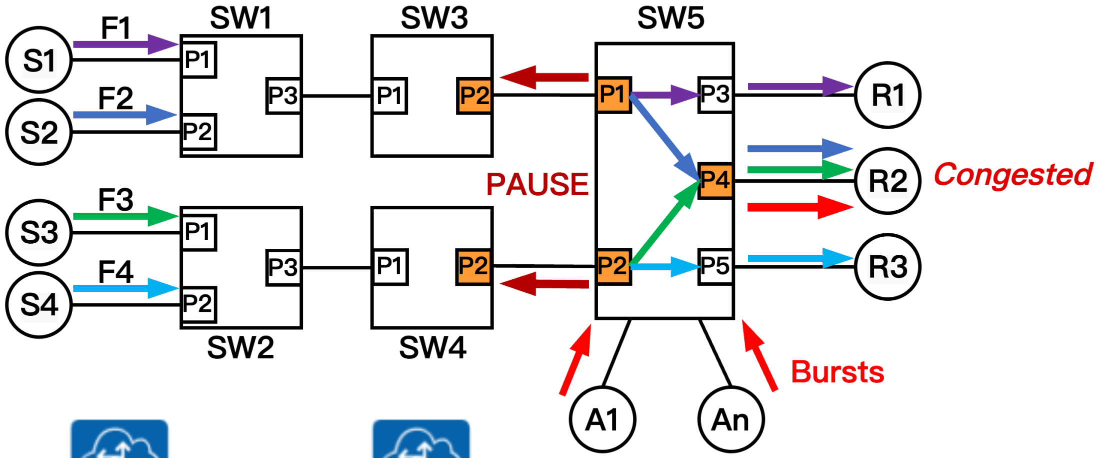
- Hop-by-hop congestion spread
  - Head-of-line blocking
  - Unfair victim flows
    - F1 and F4 are victim flows

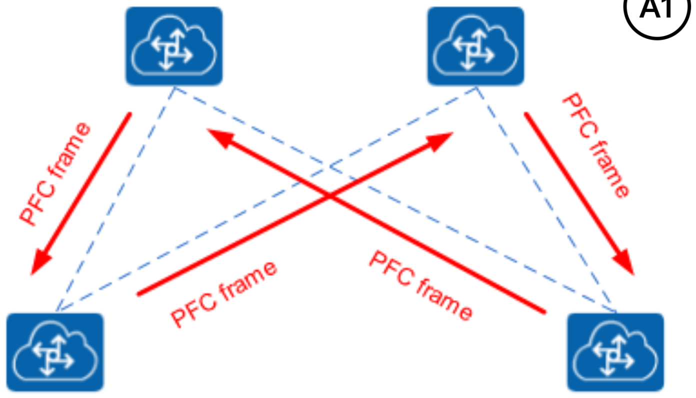


PFC issues

- Hop-by-hop congestion spread
  - Head-of-line blocking
  - Unfair victim flows
    - F1 and F4 are victim flows
  - PAUSE storm, deadlock…

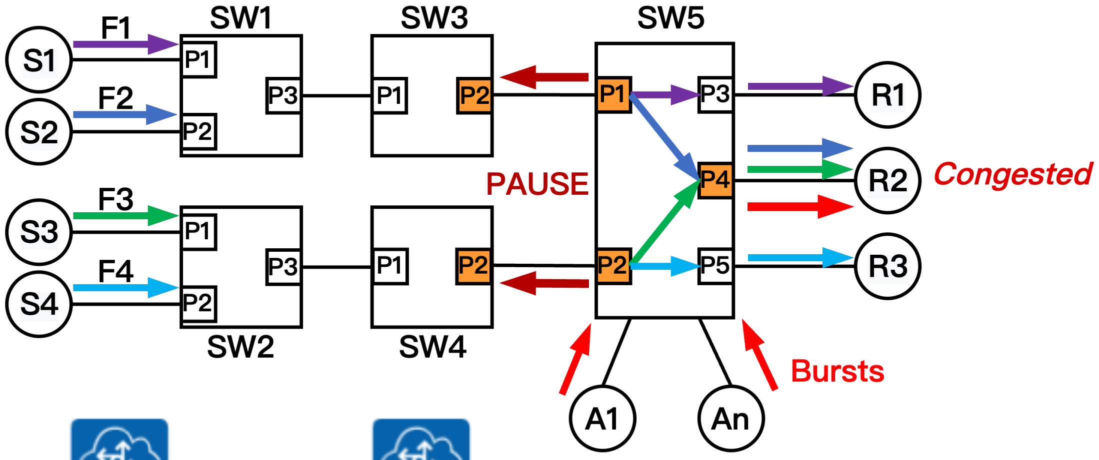


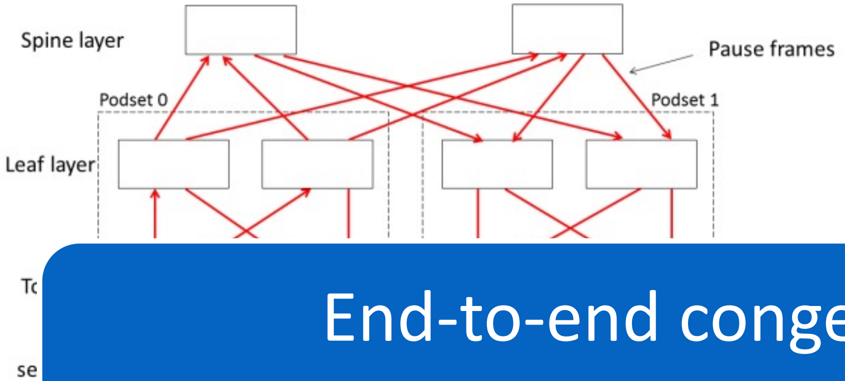


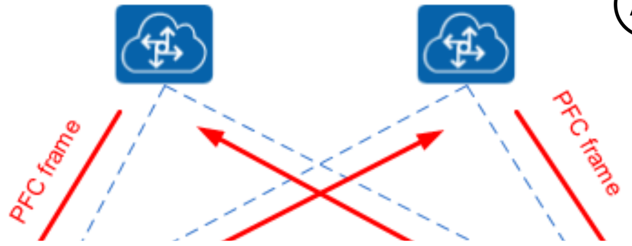


PFC issues

- Hop-by-hop congestion spread
  - Head-of-line blocking
  - Unfair victim flows
    - F1 and F4 are victim flows
  - PAUSE storm, deadlock…







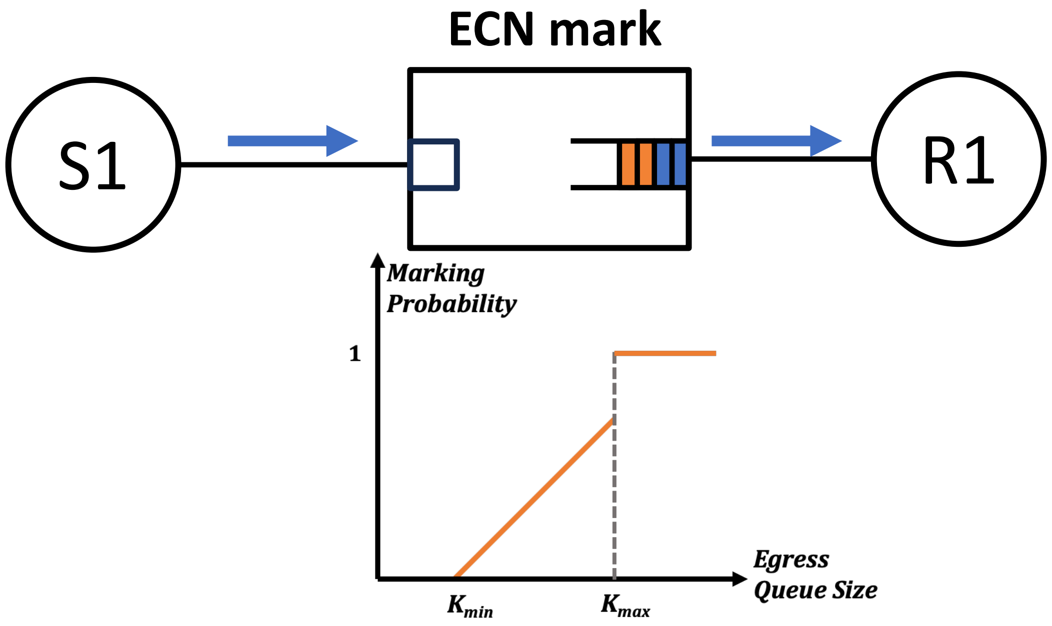
End-to-end congestion control is introduced to mitigate PFC’s side effect

RDMA congestion control

- End-to-end congestion control schemes
  - Detect the congestion and adjust the flow rate
  - **DCQCN**[SIGCOMM 2015]: the *de facto* standard

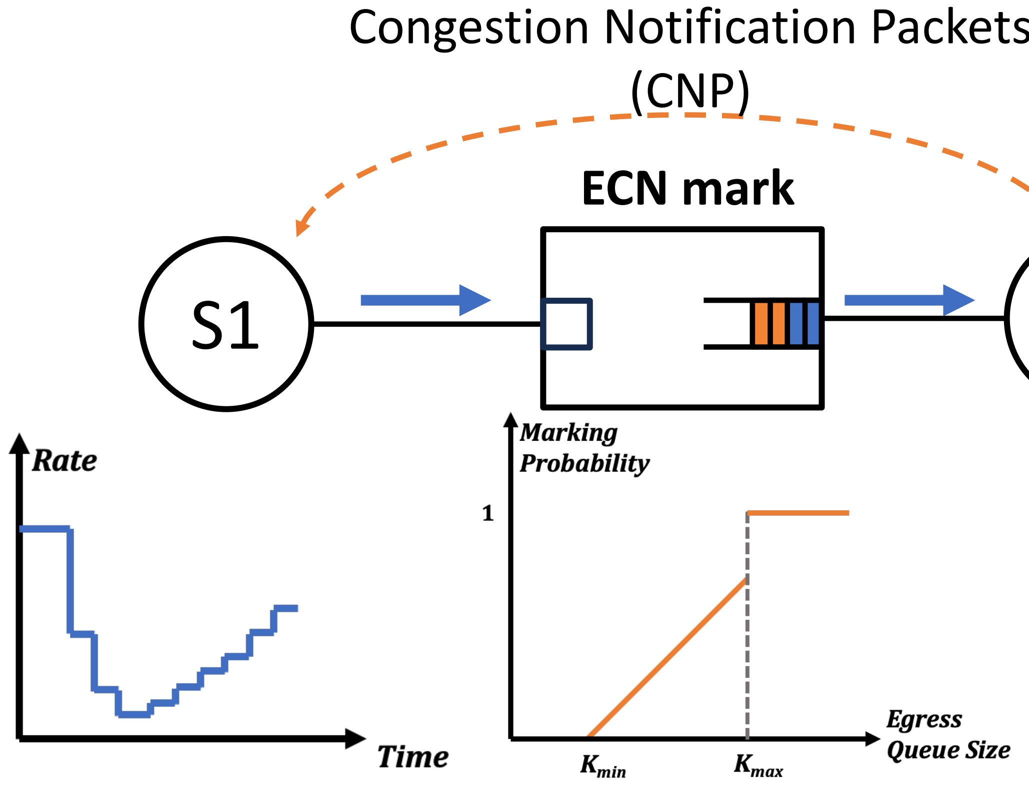
RDMA congestion control

- End-to-end congestion control schemes
  - Detect the congestion and adjust the flow rate
  - **DCQCN**[SIGCOMM 2015]: the *de facto* standard



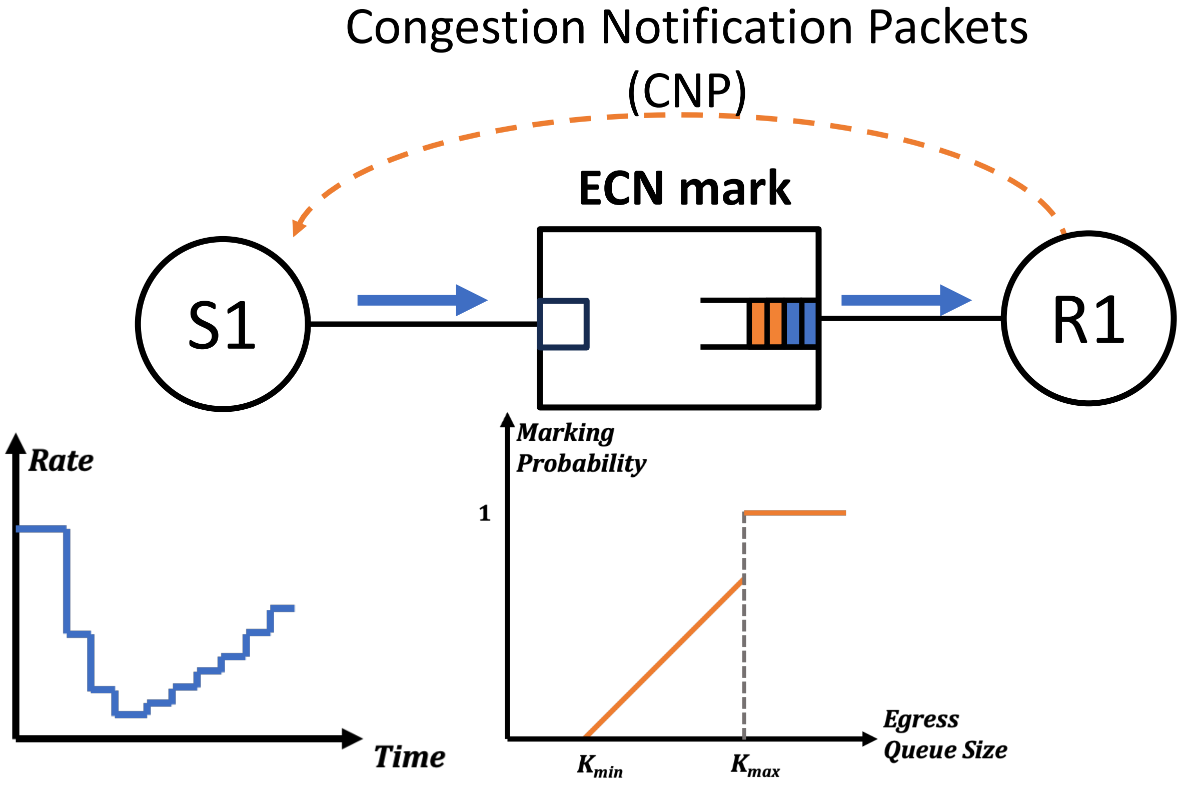
RDMA congestion control

- End-to-end congestion control schemes
  - Detect the congestion and adjust the flow rate
  - **DCQCN**[SIGCOMM 2015]: the *de facto* standard

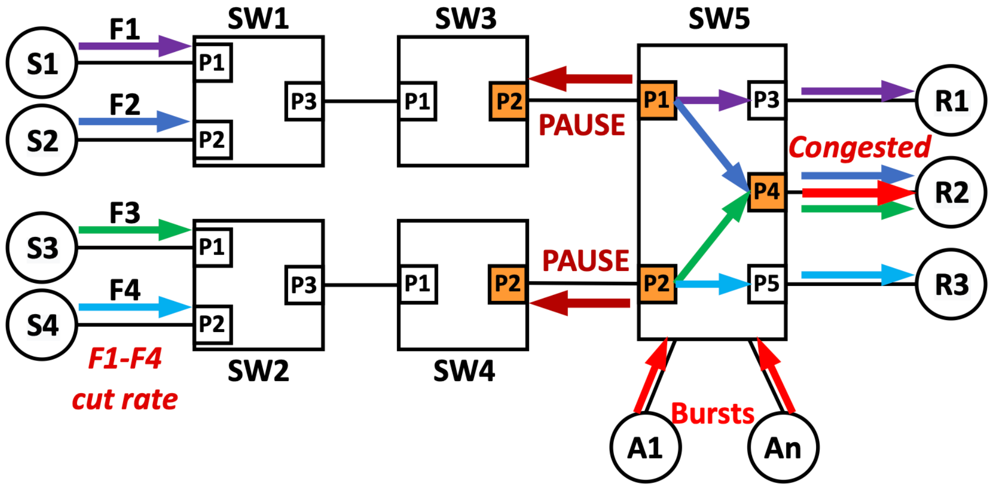


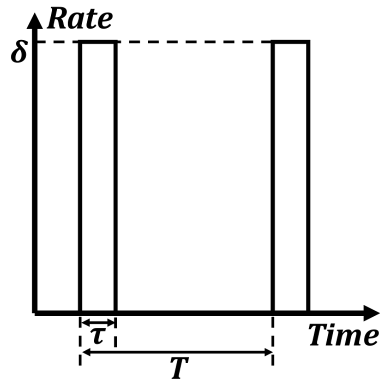
RDMA congestion control

- End-to-end congestion control schemes
  - Detect the congestion and adjust the flow rate
  - **DCQCN**[SIGCOMM 2015]: the *de facto* standard



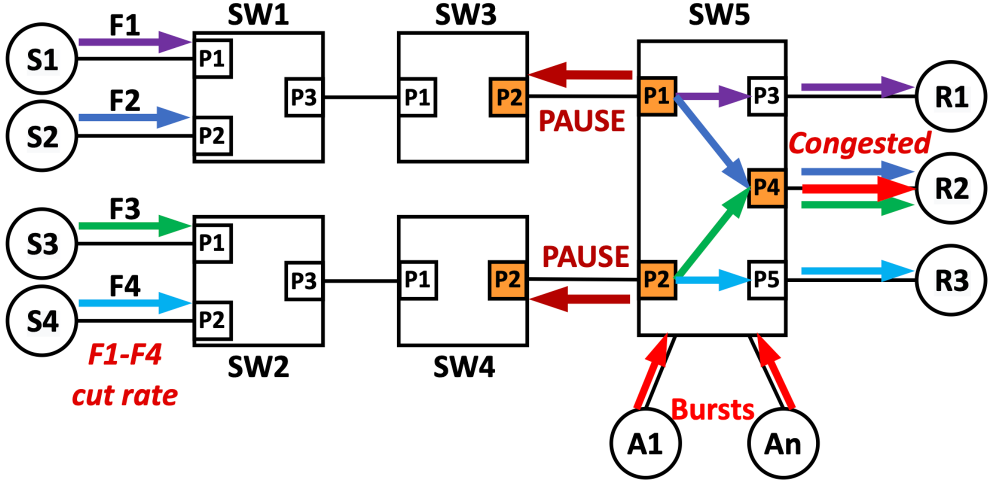
Experimental observation in RDMA traffic control

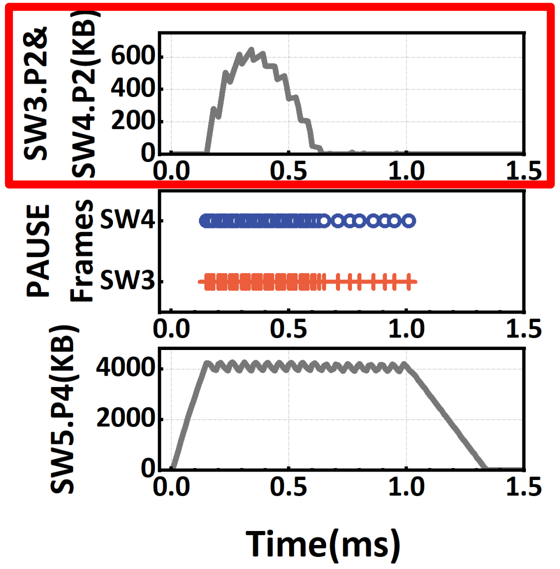




Experimental observation in RDMA traffic control

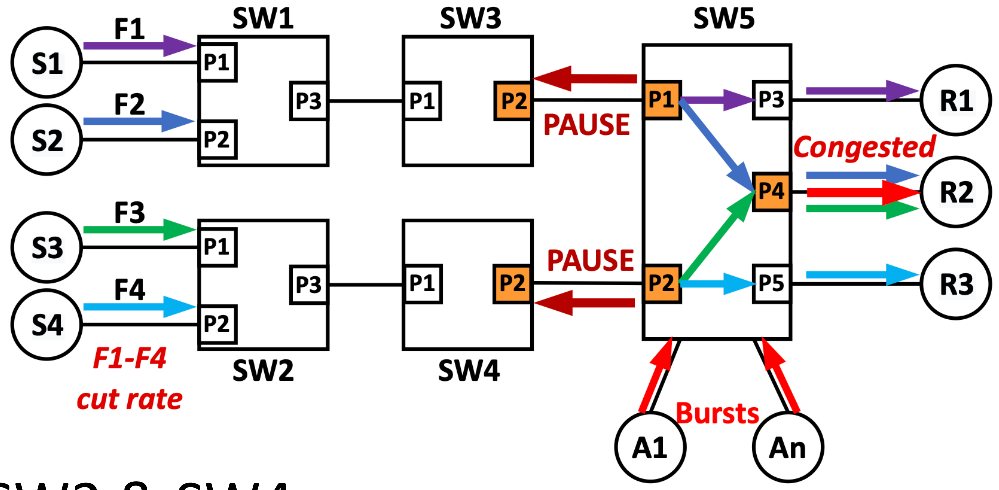
- PFC still spreads congestion
  - SW3.P2 & SW4.P2 are congested by PFC

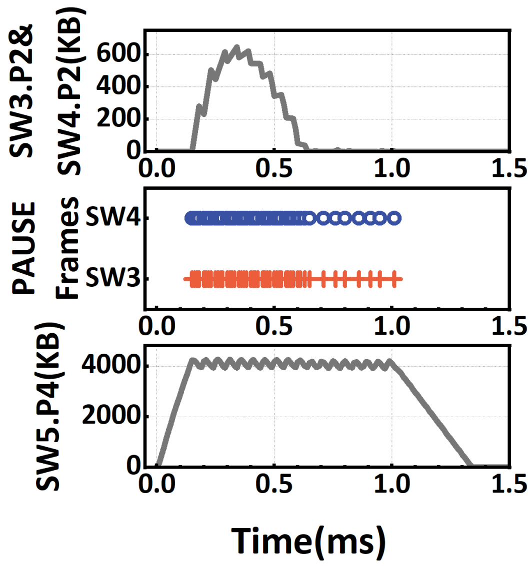


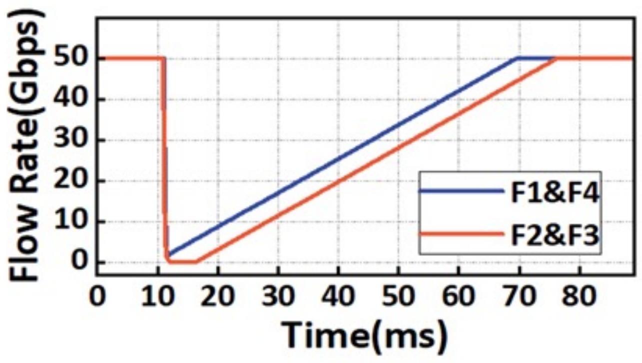


Experimental observation in RDMA traffic control

- PFC still spreads congestion
  - SW3.P2 & SW4.P2 are congested by PFC
- DCQCN is misled and cuts F1 & F4 wrongly
  - Queue length signal can be falsified by PFC
  - F1 and F4 are cut due to high queue length at SW3 & SW4

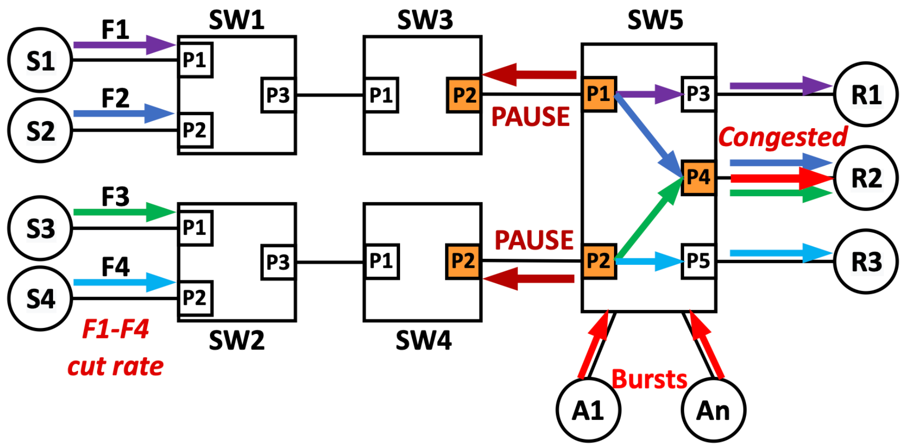






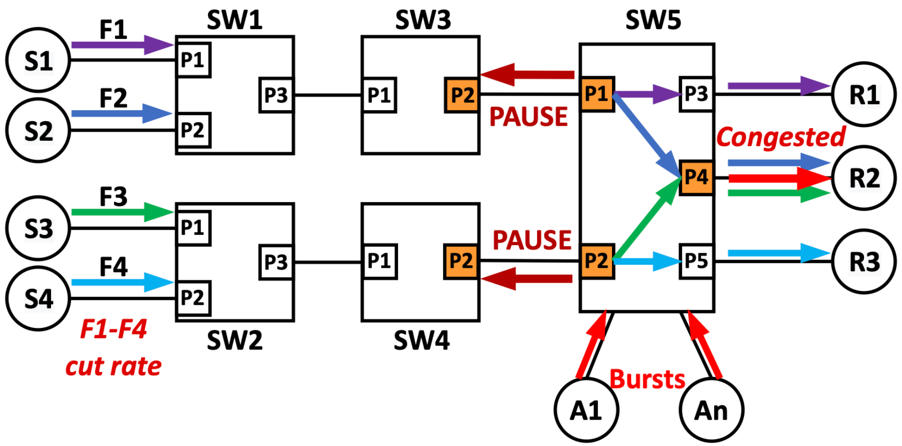
Experimental observation in RDMA traffic control

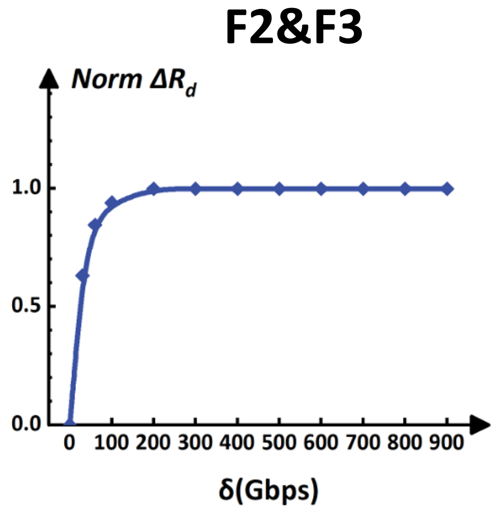
- Bursts ***indirectly*** cut flows via PFC
  - Direct victims: F2&F3 congested at SW5.P4
  - Indirect victims: F1&F4 sharing no link with bursts

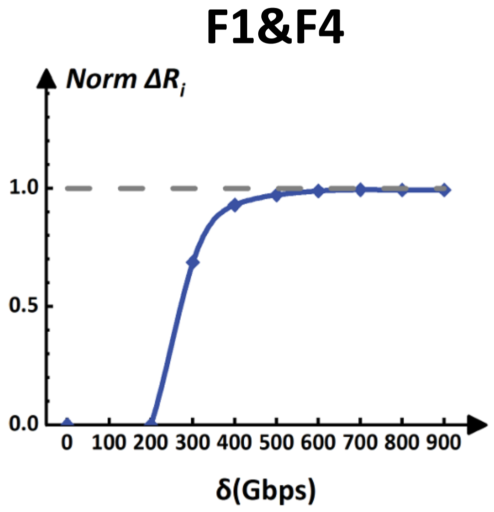


Experimental observation in RDMA traffic control

- Bursts ***indirectly*** cut flows via PFC
  - Direct victims: F2&F3 congested at SW5.P4
  - Indirect victims: F1&F4 sharing no link with bursts
- Higher burst rate δ makes severer rate cut ΔR

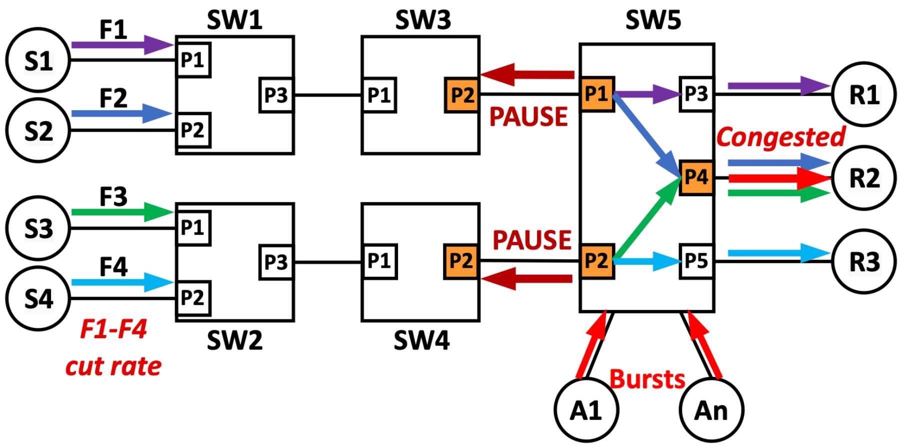




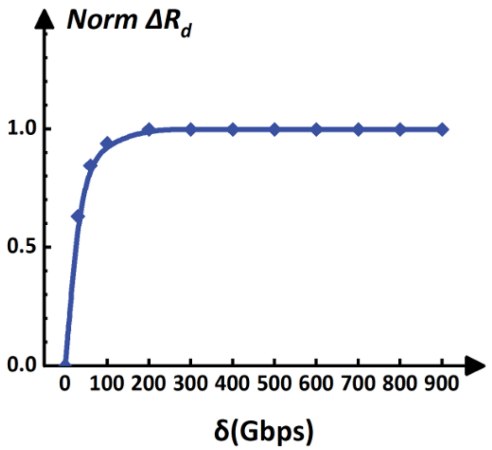


Experimental observation in RDMA traffic control

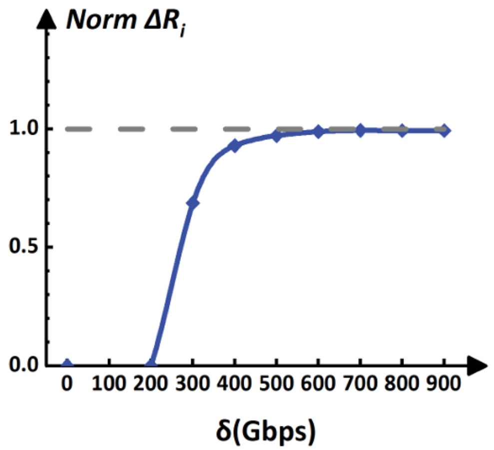
- Bursts ***indirectly*** cut flows via PFC
  - Direct victims: F2&F3 congested at SW5.P4
  - Indirect victims: F1&F4 sharing no link with bursts
- Higher burst rate δ makes severer rate cut ΔR
  - Higher δ → More PAUSE → Heavier congestion

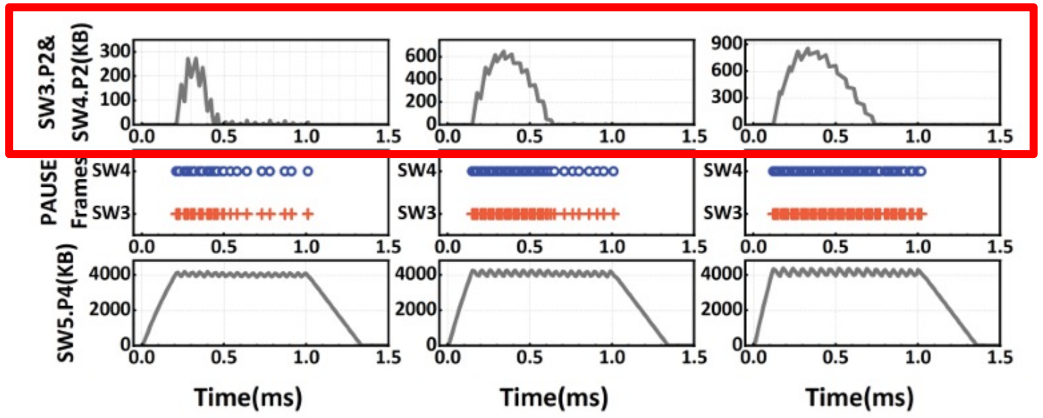


F2&F3



F1&F4

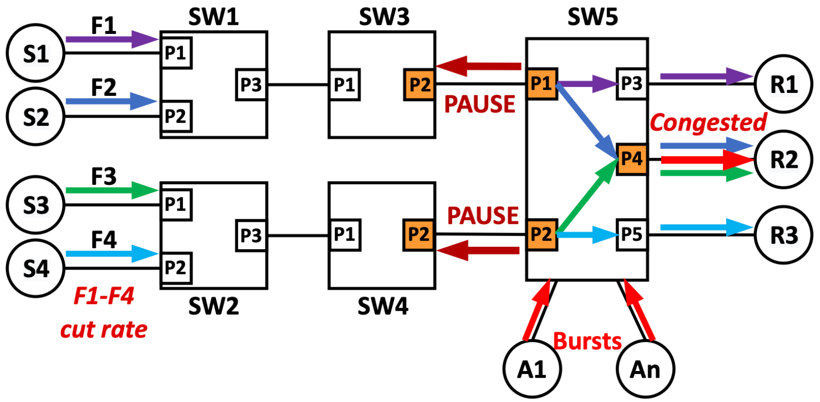


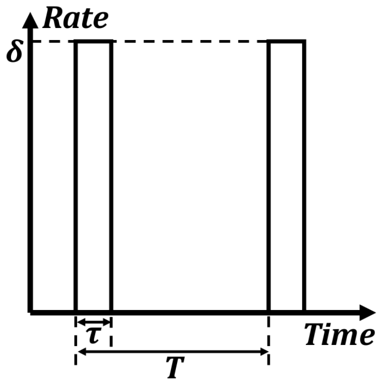


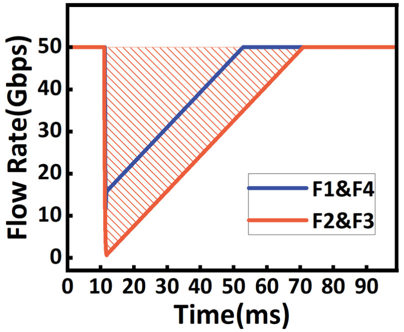
δ = 300Gbps 400Gbps 500Gbps

Experimental observation in RDMA traffic control

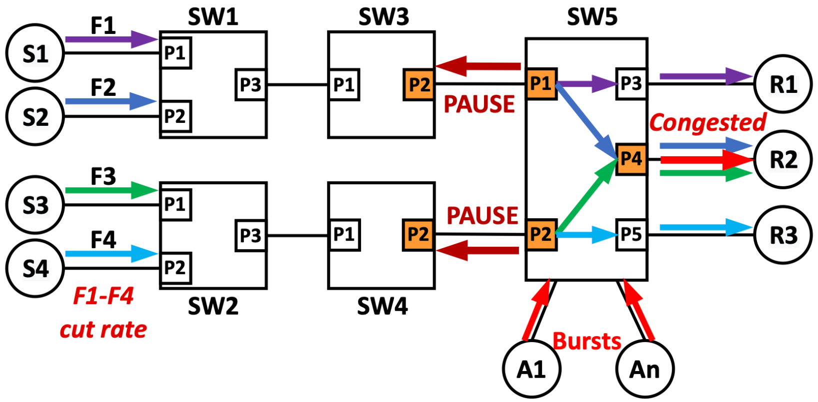
- Long performance loss due to AIMD rate control
  - ~1ms burst → 10s of ms rate recovery



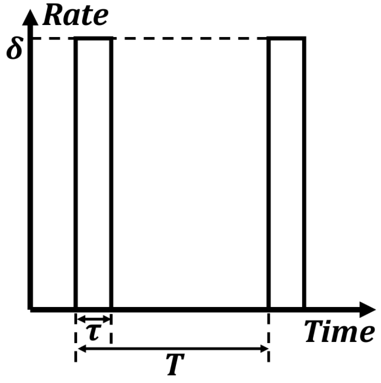


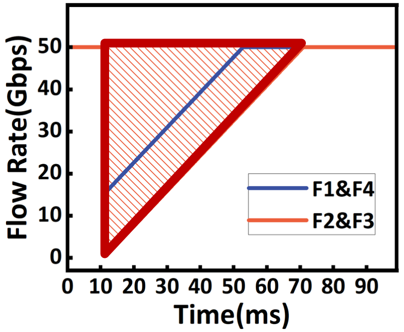


Experimental observation in RDMA traffic control



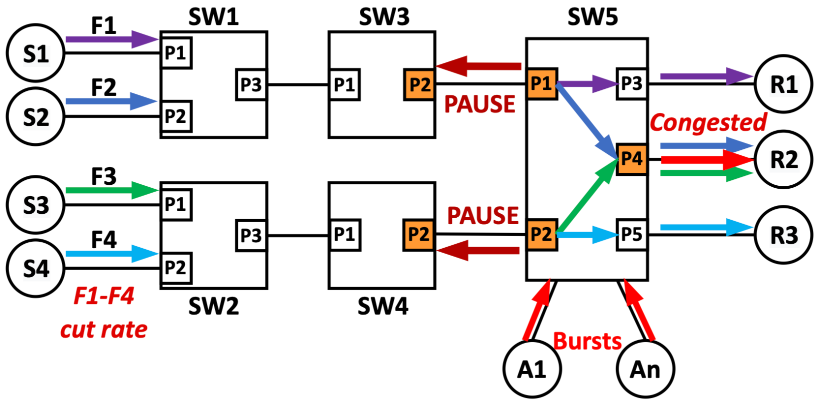
- Long performance loss due to AIMD rate control
  - ~1ms burst → 10s of ms rate recovery
  - $PL = \int_T (R_0 - R(t))\mathrm{d}t$ → Shadowed area

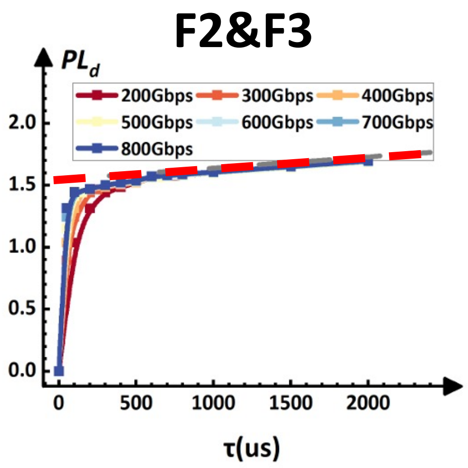


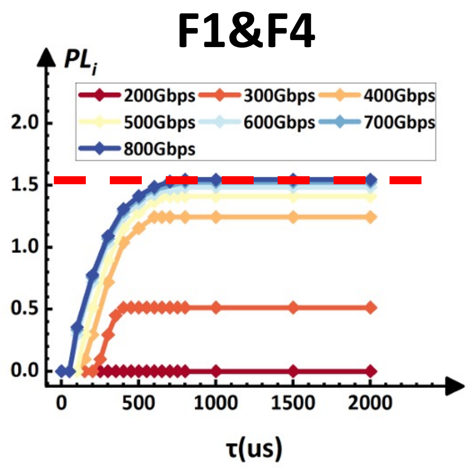


Experimental observation in RDMA traffic control

- Short bursts cause significant performance loss
- Diminishing marginal return of PL on τ
  - Direct victims: linear increase in PL
  - Indirect victims: no more increase in PL

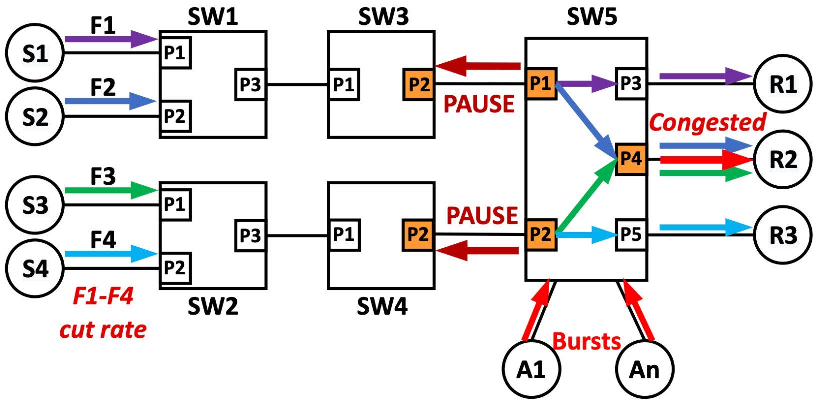


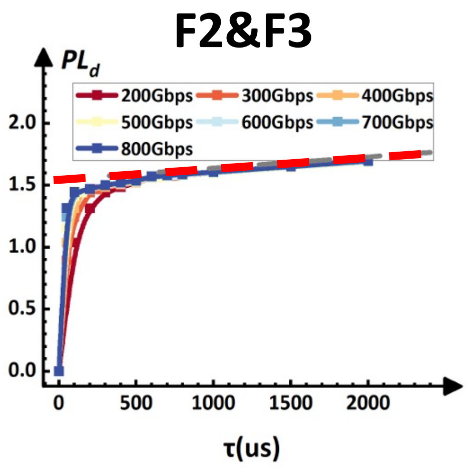


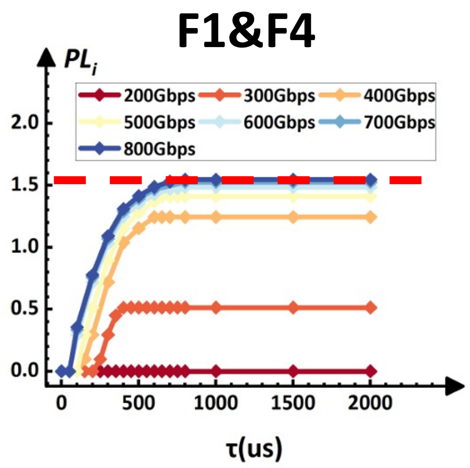


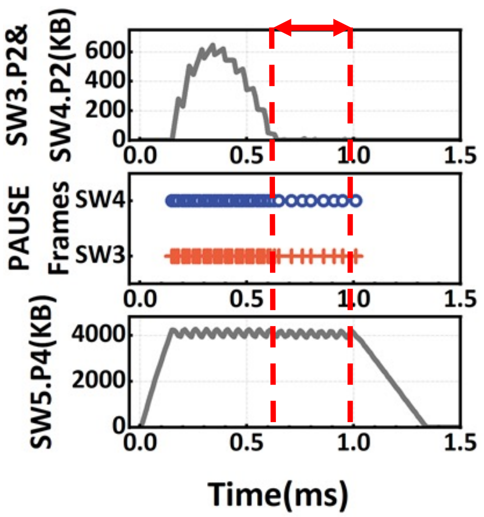
Experimental observation in RDMA traffic control

- Short bursts cause significant performance loss
- Diminishing marginal return of PL on τ
  - Direct victims: linear increase in PL
  - Indirect victims: no more increase in PL









Principles for a low-rate DoS attack

- Cover more victim flows with fewer congestion points
  - **Indirectly** cover more flows for lower direct queue contention
  - **Burst rate** δ should put sufficient **rate cut** ΔR on indirect victim flows

- A trade-off between performance loss and burst duration
  - Too long **burst duration** τ makes no further gain, but only high cost
  - Minimum **burst duration** τ for sufficient PL

Threat model

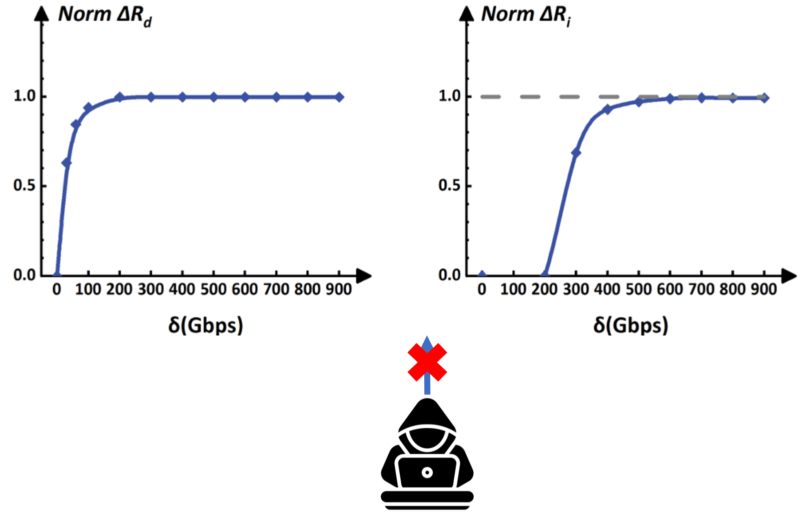
- Shared RDMA network infrastructure
  - Multiple users (malicious and benign) in the same network
- Attacker's capability
  - **Traffic crafting**: High-rate burst and probing traffic
- Attacker's knowledge
  - Network topology
  - Target flow set: A specific set of flows to cut off (Can be relaxed)
- Attacker's goal: **Efficient** attack
  - **High impact:** Cover more target flows; cause high performance loss
  - **Low cost:** Low burst rate δ and short duration τ

Challenges for an efficient attack

- Cover more target flows efficiently
  - Which target port/link to congest?
    - Generalized maximum coverage is **NP hard**

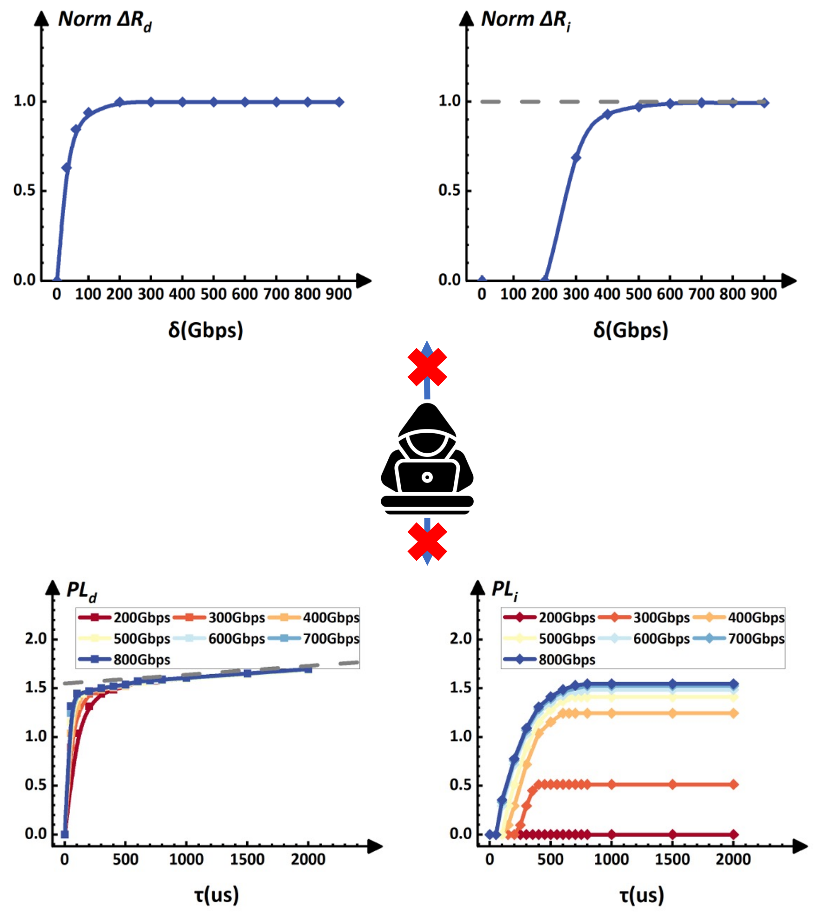
Challenges for an efficient attack

- Cover more target flows efficiently
  - Which target port/link to congest?
    - Generalized maximum coverage is **NP hard**
  - What δ should be deployed for a specific target port?
    - Relationship between ΔR and δ is unknown for attackers



Challenges for an efficient attack

- Cover more target flows efficiently
  - Which target port/link to congest?
    - Generalized maximum coverage is **NP hard**
  - What δ should be deployed for a specific target port?
    - Relationship between ΔR and δ is unknown for attackers
- Cause high performance loss efficiently
  - Too long burst duration makes attack inefficient
    - Relationship between PL and τ is unknown for attackers



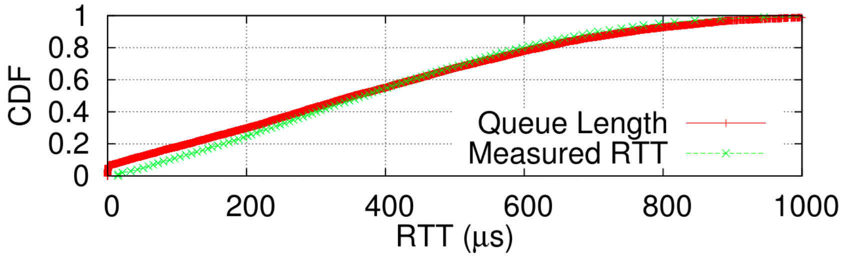
LoRDMA attack

- Coordination
  - Greedily select the highest-heuristic-value port to attack
  - **Adaptively** deploy bots until sufficient rate cut ΔR achieves

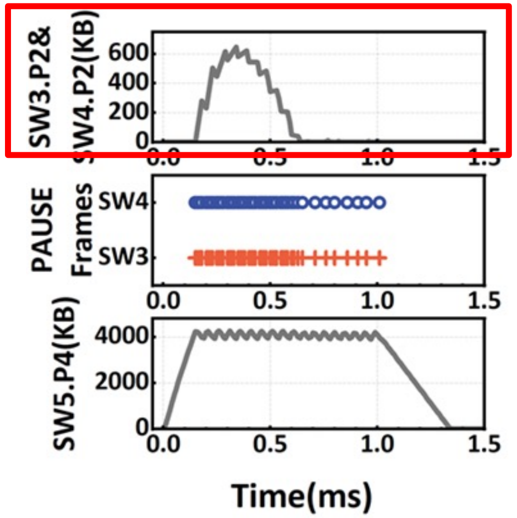
- Schedule
  - **Adaptively** adjust burst duration τ until an efficient trade-off between PL and τ achieves

RTT: A key signal reflecting congestion

- RTT is highly related to **queue length**
  - Estimate the congestion severity (ΔR) and the end-time (PL)



*TIMELY [SIGCOMM 2015]



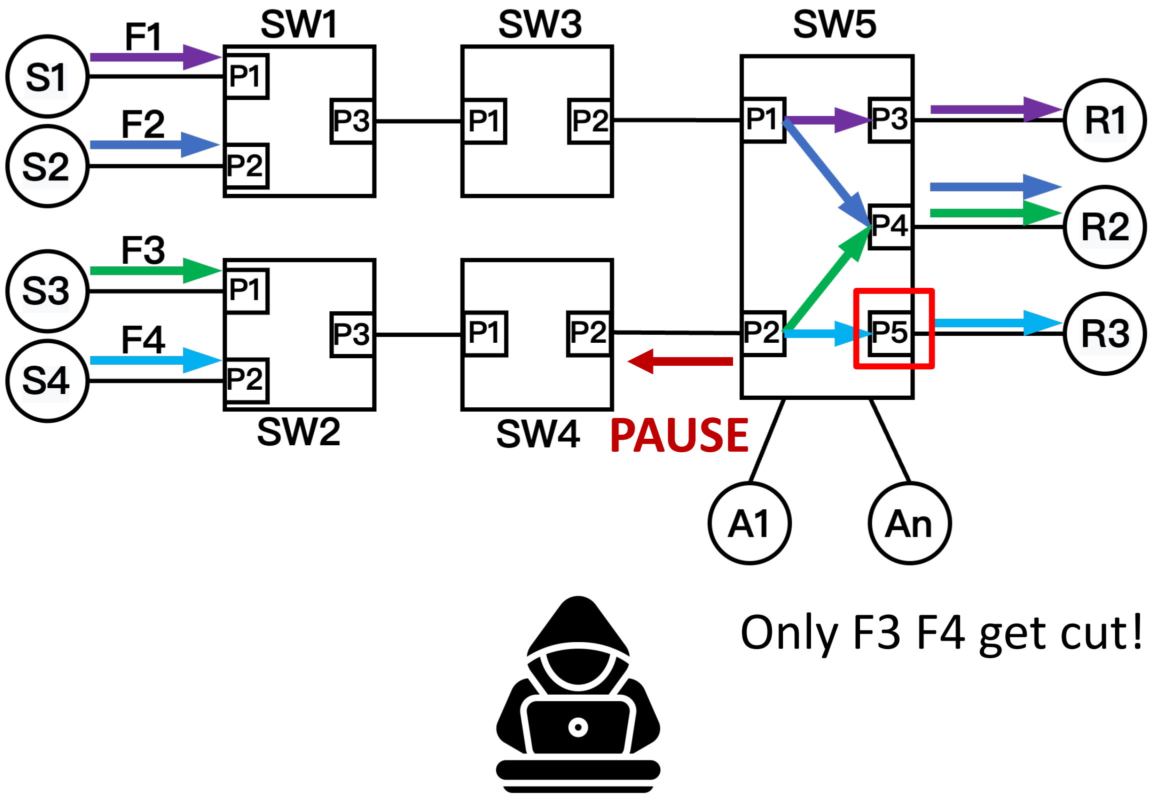
RTT: A key signal reflecting congestion

- RTT is highly related to **queue length**
  - Estimate the congestion severity (ΔR) and the end-time (PL)
- RTT prober
  - Connection request and rejection reply: A new side-channel signal
  - Monitor the long-term RTT to estimate the congestion

No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000	192.168.3.135	192.168.3.136	RRoCE	322	CM: ConnectRequest
2	0.000051	192.168.3.136	192.168.3.135	RRoCE	322	CM: ConnectReject

Coordination

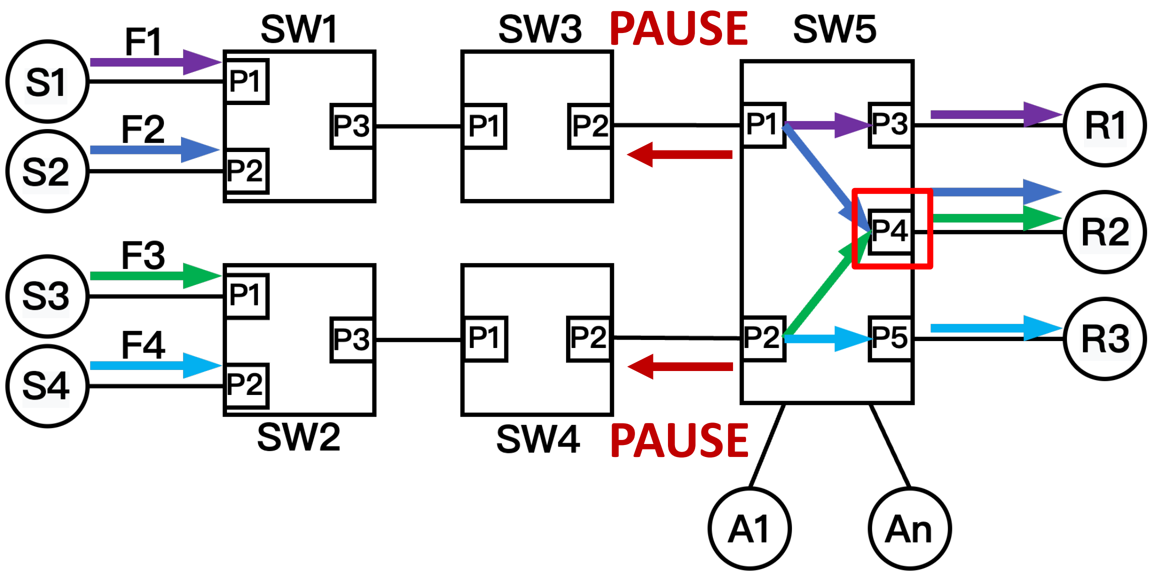
- Greedily select target port
  - Select the port to cut more flows indirectly



Only F3 F4 get cut!

Coordination

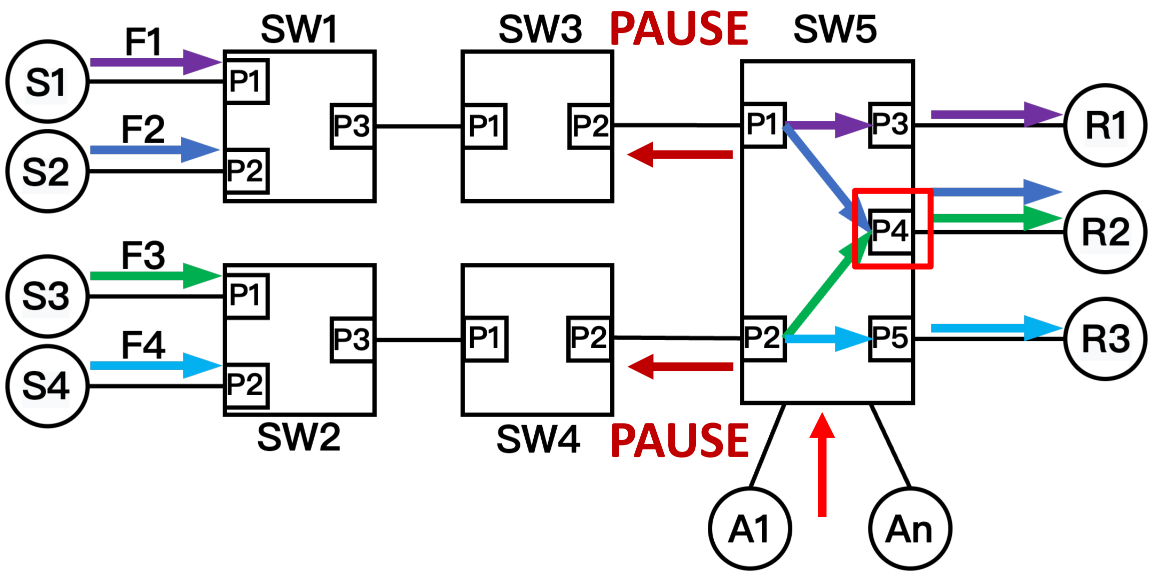
- Greedily select target port
  - Select the port to cut more flows indirectly



F1-F4 all get cut!

Coordination

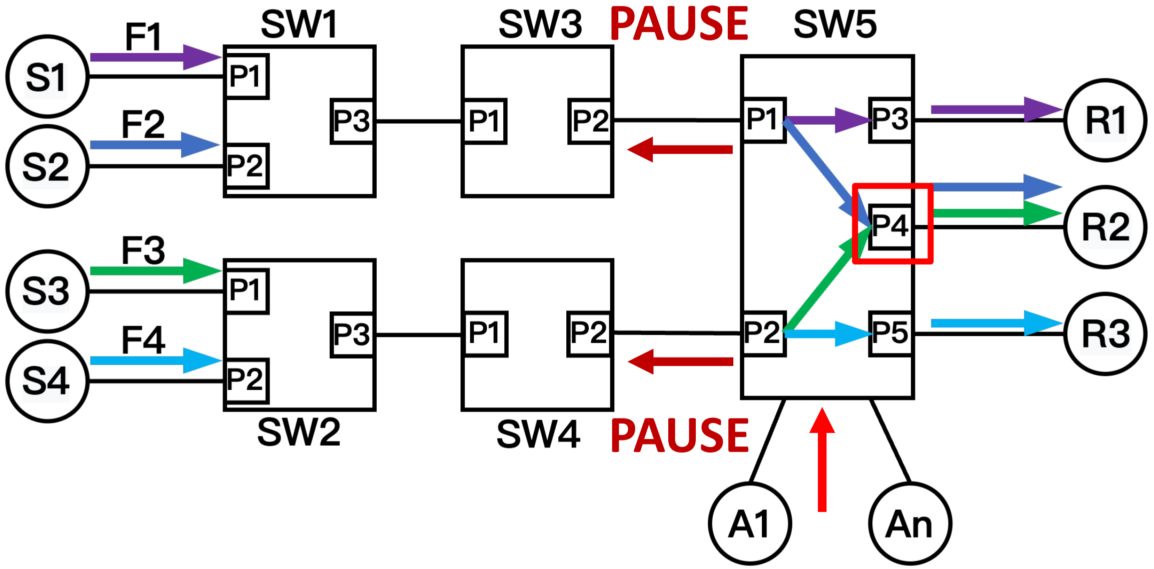
- Greedily select target port
- Adaptively add bots
  - Indirect victims (F1&F4) should suffer as ***severely*** as direct ones (F2&F3)!

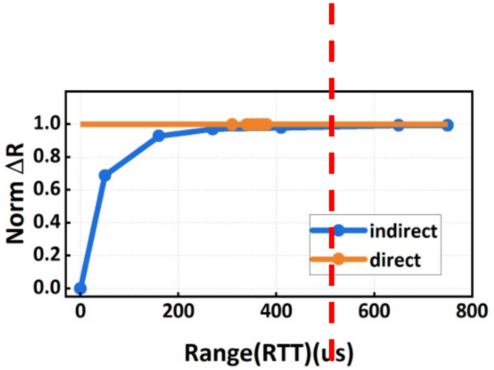


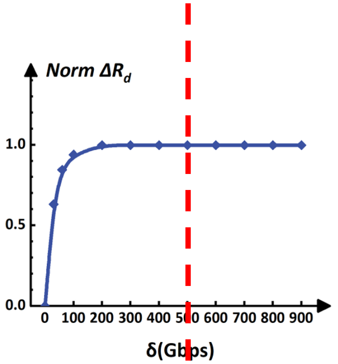
How many bots with line rate should I use?

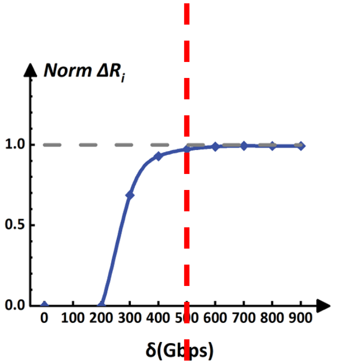
Coordination

- Greedily select target port
- Adaptively add bots
  - Indirect victims (F1&F4) should suffer as ***severely*** as direct ones (F2&F3)!
  - $\textbf{range} < RTT_i > \cong \textbf{range} < RTT_d >$





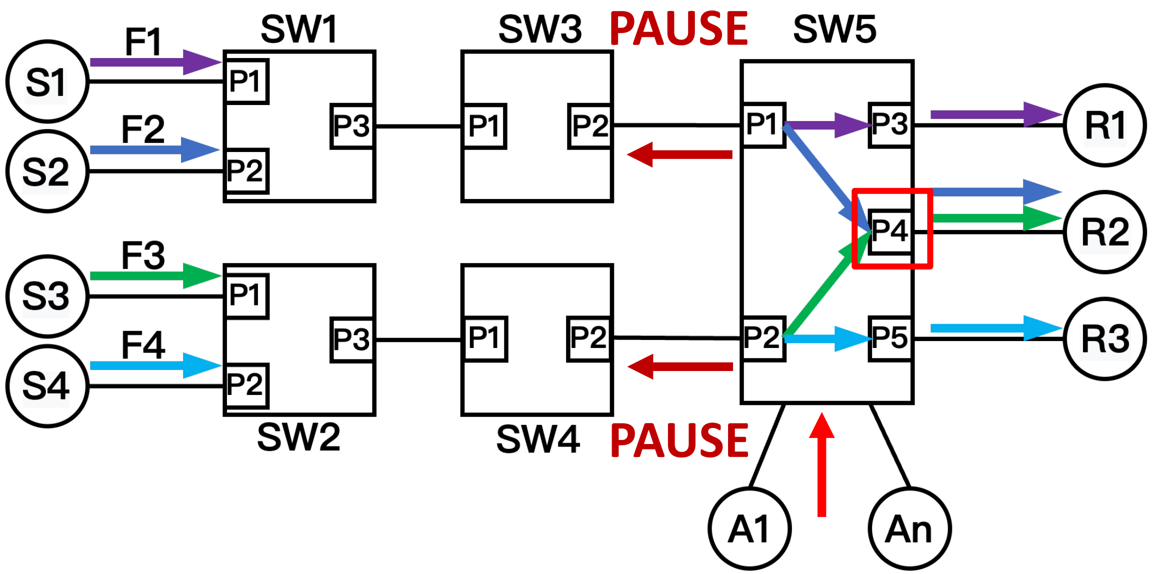


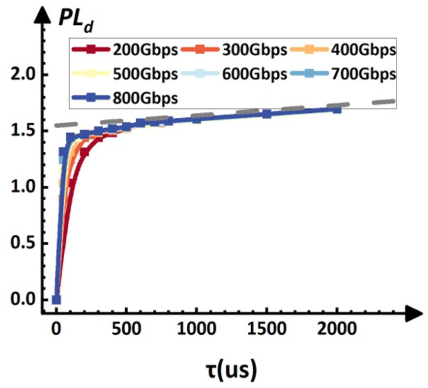


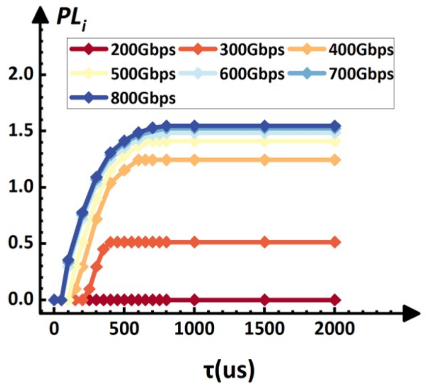
How many bots with line rate should I use?

Schedule

- Longer burst makes lower gain
  - PL_i hardly grows with burst duration τ



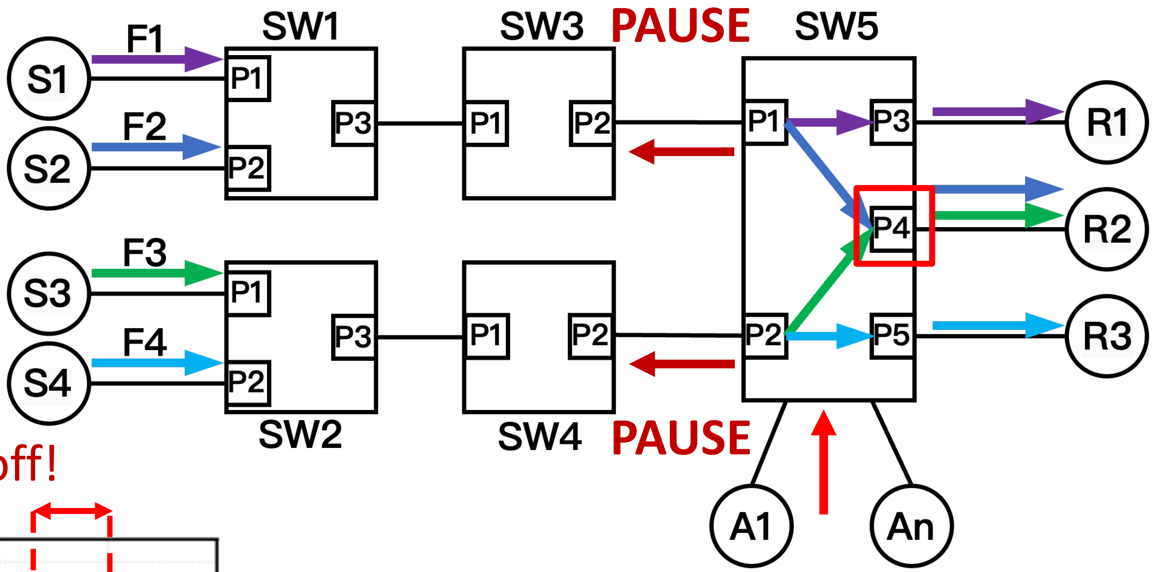




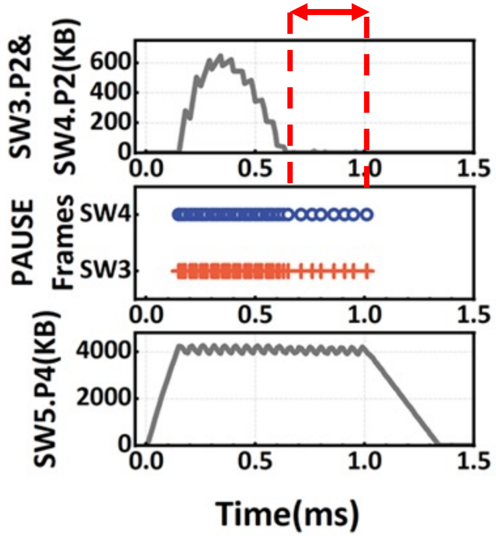
How long should the burst last?

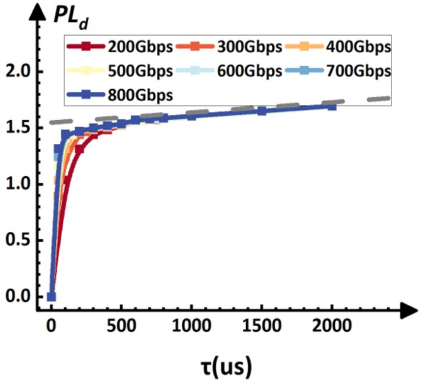
Schedule

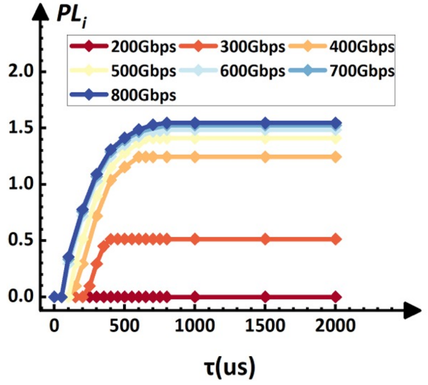
- Longer burst makes lower gain
  - PL_i hardly grows with burst duration τ



Trim it off!



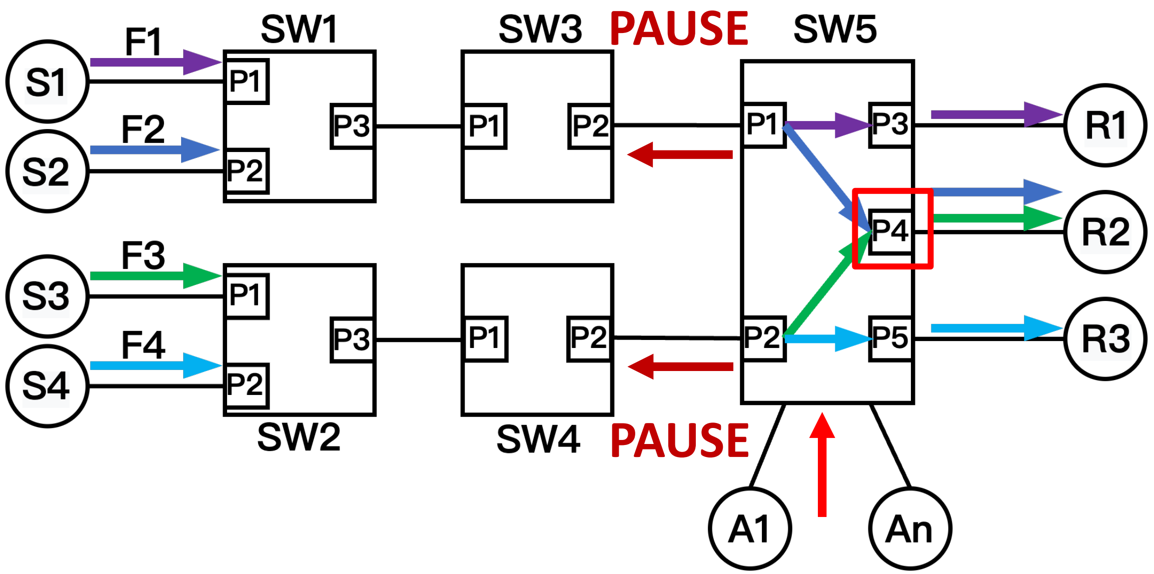




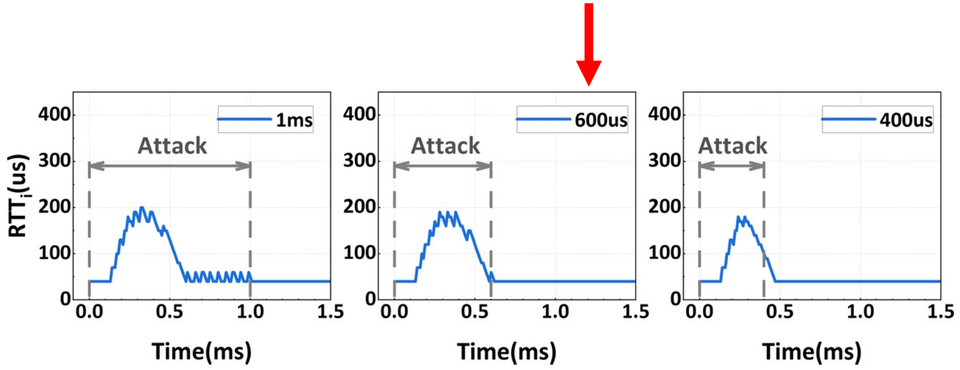
How long should the burst last?

Schedule

- Longer burst makes lower gain
  - PL_i hardly grows with burst duration τ
  - Trim off the burst duration with low $< RTT_i >$



How long should the burst last?

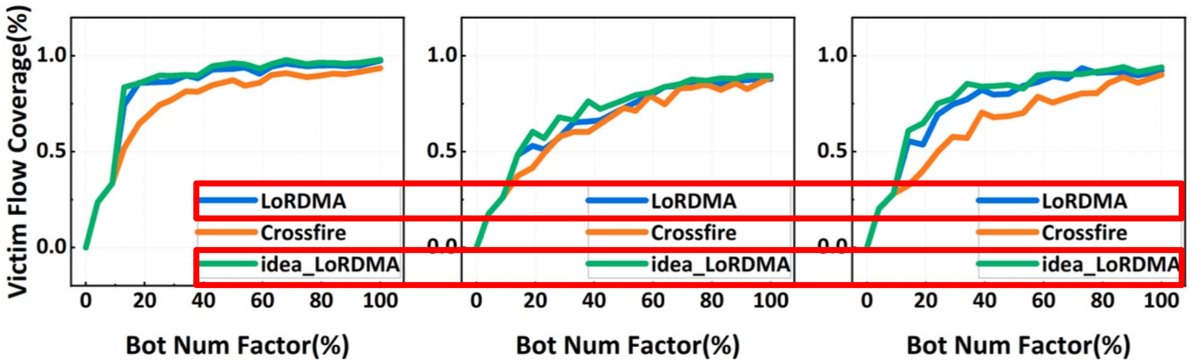


Implementation

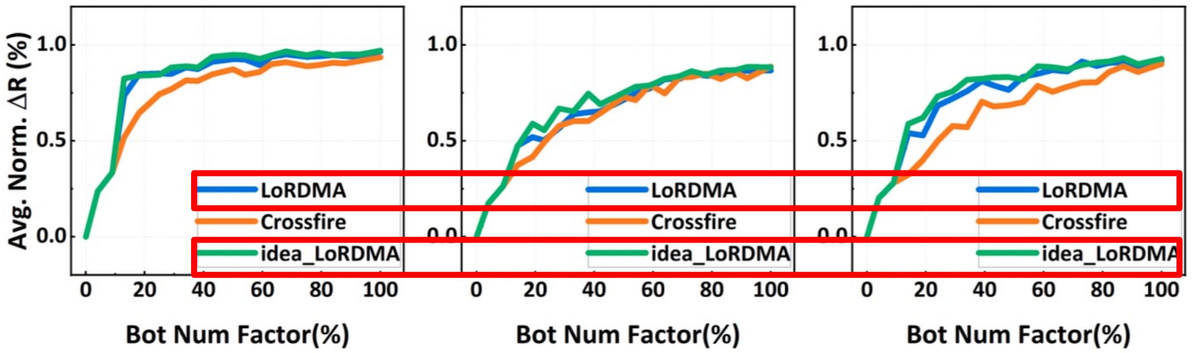
- Implementation
  - Attack tools: Burst generator, RTT prober
  - NS-3 simulation
- Experiment setup
  - Real testbed: Kuaishou cloud RDMA cluster (2 Leaf, 4 ToR, 8 RNIC 100Gbps)
  - Large-scale simulation: NS-3
- Goal of evaluation
  - Performance of the coordination and schedule
  - Attack impact on large-scale RDMA applications
  - Attack impact on real testbed

Coordination

- Higher attack performance
  - Higher victim flow coverage
  - Higher rate cut ΔR



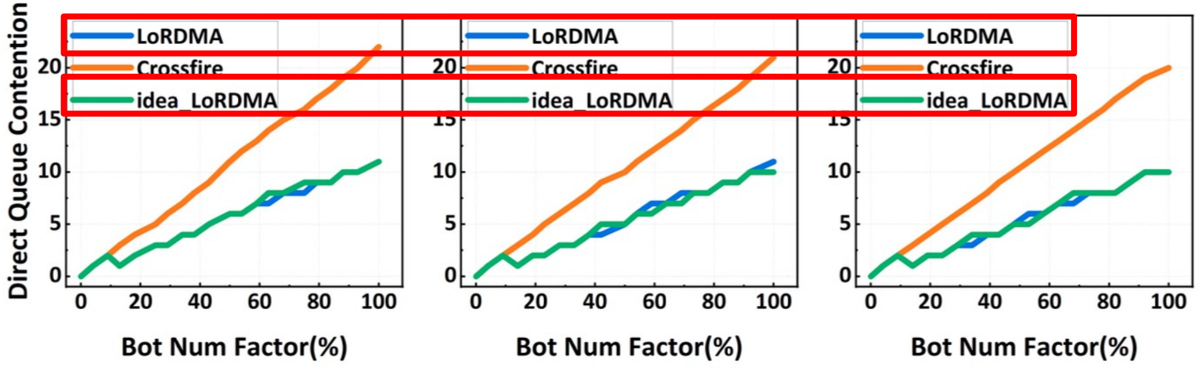
(a) Victim flow coverage at Carnet, Switch and Cernet, respectively.



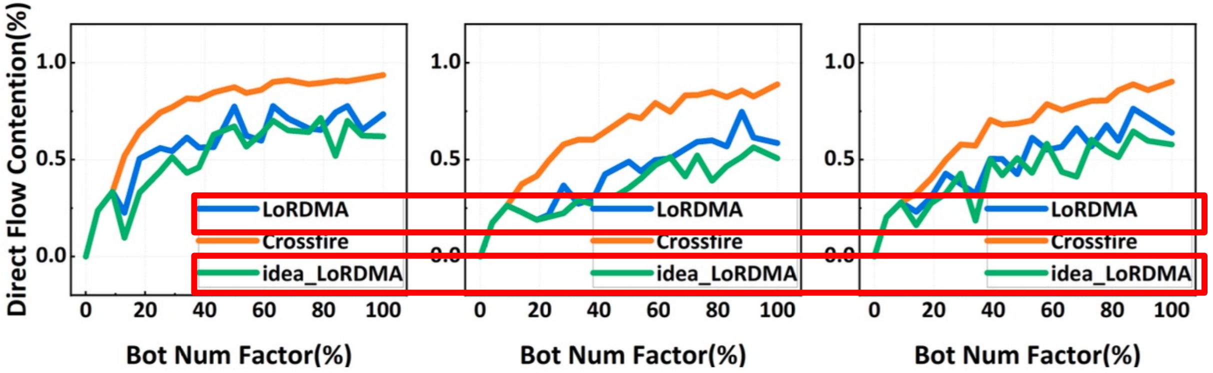
(b) Average ΔR at Carnet, Switch and Cernet, respectively.

Coordination

- Higher attack performance
  - Higher victim flow coverage
  - Higher rate cut ΔR
- Lower attack cost
  - Fewer directly congested points
  - Fewer directly congested flows



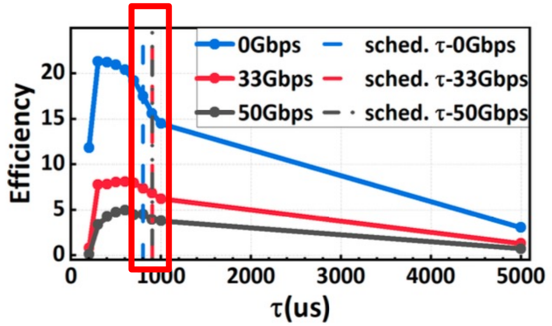
(c) Directly congested queue number at Carnet, Switch and Cernet, respectively.



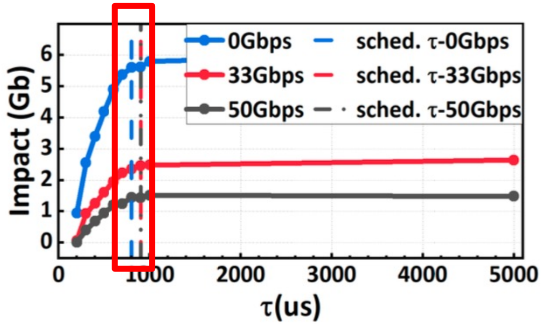
(d) Direct flow contention at Carnet, Switch and Cernet, respectively.

Schedule

- Higher attack efficiency
  - Efficient attack parameter across various background traffic
- Sufficient attack impact
  - Sufficiently high impact across various background traffic



(a) Attack efficiency as τ changes with different background traffic scenarios.



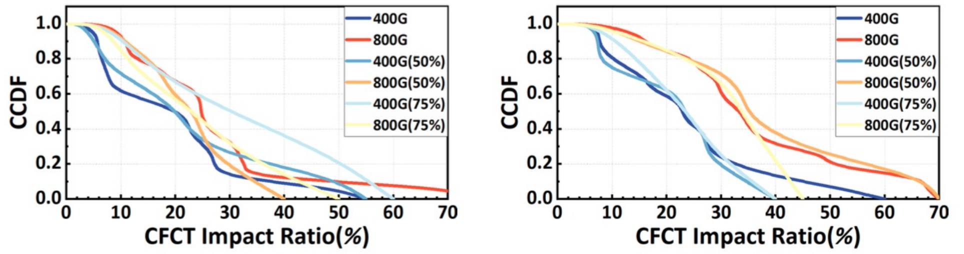
(b) Attack impact as τ changes with different background traffic scenarios.

Impact on real applications

- Simulation setup
  - Fat-tree (k=8) topology
  - Workload:
    - W1: machine learning training
    - W2: cloud storage

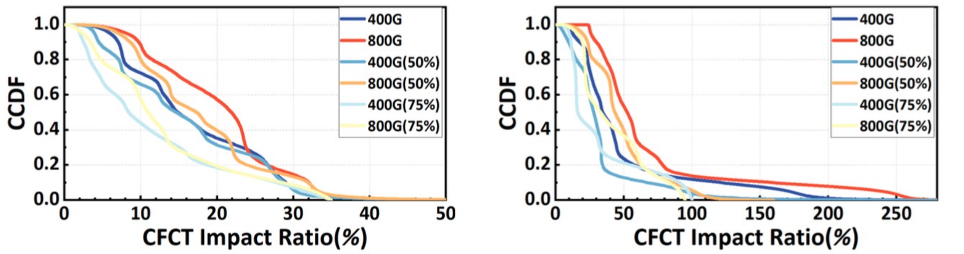
Impact on real applications

- Simulation setup
  - Fat-tree (k=8) topology
  - Workload:
    - W1: machine learning training
    - W2: cloud storage
- Impact on coflow-completion-time (CFCT)
  - Median damage on CFCT: 8.11% ~ 52.7%, averaging at 25.2%
  - Maximum damage on CFCT: 29.1% ~ 251.6%, averaging at 65.47%



(a) Co-flow completion time impact ratio of distributed machine learning training with a low flow number.

(b) Co-flow completion time impact ratio of cloud storage with a low flow number.

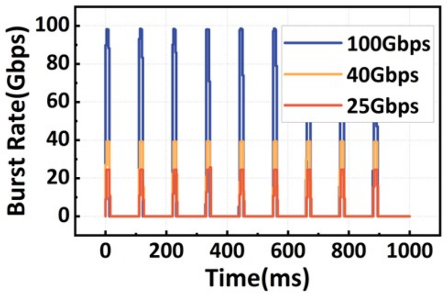


(c) Co-flow completion time impact ratio of distributed machine learning training with a high flow number.

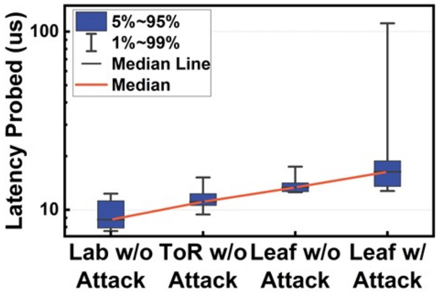
(d) Co-flow completion time impact ratio of cloud storage with a high flow number.

Real testbed

- Attack tools validation
  - Line-rate burst generation
  - RTT reflecting the congestion



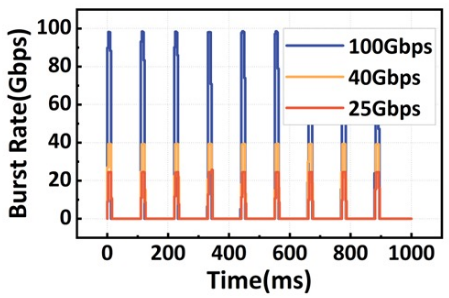
(a) Burst generator performance.



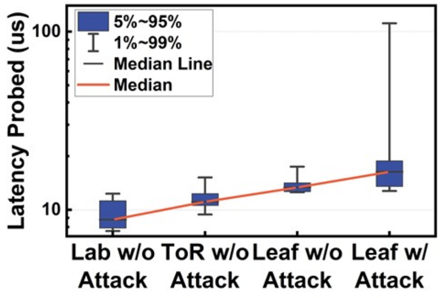
(b) Prober result.

Real testbed

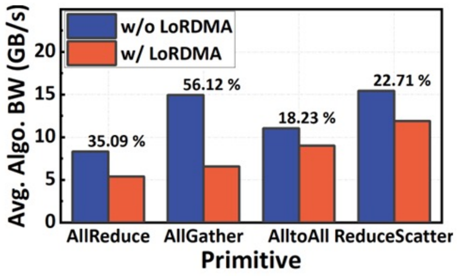
- Attack tools validation
  - Line-rate burst generation
  - RTT reflecting the congestion
- Real application impact
  - NCCL TEST:
    - 18.23% (AlltoAll) to 56.12% (AllGather)
  - PFC misleads DCQCN



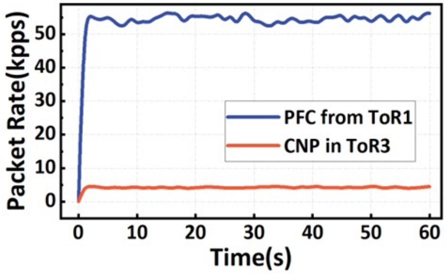
(a) Burst generator performance.



(b) Prober result.



(a) Performance of different communication primitives.



(b) PFC and CNP count over time.

Conclusions

- RDMA is less secure in transport control
  - PFC and DCQCN can be exploited to cut flows across multiple hops
  - Drastic performance loss can be caused by short-duration bursts
- **LoRDMA**: a new low-rate DoS attack
  - Coordinate & schedule for an efficient attack solution
- Evaluations demonstrate the effectiveness and efficiency
  - Large-scale simulation & real testbed

https://github.com/wangshicheng1225/LoRDMA

Thanks for your attention!
Q & A

wsc22@mails.tsinghua.edu.cn

Backup: Possible defense schemes

- PFC-driven network performance anomaly diagnosis
  - Root cause flows (bursts) are hops away from victims
    - No significant contribution to the **local** queue contention
  - Analyze the PFC spreading causality to find the culprits
- Fine-grained burst monitor
  - ms-/us-level burst requires fine granularity
  - A trade-off between effectiveness and overhead