ACM SIGCOMM 2017

Credit-Scheduled Delay-Bounded Congestion Control for Datacenters

Inho Cho, Keon Jang*, Dongsu Han

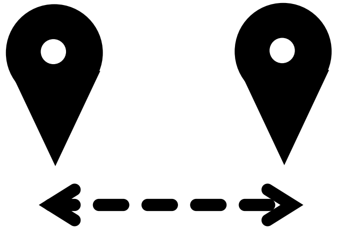
KAIST

Google*

Datacenter Network

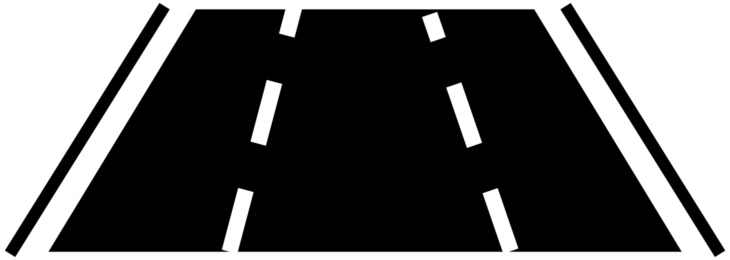
Small Latency

< 100 μs



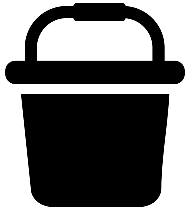
High Bandwidth

10/40 ~ 100 Gbps



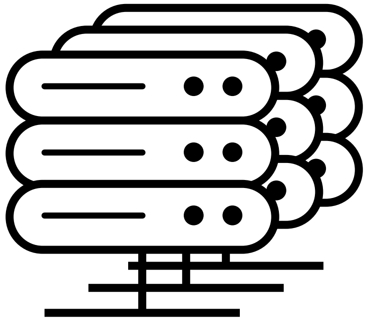
Shallow Buffer

< 30 MB for ToR



Large Scale

> 10,000 machines



Datacenter Network

Small Latency

< 100 μs

High Bandwidth

10/40 ~ 100 Gbps

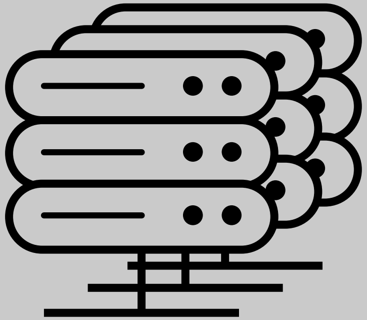
Congestion control is more challenging in datacenter.

Shallow Buffer

< 30 MB for ToR

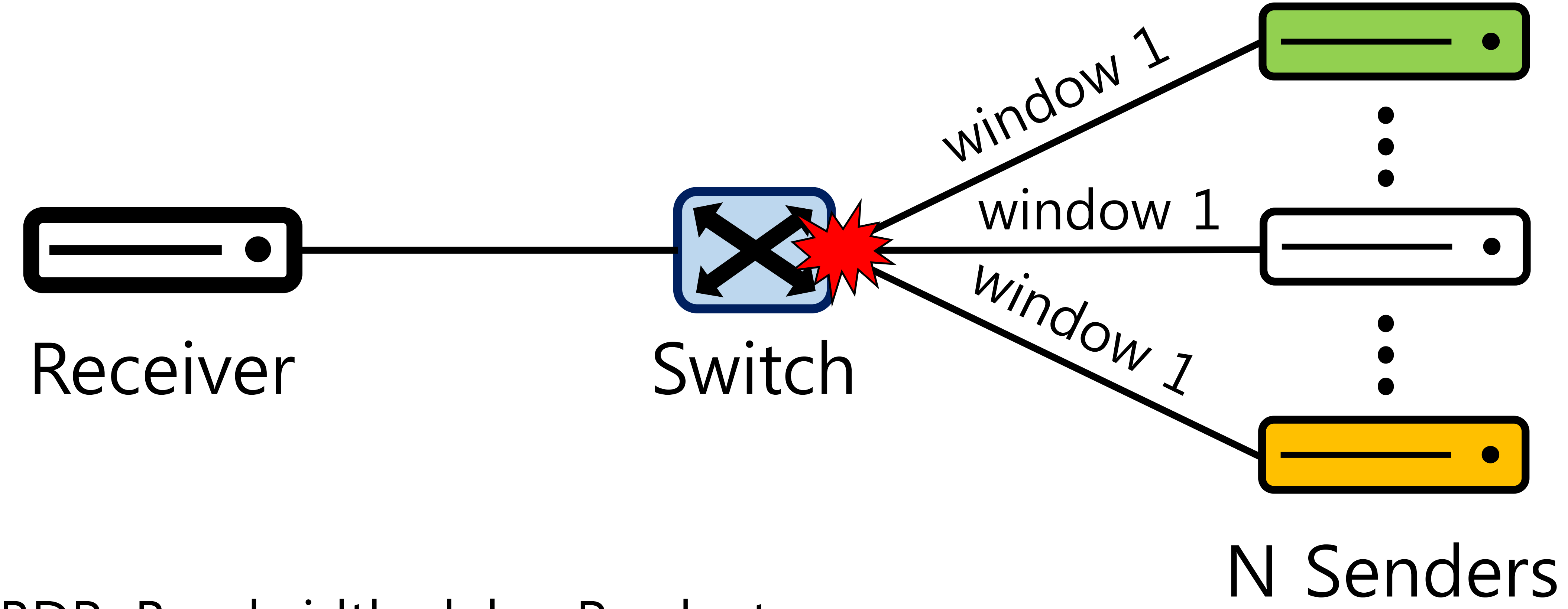
Large Scale

> 10,000 machines



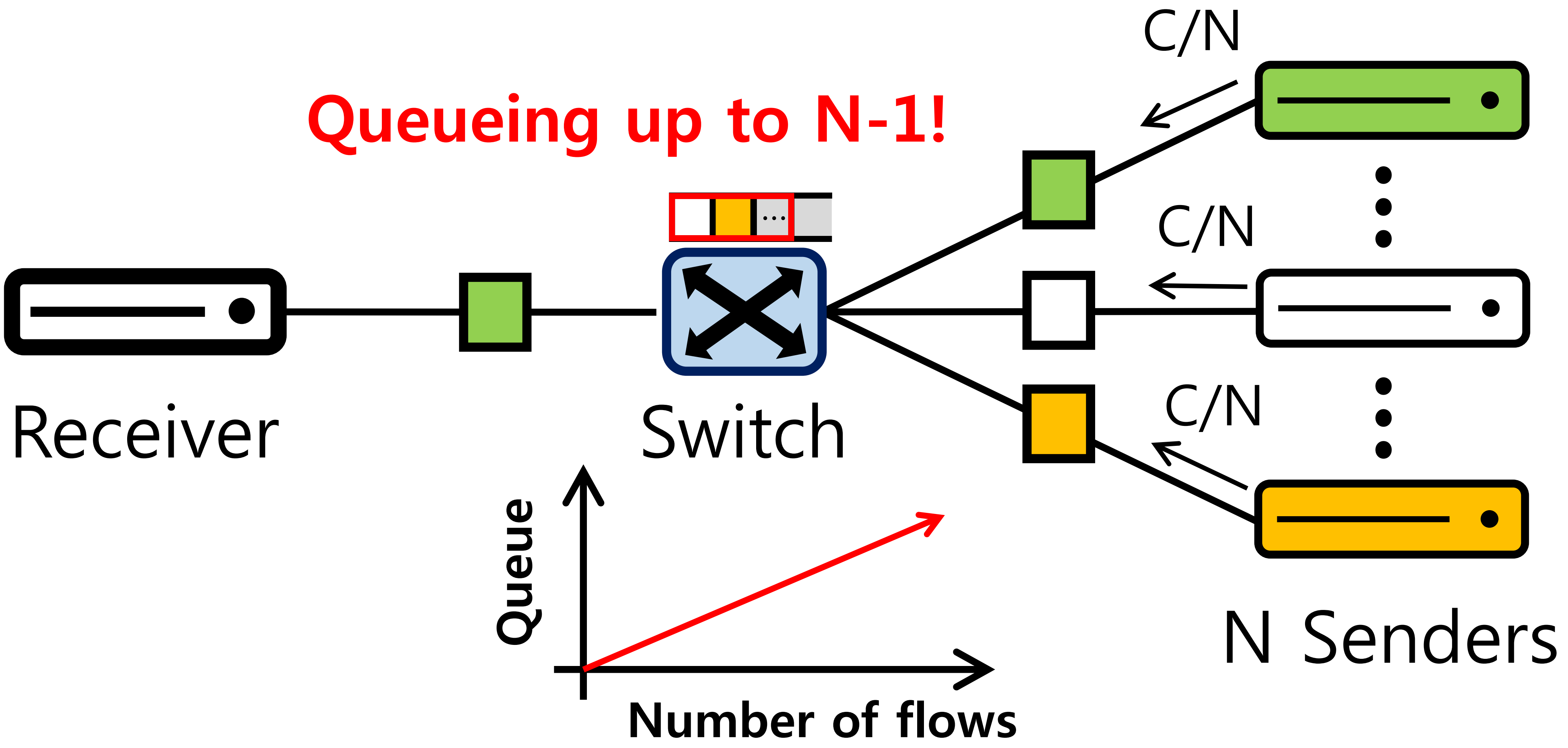
Challenge with small BDP

BDP*(100μs, 40Gbps) ≈ 300 MTUs

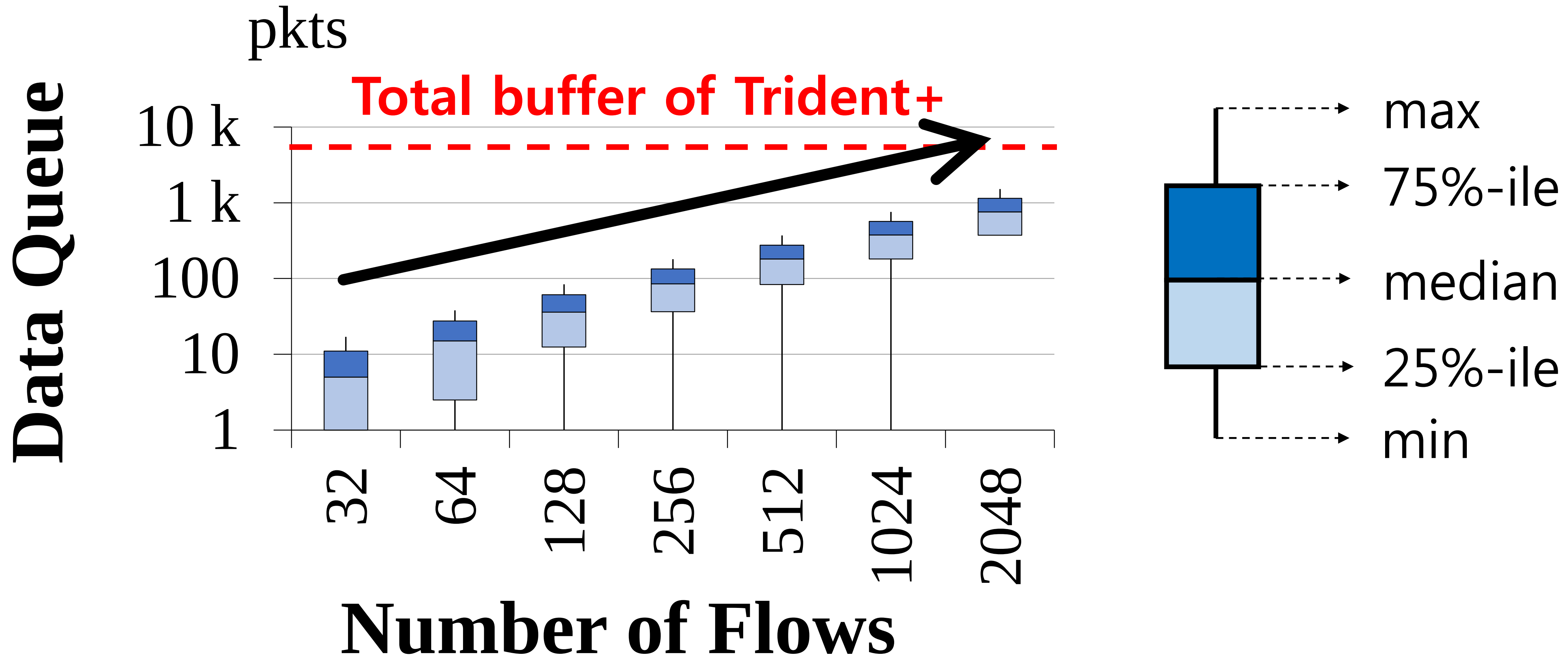


* BDP: Bandwidth-delay Product

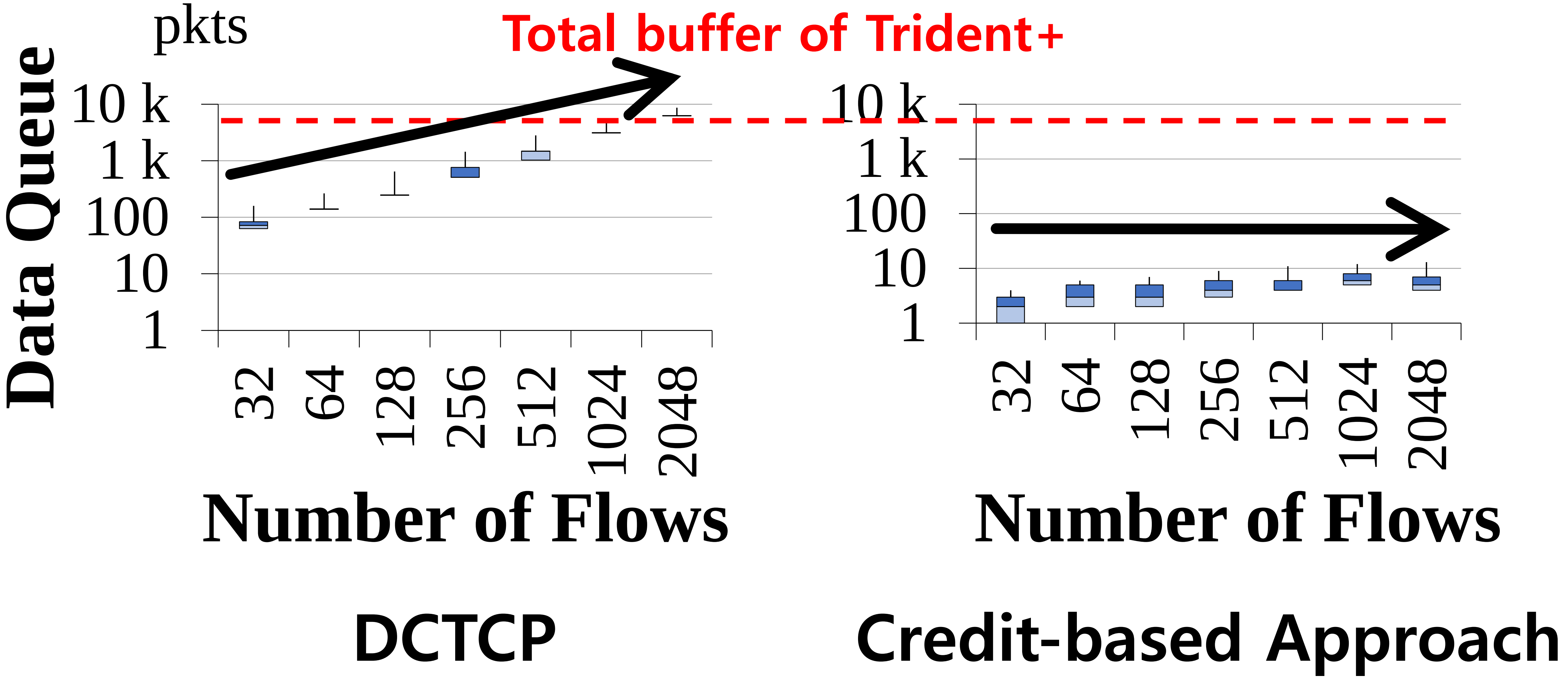
Rate-based CC + incast traffic



Rate-based CC + incast traffic



Rate-based CC vs. credit-based CC



Prior Work with Bounded Queue

Credit-based Flow Control

- InfiniBand
- ATM Network
- PCI Express

\+ Bounded queue

PFC

- RoCE/DCQCN

\+ Bounded queue

Centralized

- FastPass

\+ Bounded queue

Prior Work with Bounded Queue

Credit-based Flow Control

- InfiniBand
- ATM Network
- PCI Express

\+ Bounded queue

\- Does not scale to datacenter

\- Requires switch support

PFC

- RoCE/DCQCN

\+ Bounded queue

\- Head of line blocking

\- Possible deadlock

Centralized

- FastPass

\+ Bounded queue

\- Hard to scale

\- Global time sync

\- Single point of failure

Prior Work with Bounded Queue

Credit-based Flow Control

- InfiniBand
  - Does not scale to datacenter
  - Requires switch support

PFC

- RoCE/DCQCN
  - Head of line blocking
  - Possible deadlock

Centralized

- FastPass
  - Hard to scale
  - Global time sync
  - Single point of failure

How can we get the benefits of credit-based flow control on Ethernet?

Goal & Our Approach

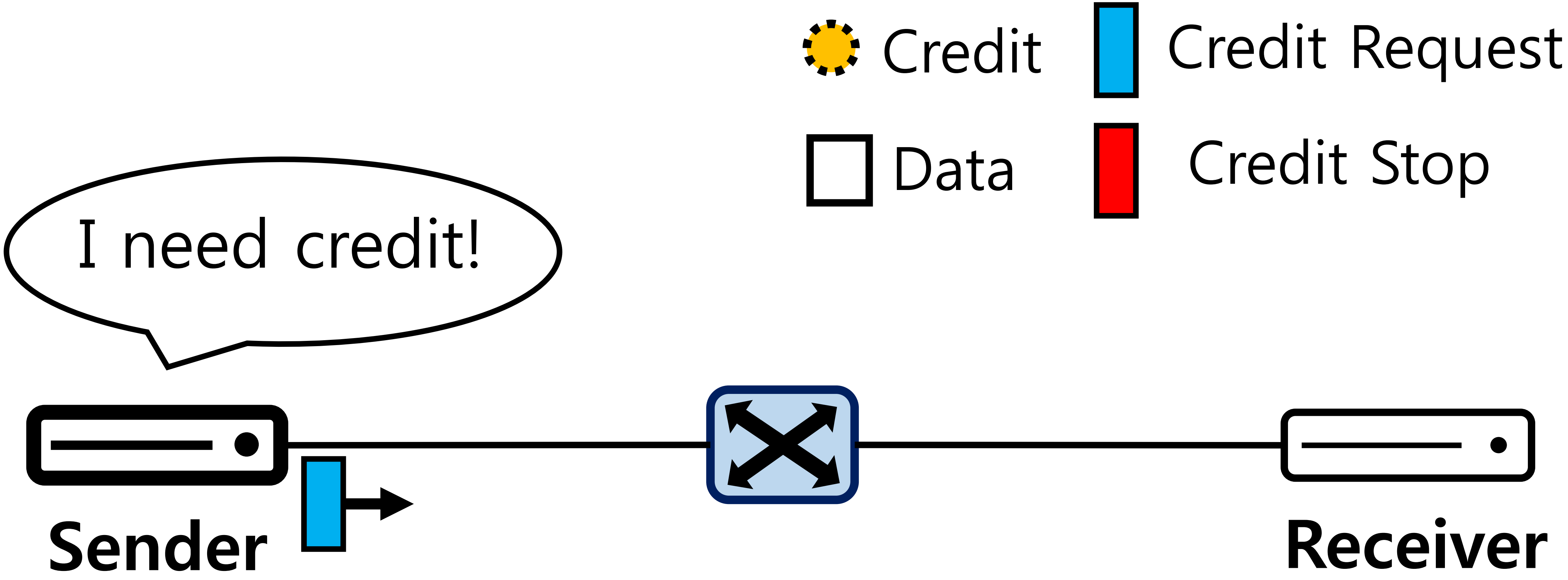
Goal

To achieve **bounded queue** even with heavy incast using **Ethernet switches.**

ExpressPass

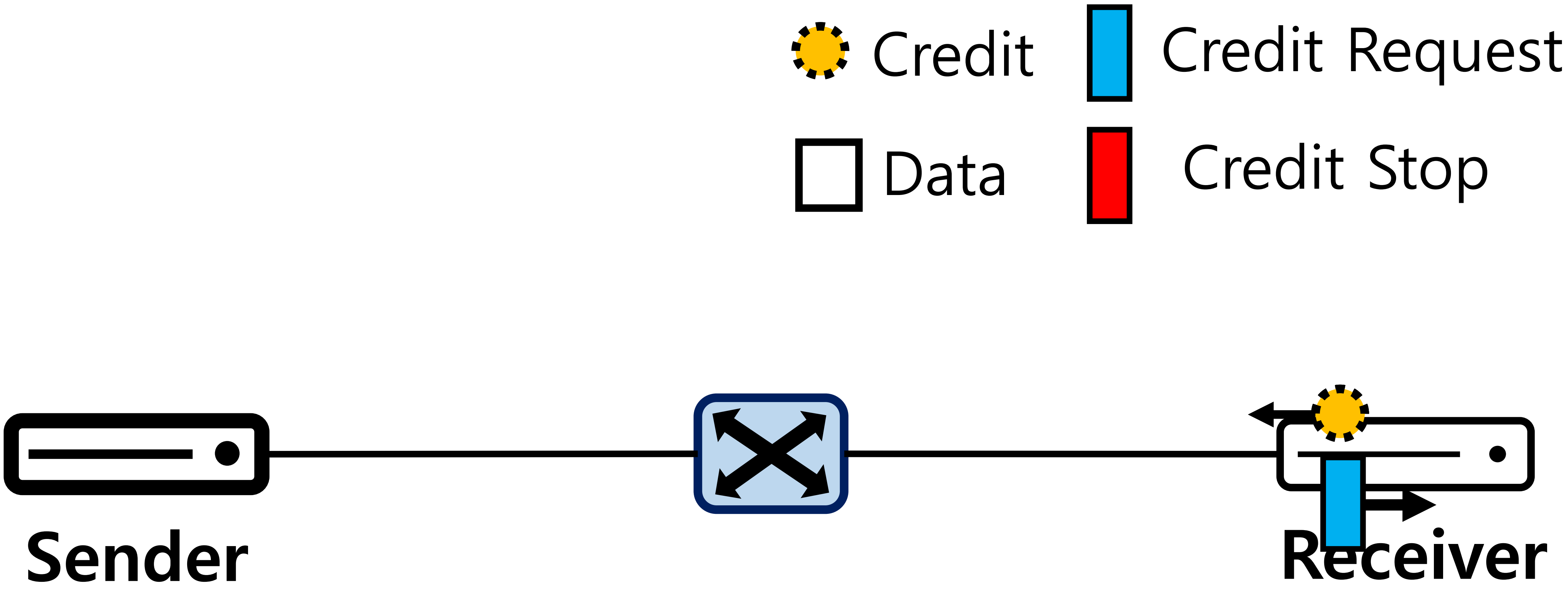
Proactive end-to-end credit-based congestion control using unreliable credits.

ExpressPass
End host behavior



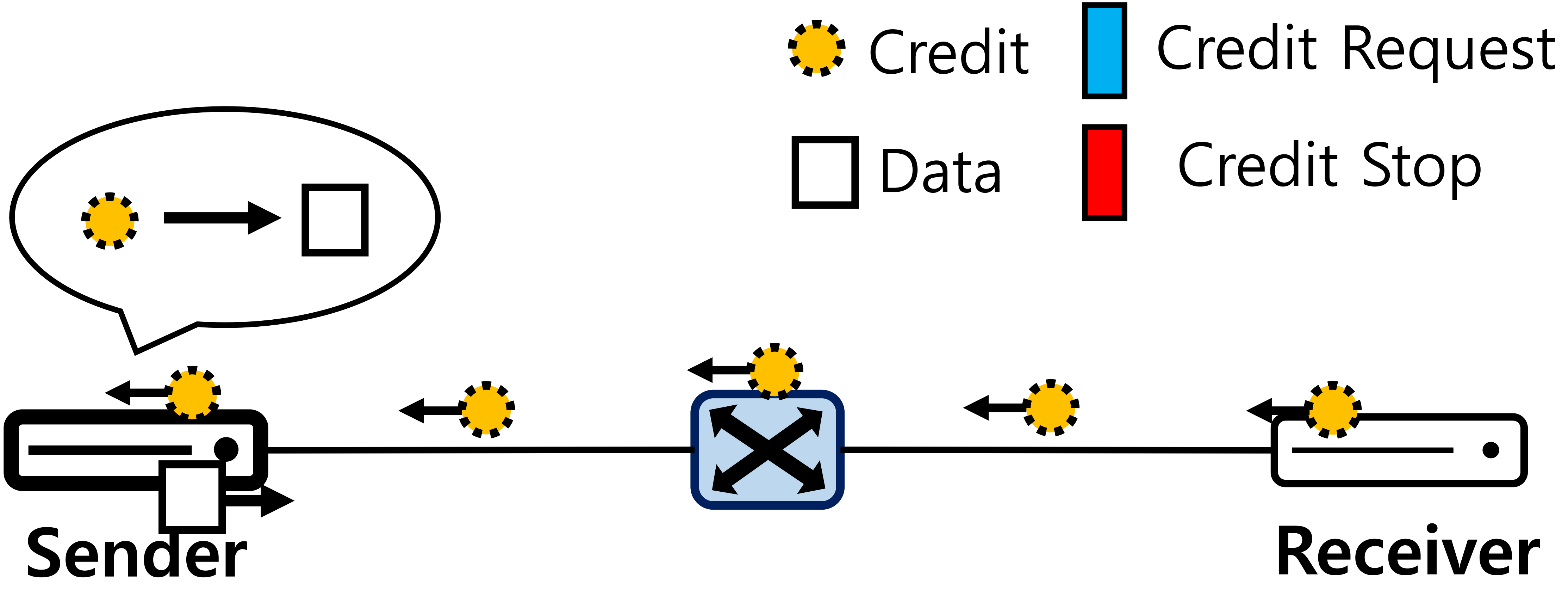
ExpressPass

End host behavior

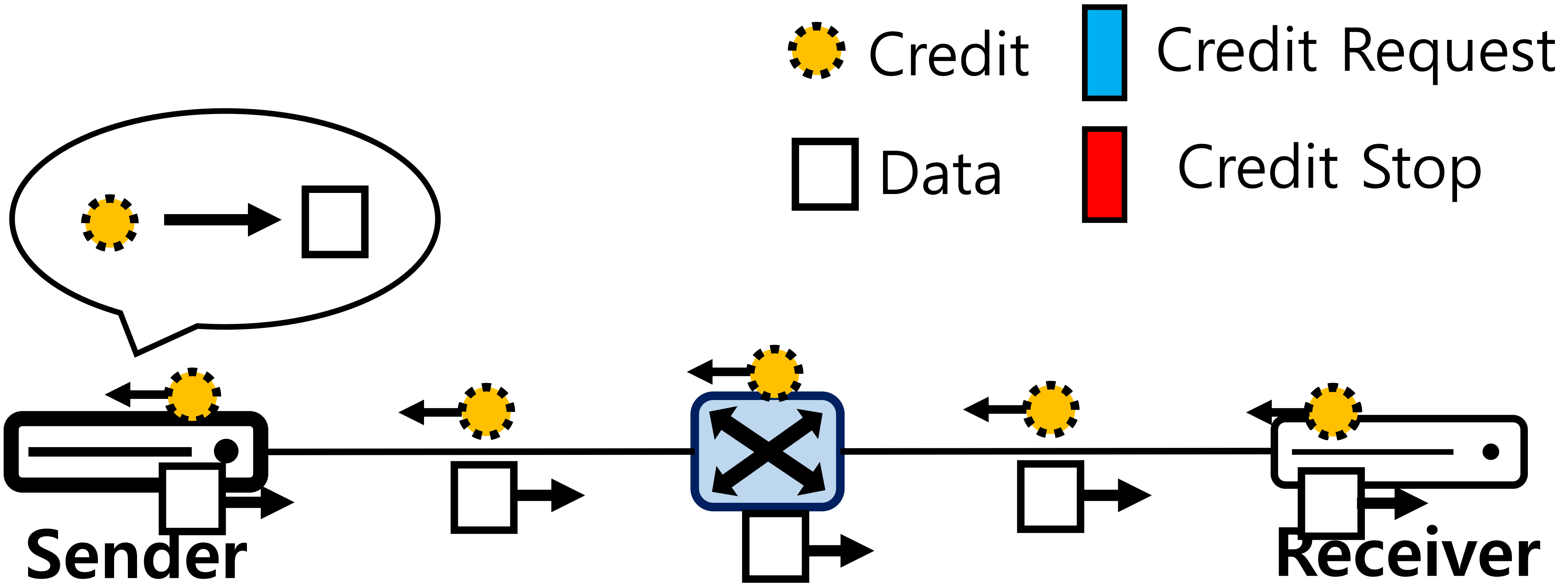


ExpressPass

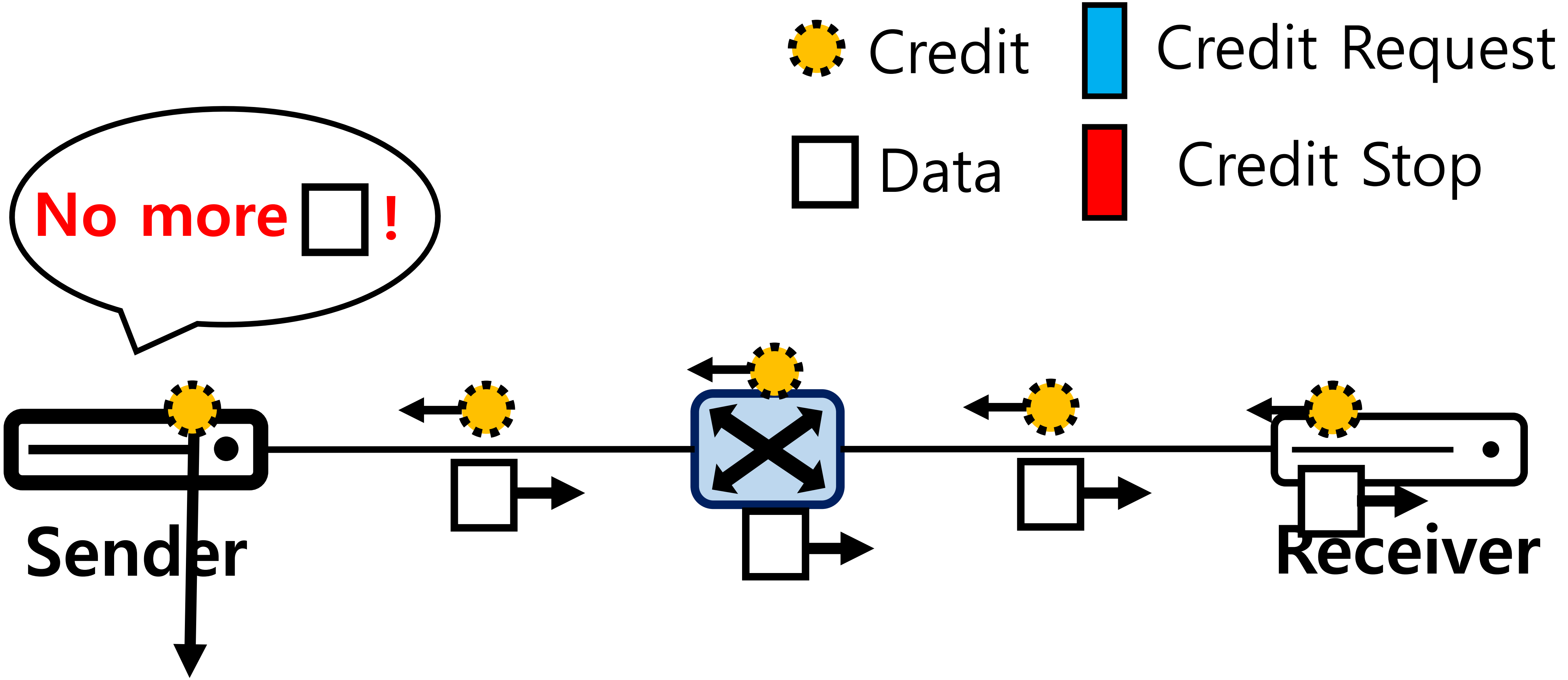
End host behavior



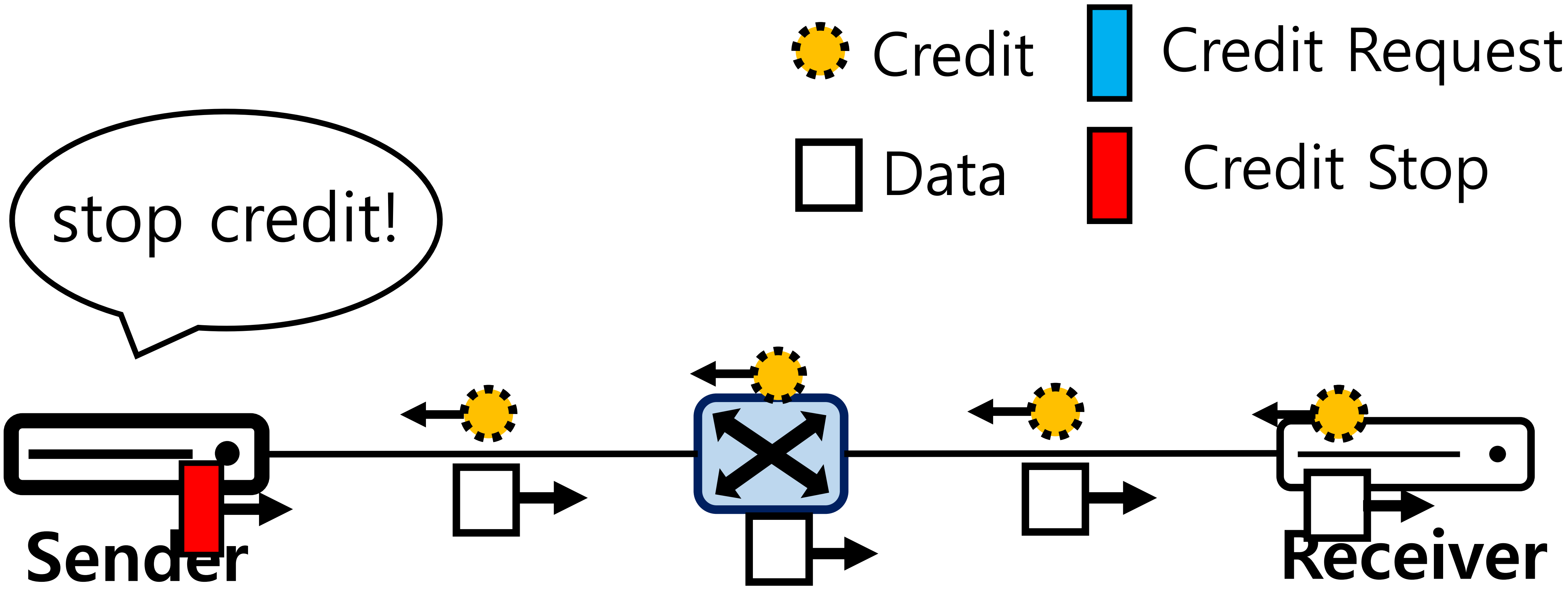
ExpressPass
End host behavior



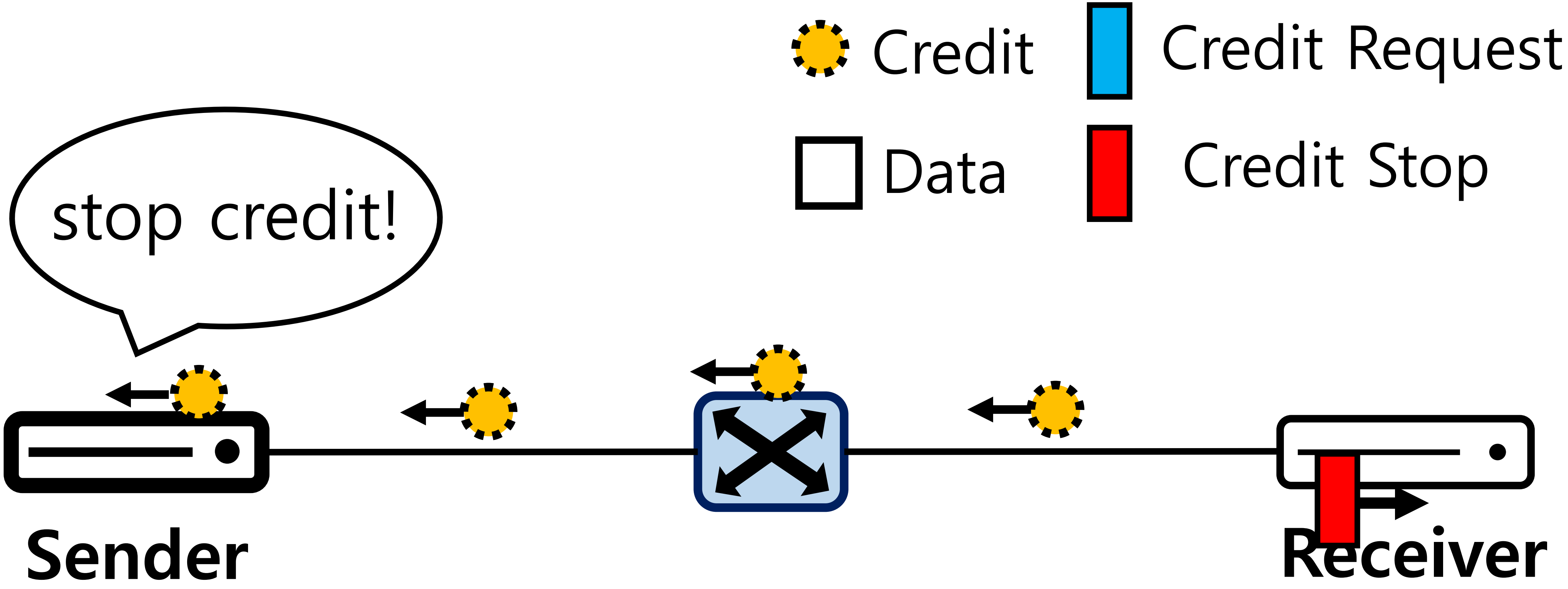
ExpressPass
End host behavior



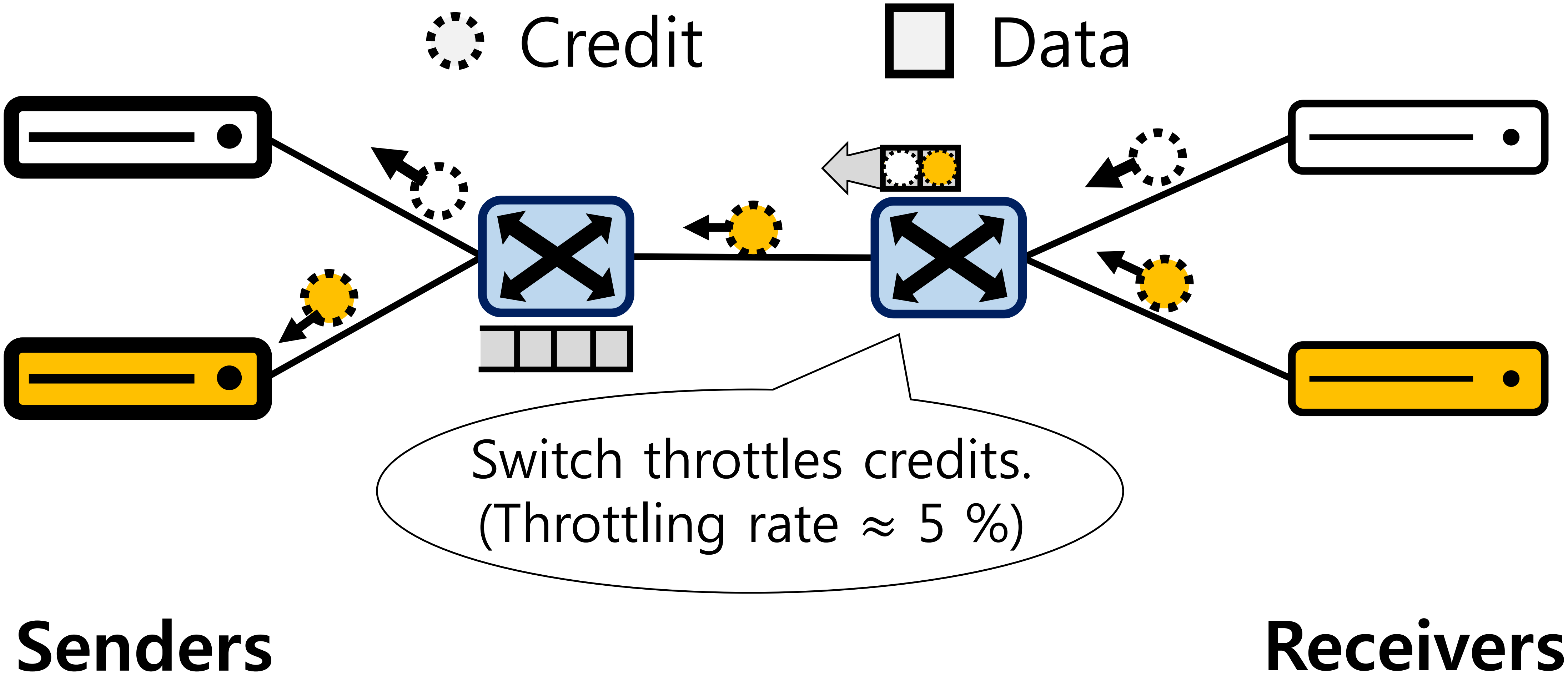
ExpressPass
End host behavior



ExpressPass
End host behavior

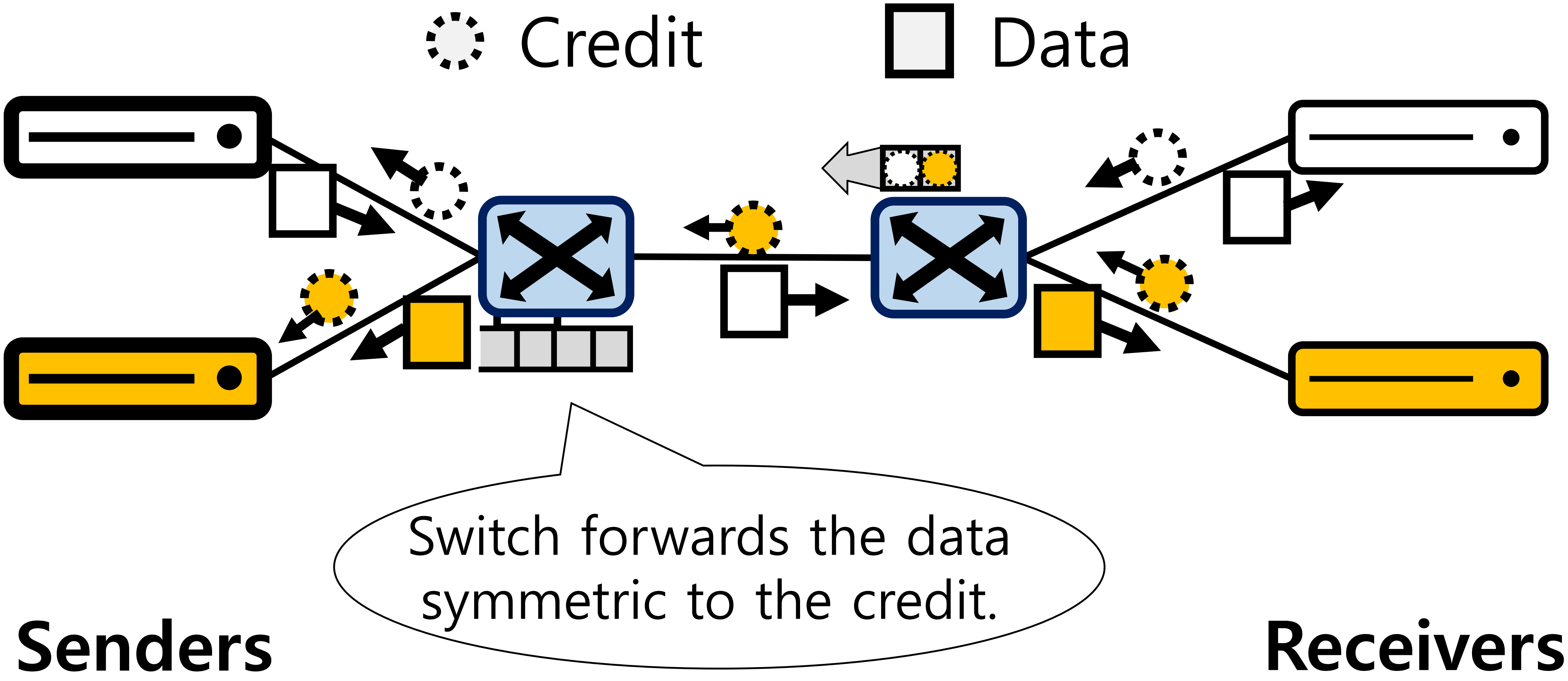


ExpressPass
Switch behavior

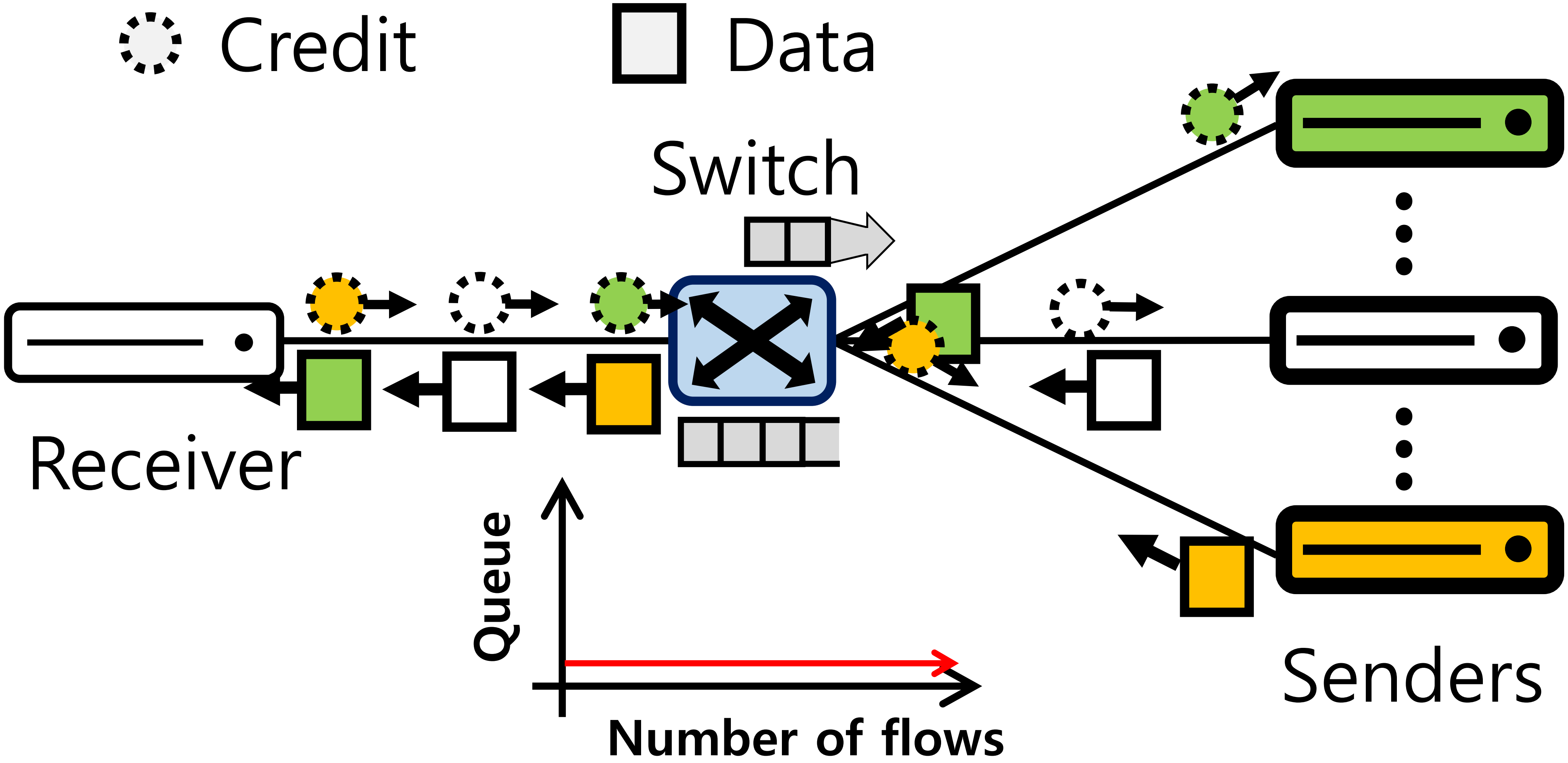


ExpressPass

Switch behavior



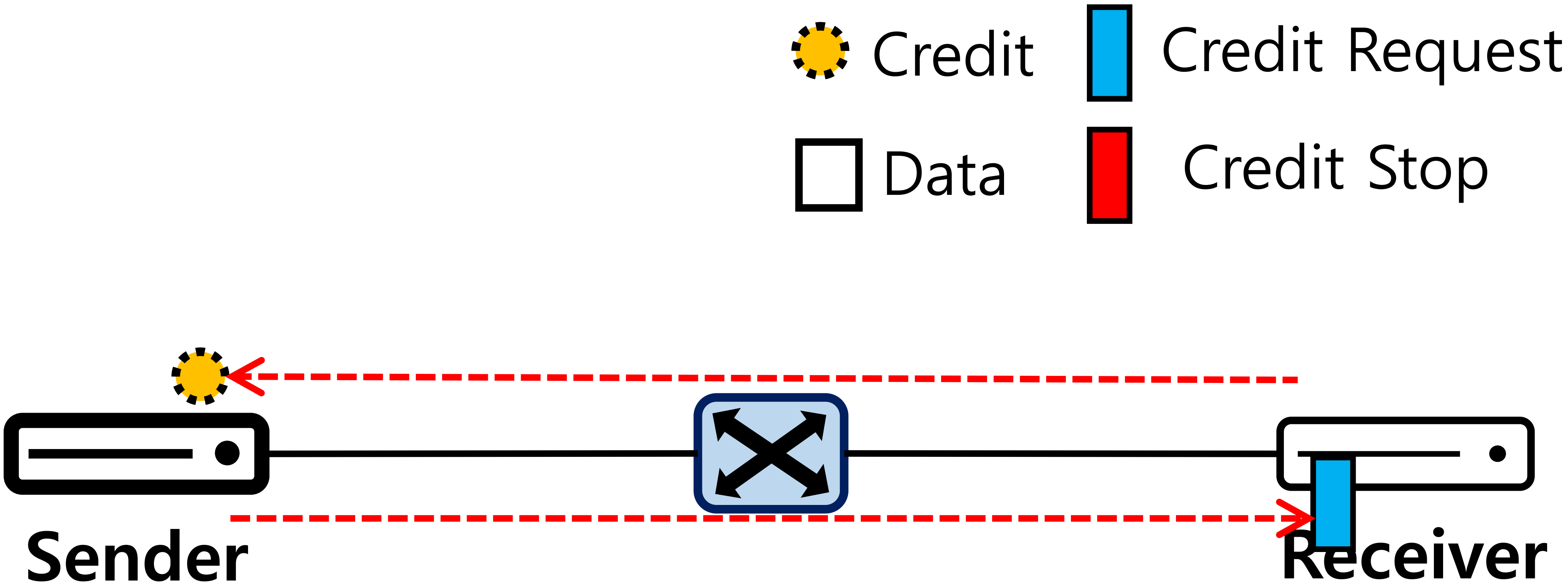
Credit-scheduled data transmission



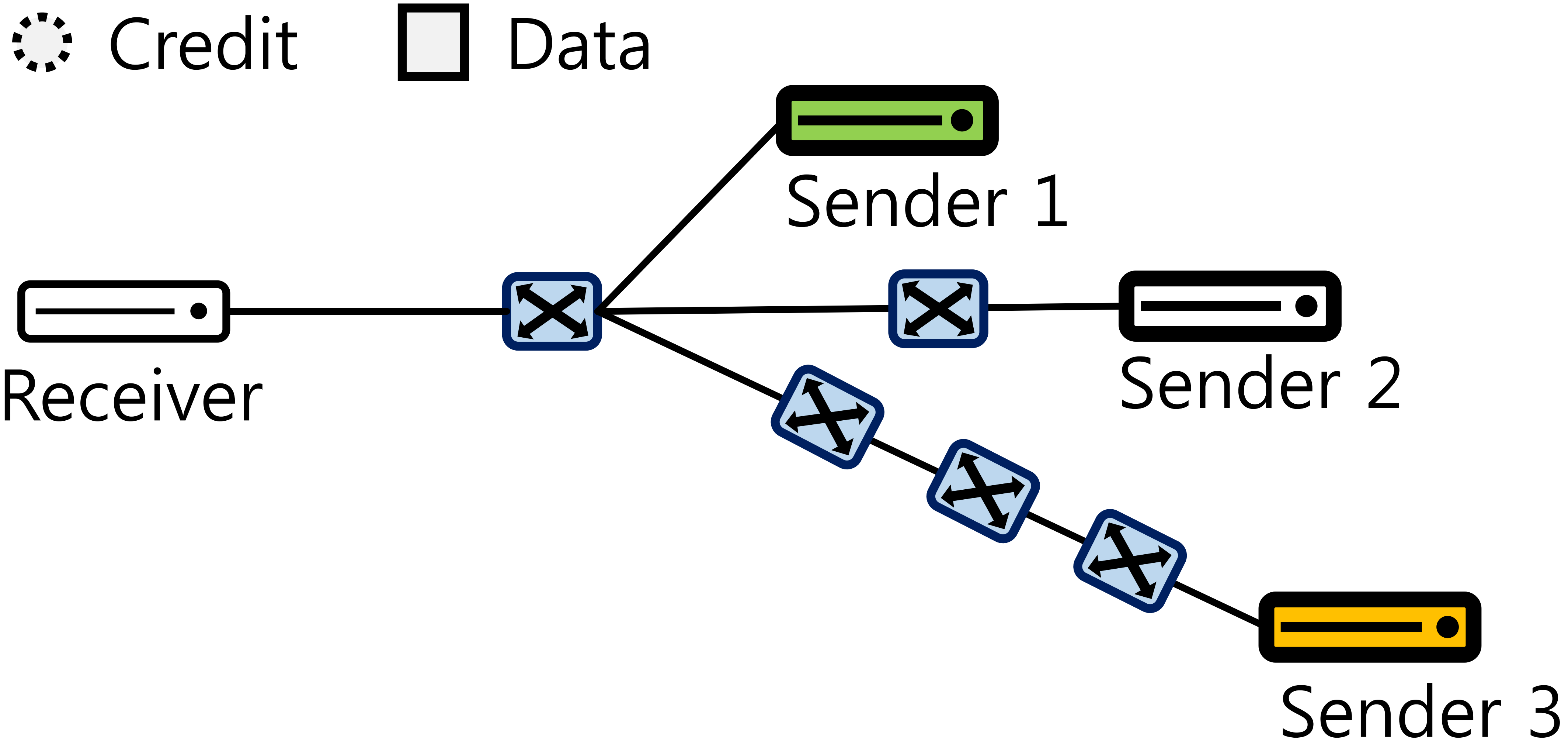
Challenges

Challenges	Techniques to address
Signaling overhead	Piggybacking to handshake packets
Non-zero queueing	Bounded queue
Credit waste	Credit feedback control
Fair drop on switch	Jitter, variable-sized credits
Path symmetry	Deterministic ECMP, packet level load balancing
Multiple traffic classes	Prioritizing credits rather than data

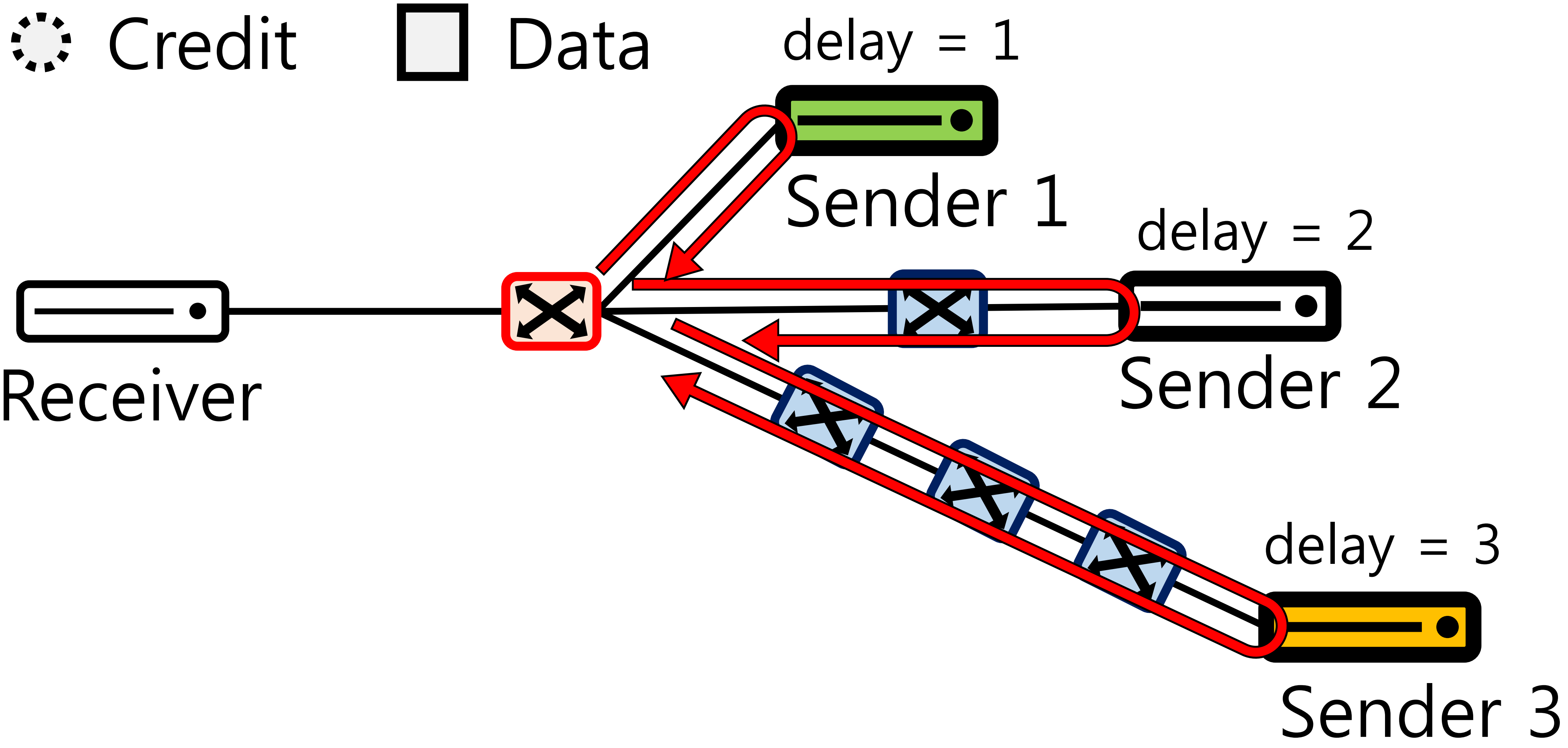
Signaling Overhead



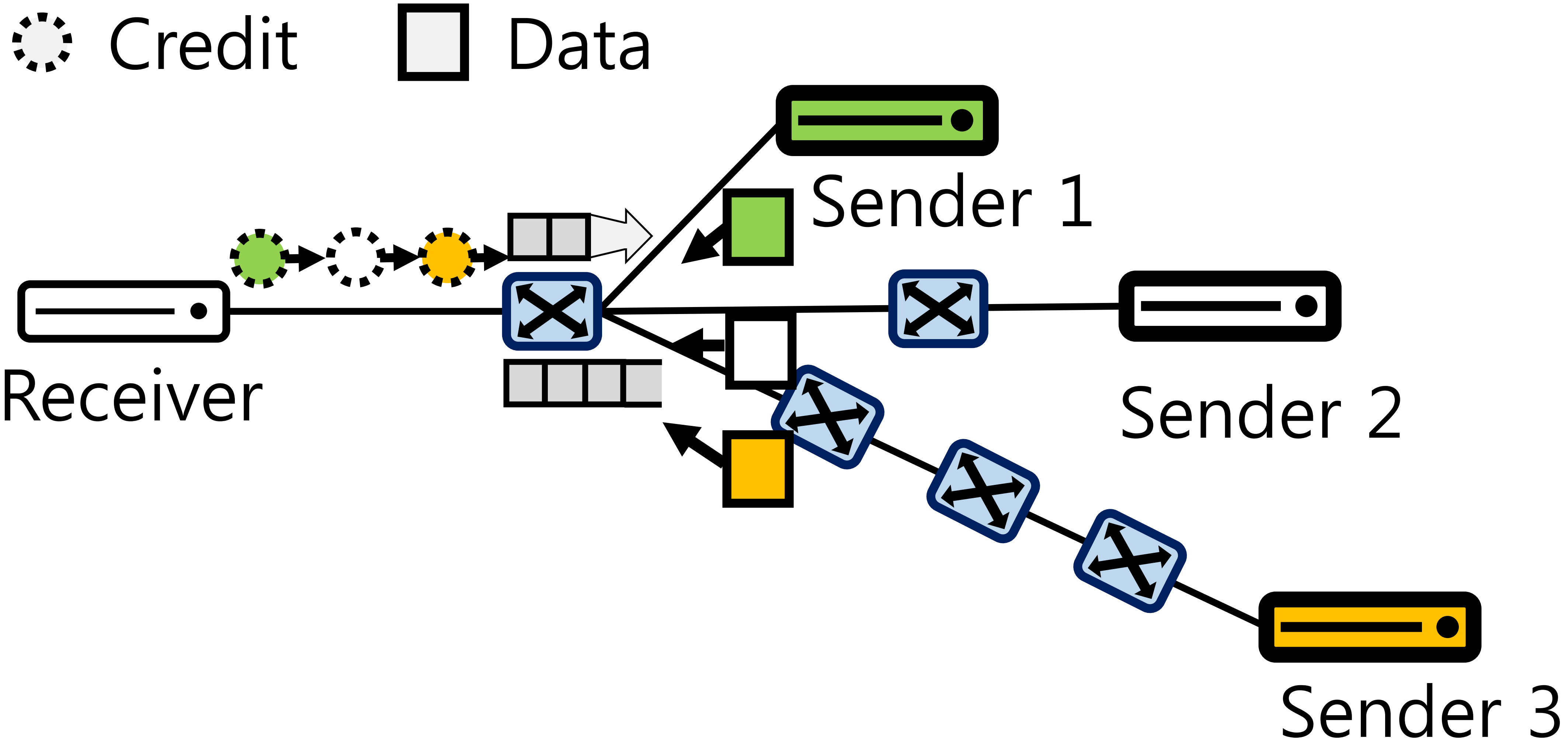
Maximum Bound of Data Queue



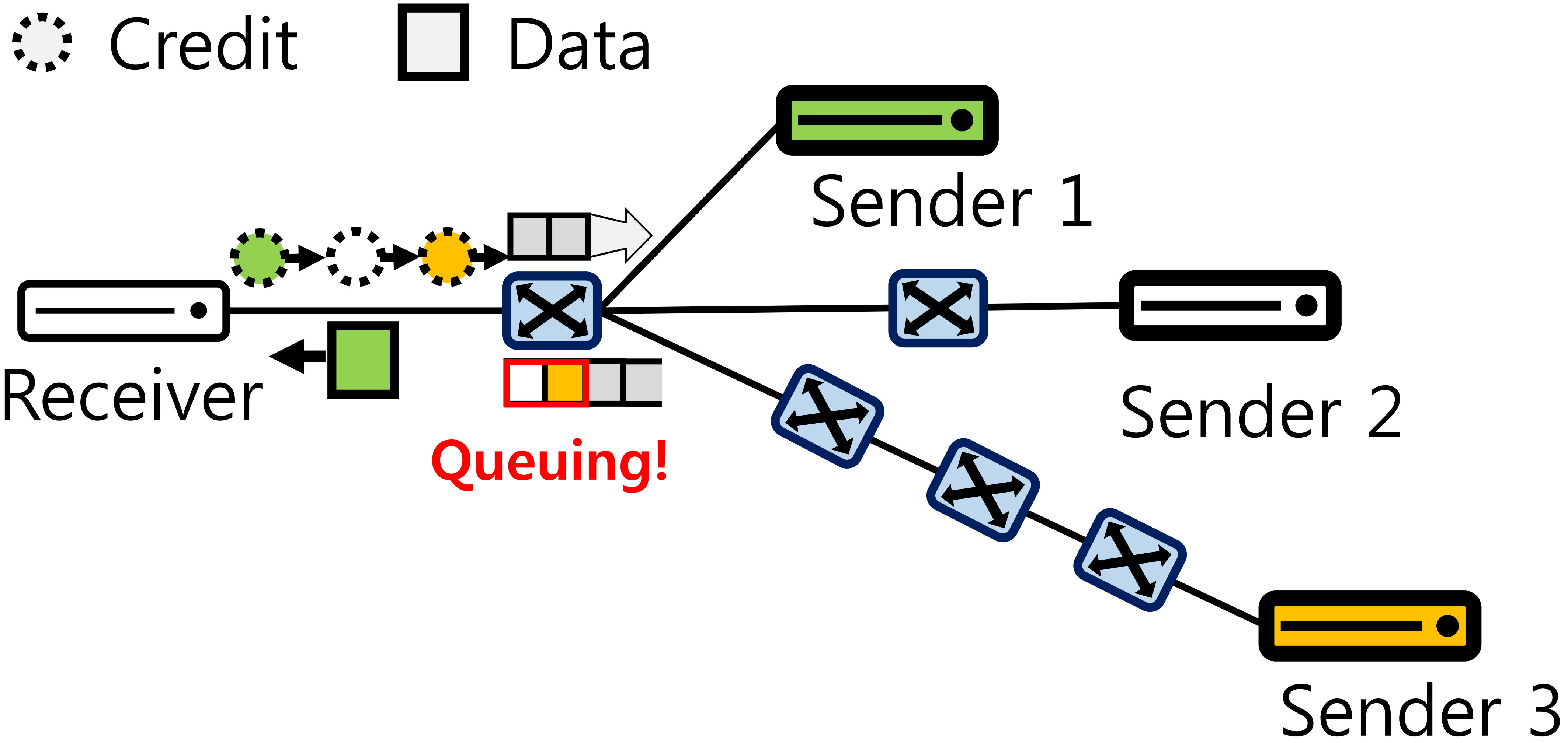
Maximum Bound of Data Queue



Maximum Bound of Data Queue

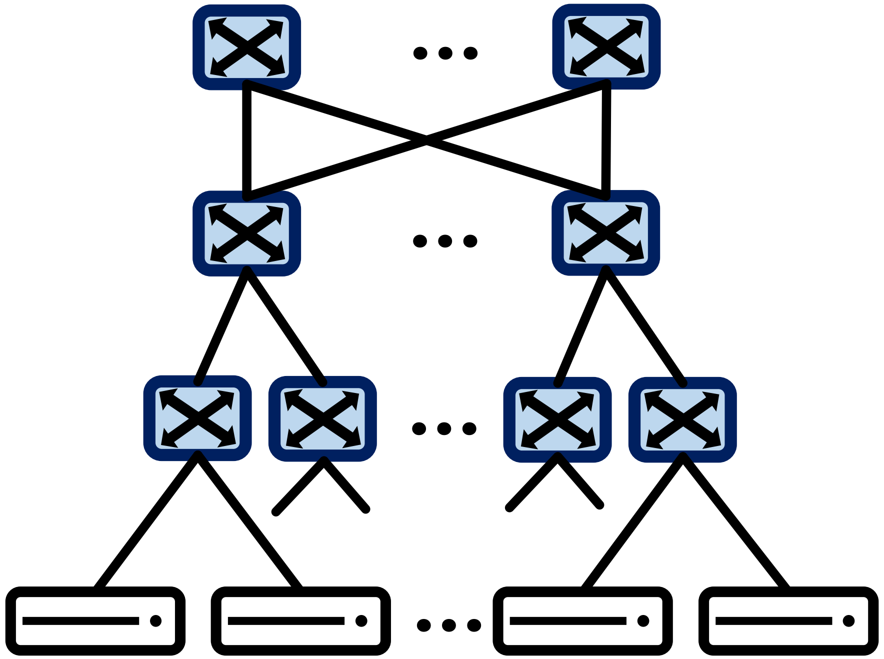


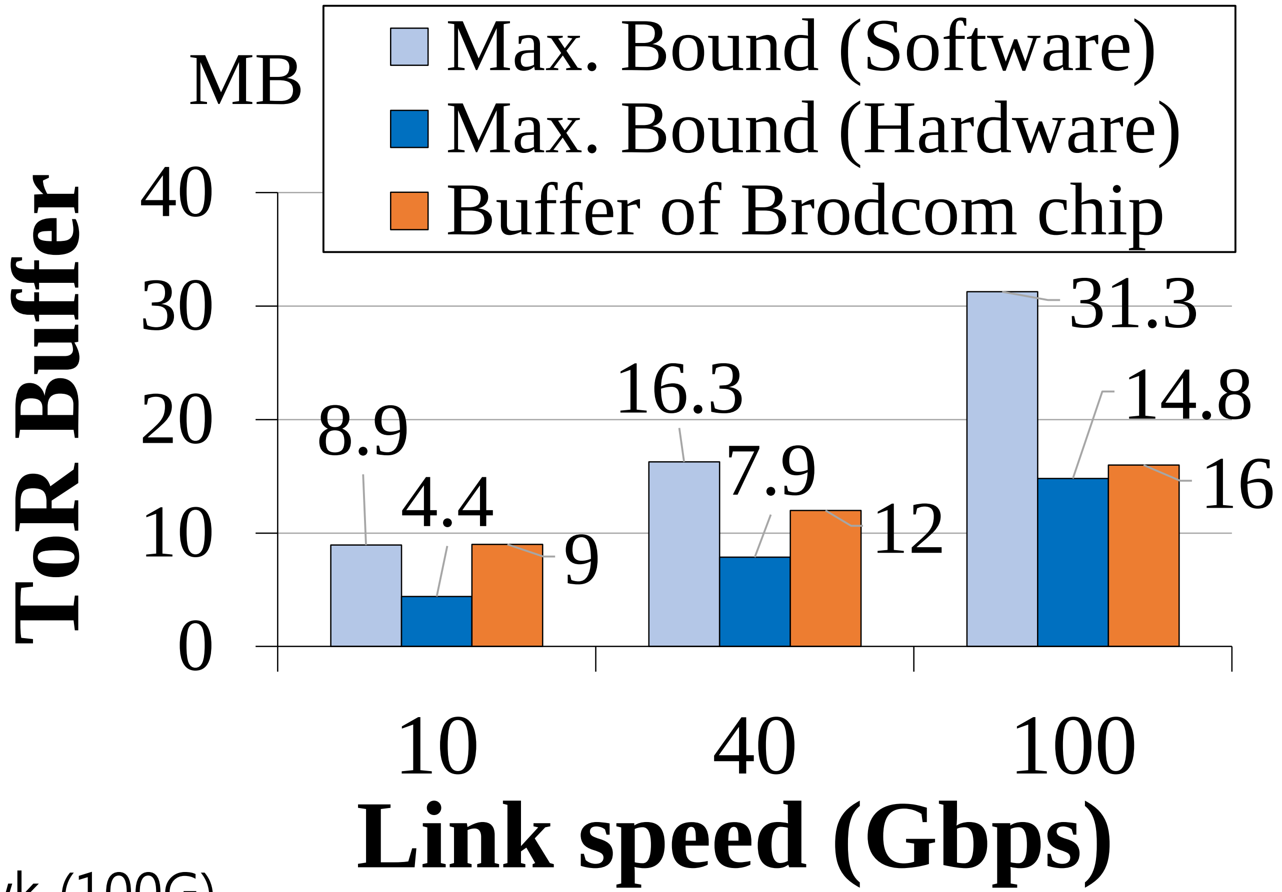
Maximum Bound of Data Queue



Maximum Bound of Data Queue

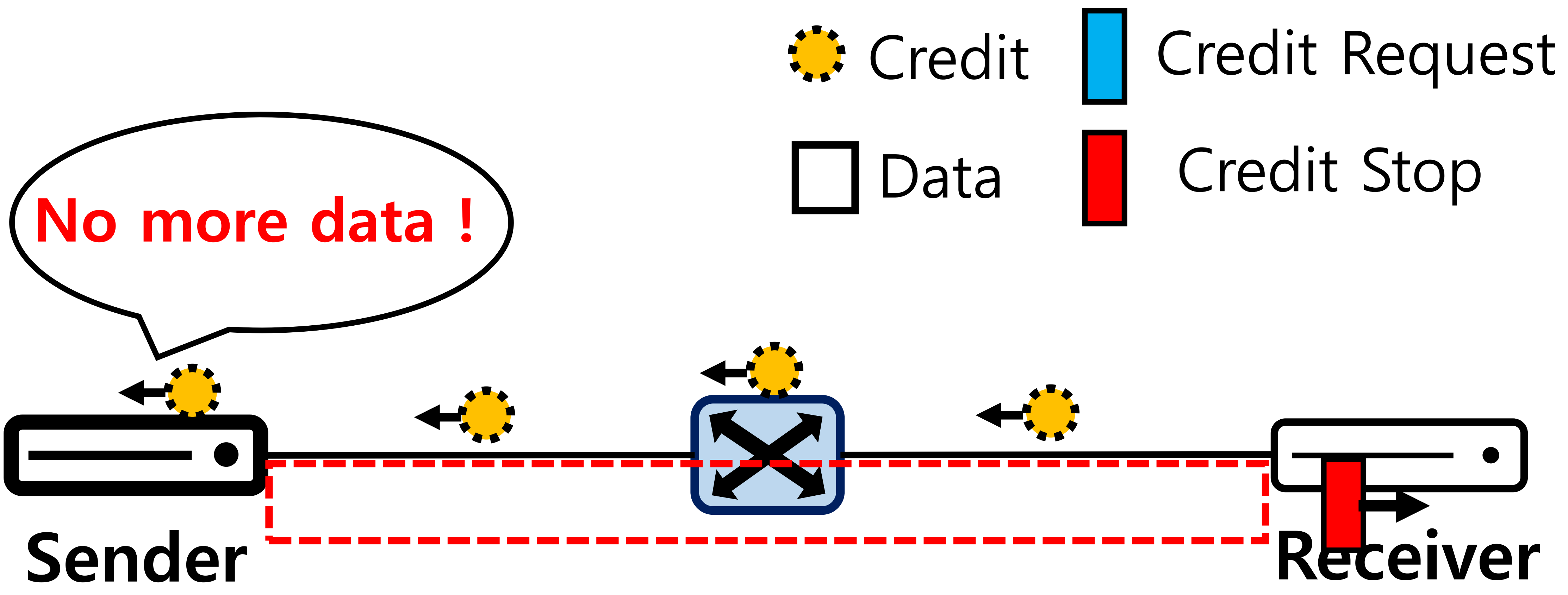
$$\max(buffer) = C * \{\max(delay) - \min(delay)\}$$





* Trident+ (10G), Trident II (40G), Tomahawk (100G)

Credit Waste



Credit Feedback Control

Proactive Congestion Control

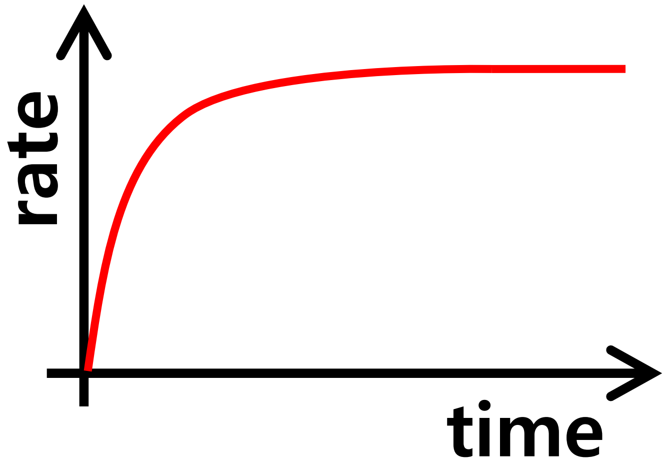
Prevents the congestion ***before*** actual congestion happens using credits.

Cheap credit drop

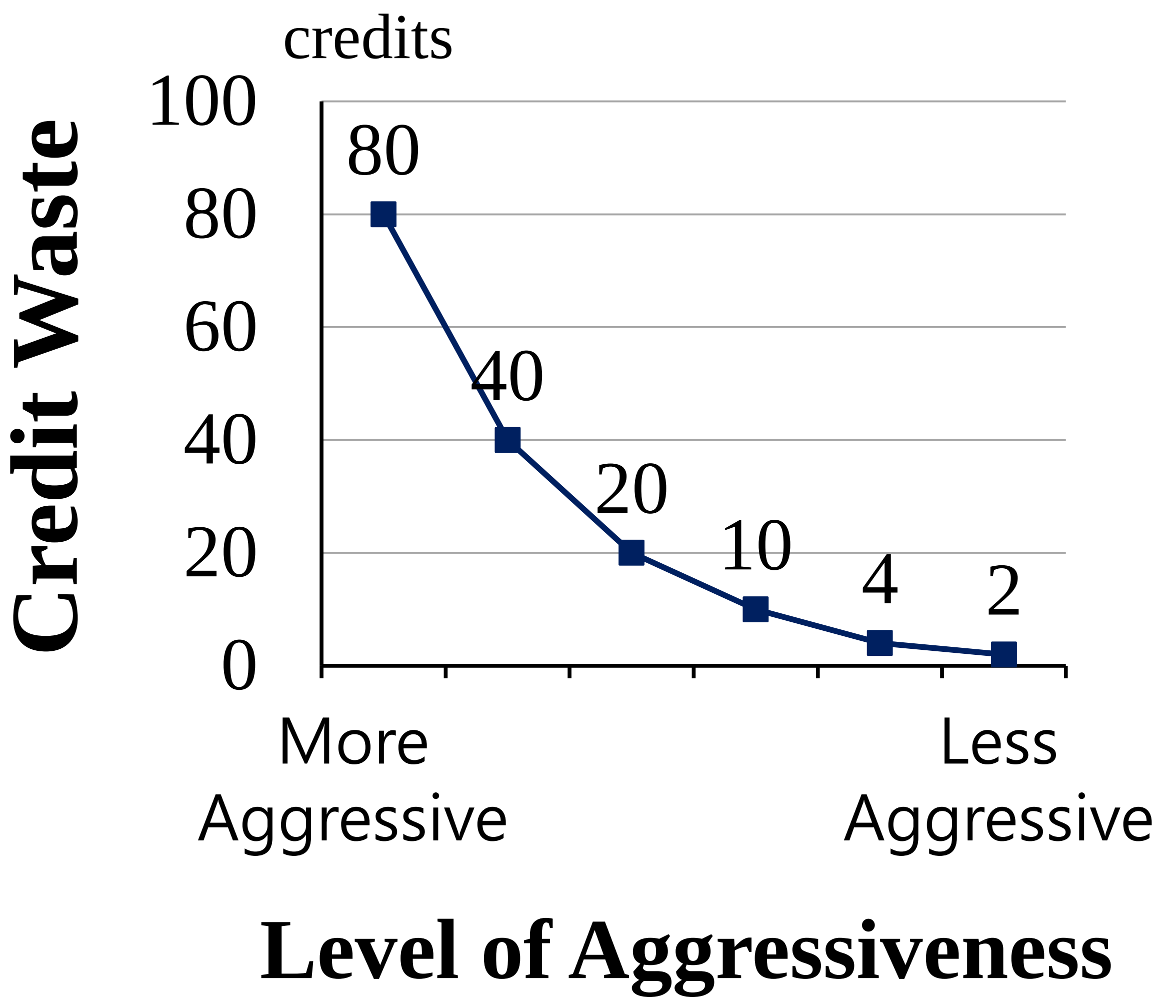
We can increase rate aggressively.

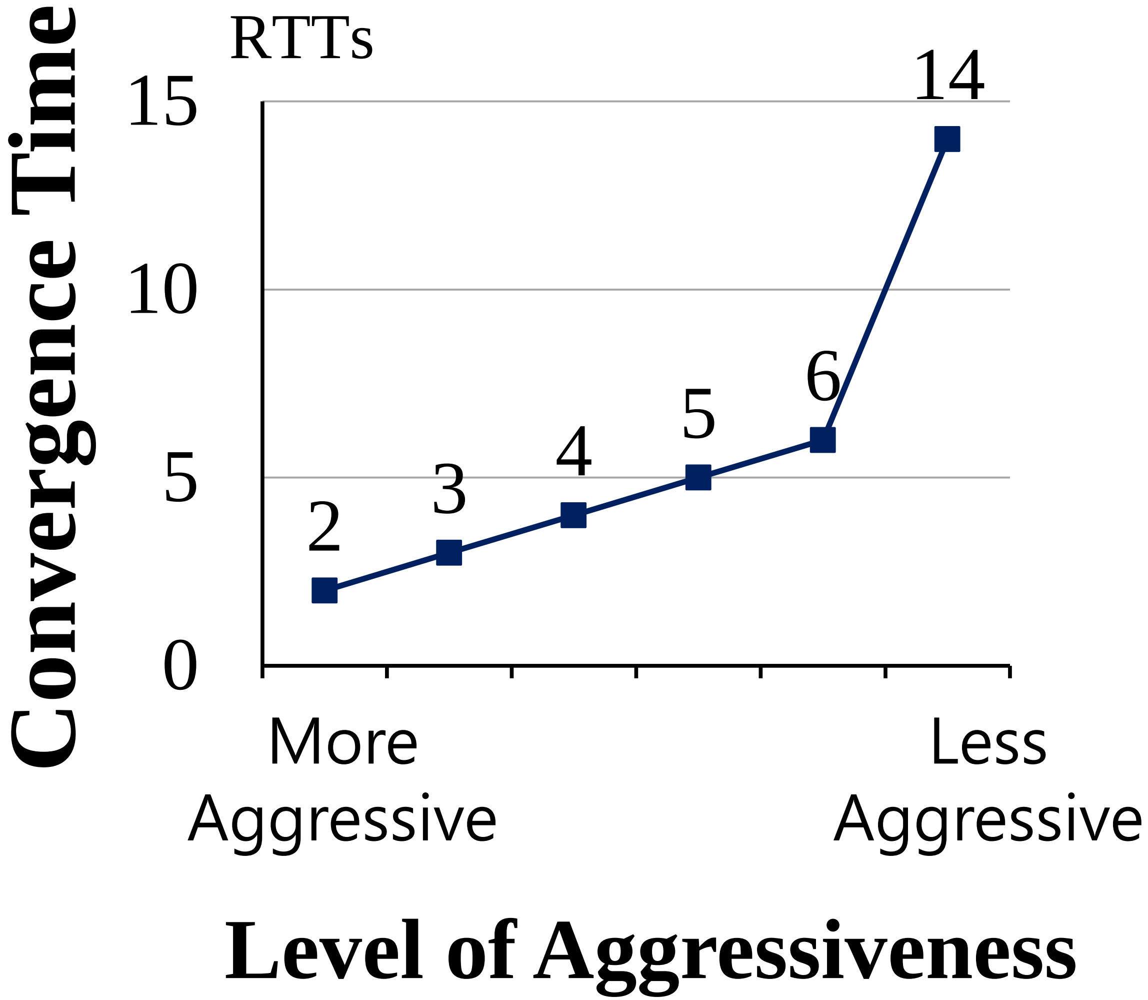
Bandwidth probing is cheap.

Convergence can be faster.

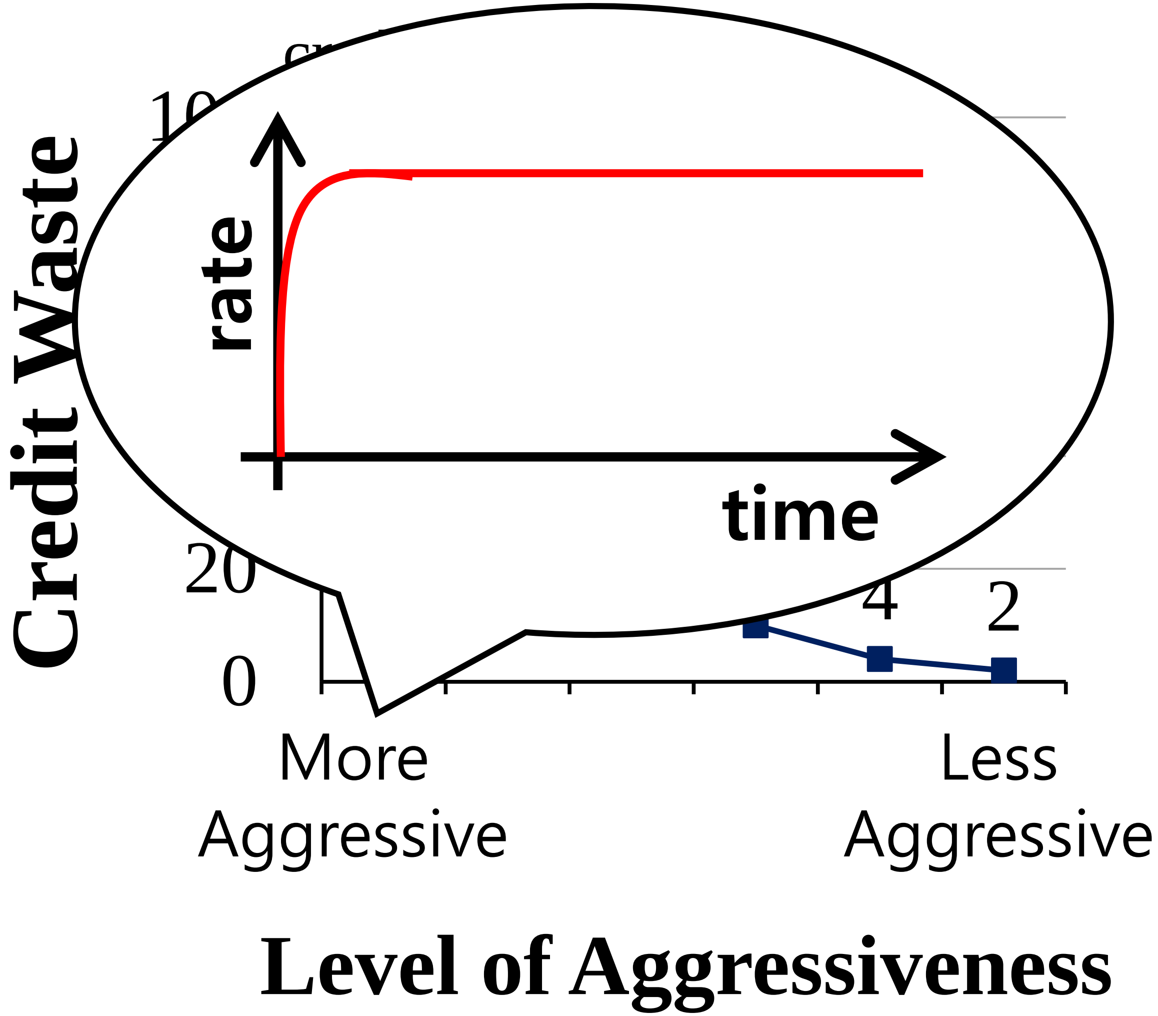


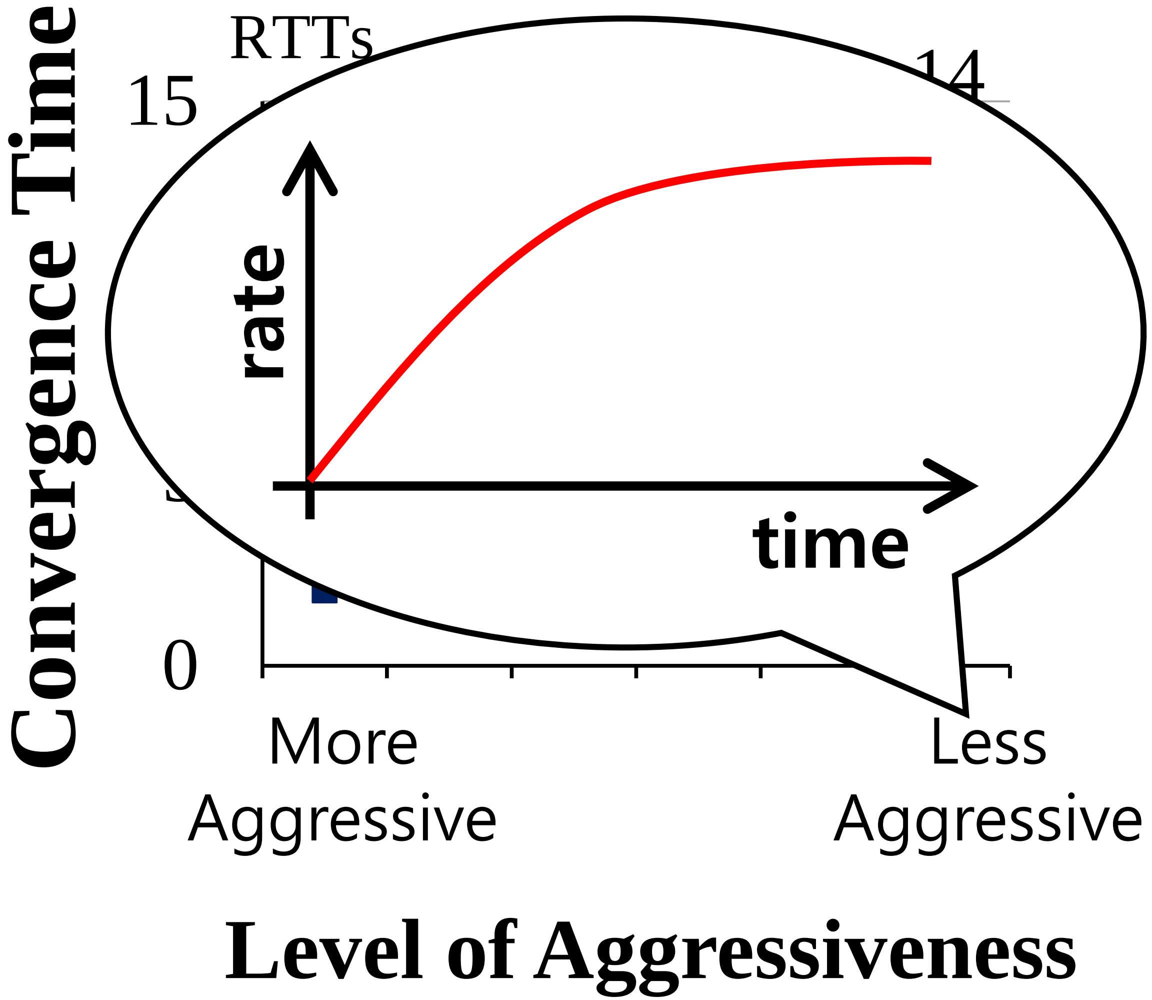
Credit Waste & Convergence Time



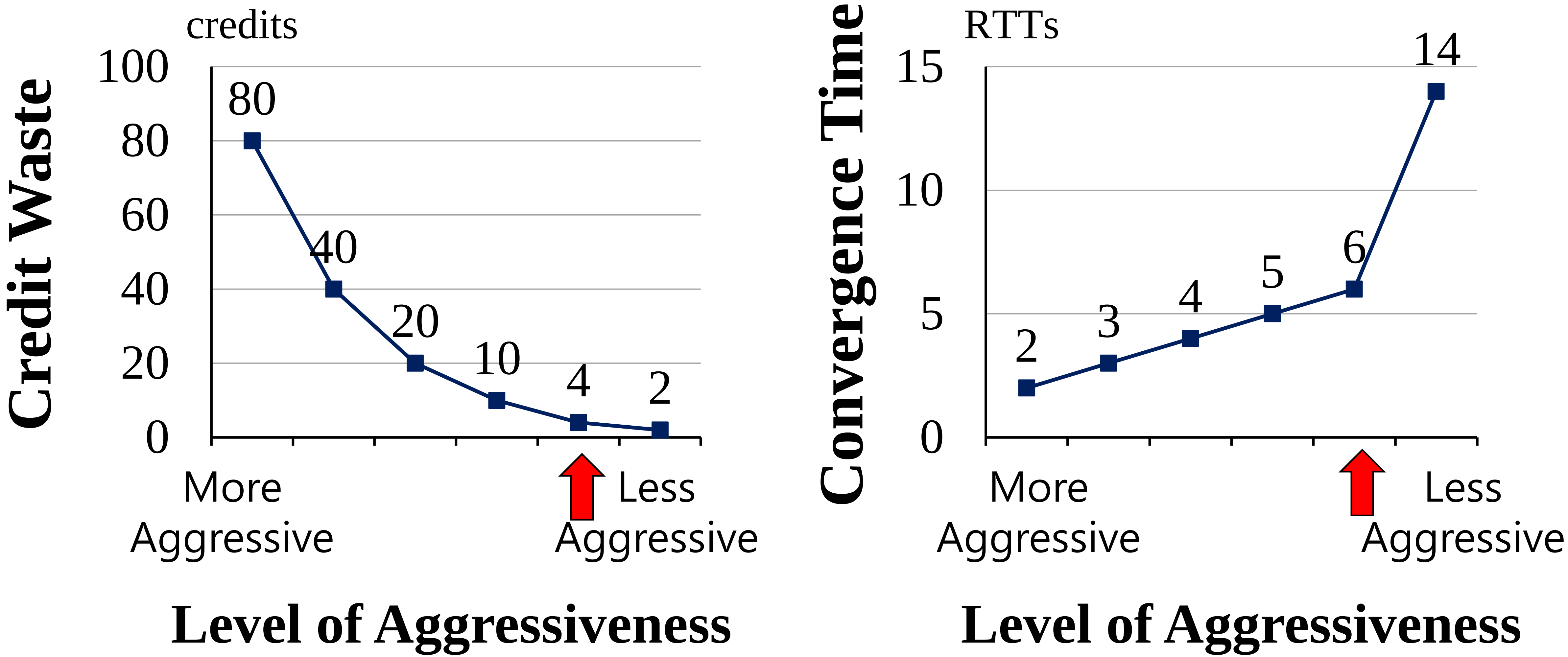


Credit Waste & Convergence Time





Credit Waste & Convergence Time



Evaluation Setup

Testbed setup

- Dumbbell topology
- Implementation on SoftNIC
- 12 hosts (Xeon E3/E5) connected to single ToR (Quanta T3048)
- Each host has 10Gbps x 1port

NS-2 Simulation Setup

- Fat-tree topology
- 192 hosts / 32 ToR / 16 aggr. / 8 core switches
- Each host has 10Gbps x 1port

Evaluation

(1) Does ExpressPass provides low & bounded queueing with realistic workloads?
(2) Is the convergence fast and stable?
(3) How low & bounded queuing and fast & stable convergence translate into the flow completion time?

Realistic Workloads

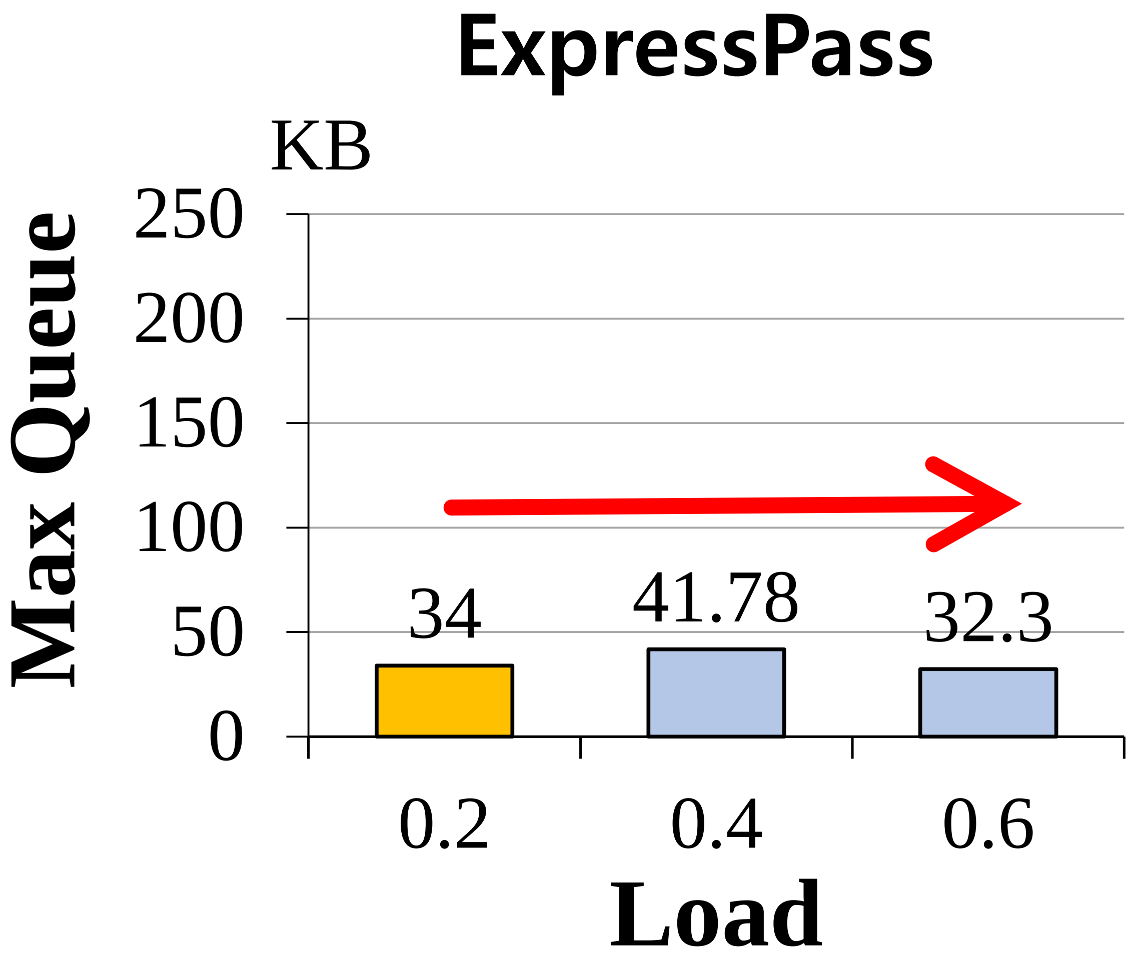
	Data Mining	Web Search	Cache Follower	Web Server
0 – 10KB (S)	78%	49%	50%	63%
10 – 100KB (M)	5%	3%	3%	18%
100KB–1MB (L)	8%	18%	18%	19%
1MB- (XL)	9%	20%	29%	-
Average flow size	7.41MB	1.6MB	701KB	64KB

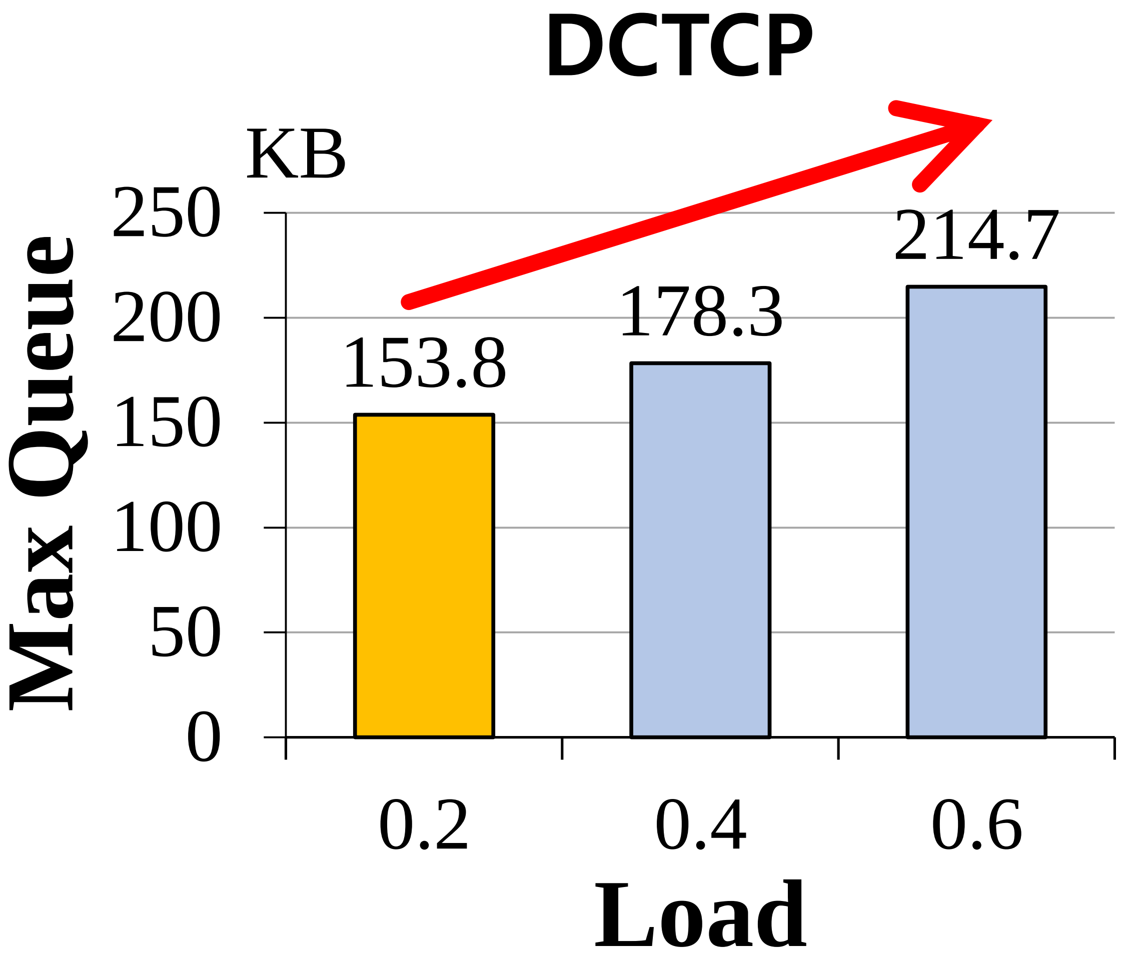
Realistic Workloads

	Data Mining	Web Search	Cache Follower	Web Server
0 – 10KB (S)	78%	49%	50%	63%
10 – 100KB (M)	5%	3%	3%	18%
100KB–1MB (L)	8%	18%	18%	19%
1MB- (XL)	9%	20%	29%	-
Average flow size	7.41MB	1.6MB	701KB	64KB

Bounded Queue

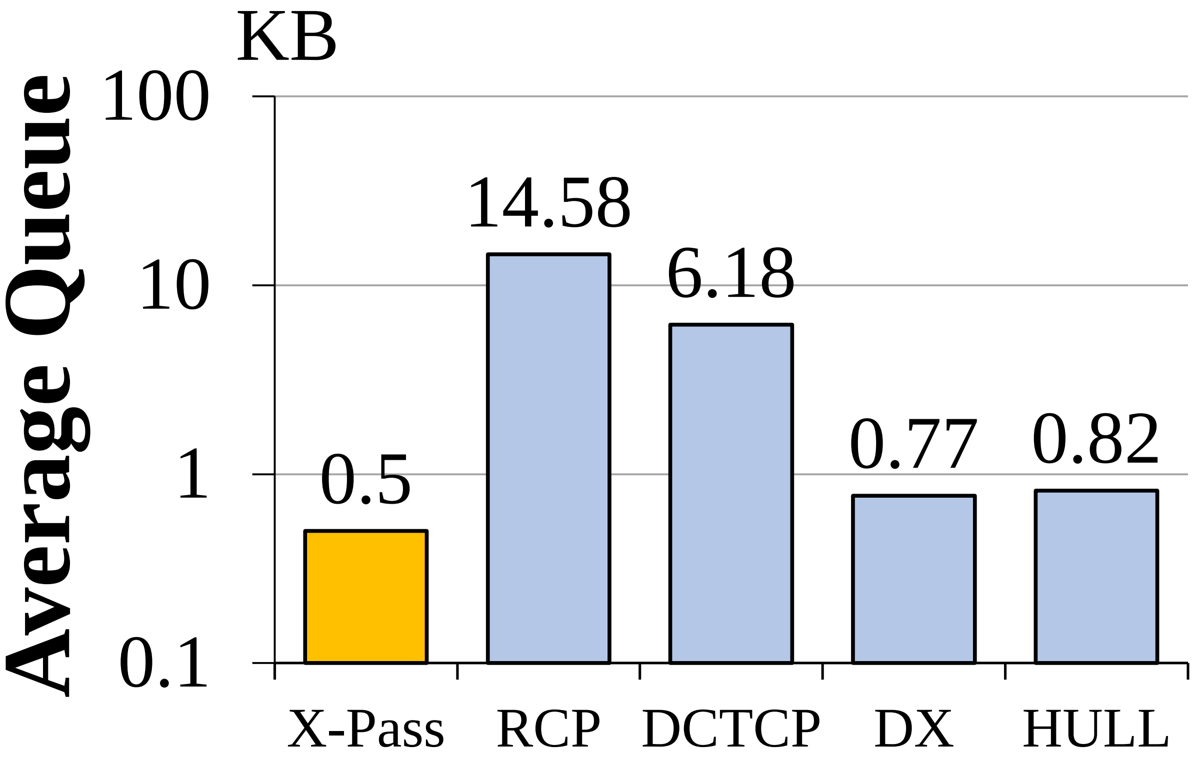
cache follower workload / load 0.2 – 0.4 / 0KB ~ (All Size)

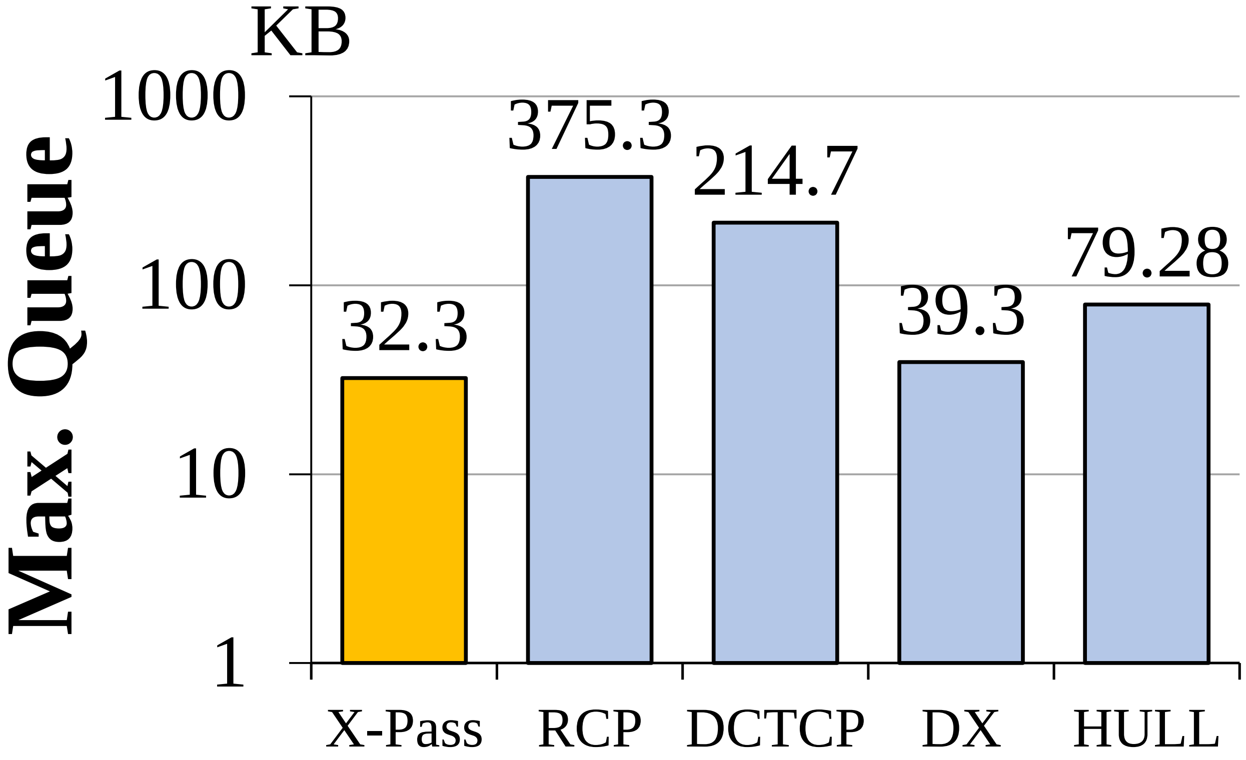




Low Average Queue

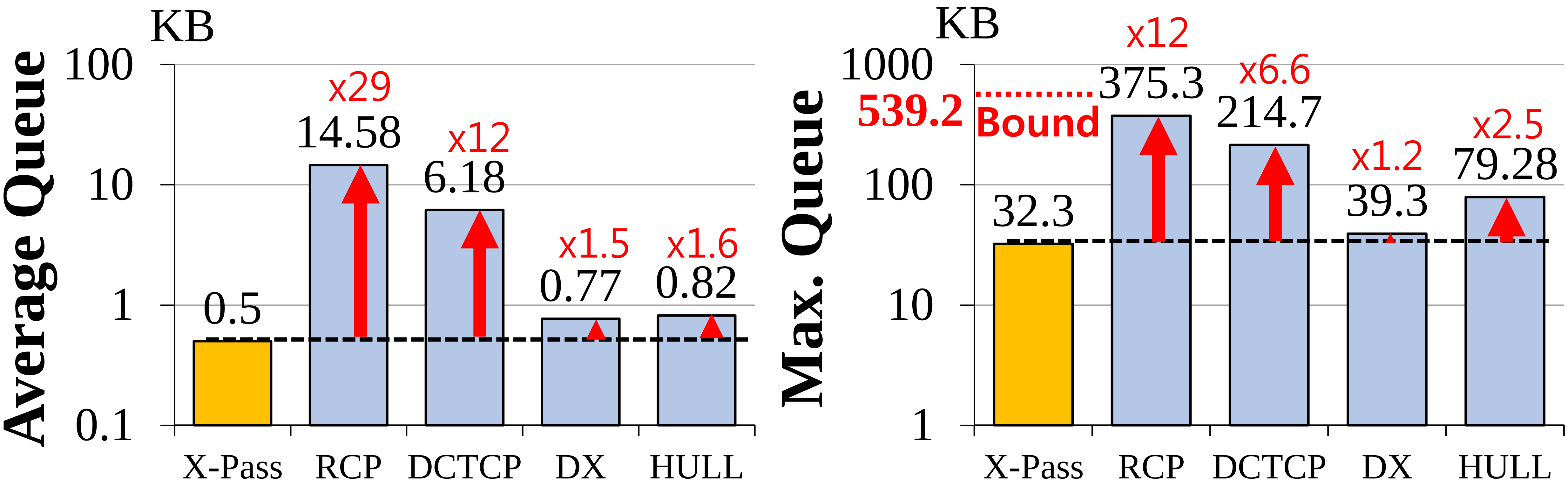
cache follower workload / load 0.6 / 0KB –



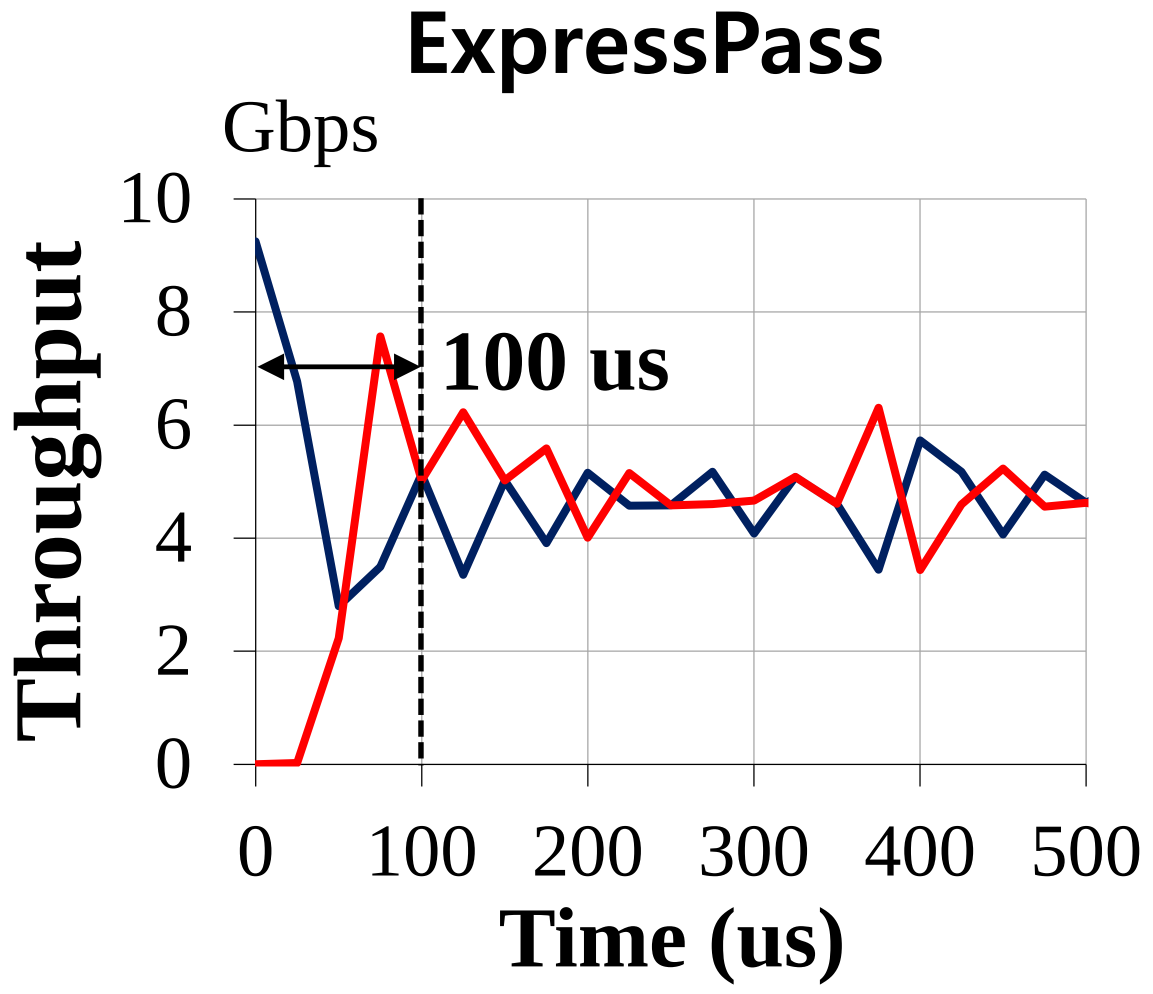


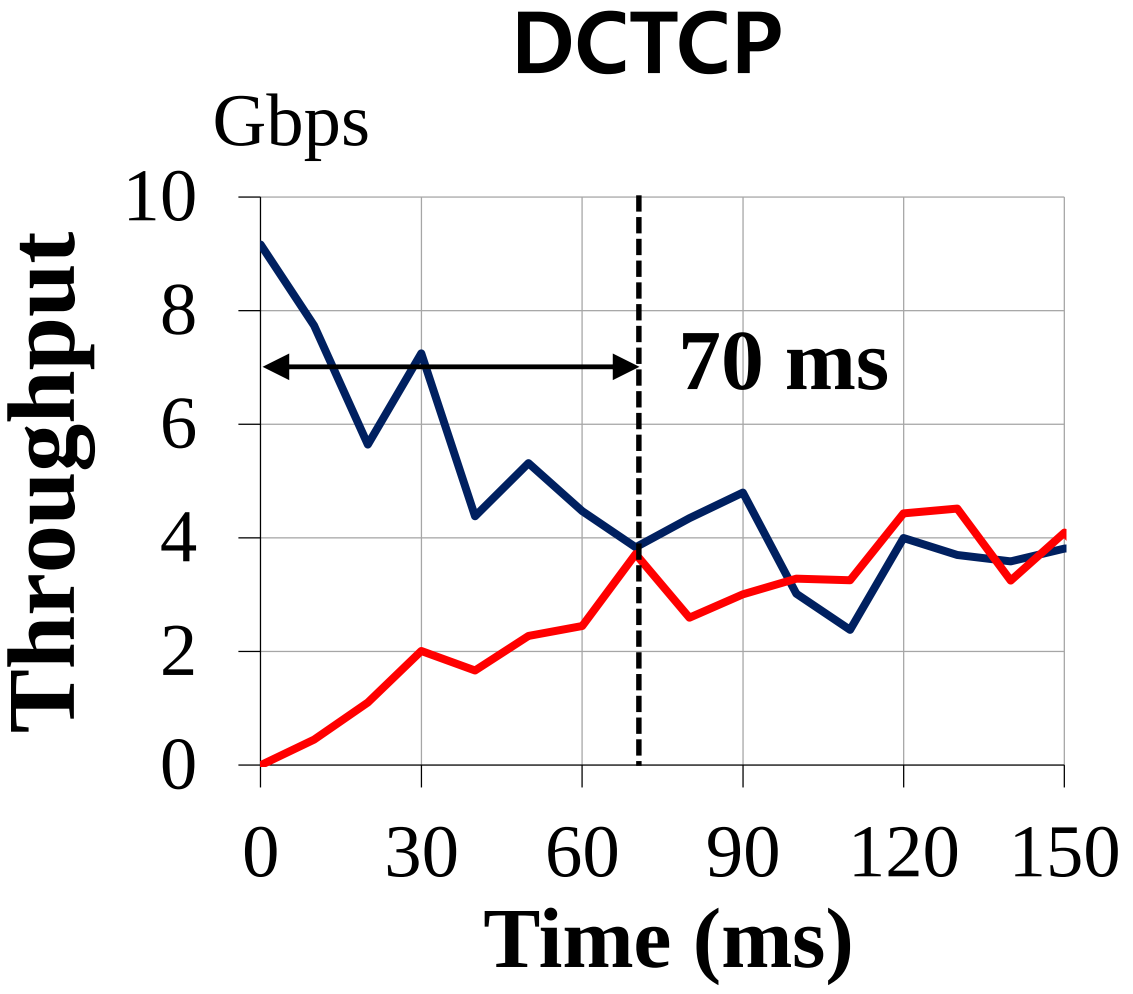
Low Average Queue

cache follower workload / load 0.6 / 0KB –

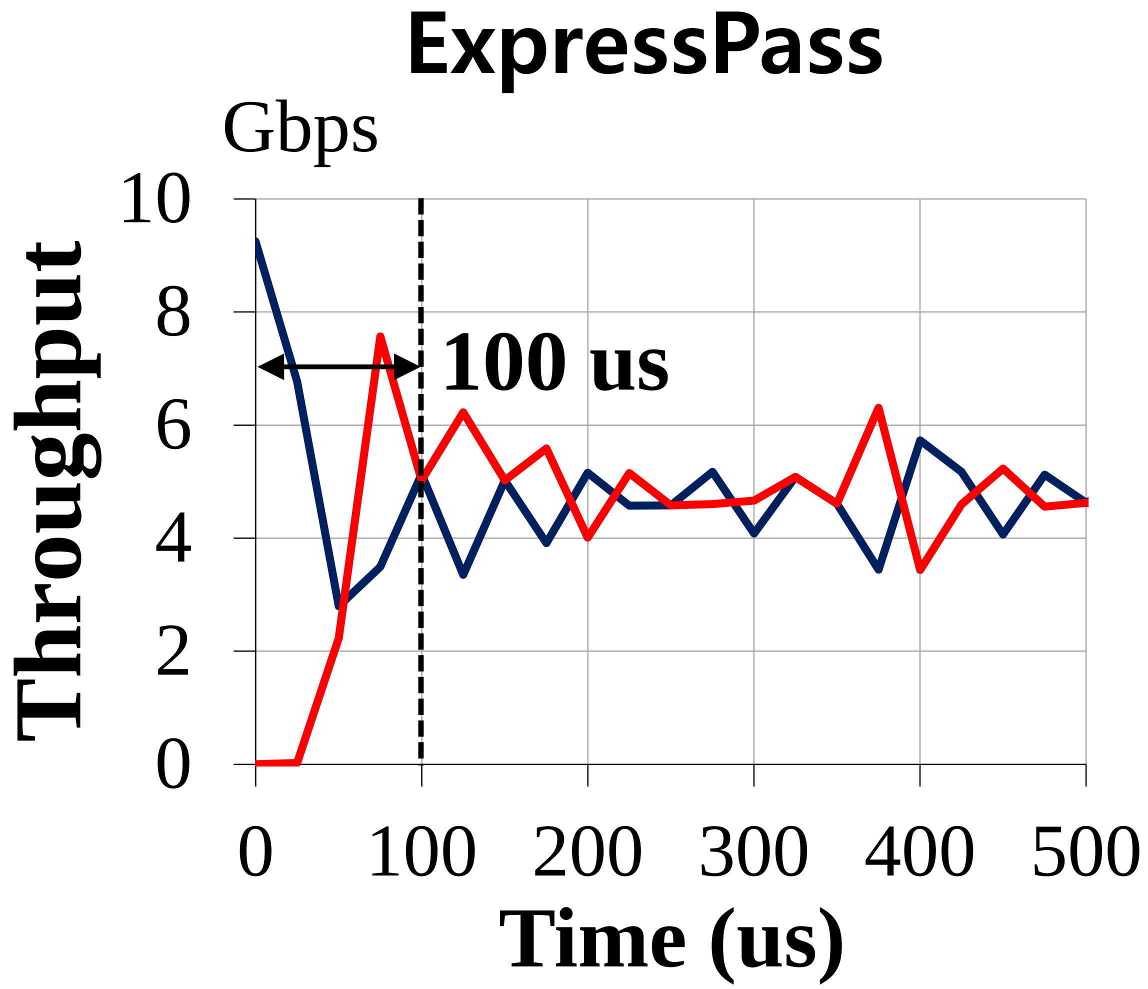


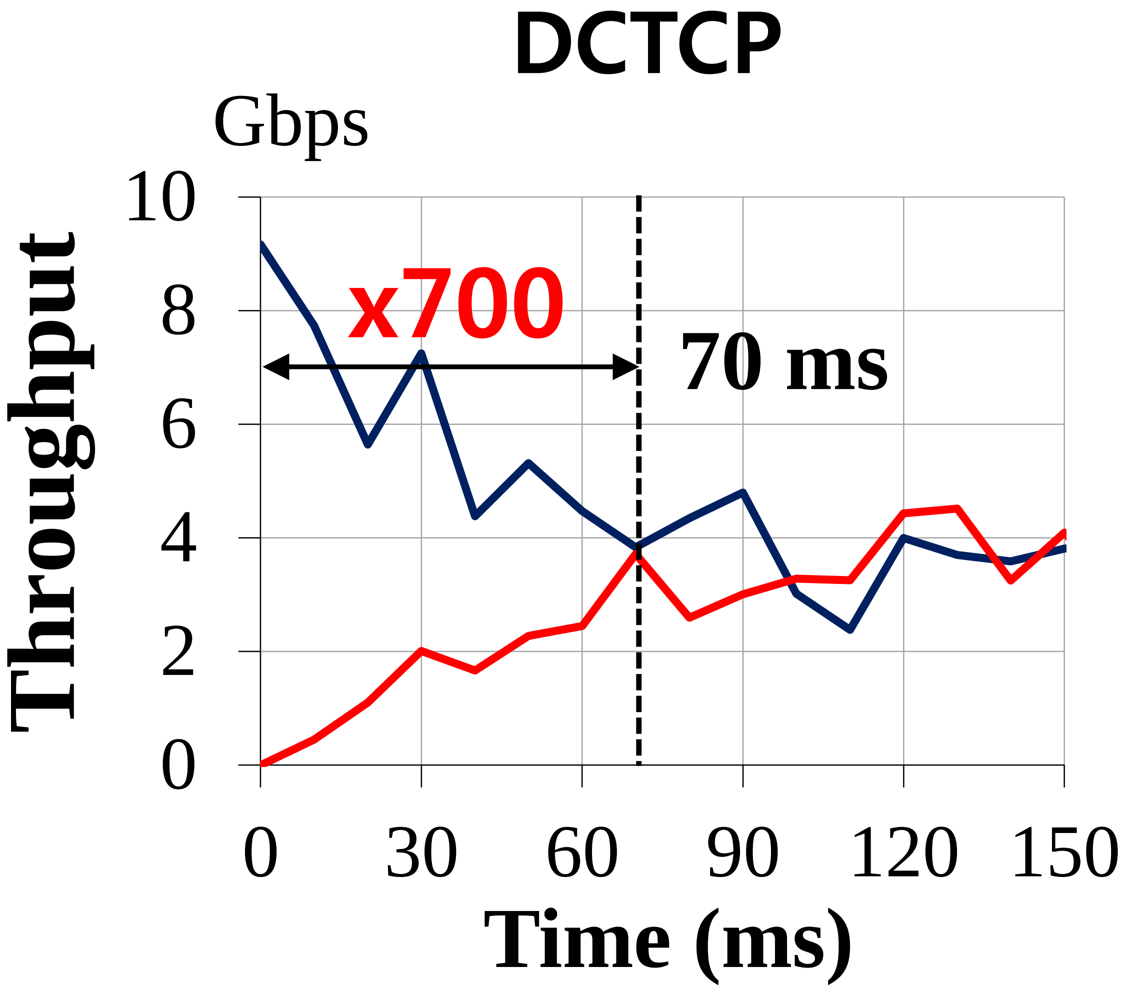
Fast & Stable Convergence





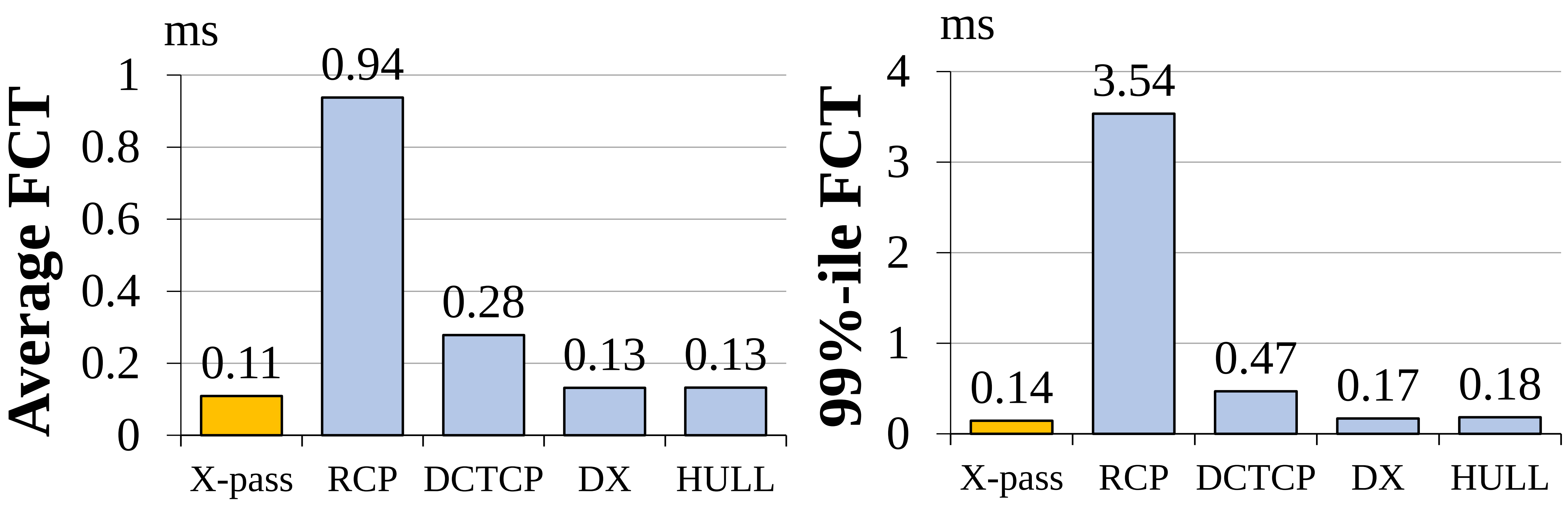
Fast & Stable Convergence





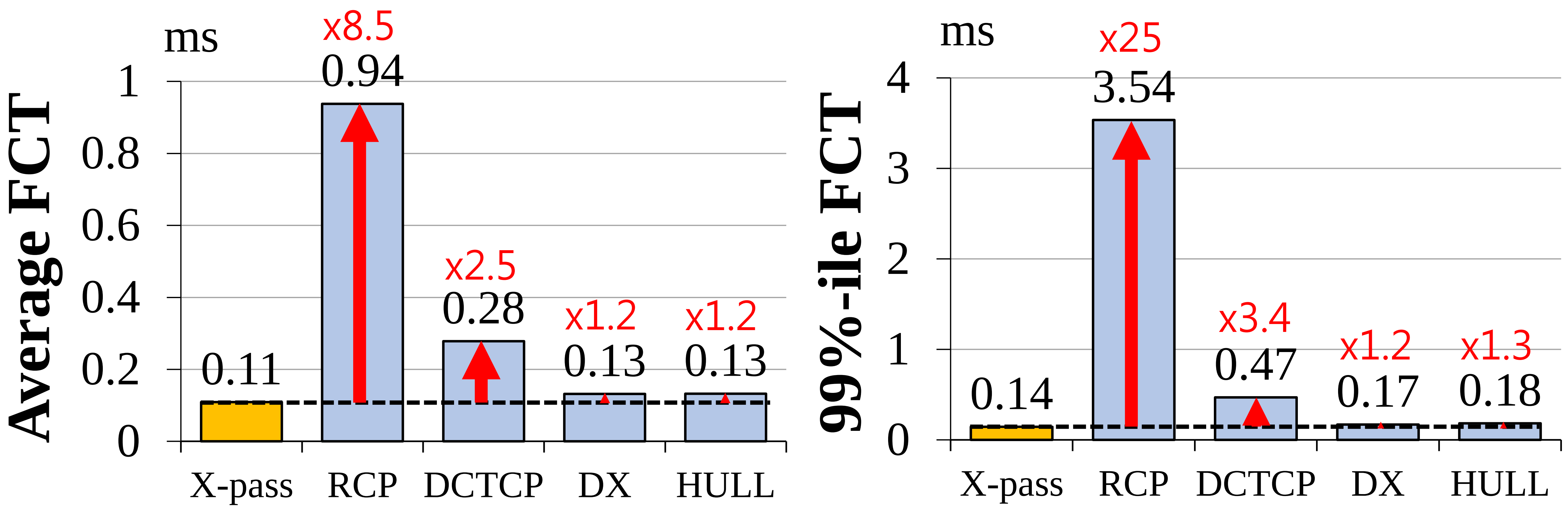
Flow Completion Time

cache follower workload / load 0.6 / 0 – 10KB



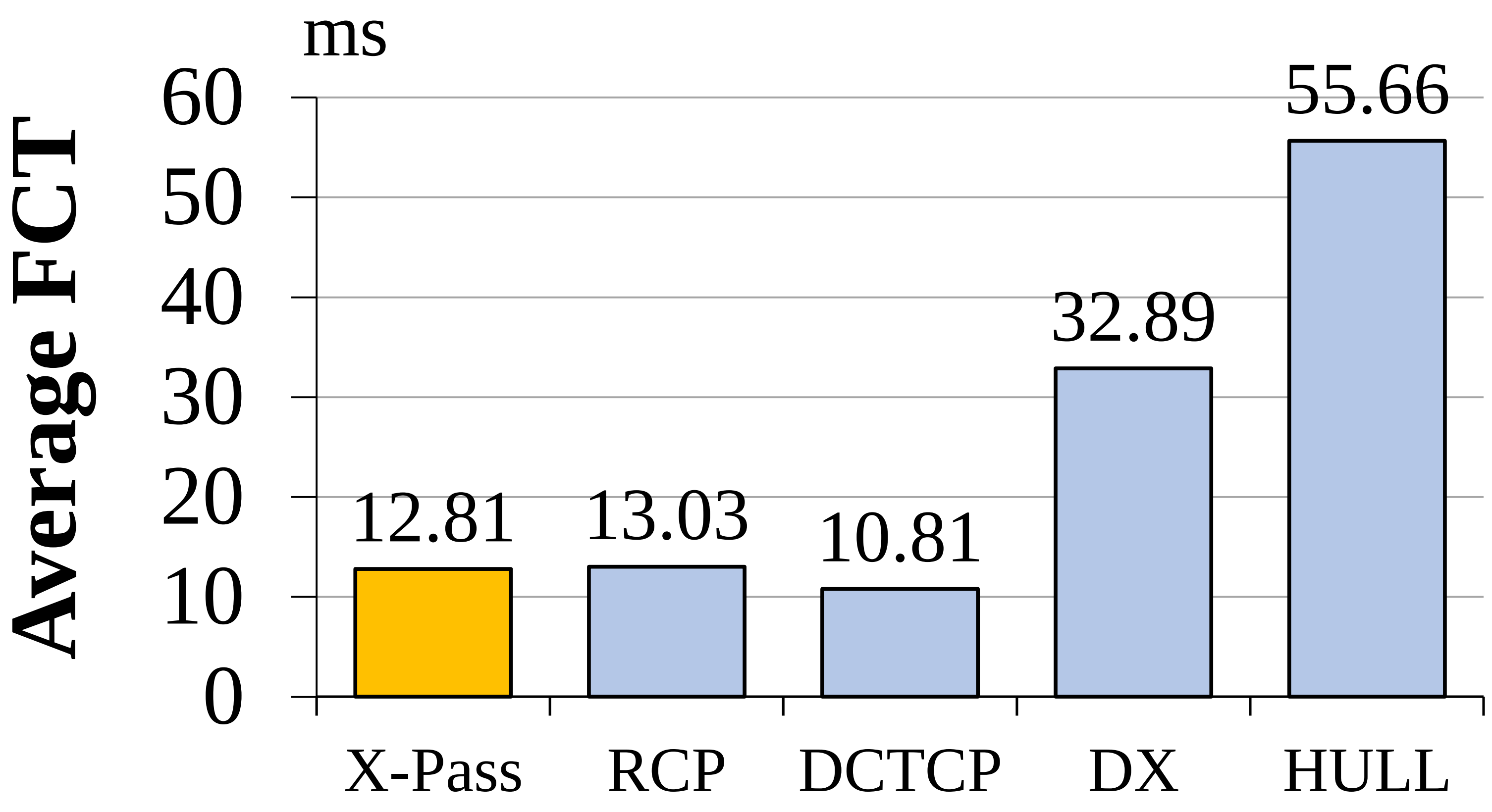
Flow Completion Time

cache follower workload / load 0.6 / 0 – 10KB



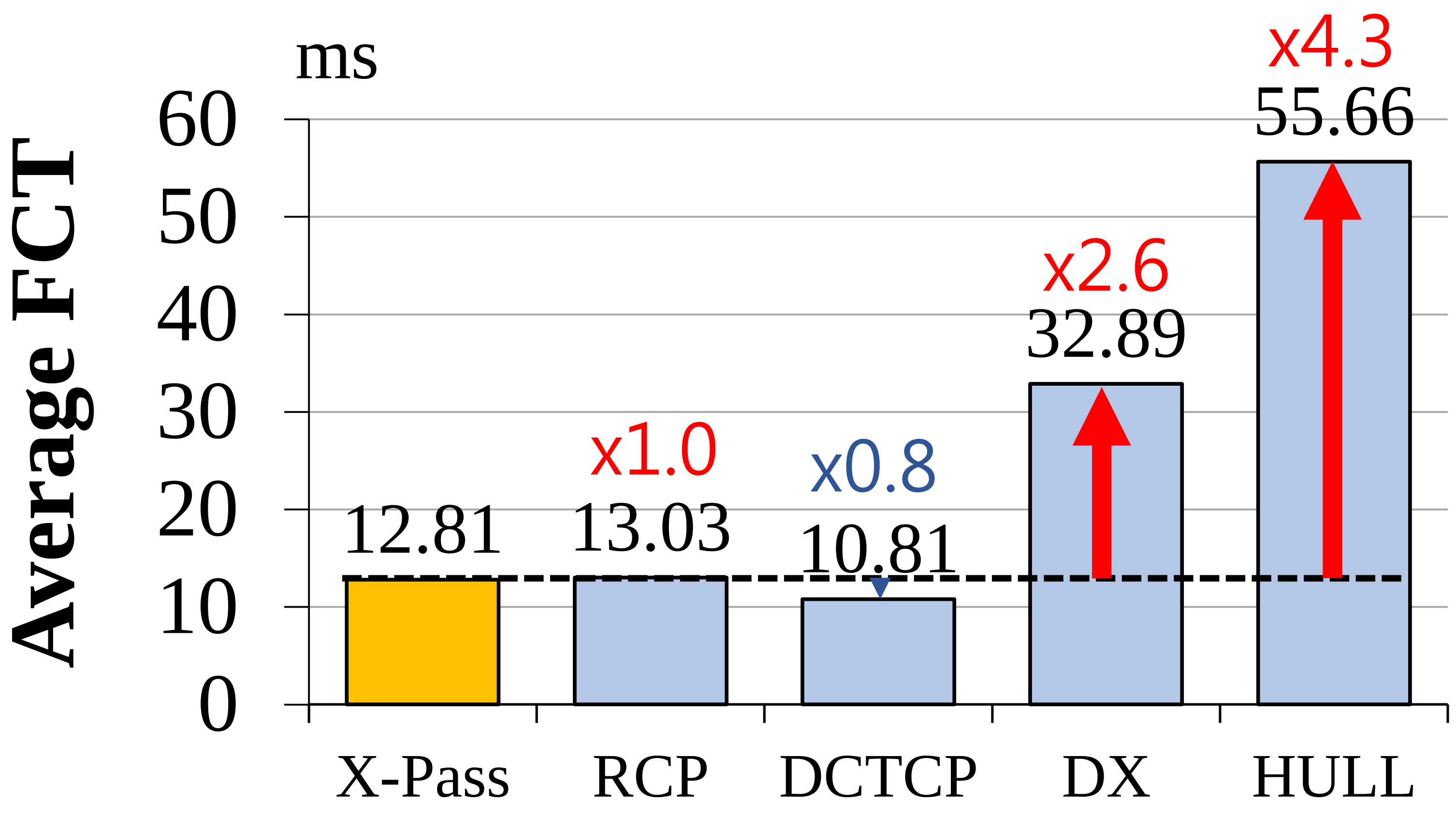
Flow Completion Time

cache follower workload / load 0.6 / 1MB –



Flow Completion Time

cache follower workload / load 0.6 / 1MB –



Conclusion

- ExpressPass is **end-to-end**, **credit-scheduled**, and **delay-bounded** congestion control for datacenter.
- ExpressPass propose a new **proactive** datacenter congestion control.
- Our evaluation on testbed and ns-2 simulation show that ExpressPass achieves
  (1) Low & bounded queueing
  (2) Fast & stable convergence
  (3) Short flow completion time especially for small flows

Thanks

Happy to answer your questions