ABM

Active Buffer Management in Datacenters

Vamsi Addanki, Maria Apostolaki, Manya Ghobadi, Stefan Schmid, Laurent Vanbever

Technische Universität Berlin

Princeton University

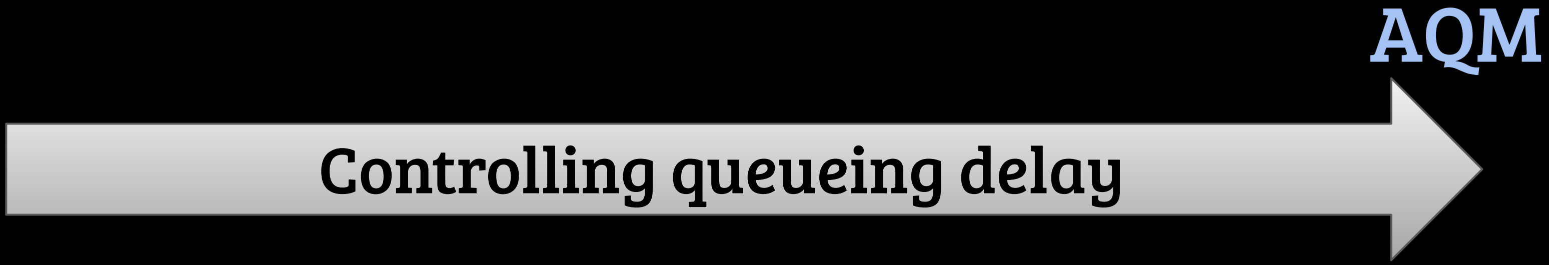
Massachusetts Institute of Technology

Technische Universität Berlin

ETH zürich

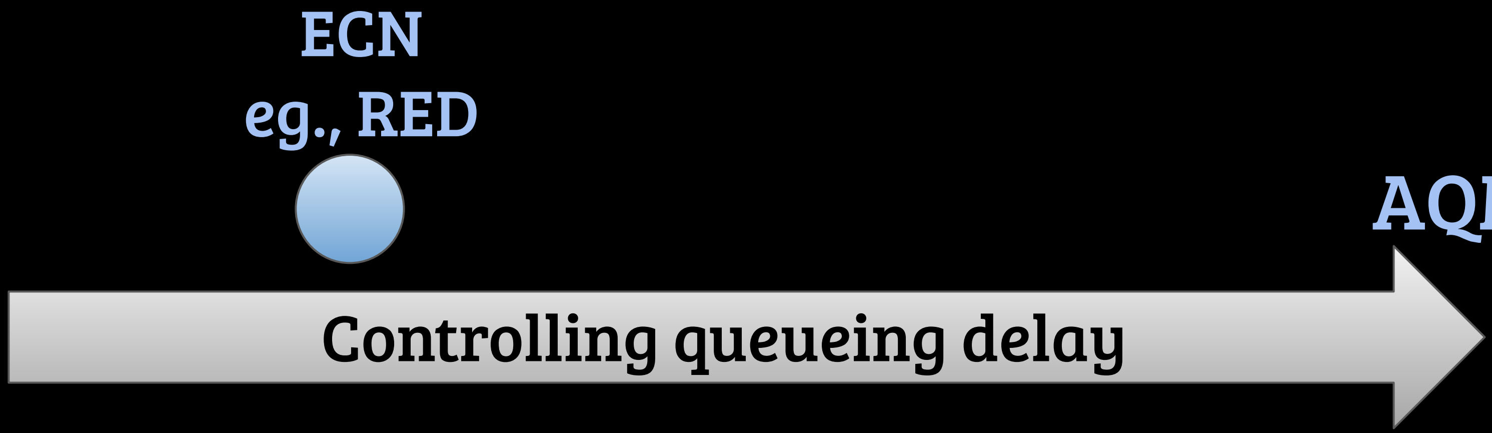
Active Queue Management (AQM)

Buffer Management (BM)



AQM

Controlling queueing delay

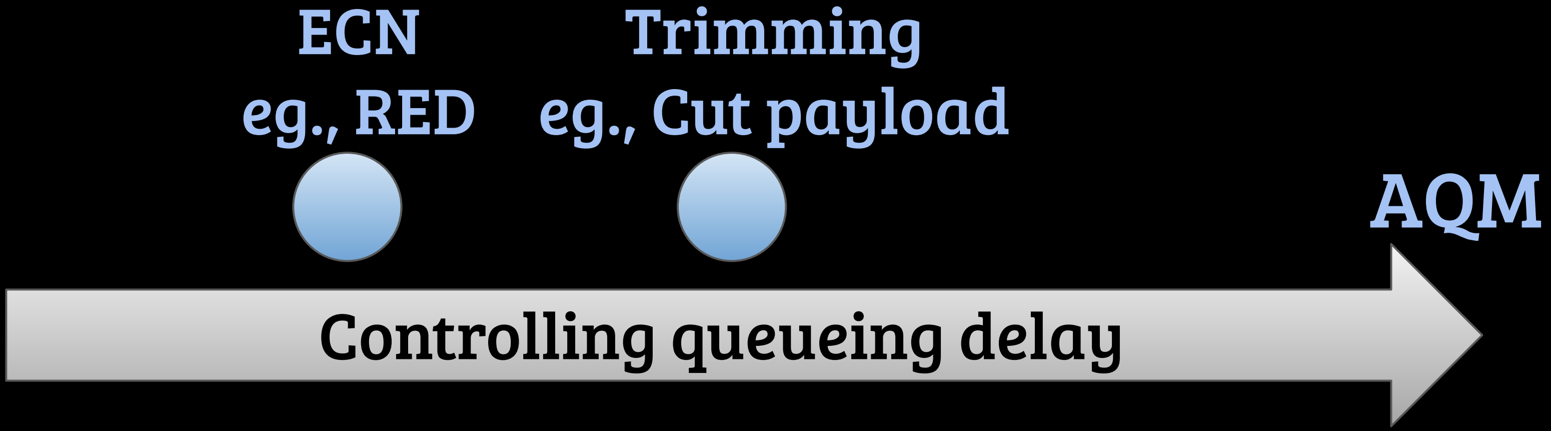


ECN

eg., RED

AQM

Controlling queueing delay

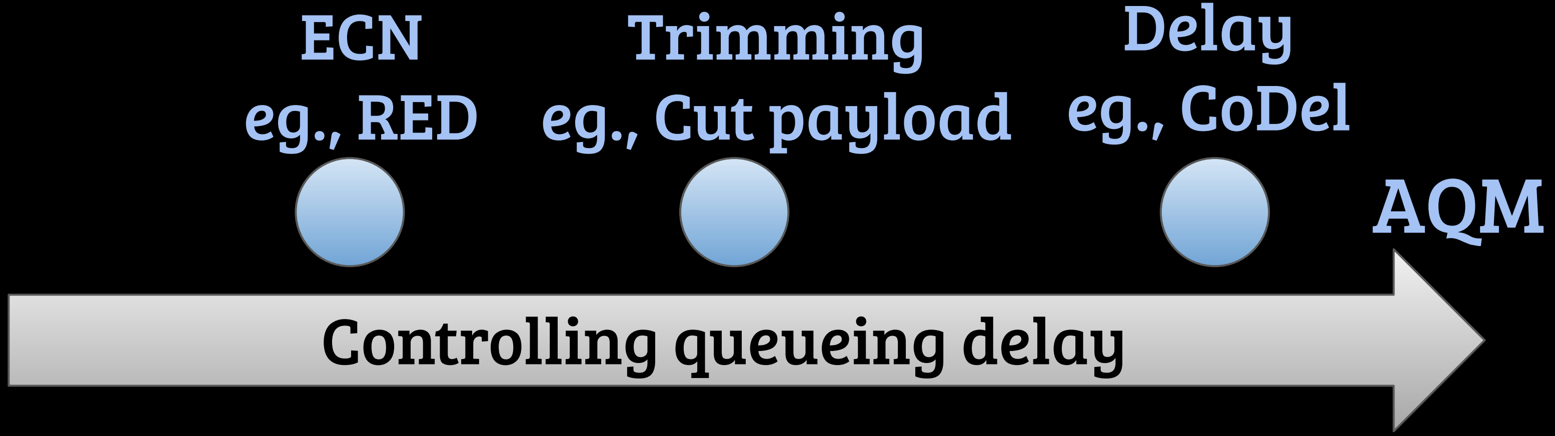


ECN
eg., RED

Trimming
eg., Cut payload

AQM

Controlling queueing delay



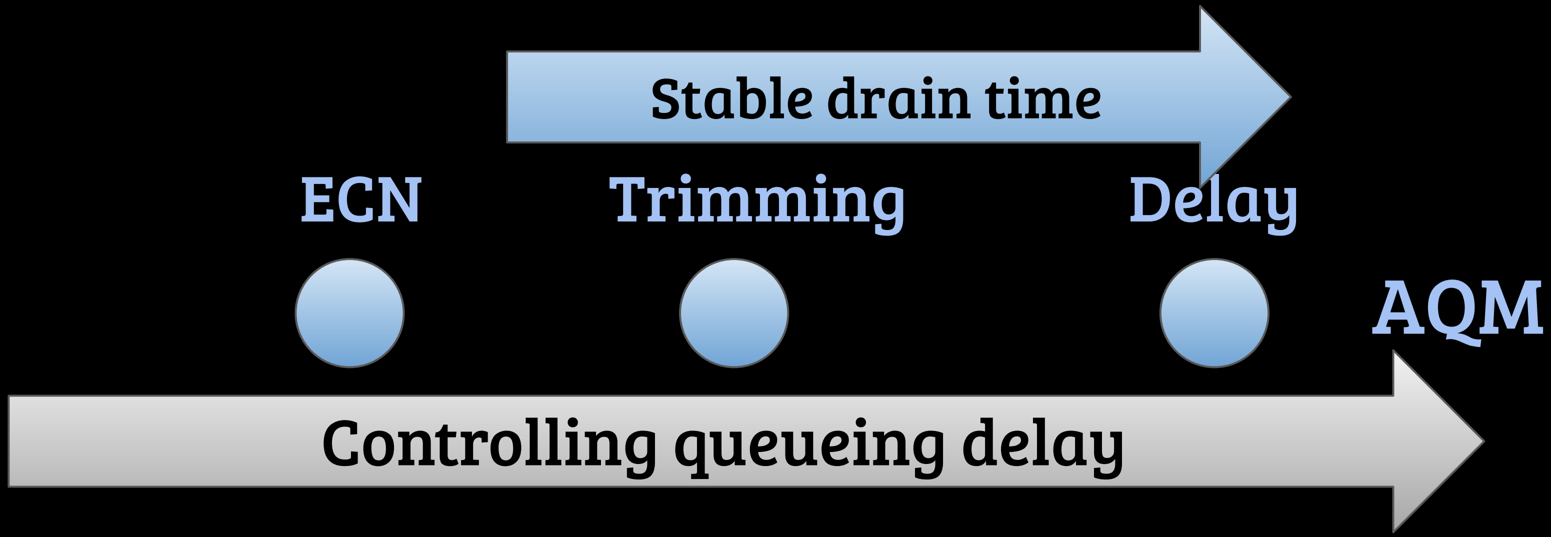
ECN
eg., RED

Trimming
eg., Cut payload

Delay
eg., CoDel

AQM

Controlling queueing delay



Stable drain time

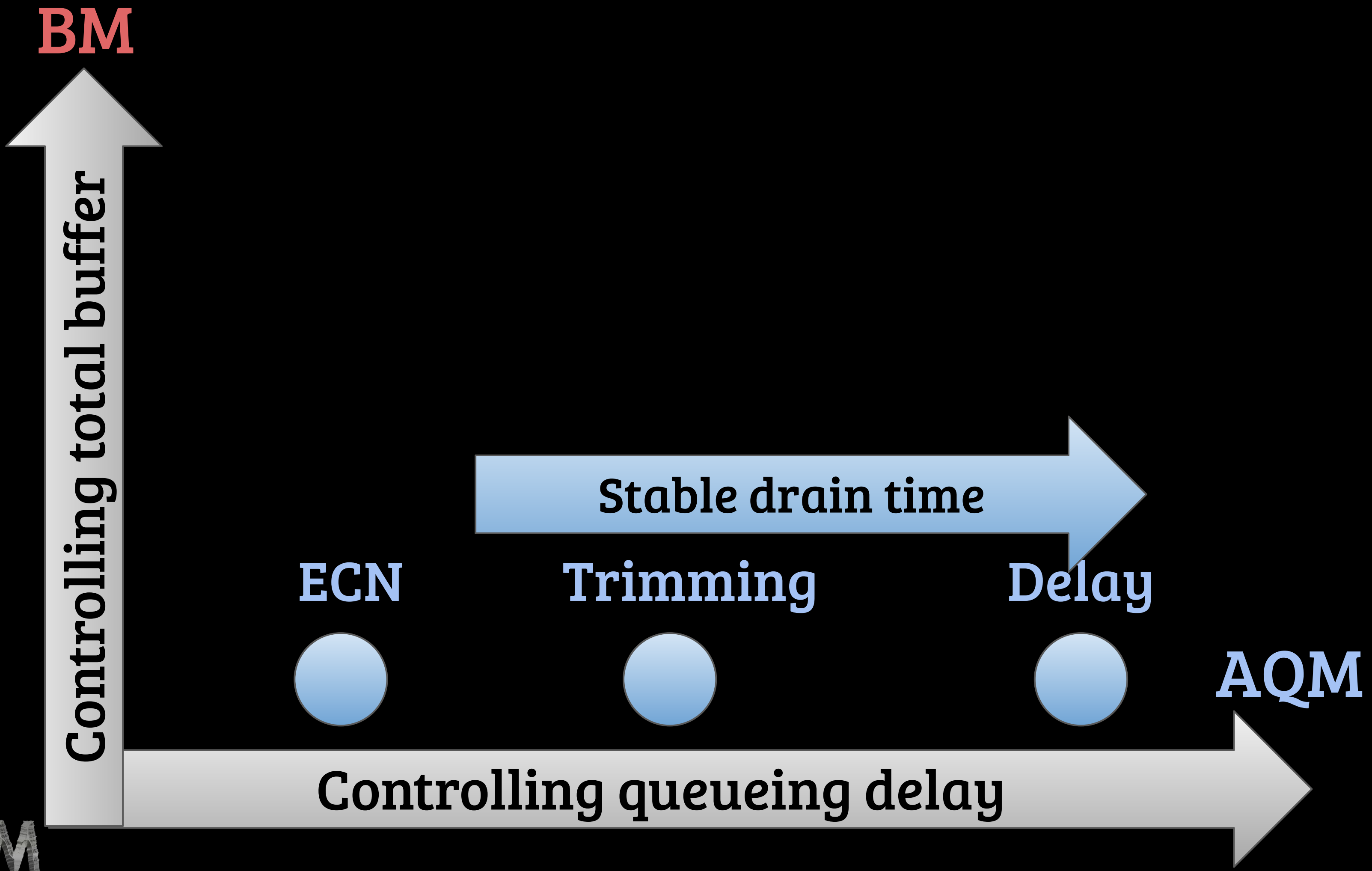
ECN

Trimming

Delay

AQM

Controlling queueing delay



BM

Controlling total buffer

Stable drain time

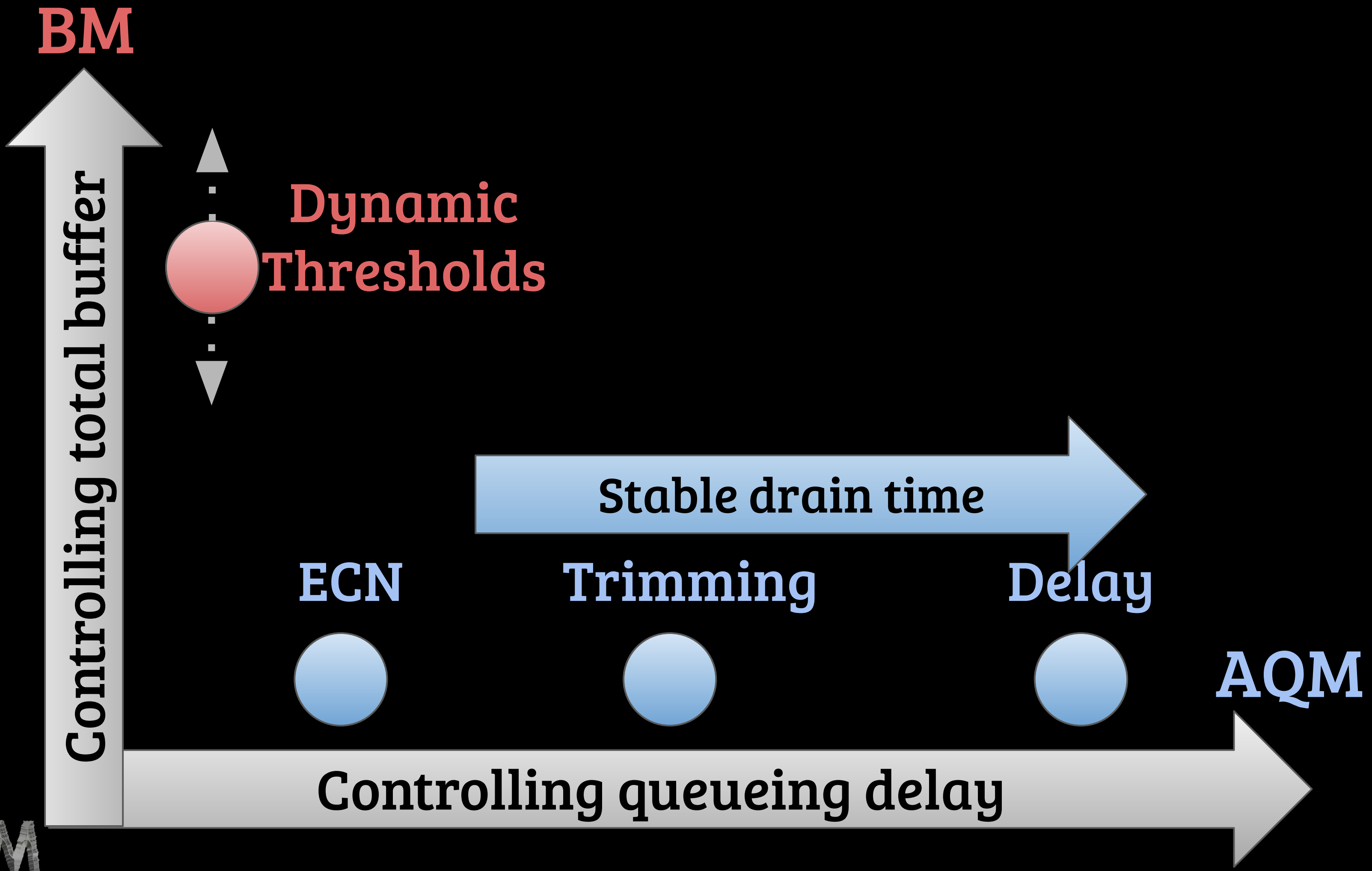
ECN

Trimming

Delay

AQM

Controlling queueing delay



BM

Dynamic Thresholds

Stable drain time

ECN

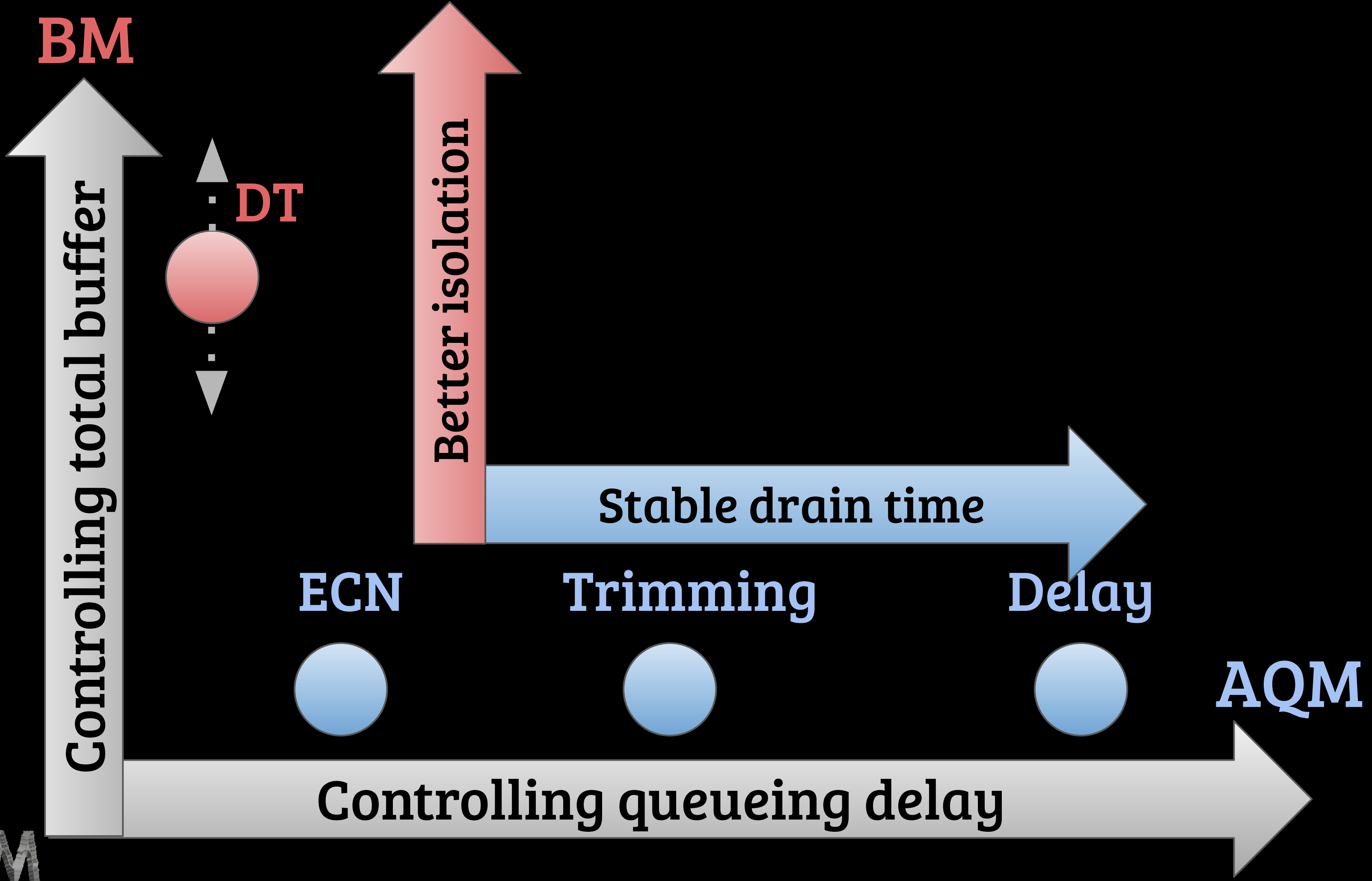
Trimming

Delay

AQM

Controlling total buffer

Controlling queueing delay



BM

DT

Controlling total buffer

Better isolation

Stable drain time

ECN

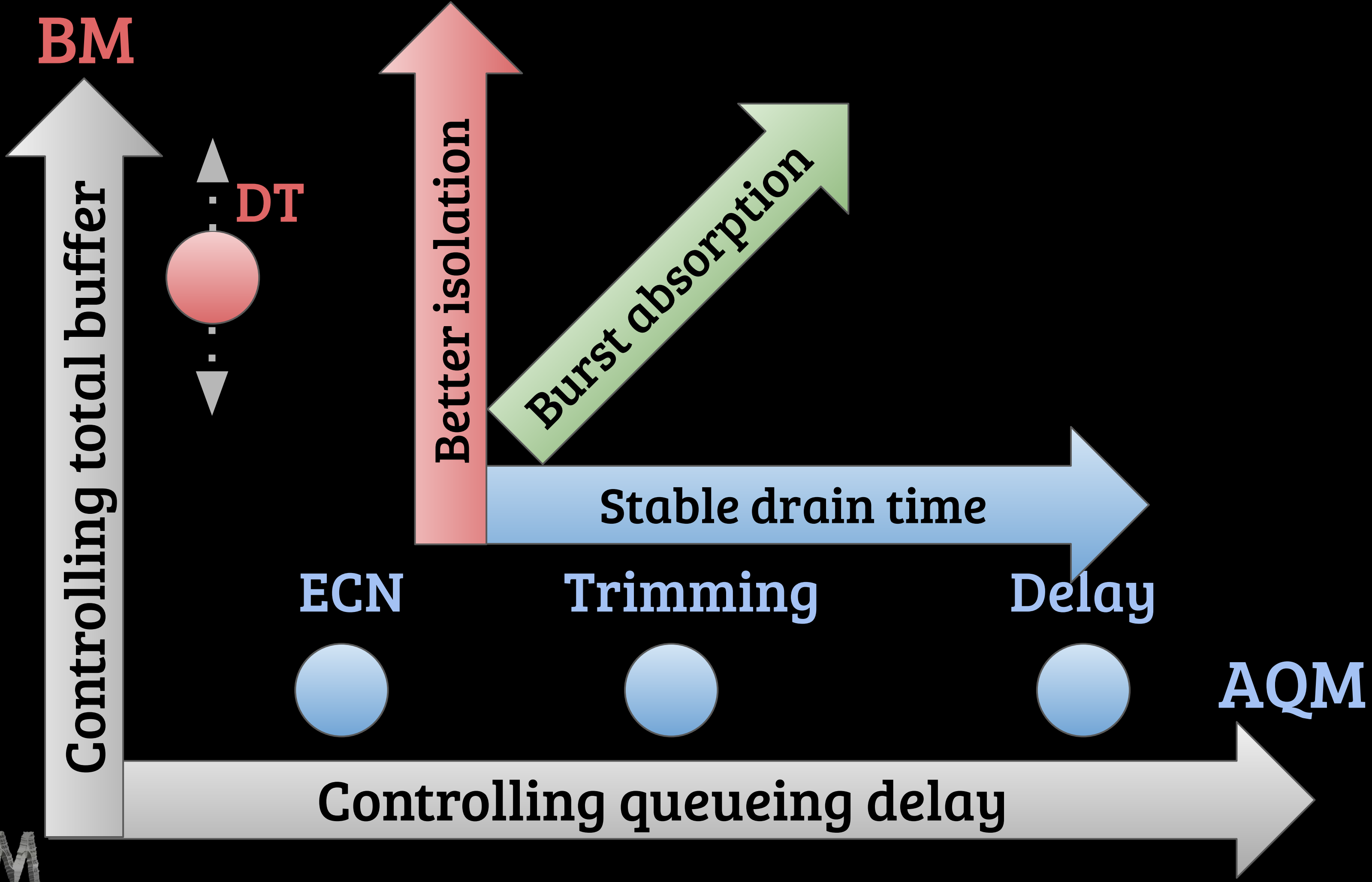
Trimming

Delay

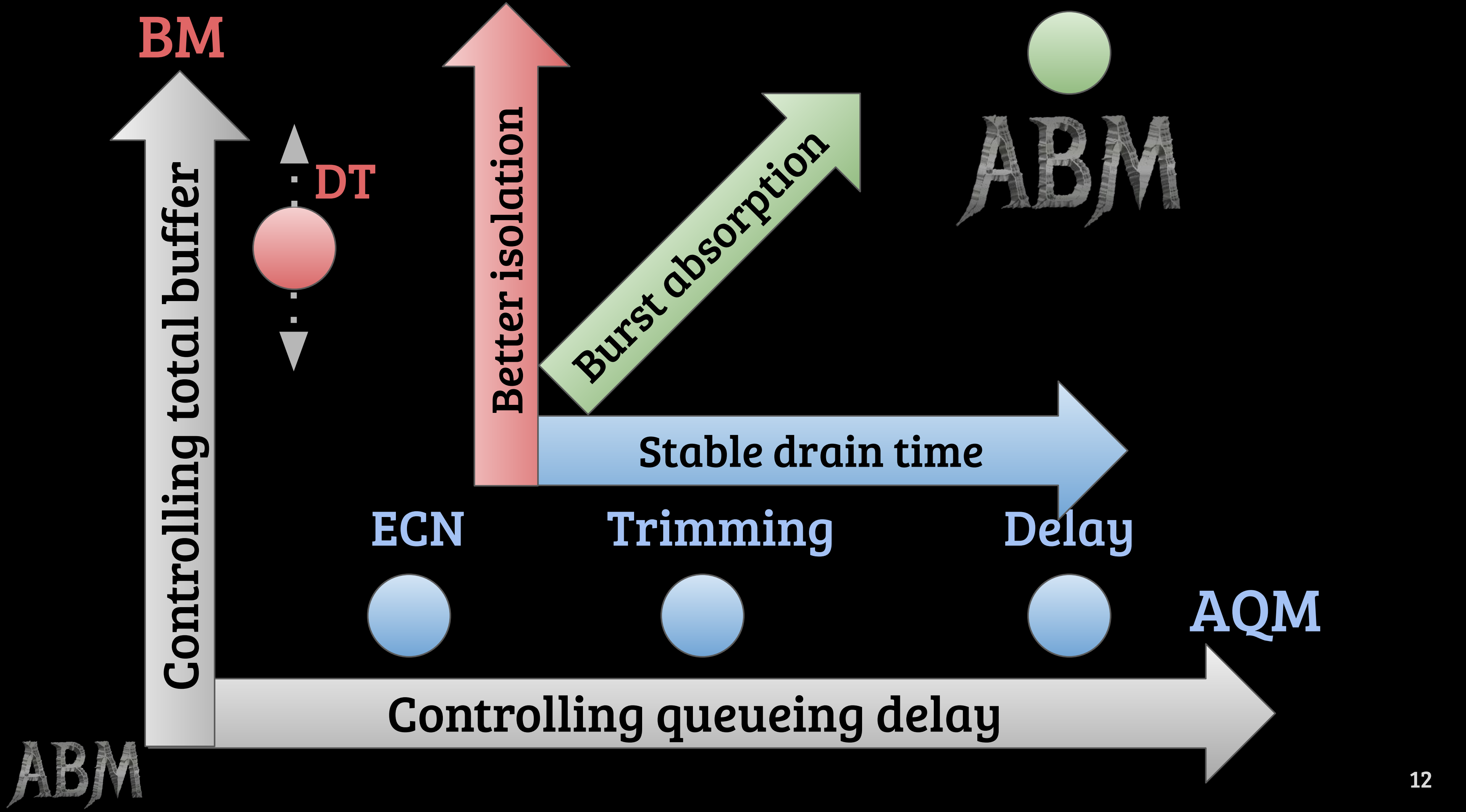
AQM

Controlling queueing delay

ABM


BM
DT
Controlling total buffer
Better isolation
Burst absorption
Stable drain time
ECN
Trimming
Delay
AQM
Controlling queueing delay

ABM


BM
DT
Controlling total buffer
Better isolation
Burst absorption
ABM
Stable drain time
ECN
Trimming
Delay
AQM
Controlling queueing delay
ABM

What is ABM?

- A novel Buffer Sharing algorithm *for datacenter switches*

What is ABM?

- A novel Buffer Sharing algorithm
- ~~Independent AQM and Buffer Management~~

What is ABM?

- A novel Buffer Sharing algorithm
- ~~Independent AQM and Buffer Management~~
- AQM that depends on Buffer Management

What is ABM?

- A novel Buffer Sharing algorithm
- ~~Independent AQM and Buffer Management~~
- Active Buffer Management

What is ABM?

- A novel Buffer Sharing algorithm
- ~~Independent AQM and Buffer Management~~
- Active Buffer Management
    - *Isolation across traffic priorities (eg., different SLAs)*

What is ABM?

- A novel Buffer Sharing algorithm
- ~~Independent AQM and Buffer Management~~
- **A**ctive **B**uffer **M**anagement
  - *Isolation across traffic priorities (eg., different SLAs)*
  - *Bounded queue drain time (Queueing delay)*

What is ABM?

- A novel Buffer Sharing algorithm
- ~~Independent AQM and Buffer Management~~
- Active Buffer Management
  - *Isolation across traffic priorities (eg., different SLAs)*
  - *Bounded queue drain time (Queueing delay)*
  - *Better burst absorption*

Background on Buffer Sharing

- Both BM and AQM calculate *drop thresholds*

Background on Buffer Sharing

- Both BM and AQM calculate *drop thresholds*
- BM calculates a threshold for every queue in a *device*
    - function of the shared buffer space

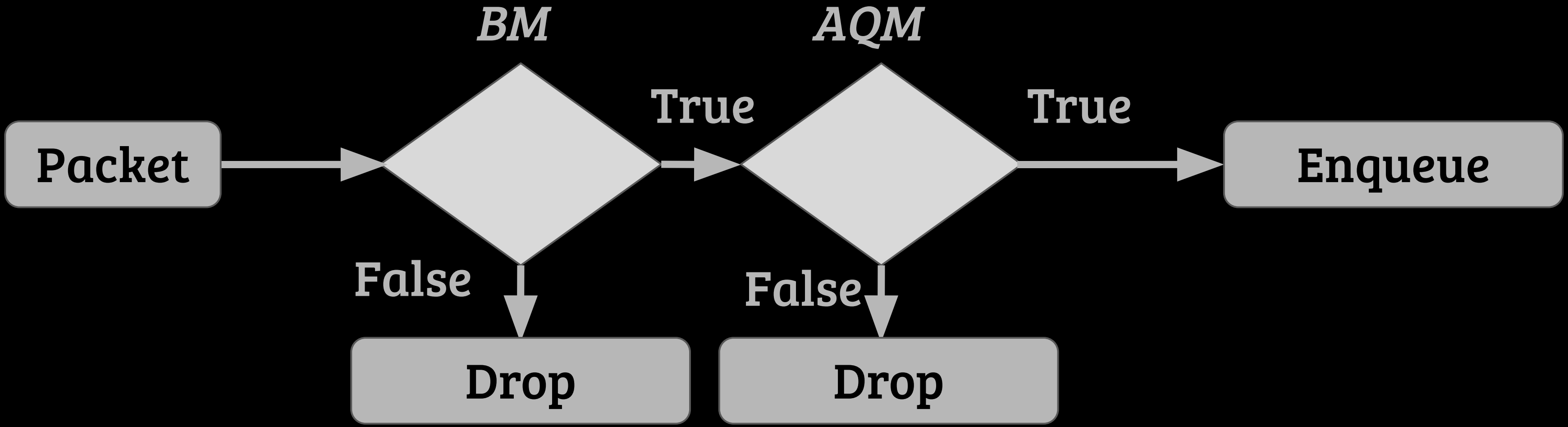
Background on Buffer Sharing

- Both BM and AQM calculate *drop thresholds*
- BM calculates a threshold for every queue in a *device*
    - function of the shared buffer space
- AQM calculates thresholds for a single queue
    - function of queue statistics

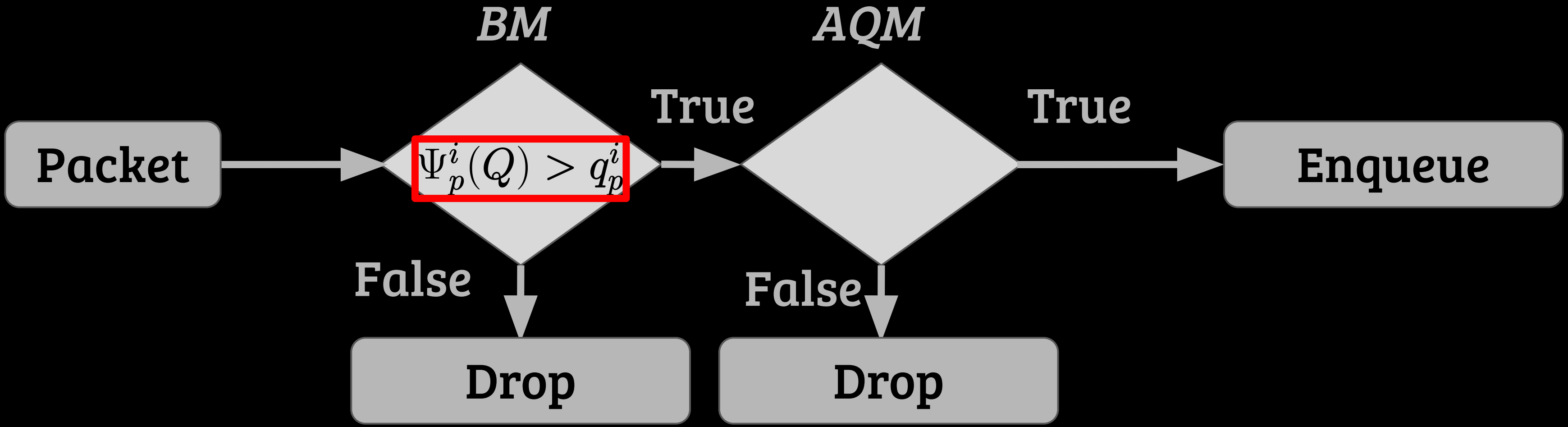
Background on Buffer Sharing

- Both BM and AQM calculate *drop thresholds*
- BM calculates a threshold for every queue in a *device*
    - function of the shared buffer space
- AQM calculates thresholds for a single queue
    - function of queue statistics
- BM and AQM act independently

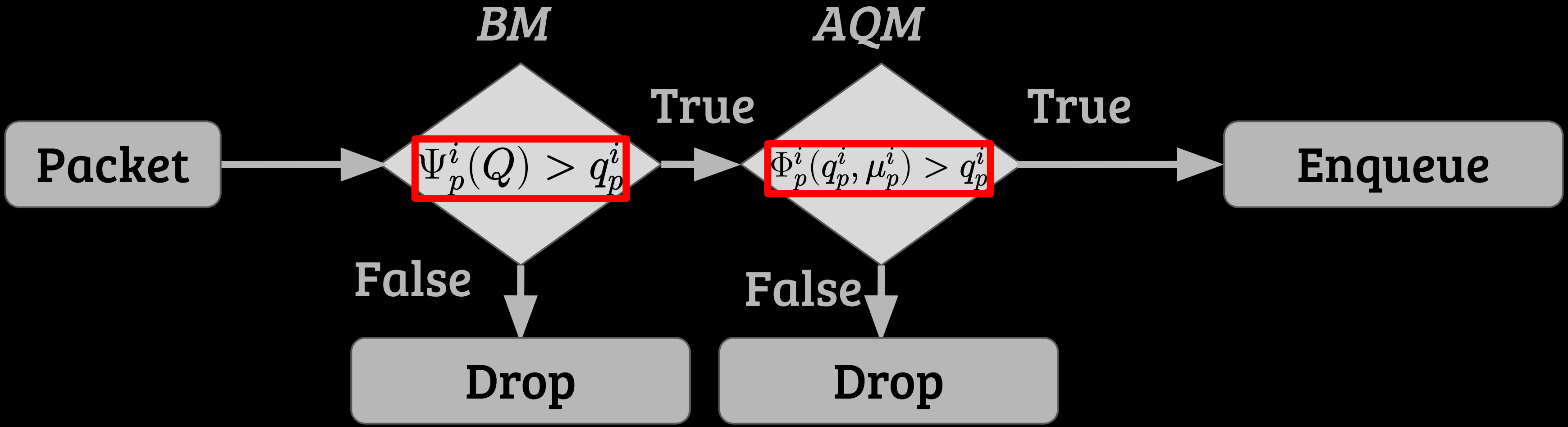
Hierarchical Admission Control Scheme



Hierarchical Admission Control Scheme



Hierarchical Admission Control Scheme



BM

AQM

Packet

$\Psi_p^i(Q) > q_p^i$

True

$\Phi_p^i(q_p^i, \mu_p^i) > q_p^i$

True

Enqueue

False

Drop

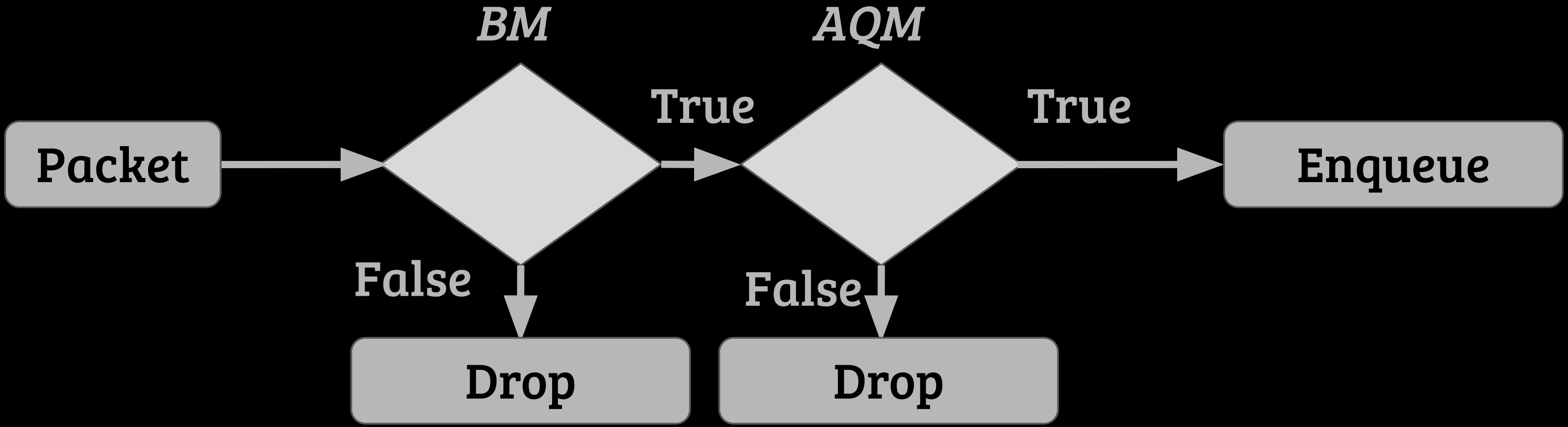
False

Drop

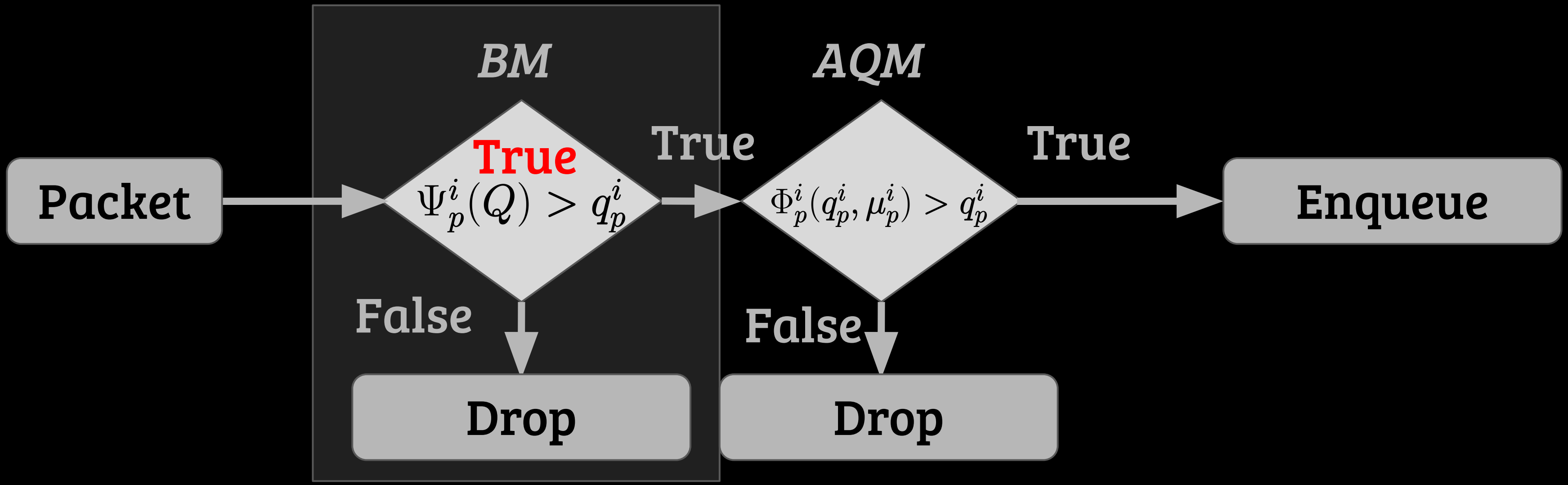
ABM

Hierarchical Admission Control Scheme

$$\boxed{min \left(\underbrace{\Psi_p^i(Q)}_{BM}, \underbrace{\Phi_p^i(q_p^i, \mu_p^i)}_{AQM} \right)}$$



Large Buffers



Packet

BM

True

$\Psi_p^i(Q) > q_p^i$

True

False

Drop

AQM

$\Phi_p^i(q_p^i, \mu_p^i) > q_p^i$

True

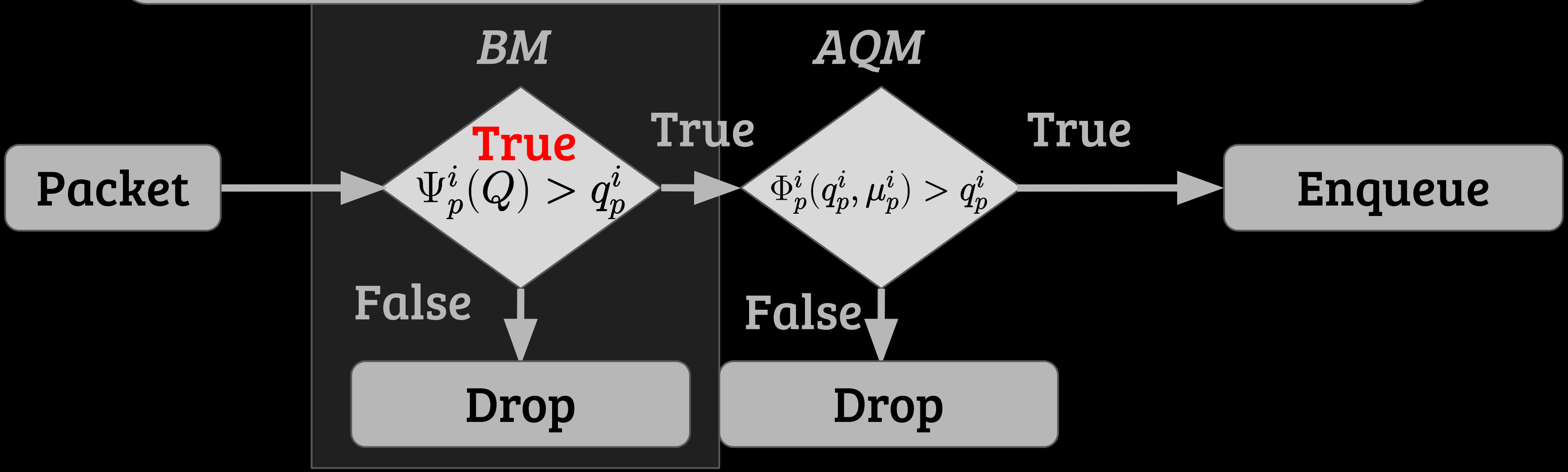
False

Drop

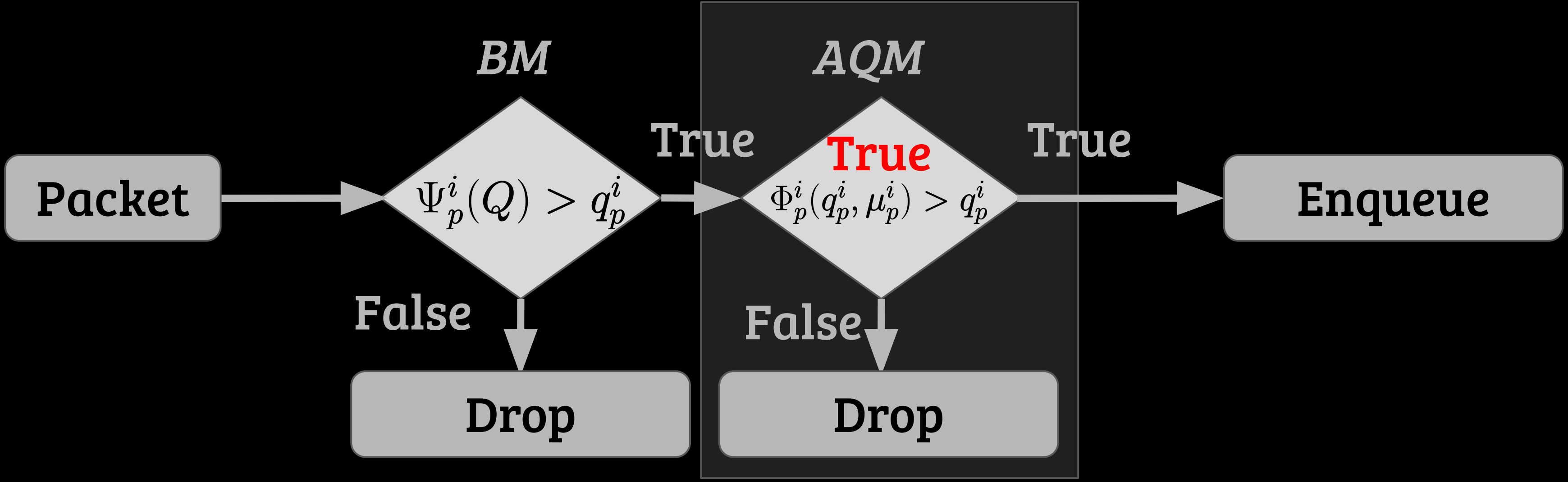
Enqueue

Large Buffers

AQM becomes more important!



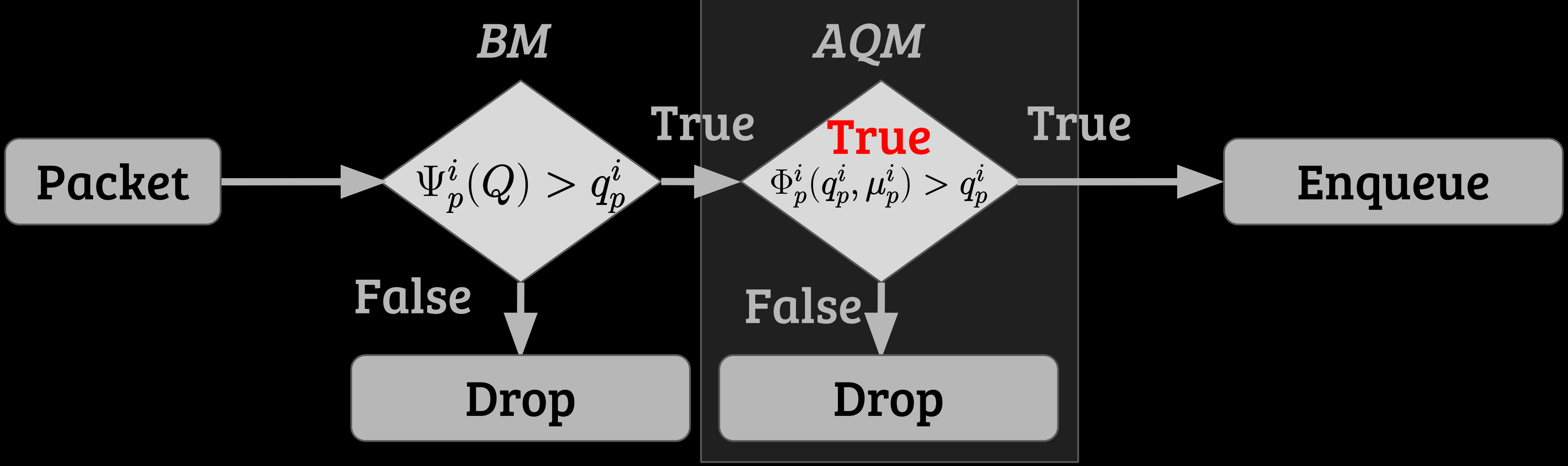
Shallow buffers



ABM

Shallow buffers

Buffer Management becomes more important!



Drawbacks of Dynamic Thresholds (State-of-the-art BM)

Threshold = alpha x (Remaining shared buffer)

$$T_p^i(t) = \alpha_p \cdot \underbrace{(B - Q(t))}_{Remaining}$$

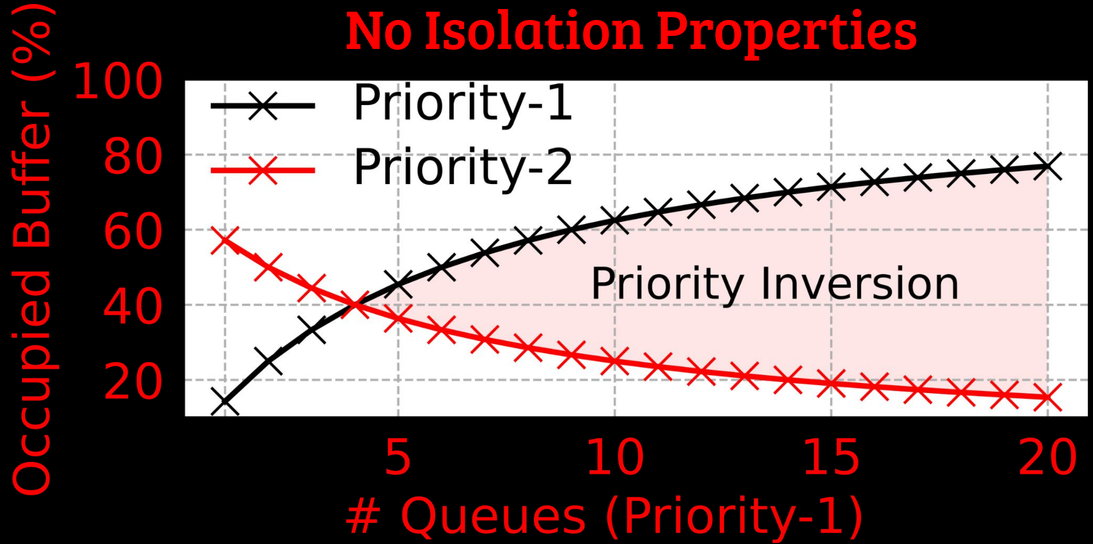
Drawbacks of Dynamic Thresholds (State-of-the-art BM)

Threshold = alpha x (Remaining shared buffer)

$$T_p^i(t) = \alpha_p \cdot \underbrace{(B - Q(t))}_{Remaining}$$

- Priority inversion (No isolation)
- Oblivious to buffer drain time

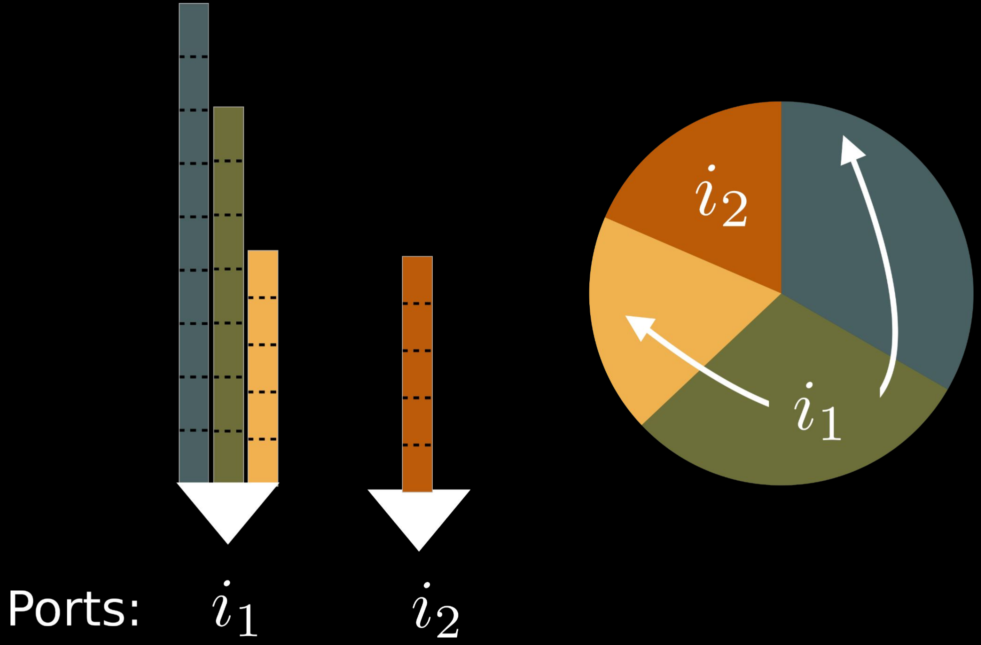
Drawbacks of Dynamic Thresholds (State-of-the-art BM)



No Isolation Properties

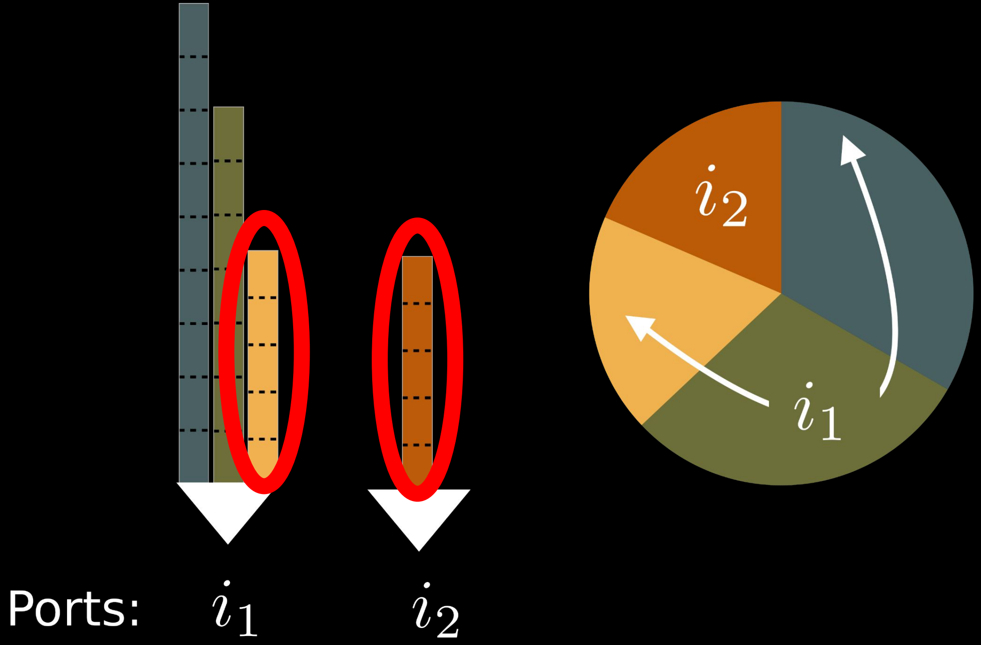
Drawbacks of Dynamic Thresholds (State-of-the-art BM)

Oblivious to drain rate



Drawbacks of Dynamic Thresholds (State-of-the-art BM)

Oblivious to drain rate

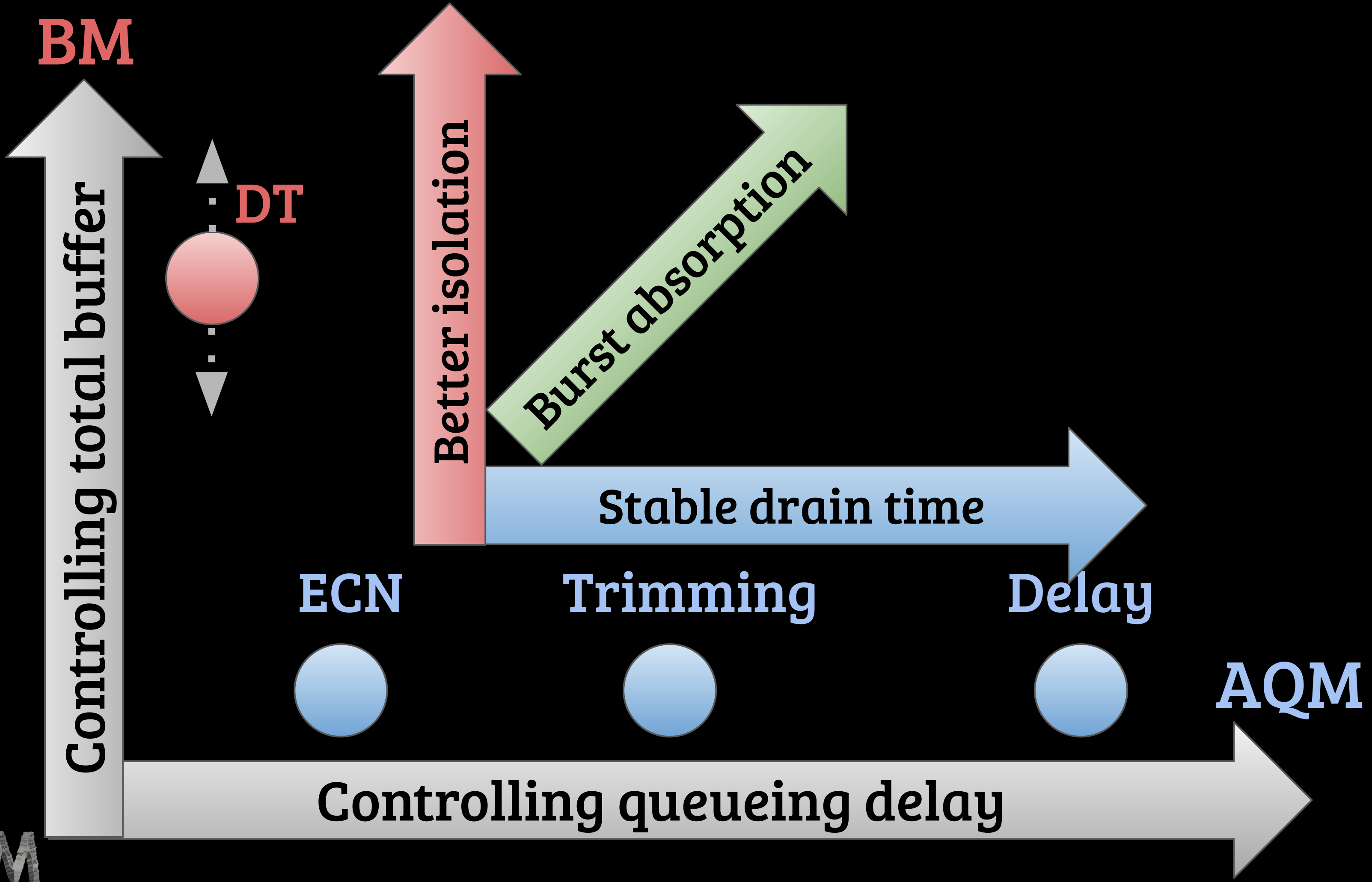


Benefits and Drawbacks of Existing Approaches

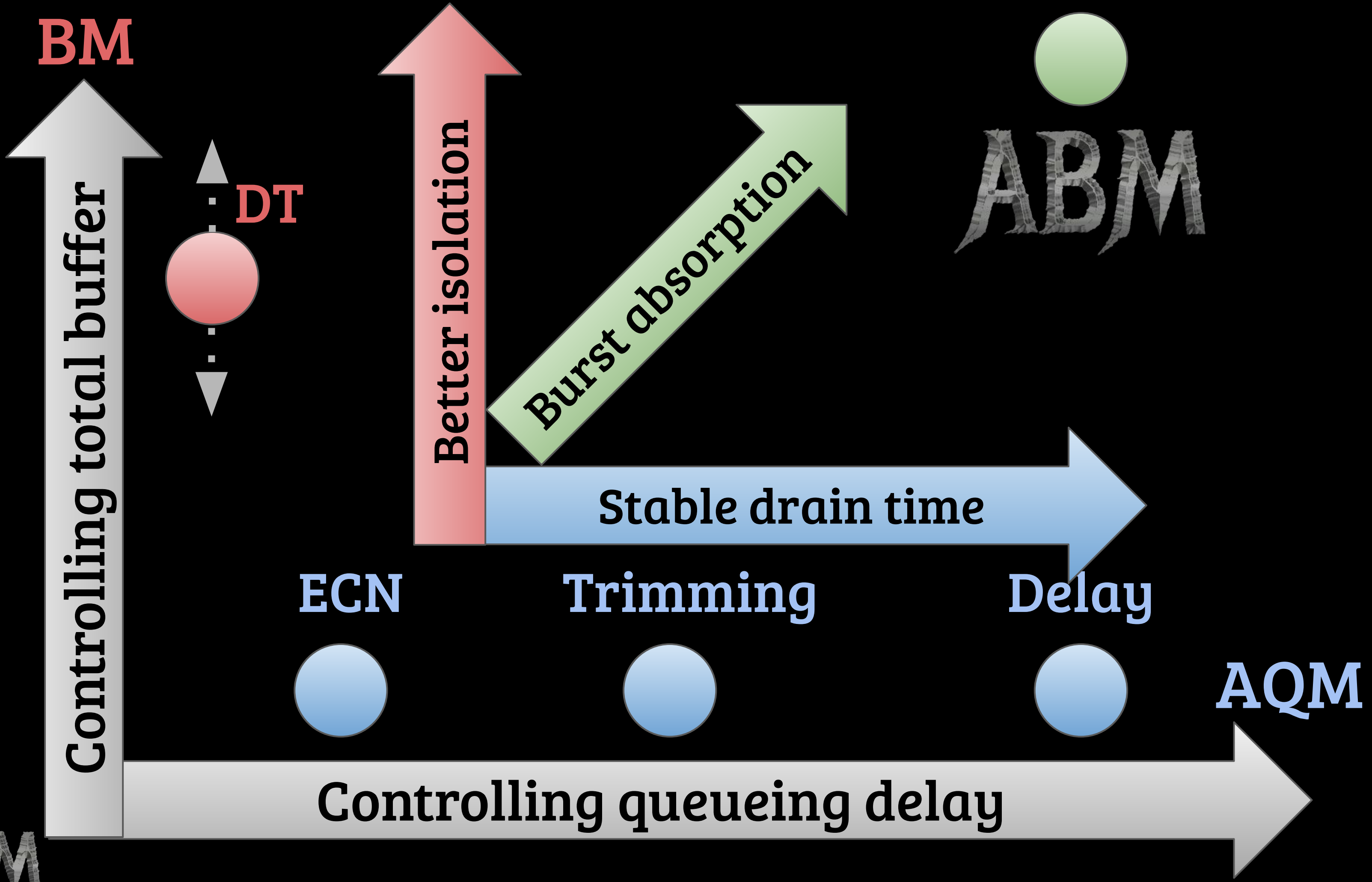
- BM can in-principle offer isolation across queues
  - oblivious to buffer drain time
- AQM can in-principle offer bounded queue drain time
  - cannot fundamentally satisfy the isolation property

Our Goals

- Isolation across traffic priorities
- Bounded drain time
- Better burst absorption
    - requires <u>both</u> isolation and bounded drain time


BM
DT
Controlling total buffer
Better isolation
Burst absorption
Stable drain time
ECN
Trimming
Delay
AQM
Controlling queueing delay

ABM



BM

DT

Controlling total buffer

Better isolation

Burst absorption

ABM

Stable drain time

ECN

Trimming

Delay

AQM

Controlling queueing delay

ABM

Active Buffer Management

$$\boxed{T_p^i(t)} = \alpha_p \cdot \frac{1}{n_p} \cdot (B - Q(t)) \cdot \frac{\mu_p^i}{b}$$

Threshold per queue
port i, priority p

ABM

Active Buffer Management

$$T_p^i(t) = \boxed{\alpha_p} \cdot \frac{1}{n_p} \cdot (B - Q(t)) \cdot \frac{\mu_p^i}{b}$$

Parameter

To be set for each priority

ABM

Active Buffer Management

$$T_p^i(t) = \alpha_p \cdot \frac{1}{\boxed{n_p}} \cdot (B - Q(t)) \cdot \frac{\mu_p^i}{b}$$

# congested queues of priority p

ABM

Active Buffer Management

$$T_p^i(t) = \alpha_p \cdot \frac{1}{n_p} \cdot \boxed{(B - Q(t))} \cdot \frac{\mu_p^i}{b}$$

Remaining shared buffer

ABM

Active Buffer Management

$$T_p^i(t) = \alpha_p \cdot \frac{1}{n_p} \cdot (B - Q(t)) \cdot \boxed{\frac{\mu_p^i}{b}}$$

Normalized dequeue rate

ABM

Active Buffer Management

$$T_p^i(t) = \underbrace{\alpha_p \cdot \frac{1}{n_p} \cdot (B - Q(t))}_{Buffer\ \ Management} \cdot \underbrace{\frac{\mu_p^i}{b}}_{AQM}$$

Properties of ABM

- Upper bounds the buffer allocated to a priority **(Prevents monopoly)**

$$B_p^{max} \leq \frac{B \cdot \alpha_p}{1 + \alpha_p}$$

Depends only on the parameter set for the corresponding priority

Properties of ABM

- Lower bounds the buffer allocated to a priority **(Minimum buffer guarantee)**

$$B_p^{min} \geq \frac{B \cdot \alpha_p}{1 + \sum\limits_{p \in \mathcal{P}} \alpha_p}$$

Depends only on the parameter set *for all priorities*

Properties of ABM

- Upper bounds the drain time for each priority **(Bounded queuing delays)**

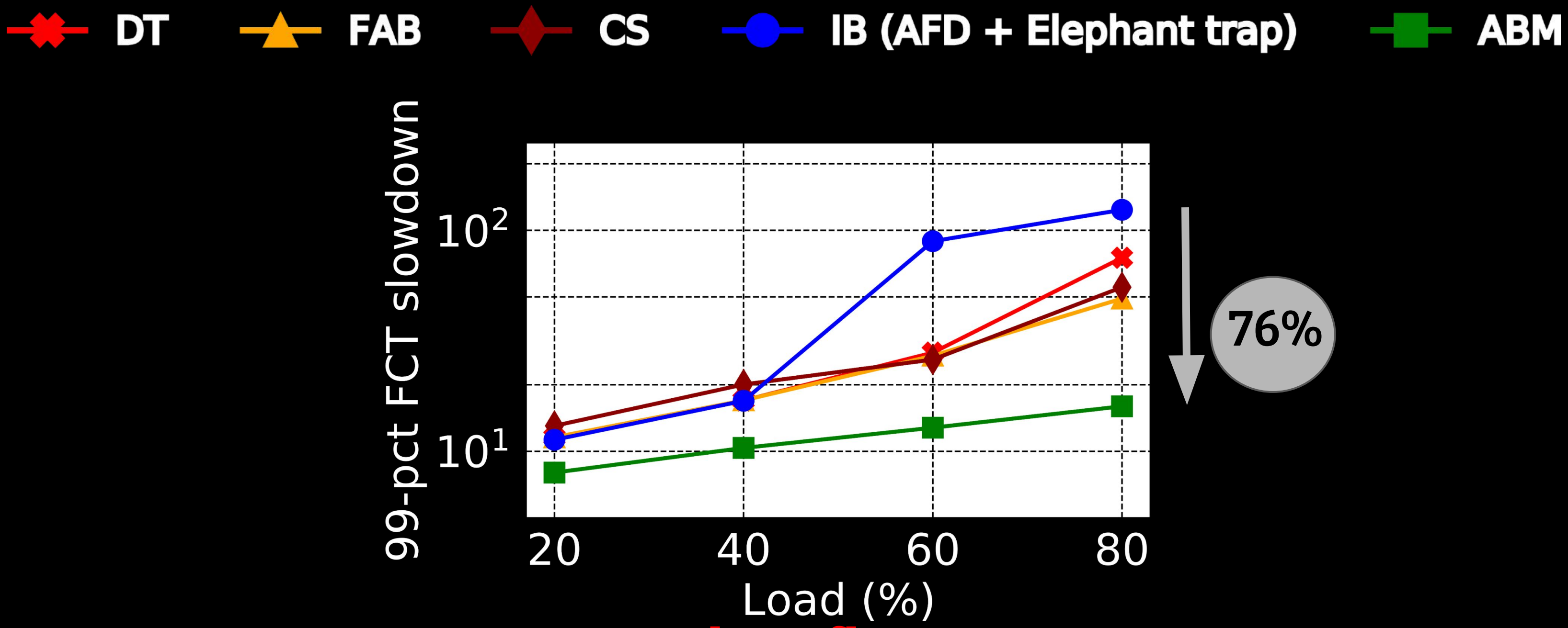
$$\Gamma \leq \frac{B \cdot \alpha_p}{(1+\alpha_p) \cdot b}$$

Depends only on the parameter set for the corresponding priority and the port bandwidth

Evaluation

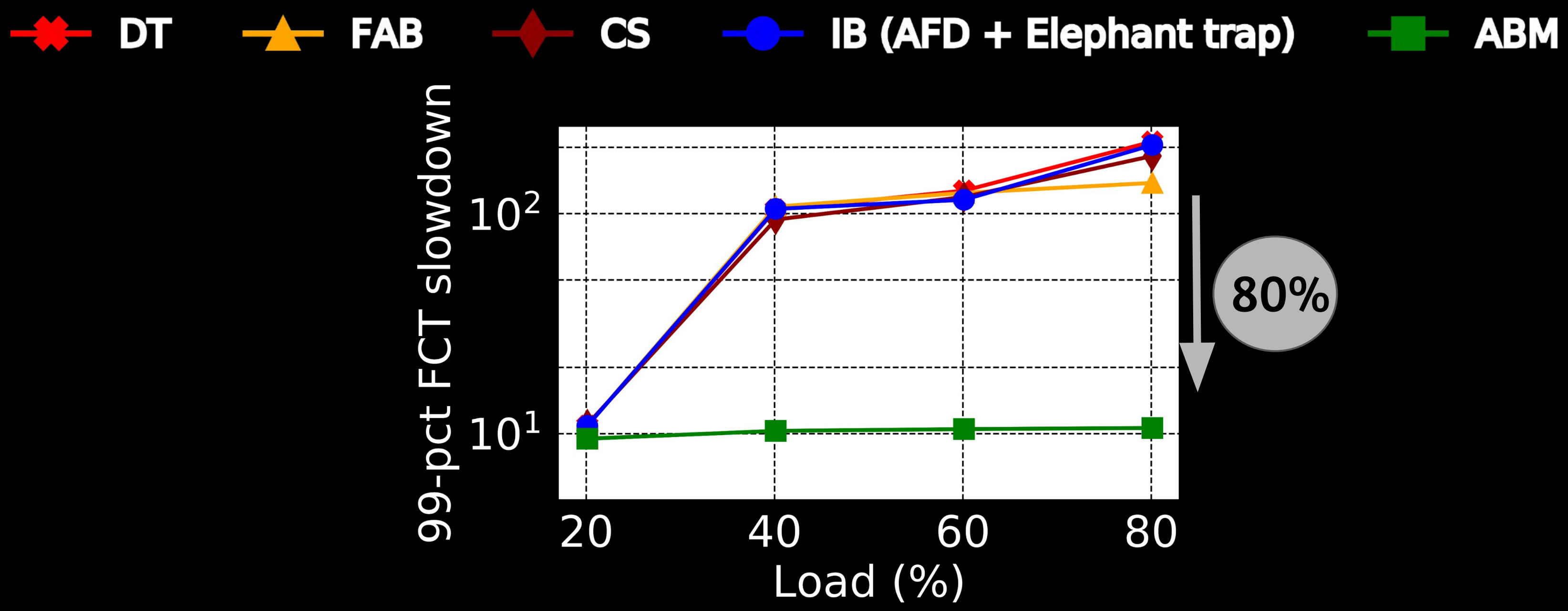
- NS3 simulations
- Leaf-Spine topology (4:1 oversubscription)
- 9.6KB buffer-per-port-per-Gbps for all switches
  - Similar to Broadcom TridentII switch
- Websearch + incast workload

ABM Improves Short Flows FCTs



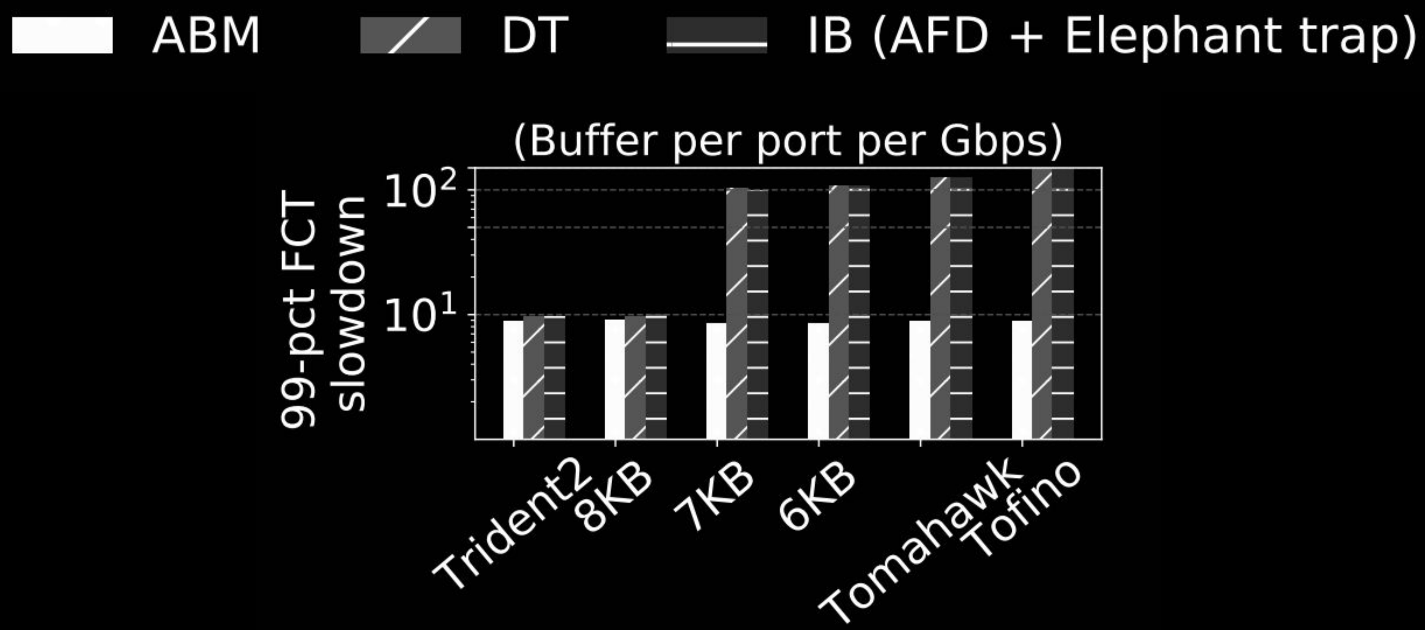
Short flows

ABM Improves Incast Flows FCTs



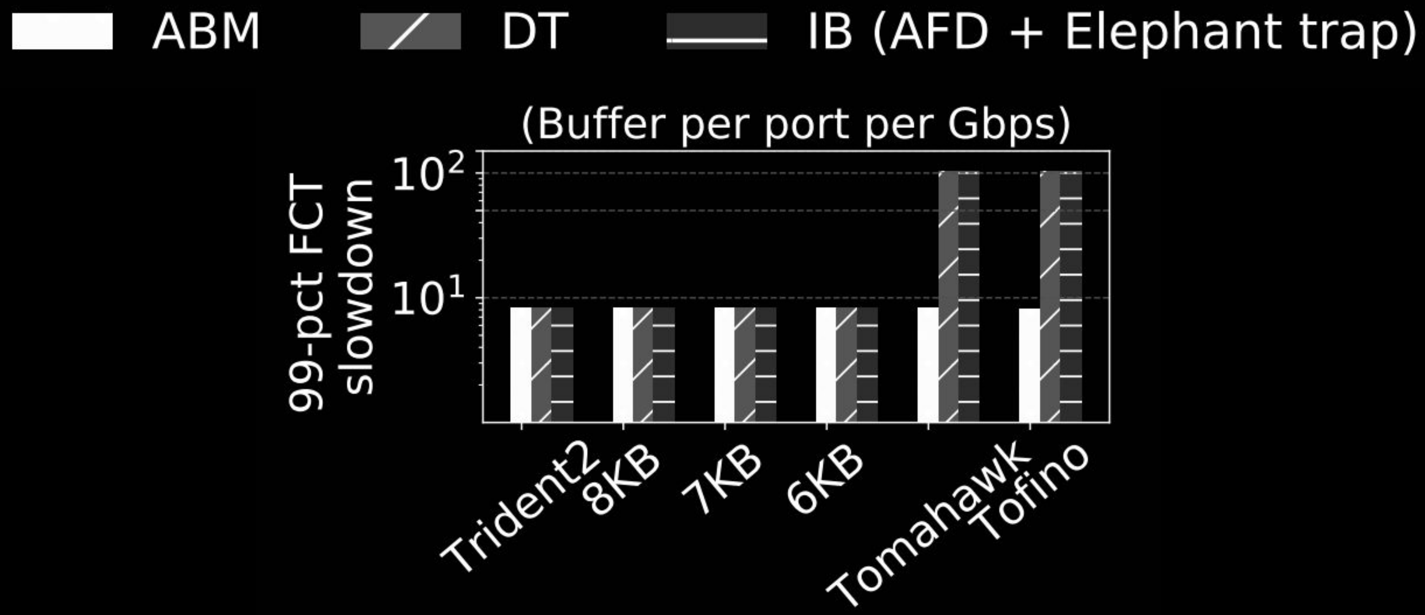
Incast flows

Evaluation under Shallow Buffers and Advanced CC



DCTCP

Evaluation under Shallow Buffers and Advanced CC



PowerTCP

Conclusion

- Existing approach of hierarchical buffer sharing is fundamentally limited to a single dimension
- ABM offers both isolation and stable drain time; and improves burst absorption
- ABM significantly improves the performance of incast flows
- ABM works well even under shallow buffers

https://github.com/inet-tub/ns3-datacenter

Thank you

https://github.com/inet-tub/ns3-datacenter