POWERTCP

Pushing the Performance Limits of Datacenter Networks

Vamsi Addanki, Oliver Michel, Stefan Schmid

Technische Universität Berlin

PRINCETON UNIVERSITY

Technische Universität Berlin

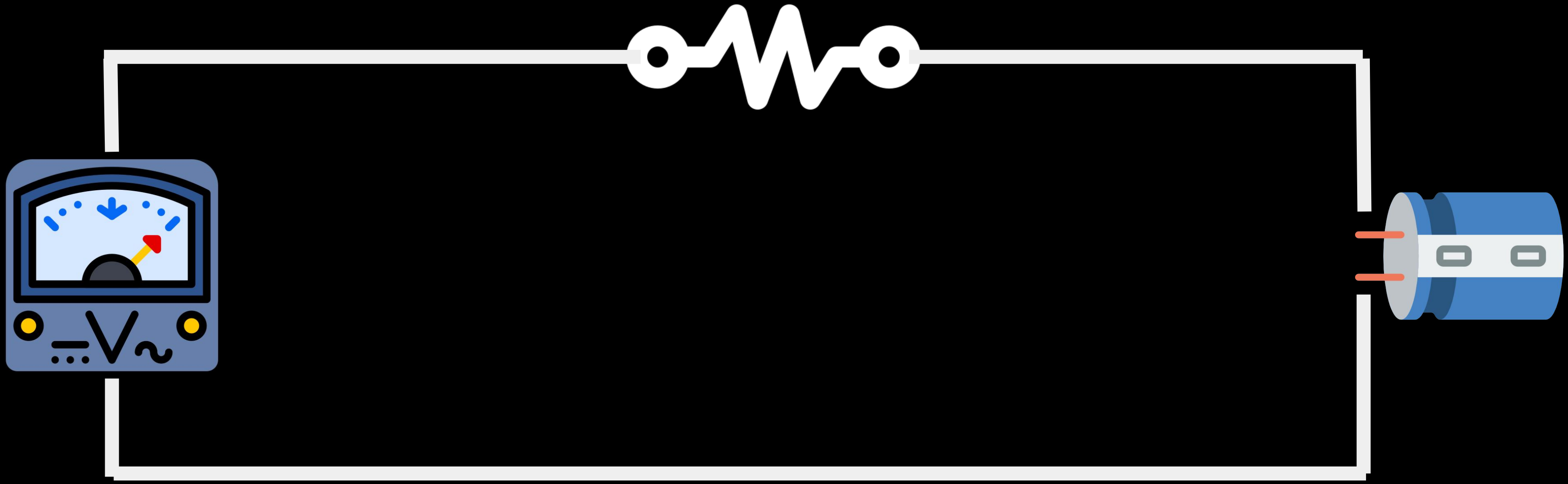
universität wien

erc

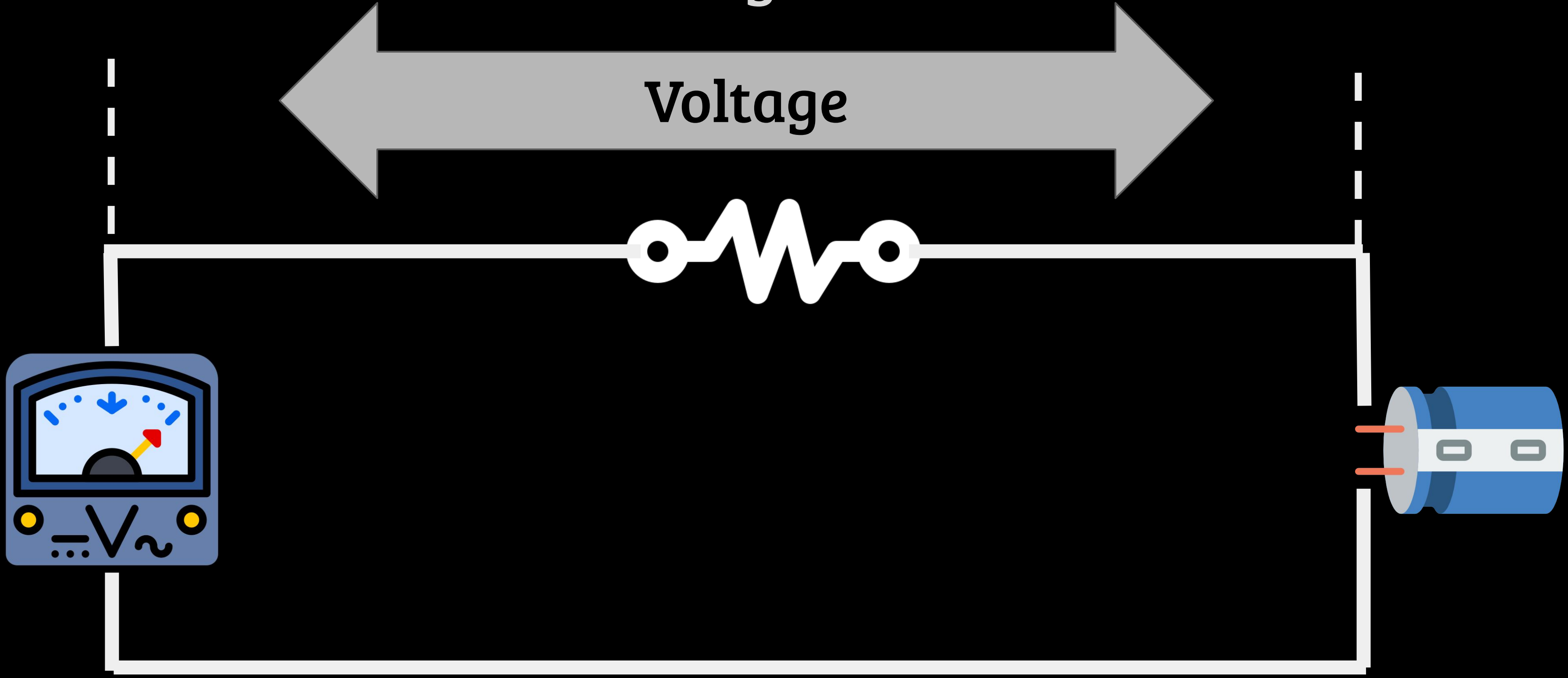
European Research Council

Established by the European Commission

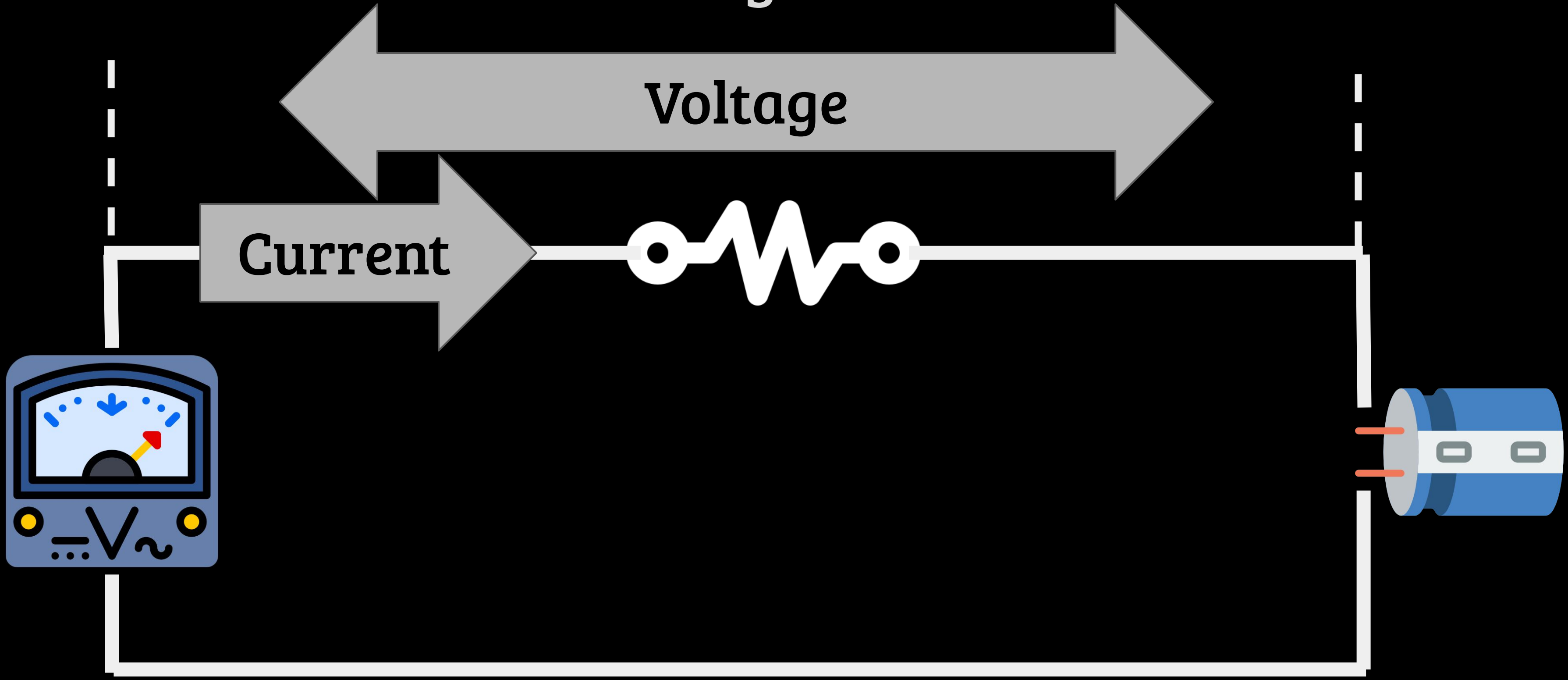
Brief context of electrical systems



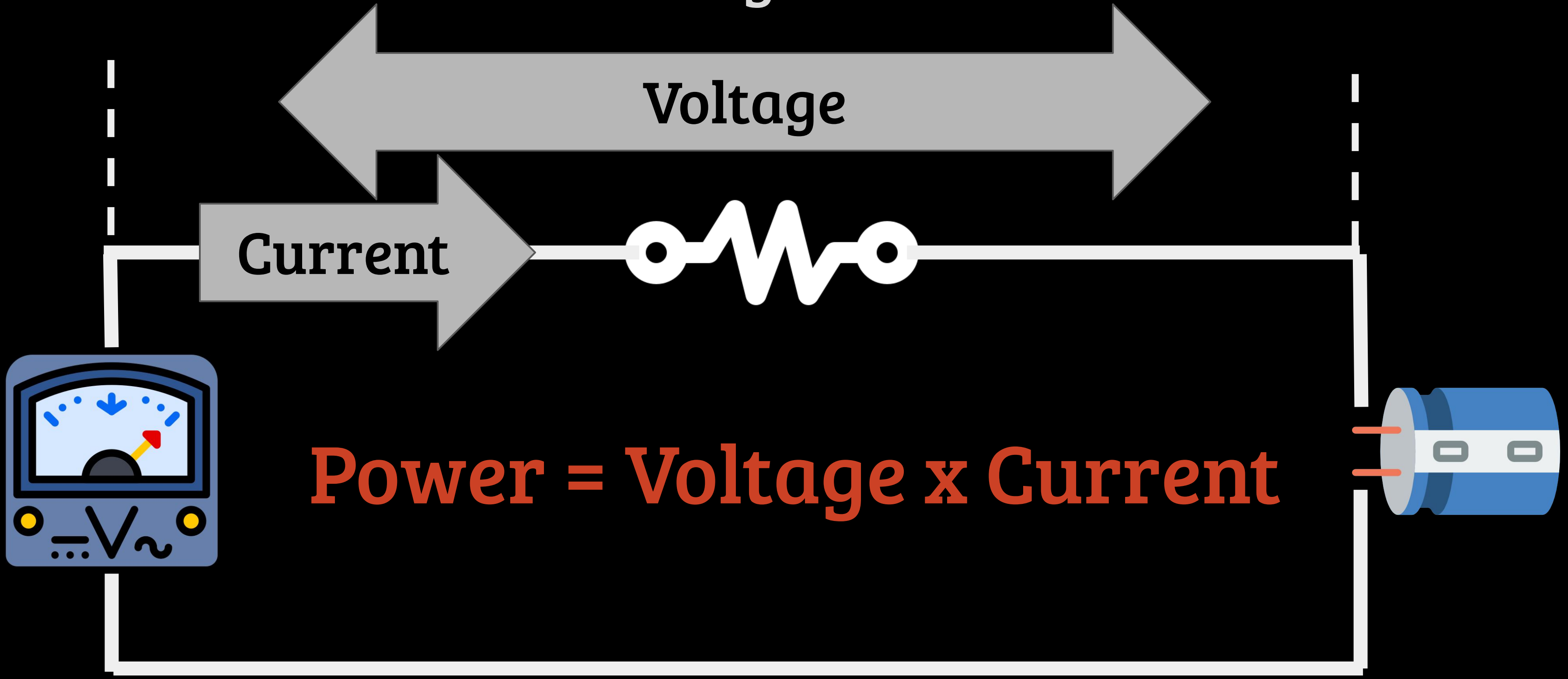
Brief context of electrical systems



Brief context of electrical systems

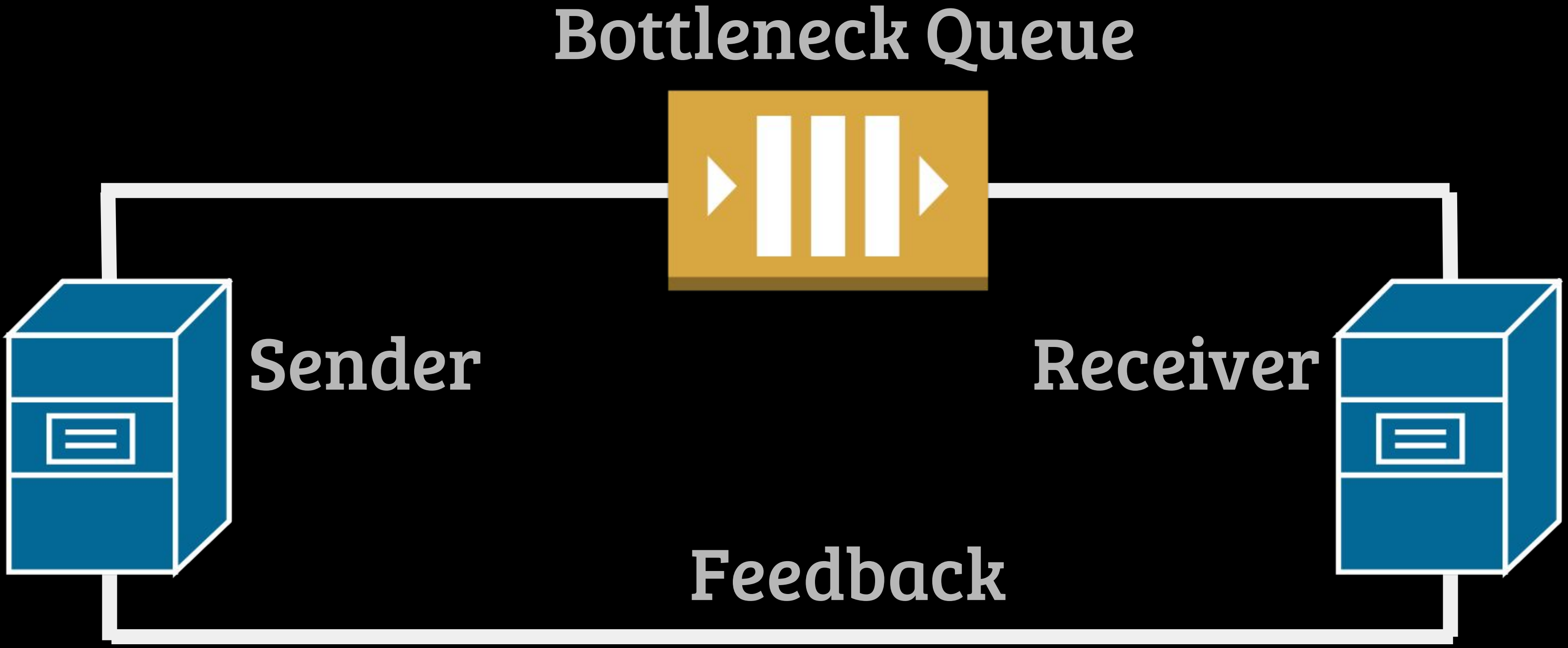


Brief context of electrical systems

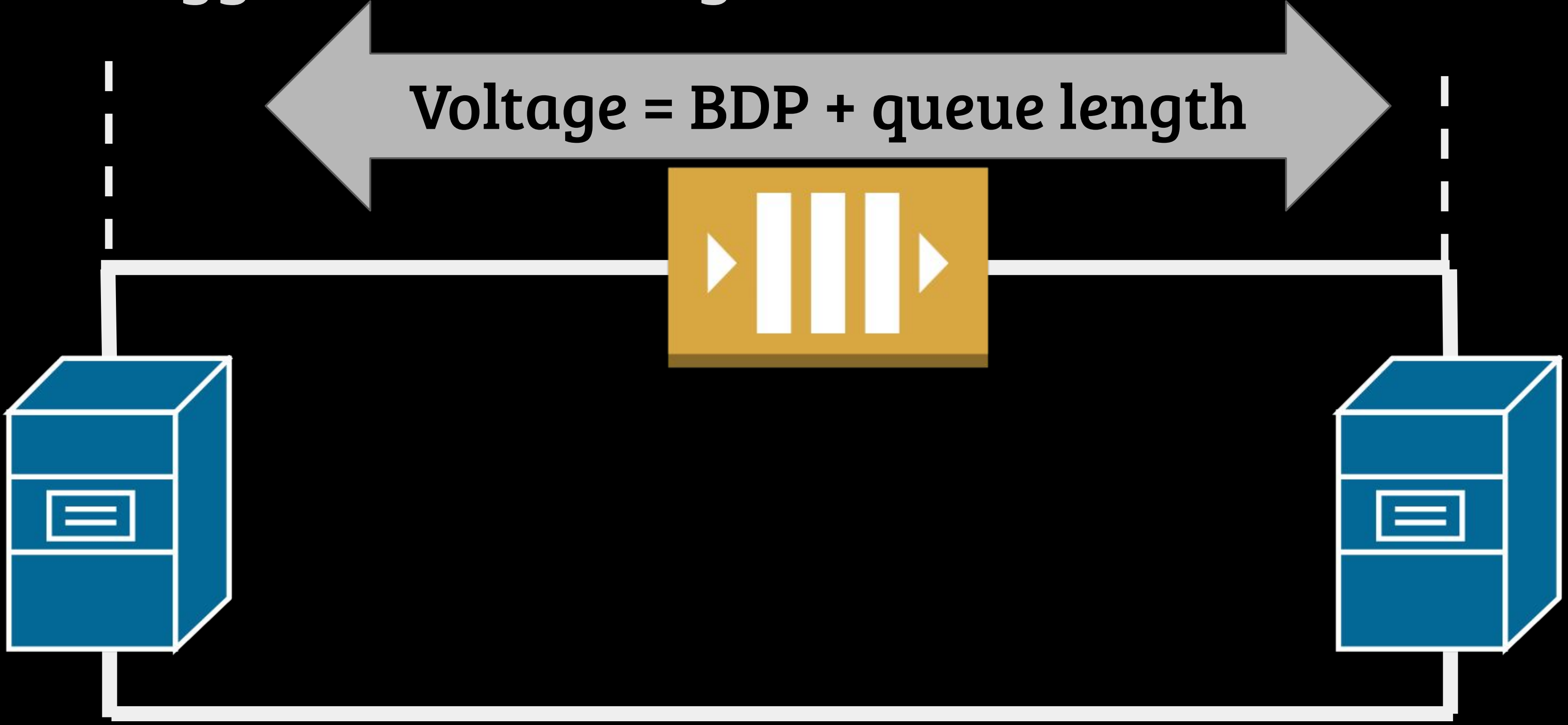


Power = Voltage x Current

Analogy to networked systems



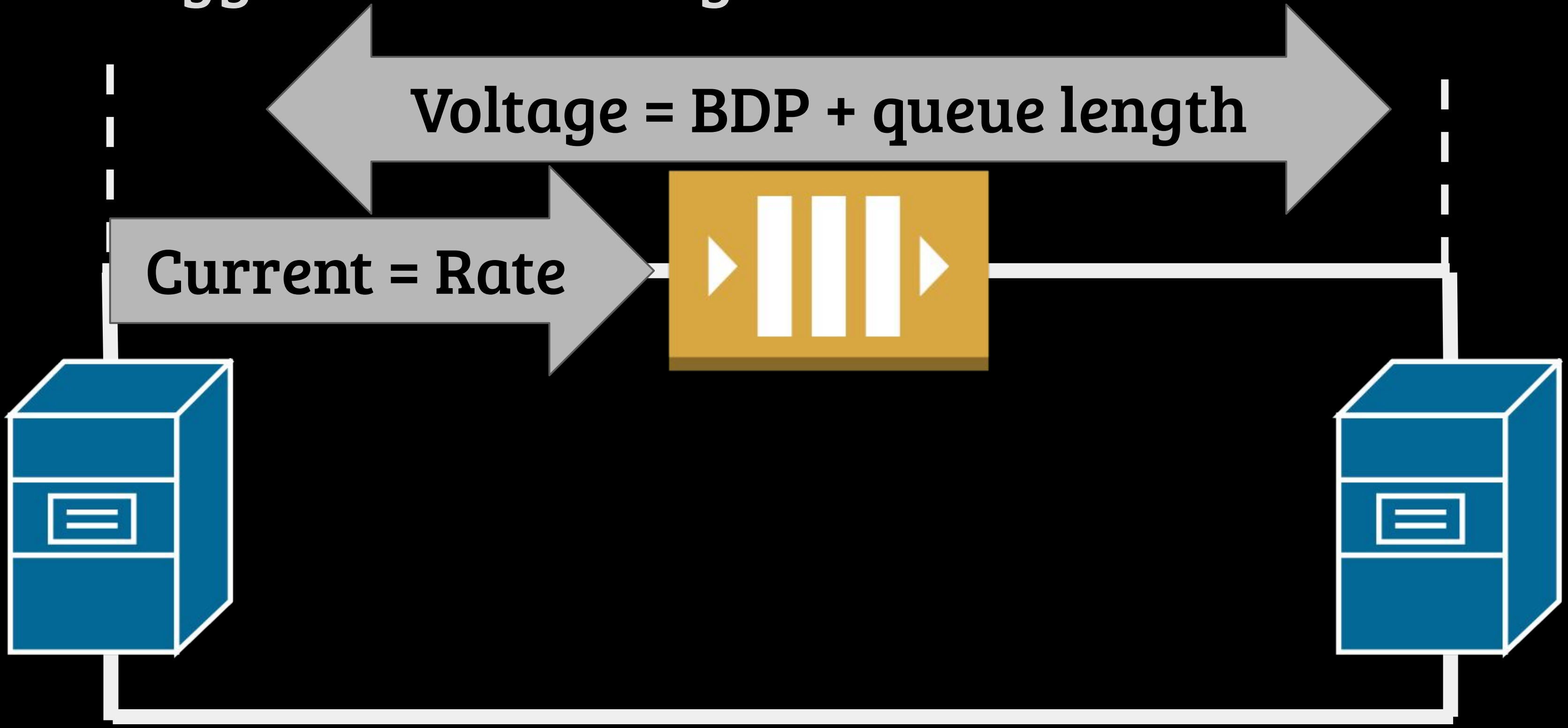
Analogy to networked systems



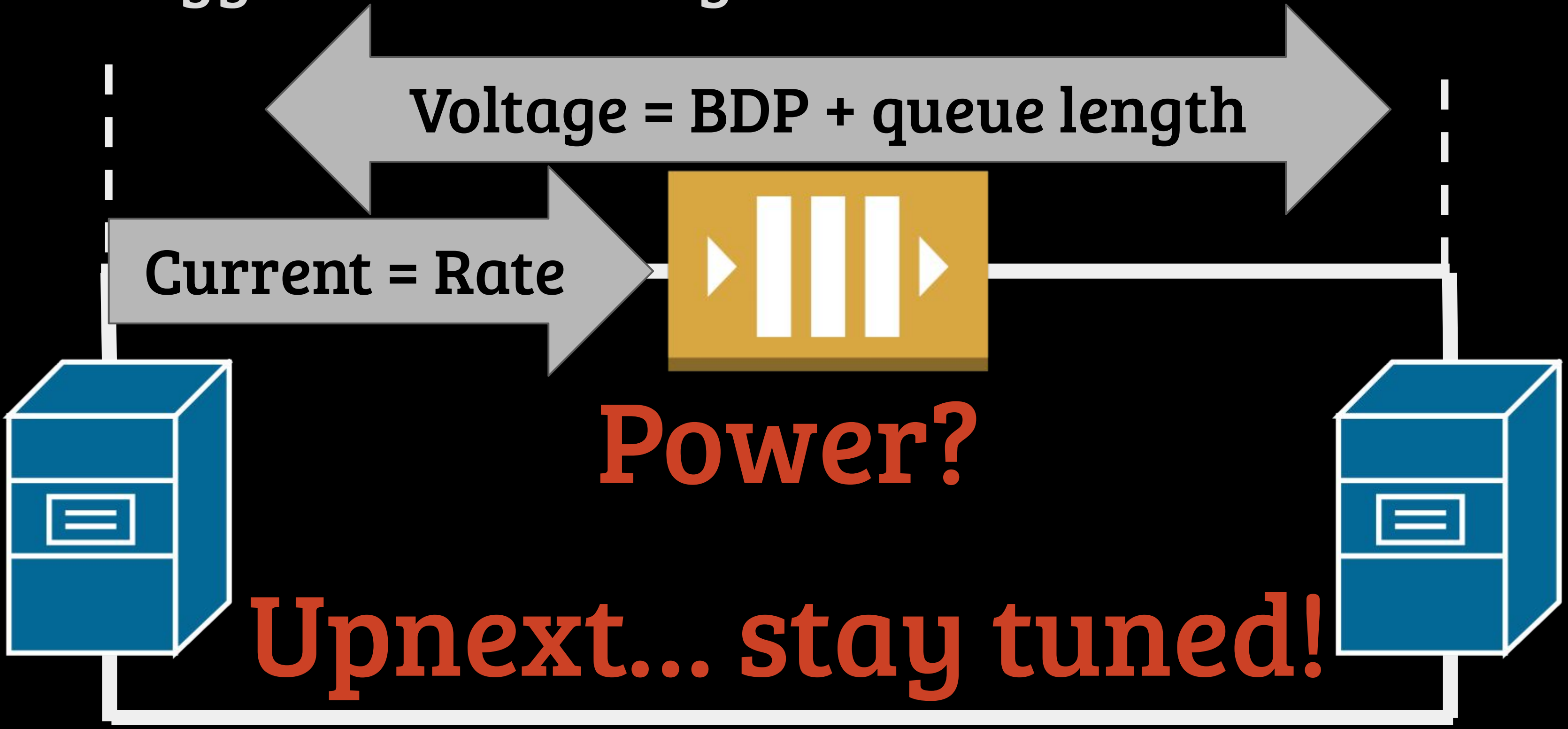
Voltage = BDP + queue length

POWERTCP

Analogy to networked systems



Analogy to networked systems



Voltage = BDP + queue length

Current = Rate

Power?

Upnext... stay tuned!

PowerTCP in a Nutshell

- Power-based congestion control
- Quickly reacts to congestion without losing throughput
- Rapidly converges within 1 RTT
- Fair and asymptotically stable
- Reduces FCTs for short flows by up to 90%

How do we measure Power?

The debate over congestion signals

Microsoft says ECN is better [dctcp]

Google says delay is simple and effective [Timely, Swift]

Alibaba says INT is accurate [HPCC]

ECN, Delay or INT are essential

What matters more: what we do with it

~~The debate over feedback signals~~

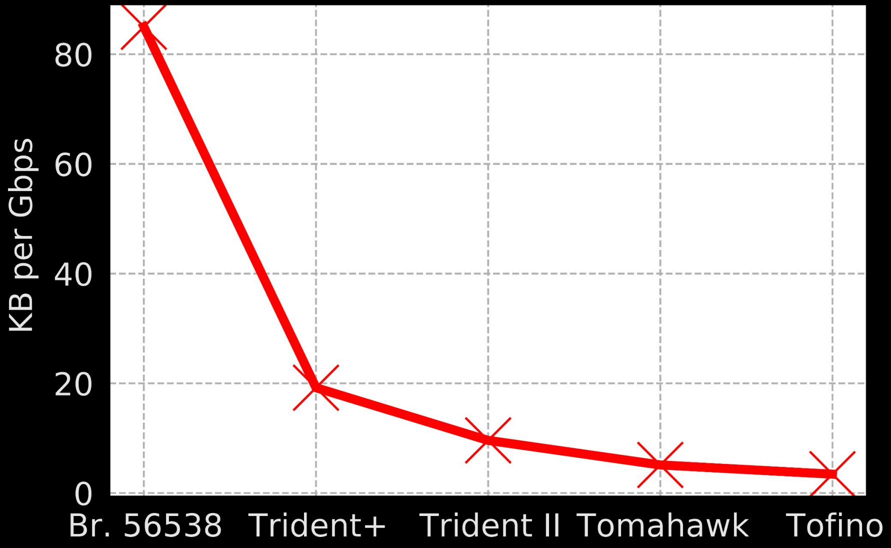
A debate over how to use the feedback

Rare glimpse of Google datacenter

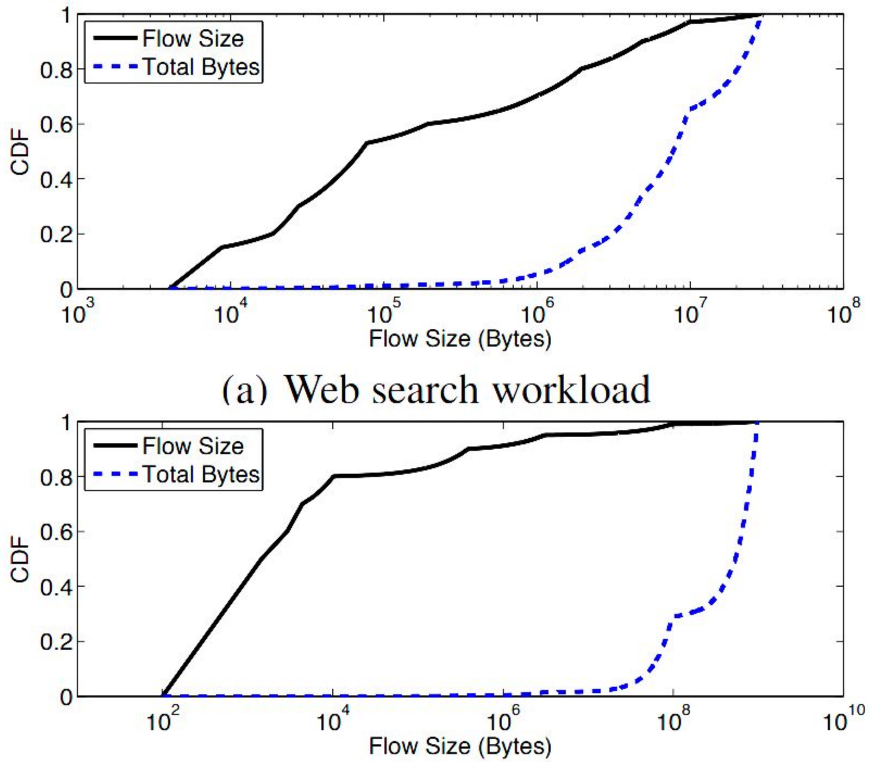
Rare glimpse of Google datacenter

Fear of the buffer

Buffer per unit capacity (KB/Gbps)



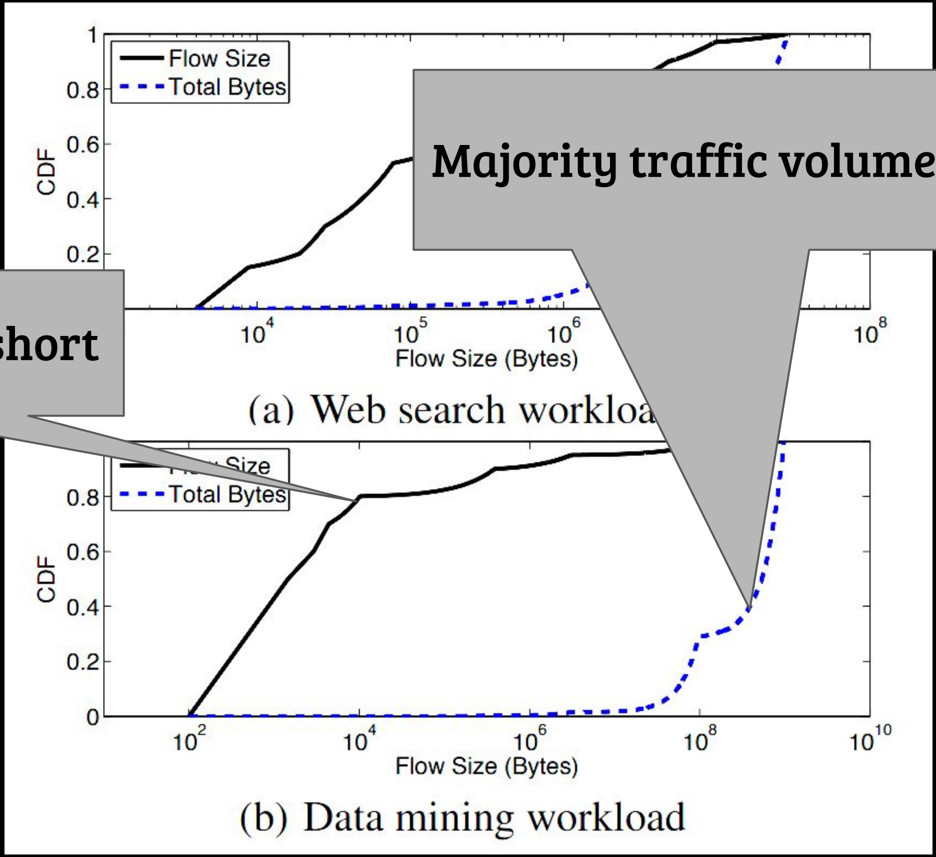
DC workloads and short flows



(a) Web search workload

(b) Data mining workload

DC workloads and short flows



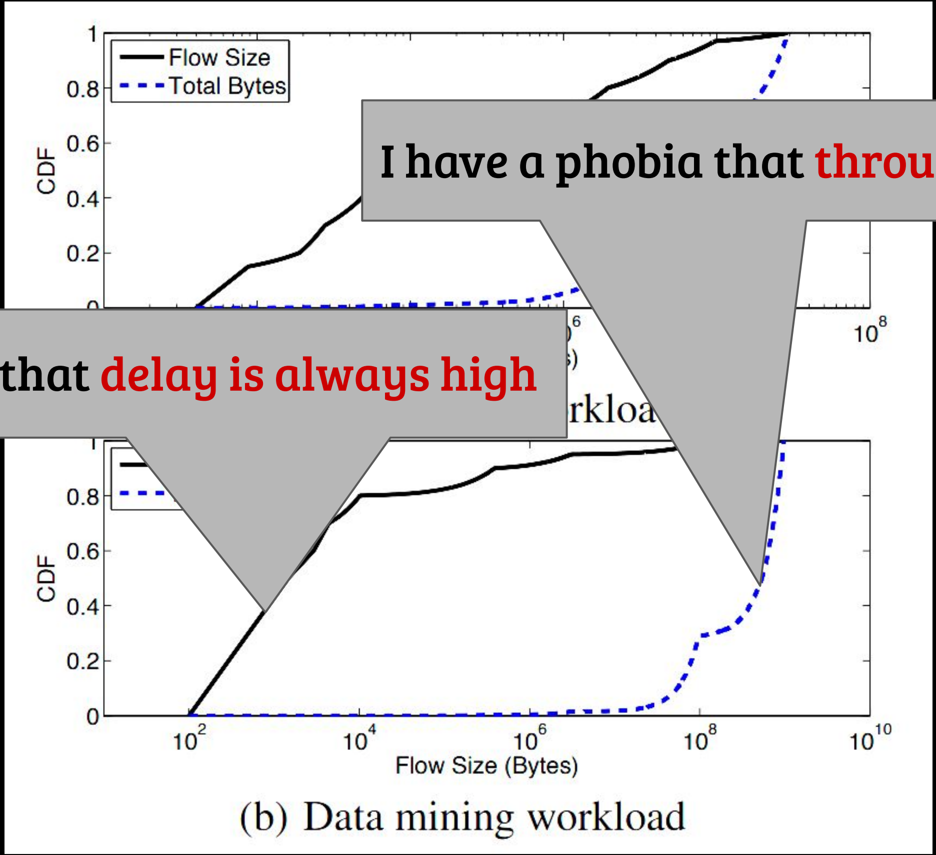
(a) Web search workloa

(b) Data mining workload

Majority traffic volume is from long flows

Majority Flows are short

DC workloads and short flows



(b) Data mining workload

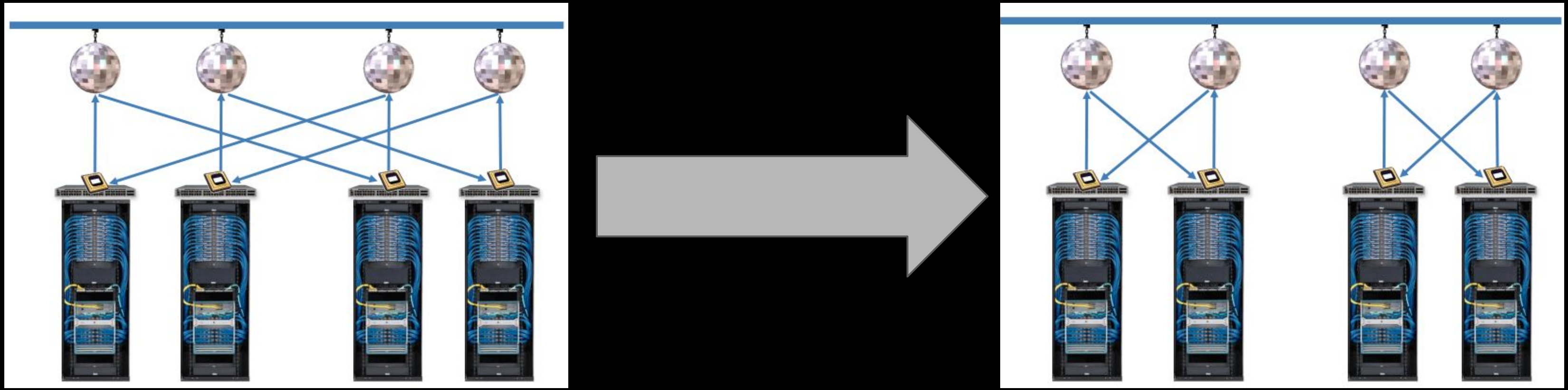
I have a phobia that throughput is always low

I have a constant fear that delay is always high

Emerging technologies and challenges

Not just queueing but quickly utilizing available bandwidth is important too

eg., Emerging Reconfigurable Datacenter Networks (RDCNs)



Fine-grained congestion control is important for datacenter performance

Timeline of congestion control in datacenters

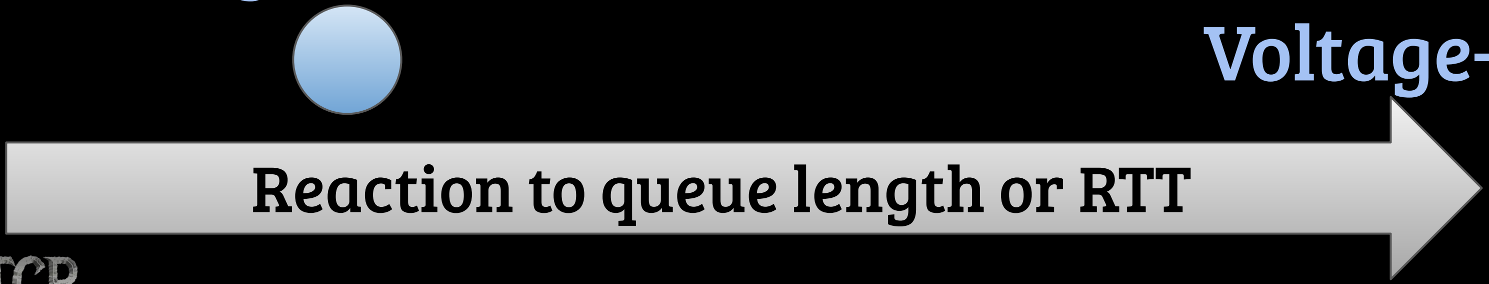
- Reno, Cubic
- DCTCP, DCQCN
- Timely
- HPCC
- Swift

Timeline of congestion control in datacenters

- Voltage-based (BDP + Queue Length)
  - ECN/Loss (eg., DCTCP)
  - RTT based (eg., Swift)
  - Inflight based (eg., HPCC)
- Current-based (Total transmission rate)
  - RTT-gradient based (Eg., Timely)

Voltage-based

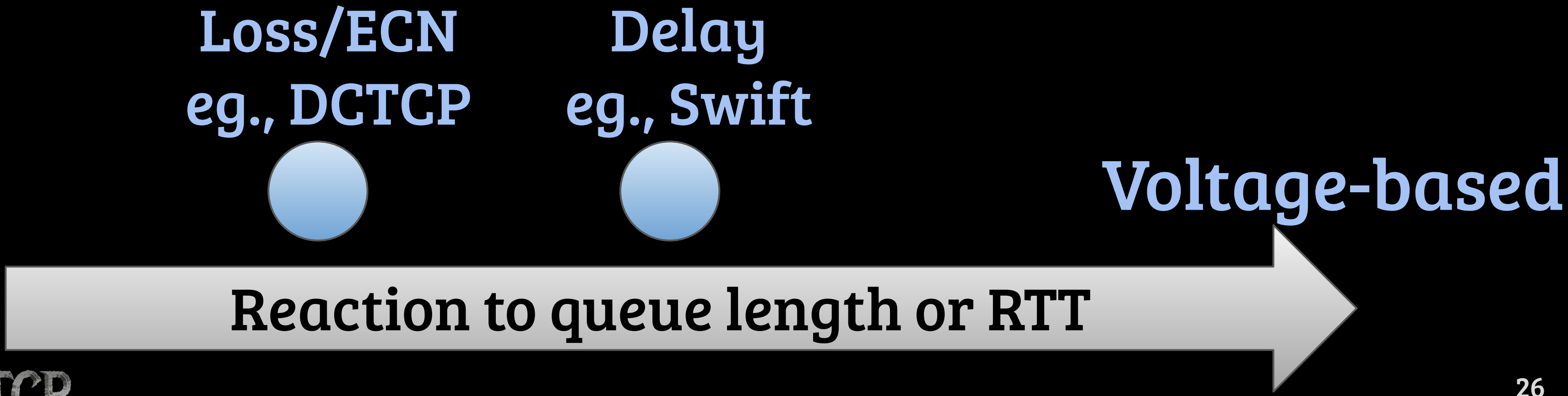
Reaction to queue length or RTT



Loss/ECN
eg., DCTCP

Voltage-based

Reaction to queue length or RTT

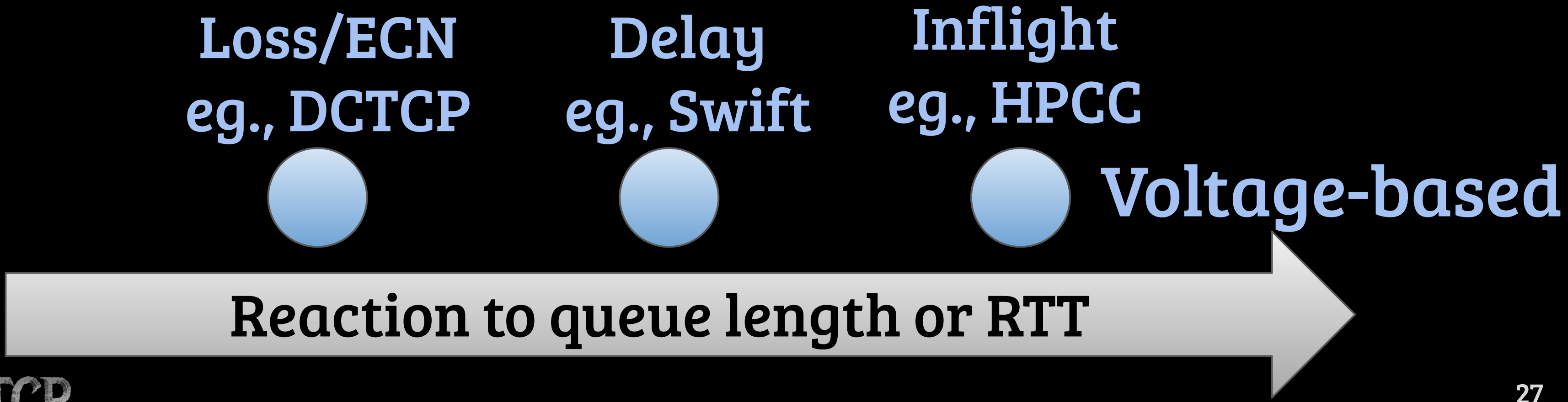


Loss/ECN
eg., DCTCP

Delay
eg., Swift

Voltage-based

Reaction to queue length or RTT



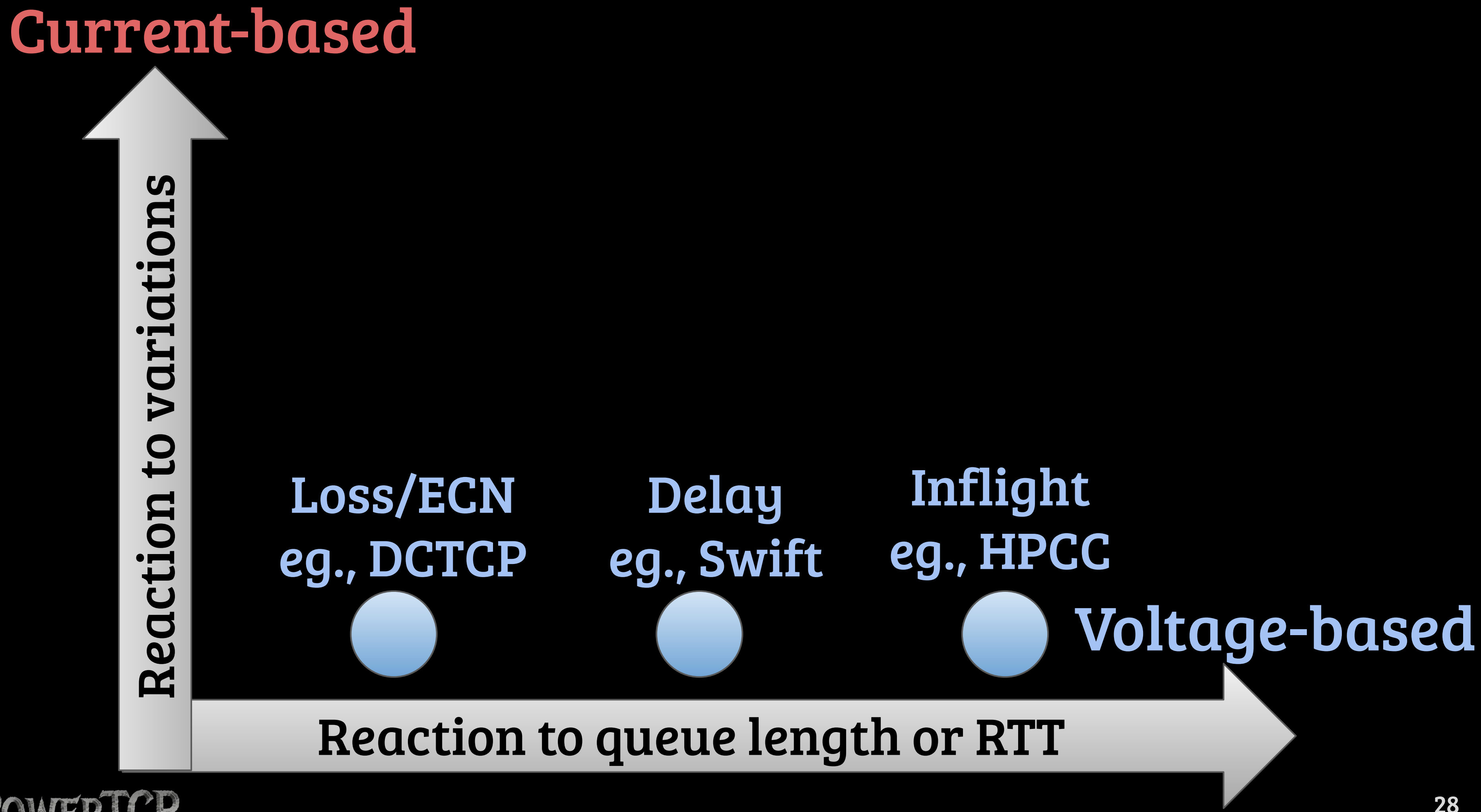
Loss/ECN
eg., DCTCP

Delay
eg., Swift

Inflight
eg., HPCC

Voltage-based

Reaction to queue length or RTT



Current-based

Reaction to variations

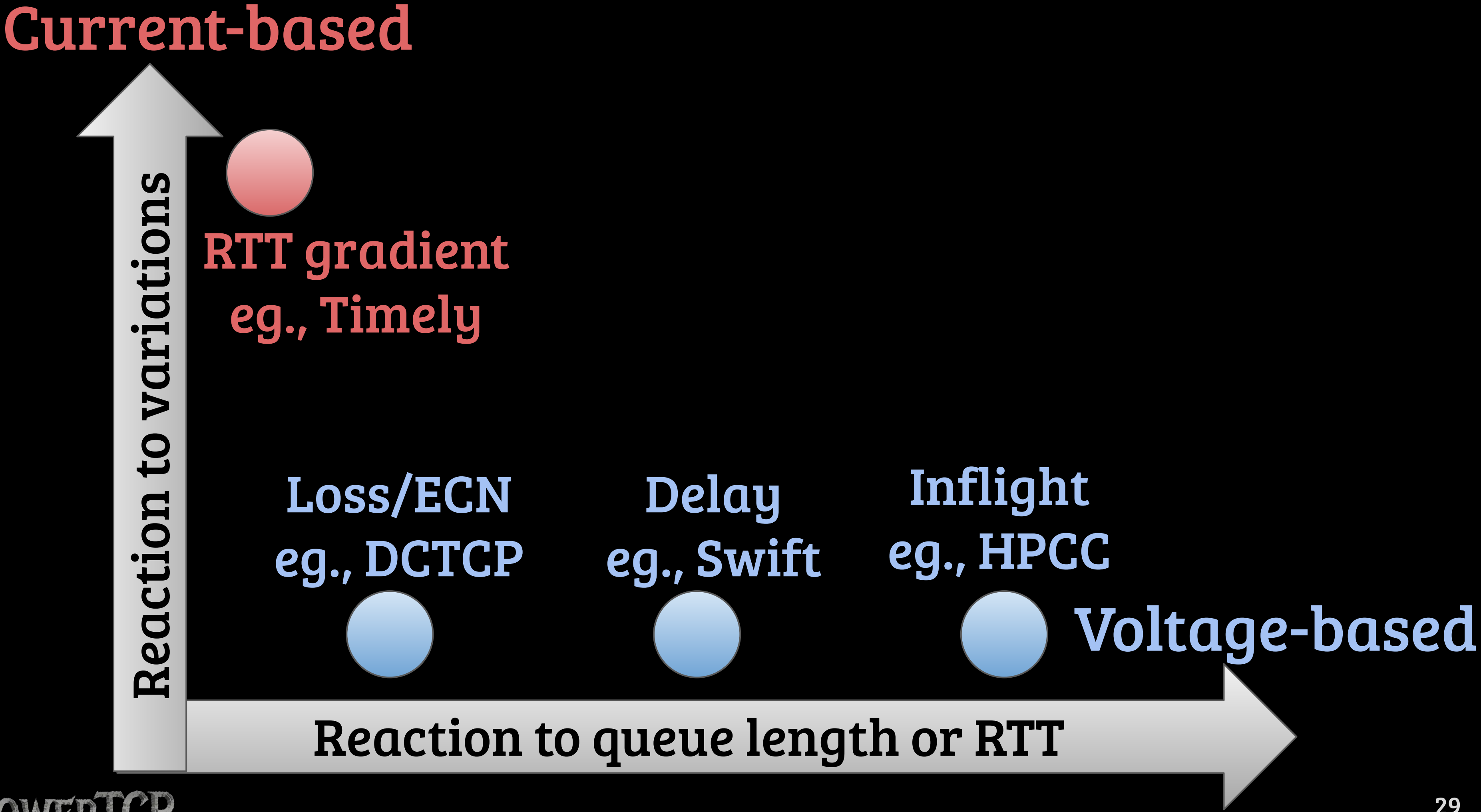
Loss/ECN
eg., DCTCP

Delay
eg., Swift

Inflight
eg., HPCC

Voltage-based

Reaction to queue length or RTT



Current-based

RTT gradient
eg., Timely

Loss/ECN
eg., DCTCP

Delay
eg., Swift

Inflight
eg., HPCC

Voltage-based

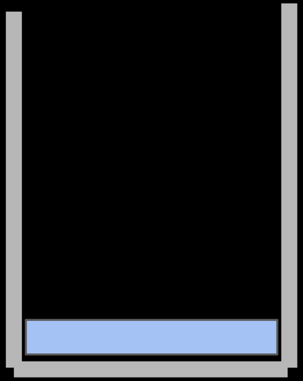
Reaction to variations

Reaction to queue length or RTT

Problems of existing approaches

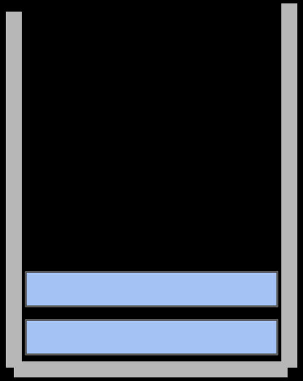
Fundamentally limited to a single dimension

Problems of existing approaches



5 Packets

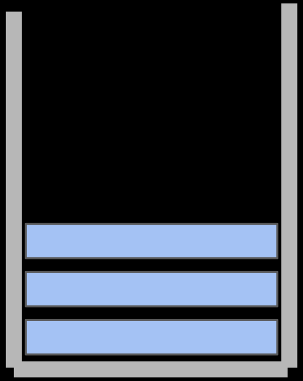
Problems of existing approaches



10 Packets

PowerTCP

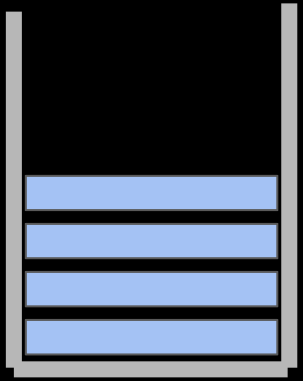
Problems of existing approaches



15 Packets

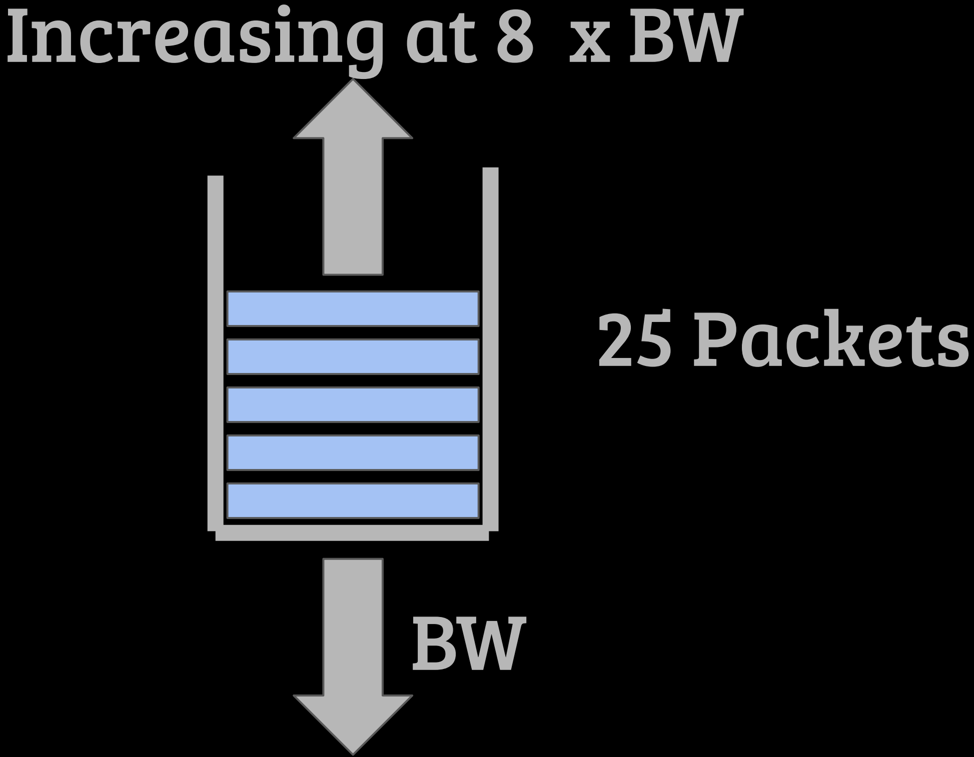
POWERTCP

Problems of existing approaches



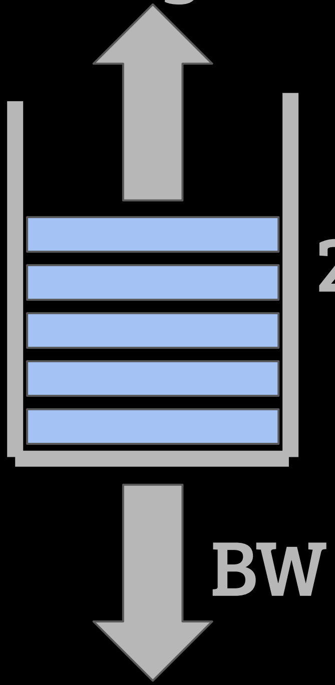
20 Packets

Problems of existing approaches



Problems of existing approaches

Increasing at 8x BW

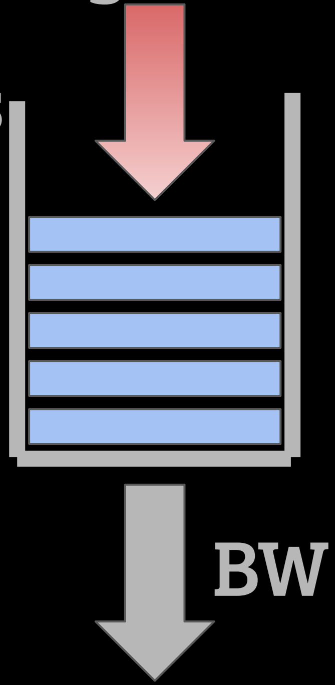


25 Packets

BW

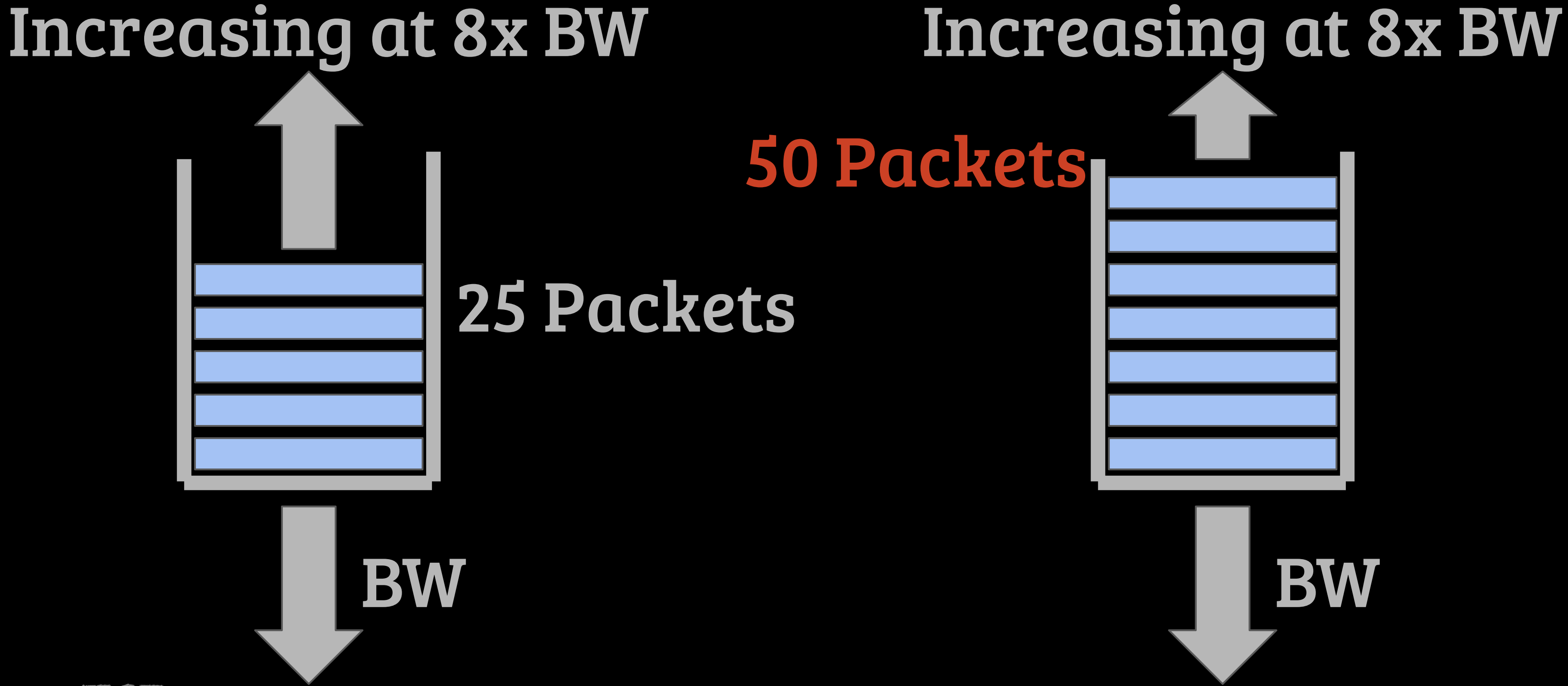
Draining at max rate

25 Packets



BW

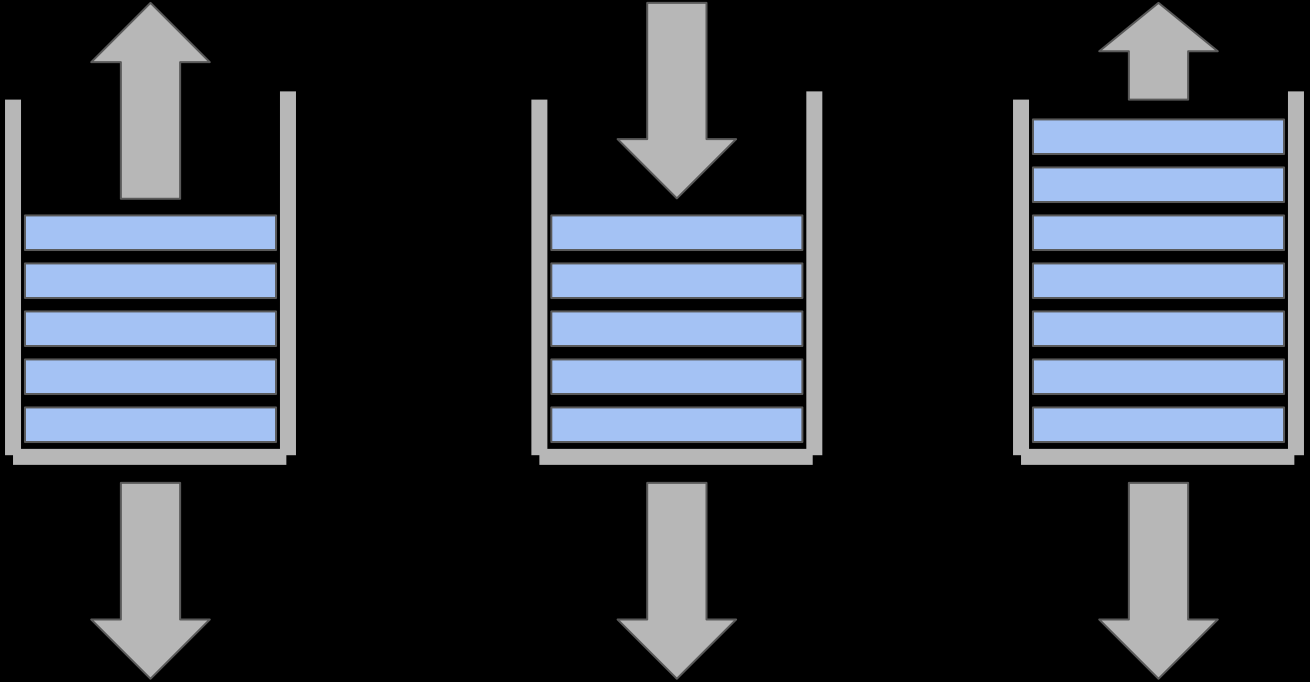
Problems of existing approaches



POWERTCP

Problems of existing approaches

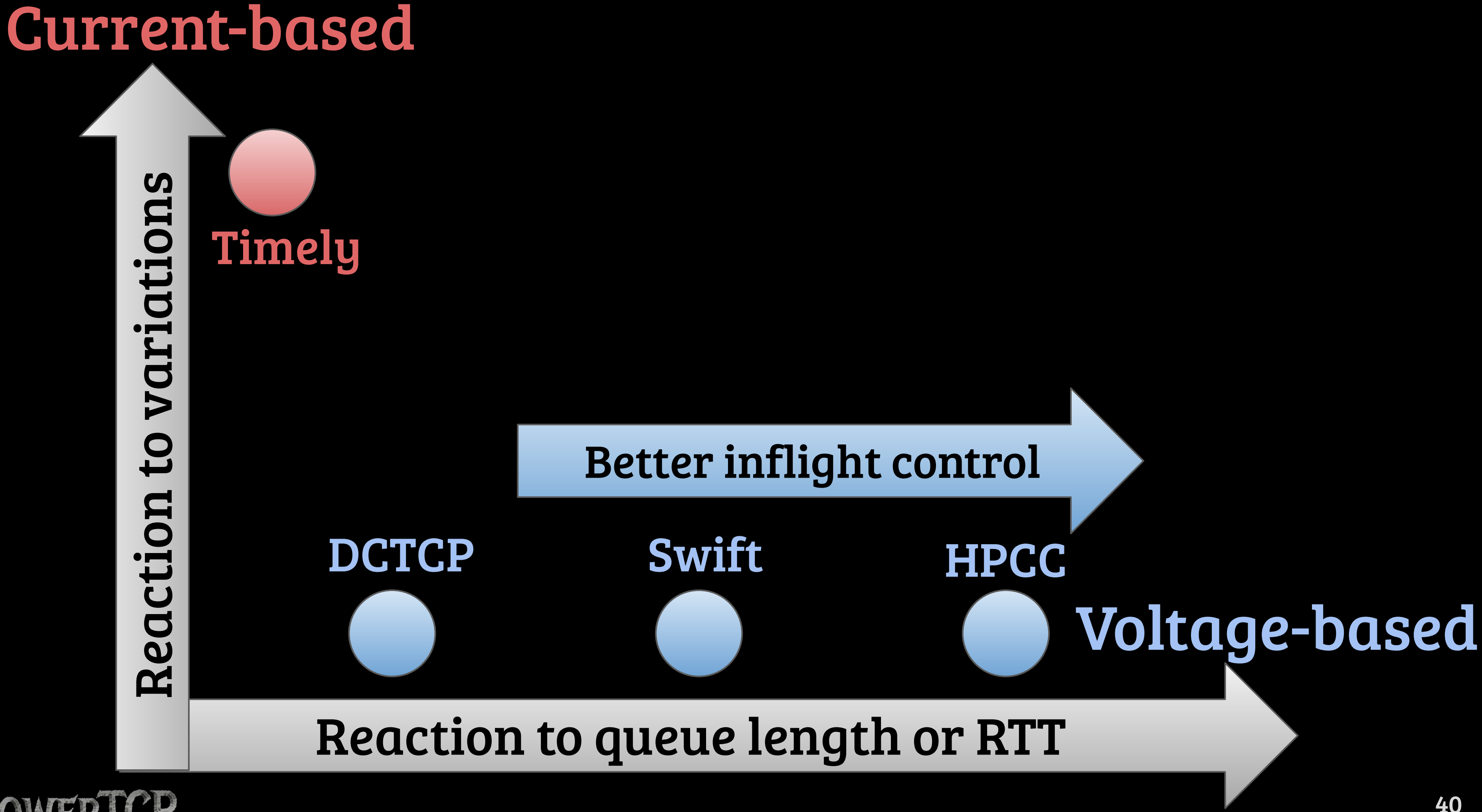
Fundamentally limited to a single dimension



POWERTCP

Summary of Our Analysis

- Voltage-based
  - Can in-principle achieve near-zero queue equilibrium
  - Slow reaction
- Current-based
  - Unstable with no equilibrium
  - Fast Reaction



Current-based

Reaction to variations

Timely

Better inflight control

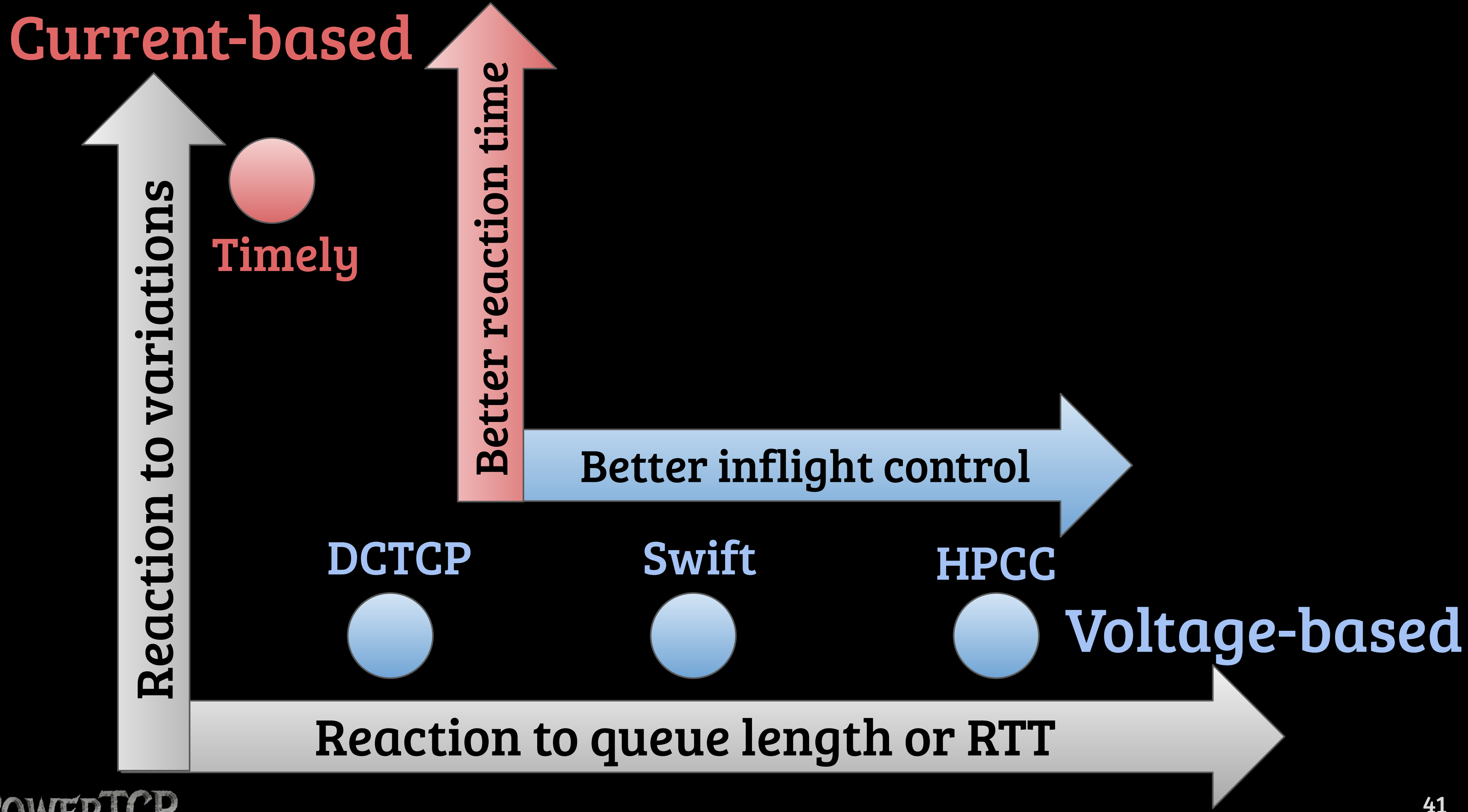
DCTCP

Swift

HPCC

Voltage-based

Reaction to queue length or RTT



Current-based

Timely

Better reaction time

Better inflight control

Reaction to variations

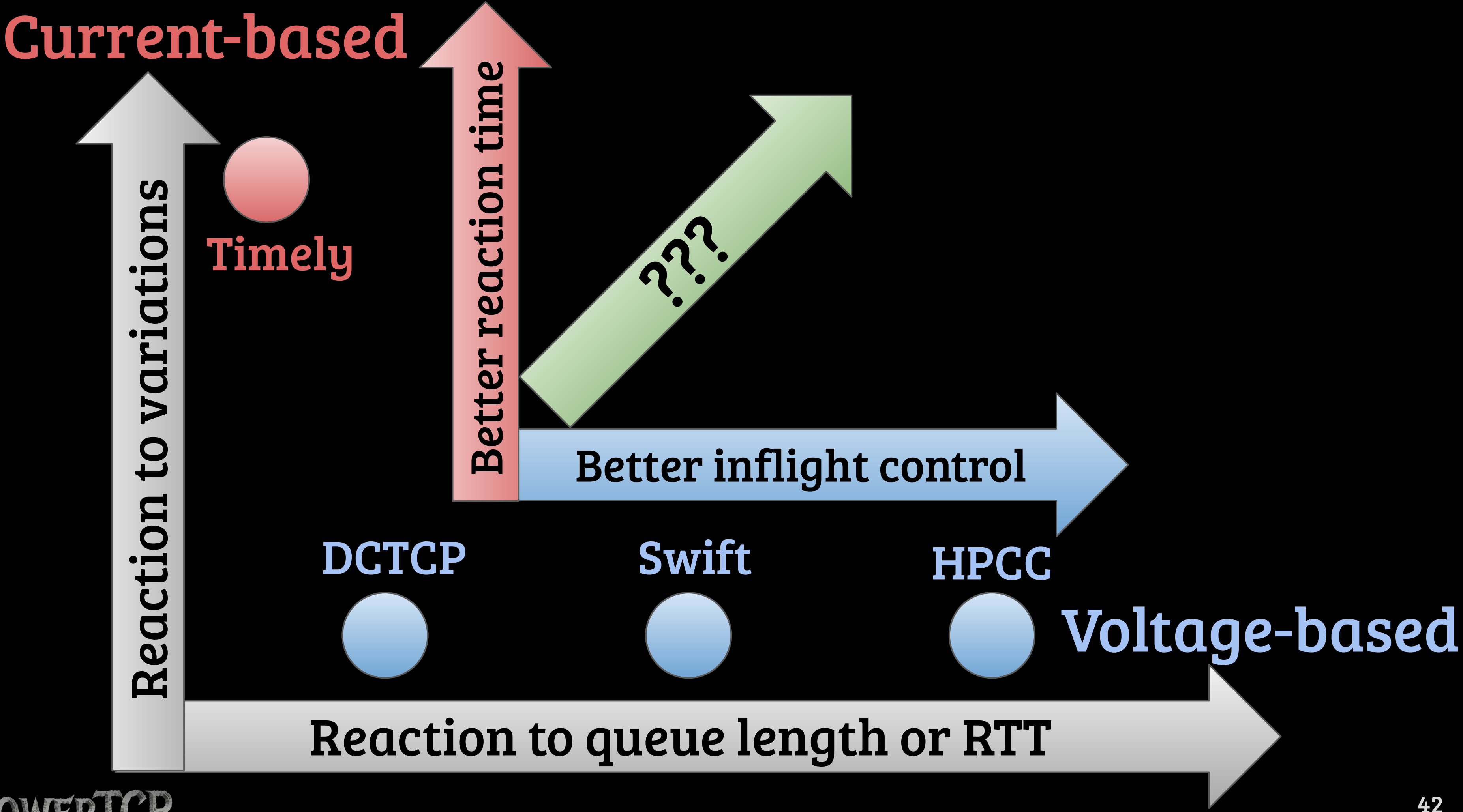
DCTCP

Swift

HPCC

Voltage-based

Reaction to queue length or RTT


Current-based
Timely
Reaction to variations
Better reaction time
???
Better inflight control
DCTCP
Swift
HPCC
Voltage-based
Reaction to queue length or RTT

The notion of power

Power = Voltage x Current

$$\underbrace{\Gamma}_{\text{Power}} = \underbrace{\left(q(t) + b \times \tau\right)}_{\substack{\text{Voltage} \\ \Big\downarrow \\ \text{BDP+queue bytes}}} \times \underbrace{\left(\dot{q}(t) + \mu(t)\right)}_{\substack{\text{Current} \\ \Big\downarrow \\ \text{Total rate}}}$$

The notion of power

Enqueue rate = queue-gradient + Dequeue rate

$$\lambda(t - t^f) = \dot{q}(t) + \mu(t)$$

Sending rate = Window per RTT

$$\lambda(t) = \frac{w(t)}{\theta(t)}$$

RTT = queueing delay + base RTT

$$\theta(t - t^f) = \frac{q(t)}{b} + \tau$$

The notion of power

$$b \times w(t - t^f) = \underbrace{(q(t) + b \times \tau)}_{\text{Voltage}} \times \underbrace{(\dot{q}(t) + \mu(t))}_{\text{Current}}$$

The notion of power

A function of both queue length and variations

The notion of power

A function of both queue length and variations

- Detects increased queue lengths

The notion of power

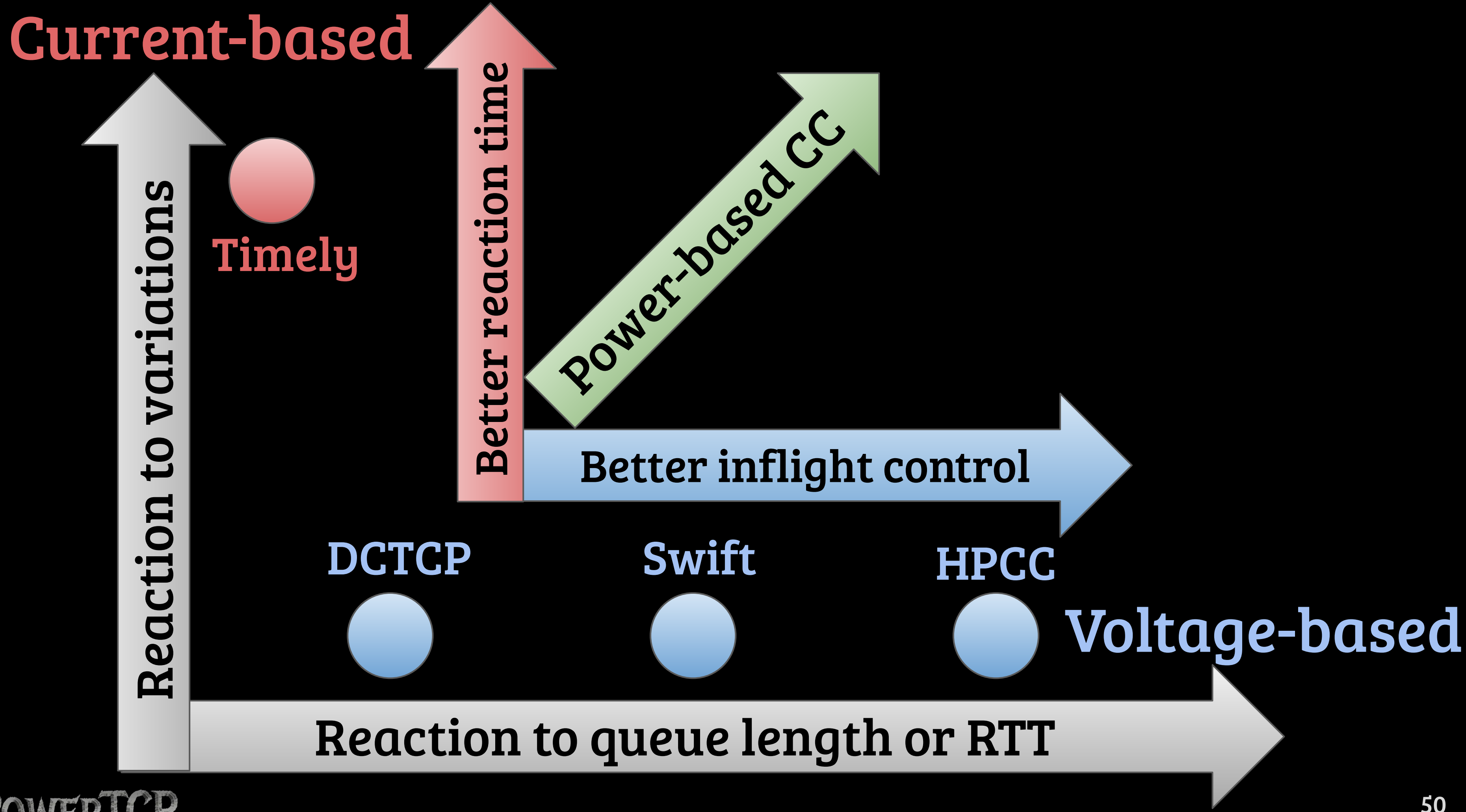
A function of both queue length and variations

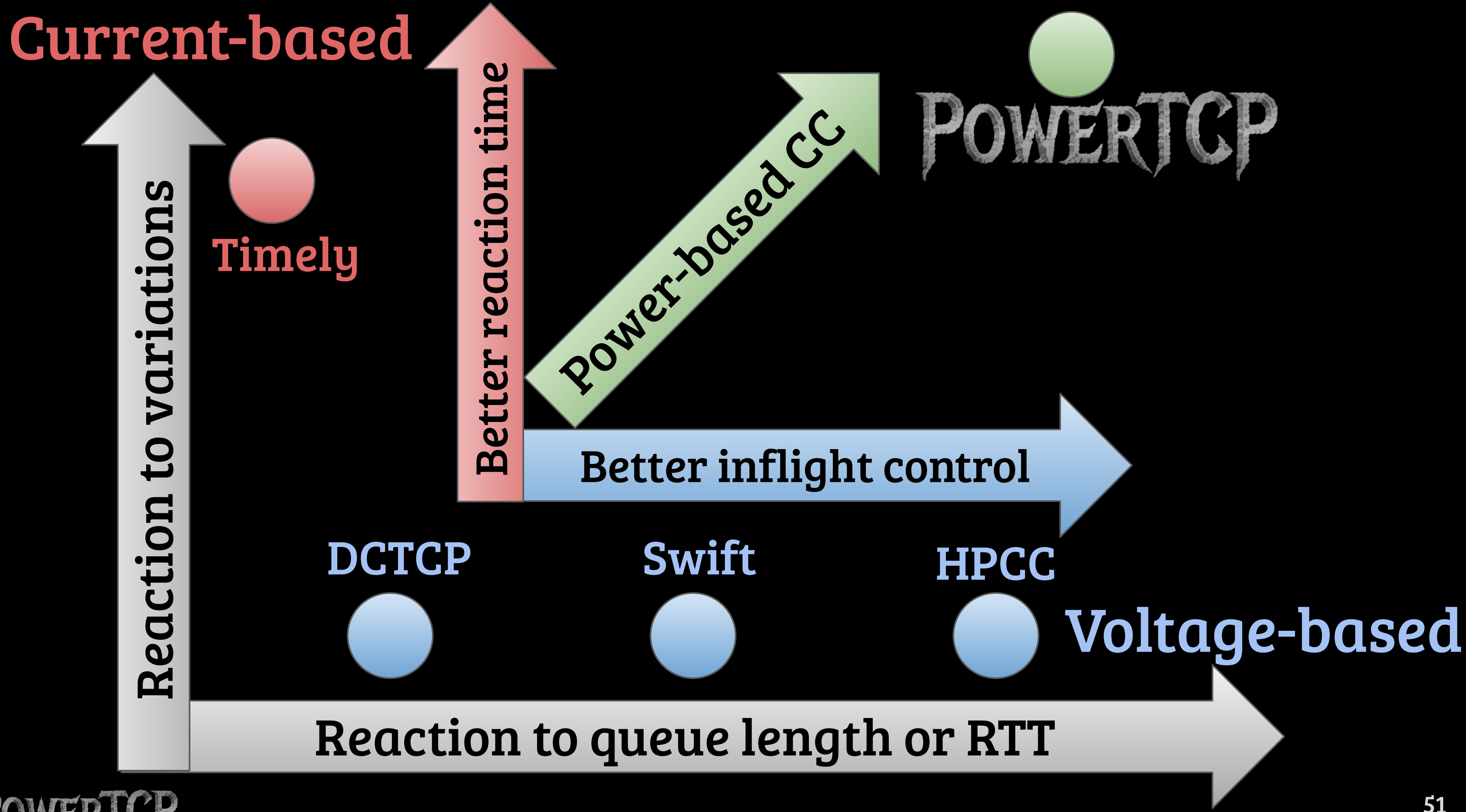
- Detects increased queue lengths
- Detects congestion onset and intensity

The notion of power

A function of both queue length and variations

- Detects increased queue lengths
- Detects congestion onset and intensity
- Detects rapid drop in queue lengths


Current-based
Reaction to variations
Timely
Better reaction time
Power-based CC
Better inflight control
DCTCP
Swift
HPCC
Voltage-based
Reaction to queue length or RTT


Current-based
Timely
Reaction to variations
Better reaction time
Power-based CC
POWERTCP
Better inflight control
DCTCP
Swift
HPCC
Voltage-based
Reaction to queue length or RTT

PowerTCP control law

$$\boxed{w_i(t+\delta t)} = \gamma \cdot \left(w_i(t) \cdot \tfrac{e}{f(t)} + \beta \right) + (1-\gamma) \cdot w_i(t)$$

New window size

PowerTCP control law

$$w_i(t+\delta t) = \gamma \cdot \left(\boxed{w_i(t)} \cdot \frac{e}{f(t)} + \beta\right) + (1-\gamma) \cdot w_i(t)$$

Old window size

PowerTCP control law

$$w_i(t+\delta t) = \gamma \cdot \left(w_i(t) \cdot \boxed{\frac{e}{f(t)}} + \beta \right) + (1-\gamma) \cdot w_i(t)$$

MIMD based on Power

(Multiplicative increase - multiplicative decrease)

PowerTCP control law

$$w_i(t+\delta t) = \gamma \cdot \left(w_i(t) \cdot \frac{e}{f(t)} + \boxed{\beta} \right) + (1-\gamma) \cdot w_i(t)$$

Additive increase

PowerTCP control law

$$w_i(t+\delta t) = \boxed{\gamma \cdot} \left(w_i(t) \cdot \frac{e}{f(t)} + \beta \right) + \boxed{(1-\gamma)} \cdot w_i(t)$$

Exponential Weighted Moving Average (EWMA)

PowerTCP feedback

Power is measured via Inband Network Telemetry (INT)

- Queue lengths
- Timestamps
- Tx bytes
- Bandwidth

PowerTCP without switch support

- Power can be measured via delay signal

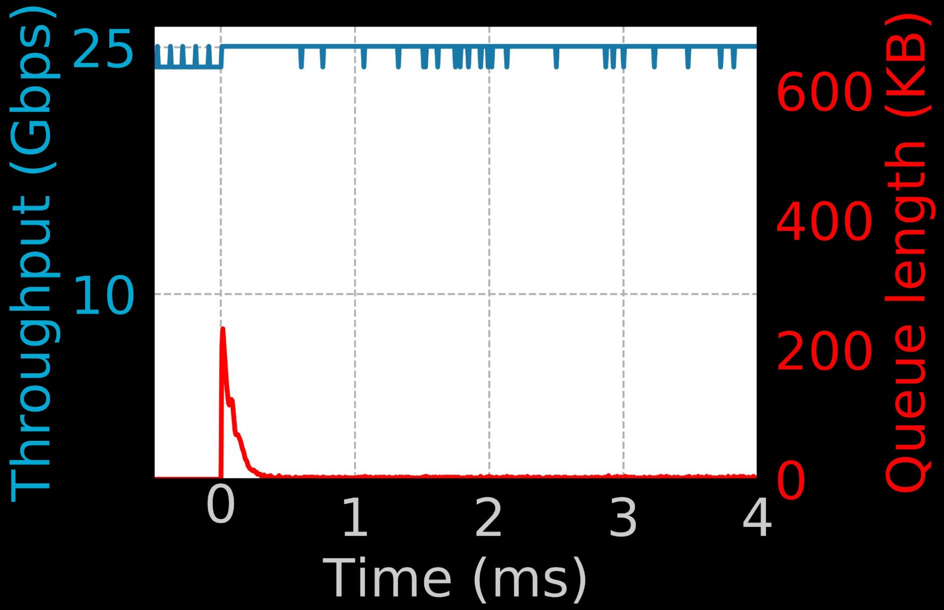
PowerTCP without switch support

- Power can be measured via delay signal

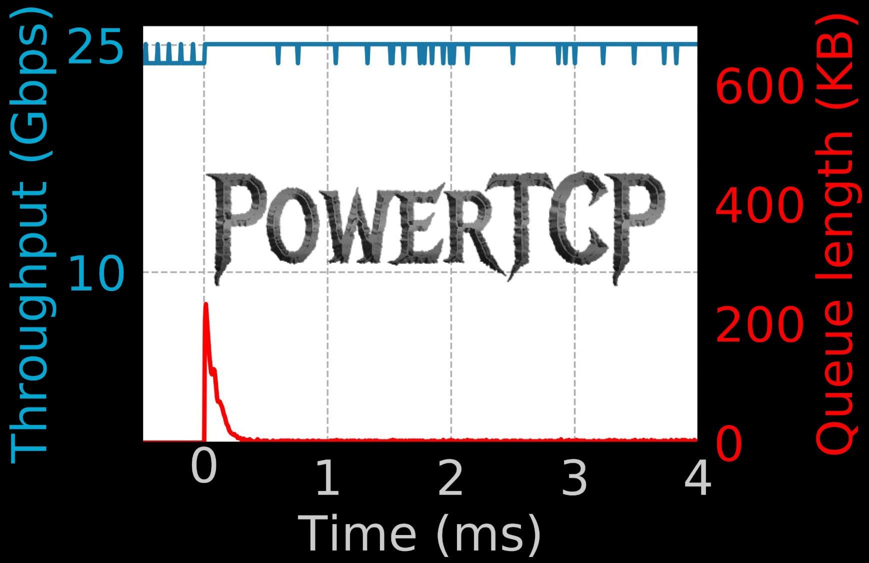
$$\underbrace{\Gamma}_{\text{Power}} = b^2 \times \underbrace{\theta}_{\substack{\text{Voltage} \\ \big\downarrow \\ \text{RTT}}} \times \underbrace{(\dot{\theta} + 1)}_{\substack{\text{Current} \\ \big\downarrow \\ \text{RTT gradient}}}$$

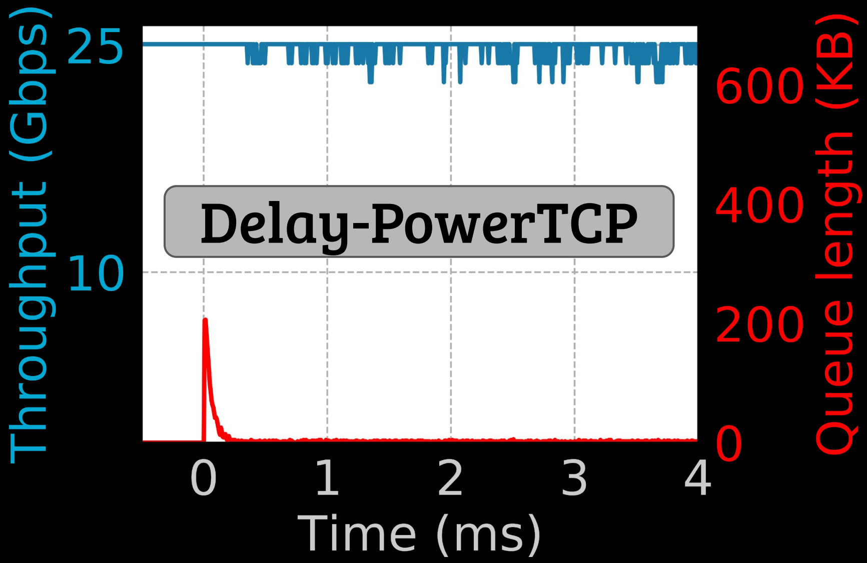
Evaluation

Evaluation - Incast

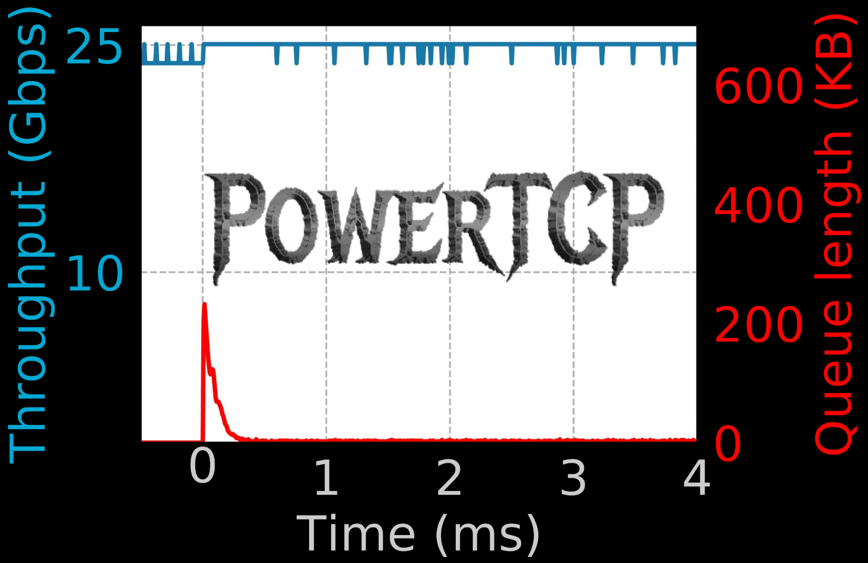


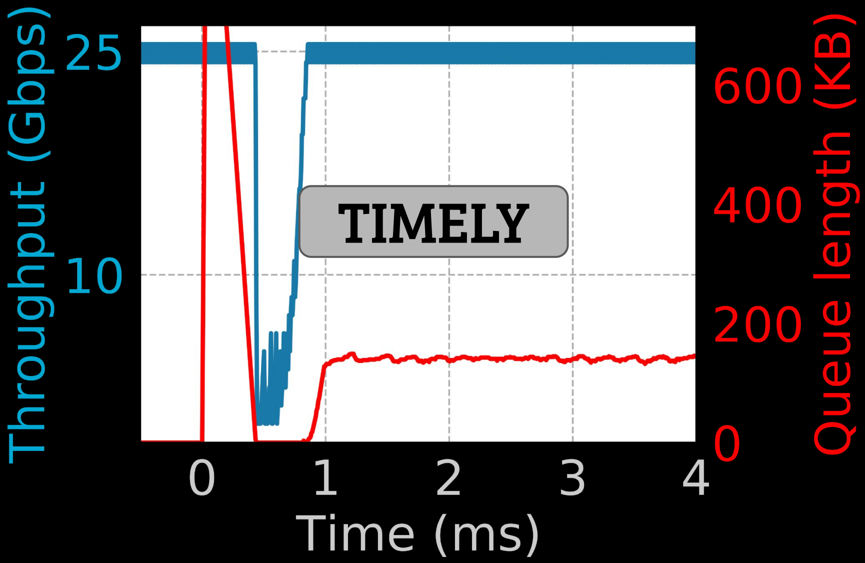
Evaluation - Incast



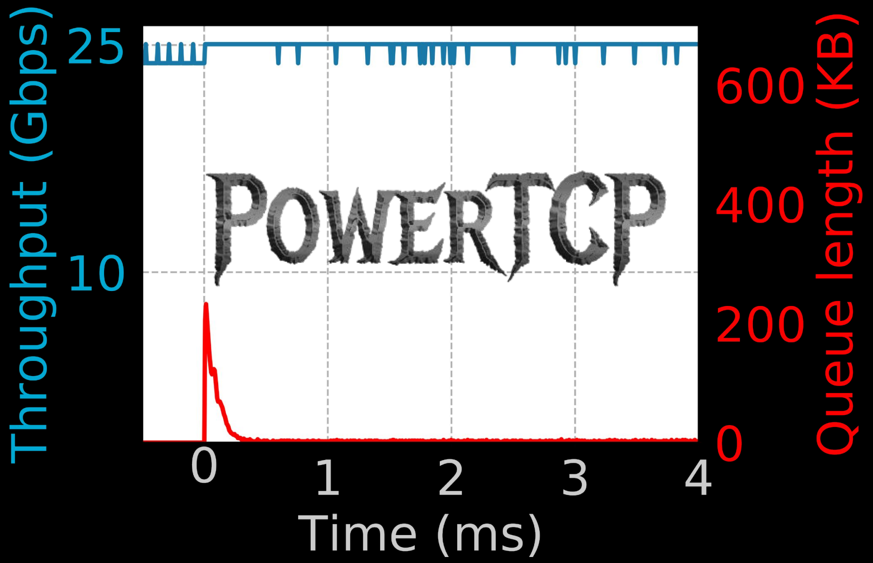


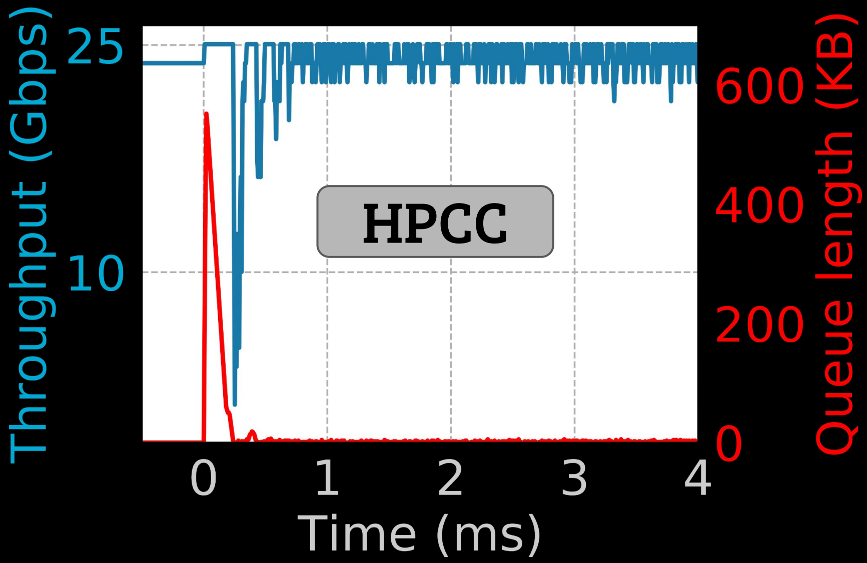
Evaluation - Incast



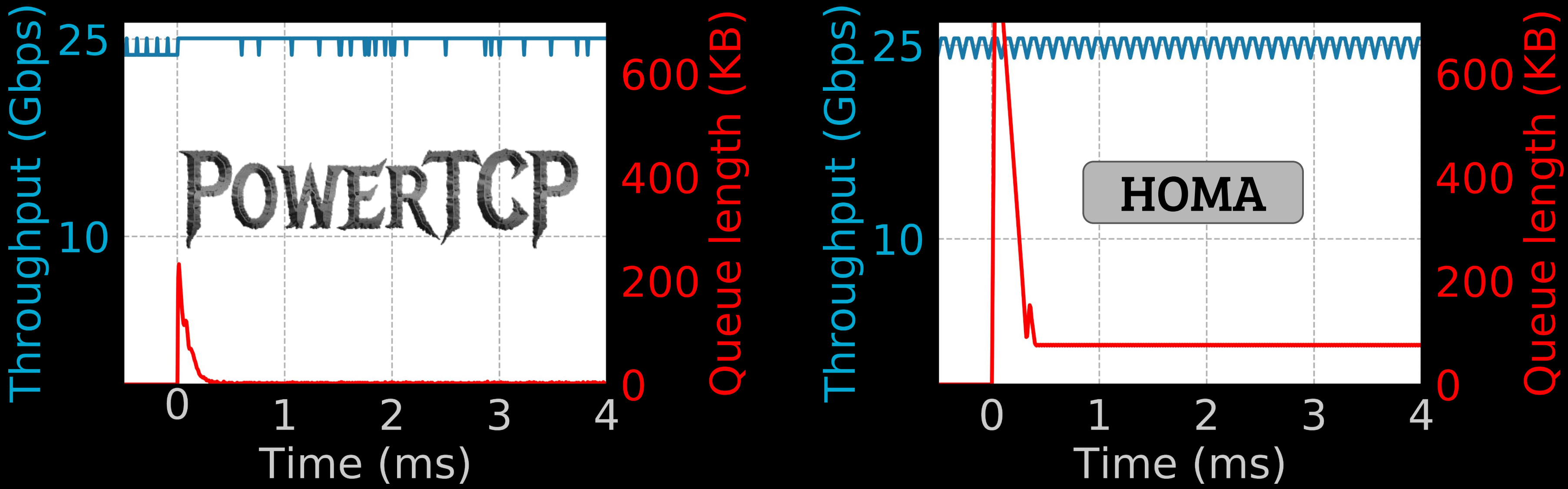


Evaluation - Incast

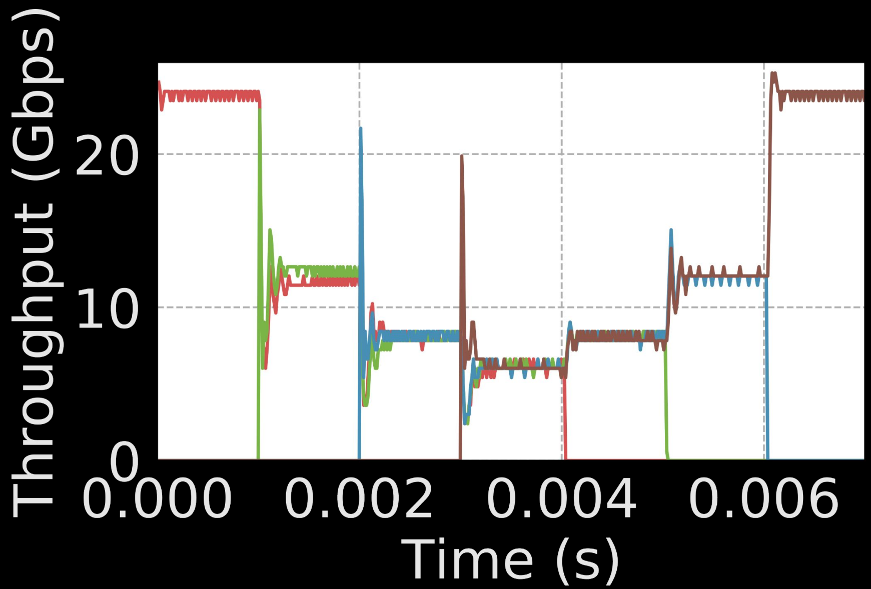




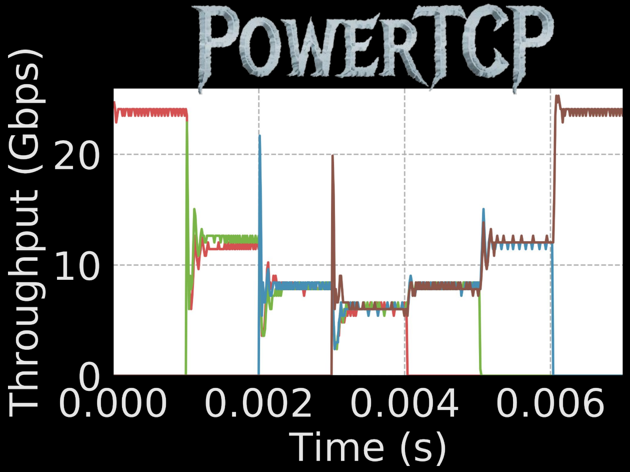
Evaluation - Incast

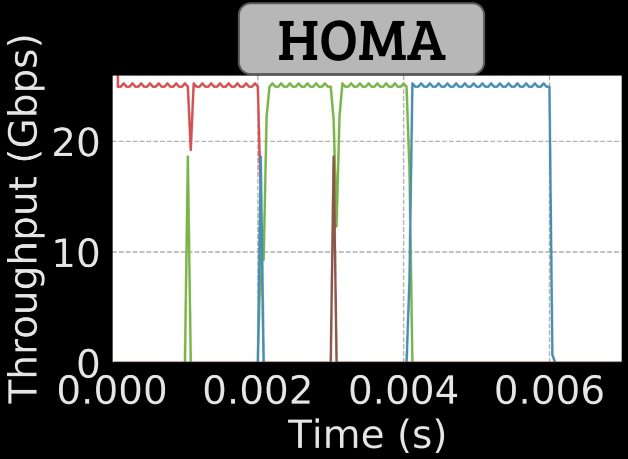


Evaluation - Fairness & Stability



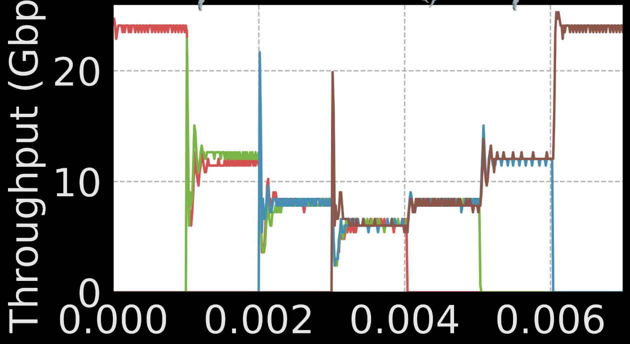
Evaluation - Fairness & Stability



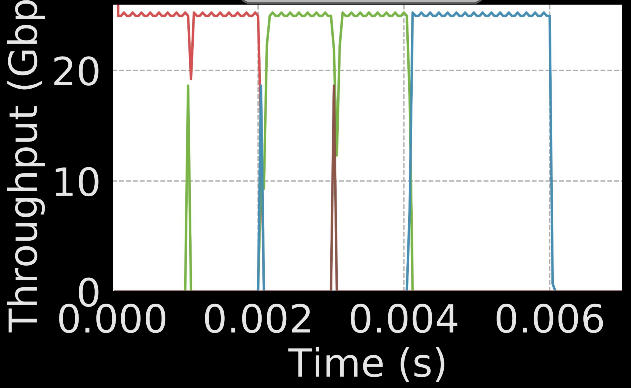


Evaluation - Fairness & Stability

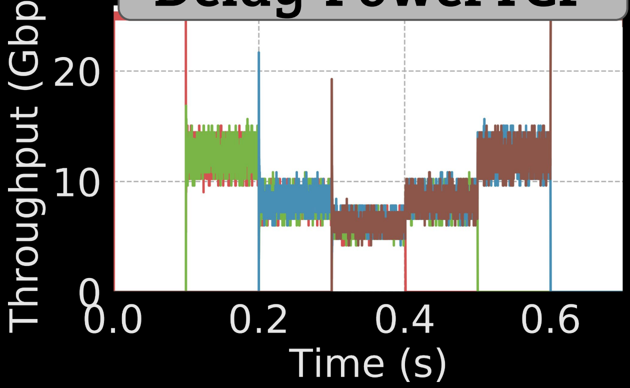
PowerTCP



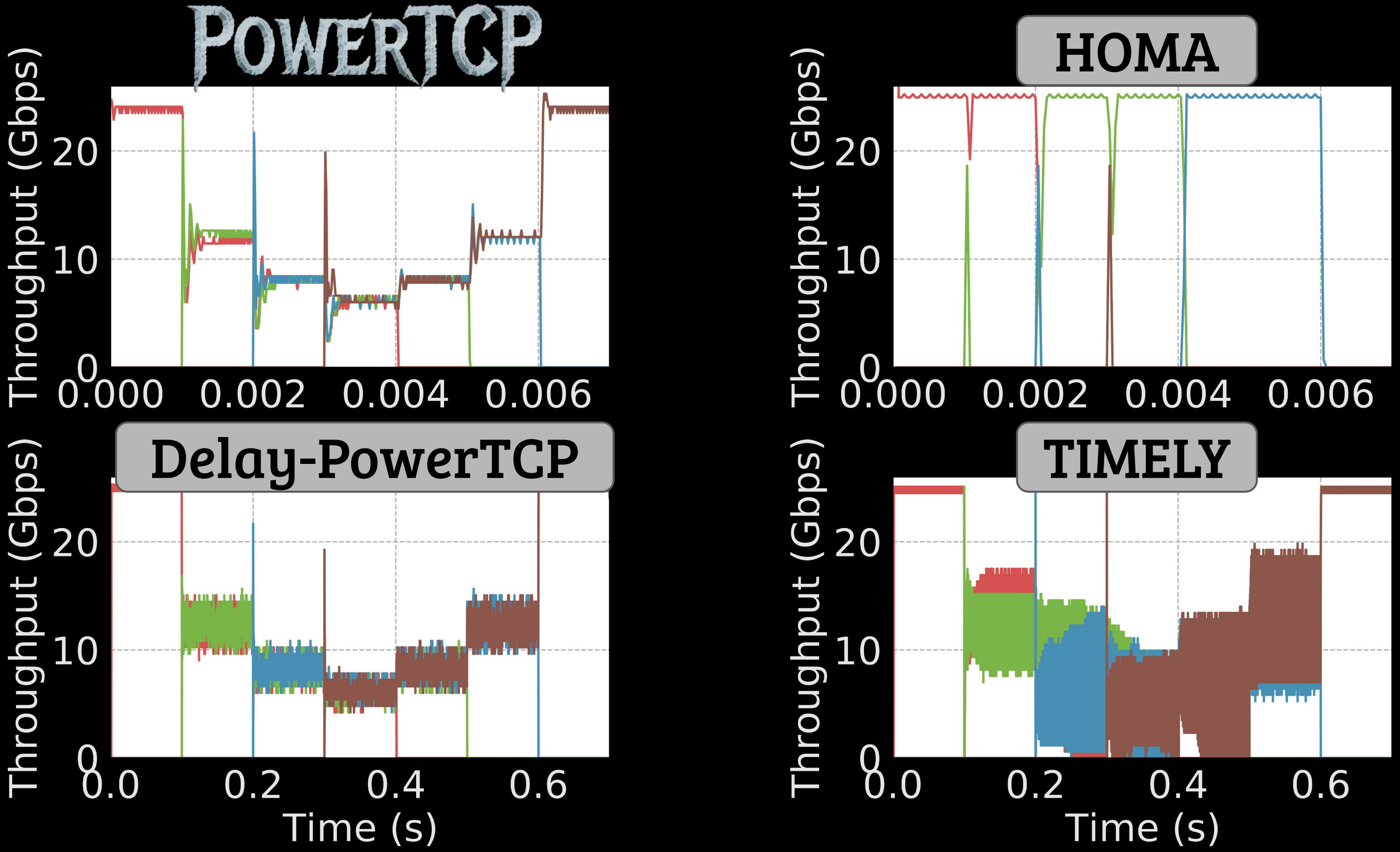
HOMA



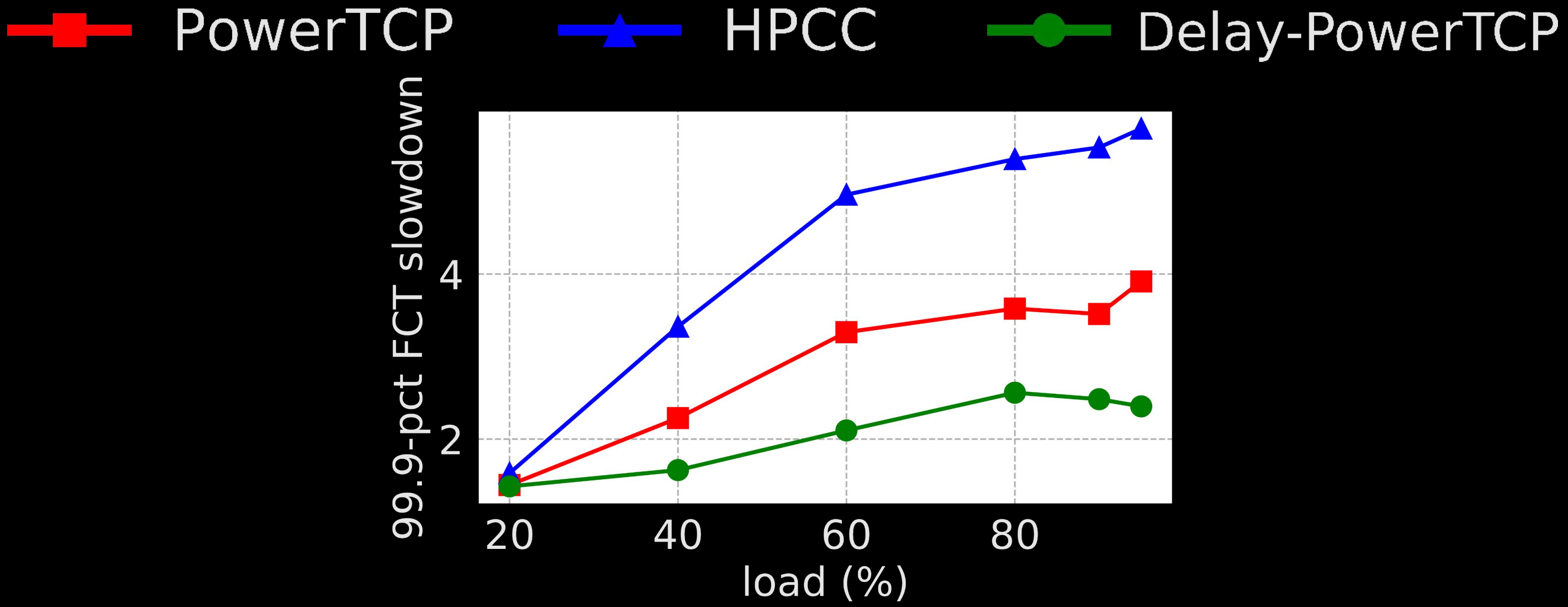
Delay-PowerTCP



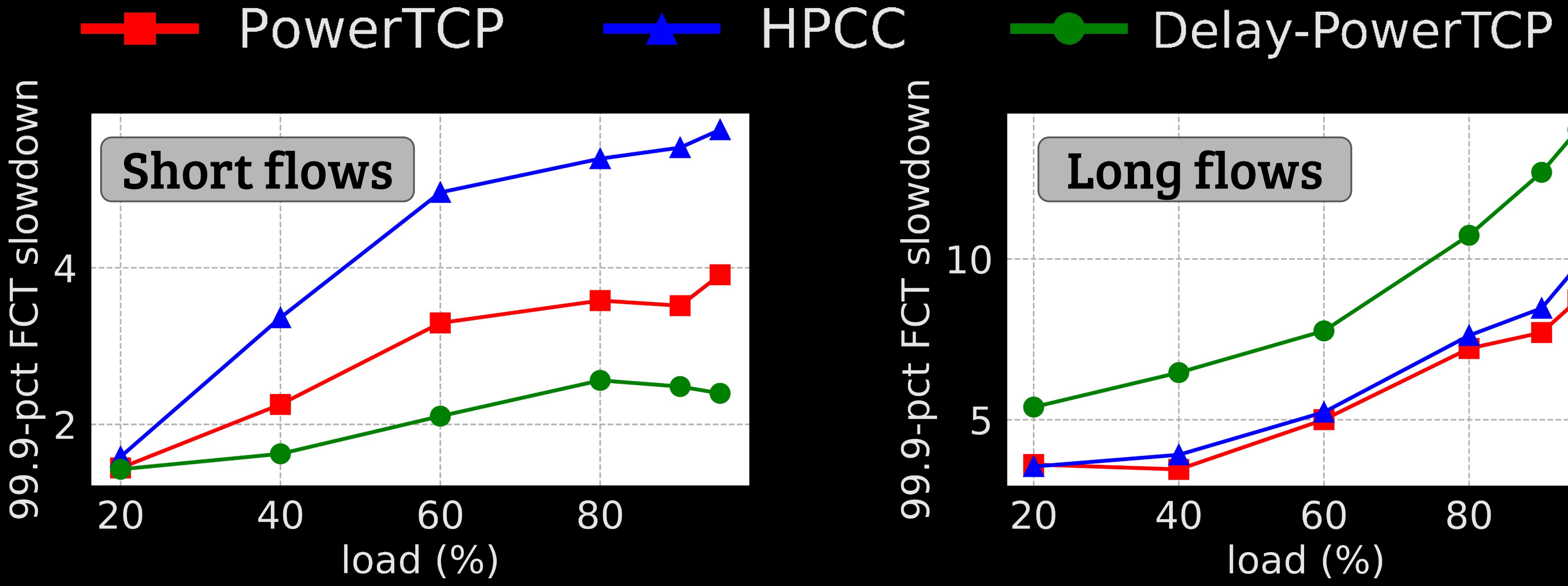
Evaluation - Fairness & Stability



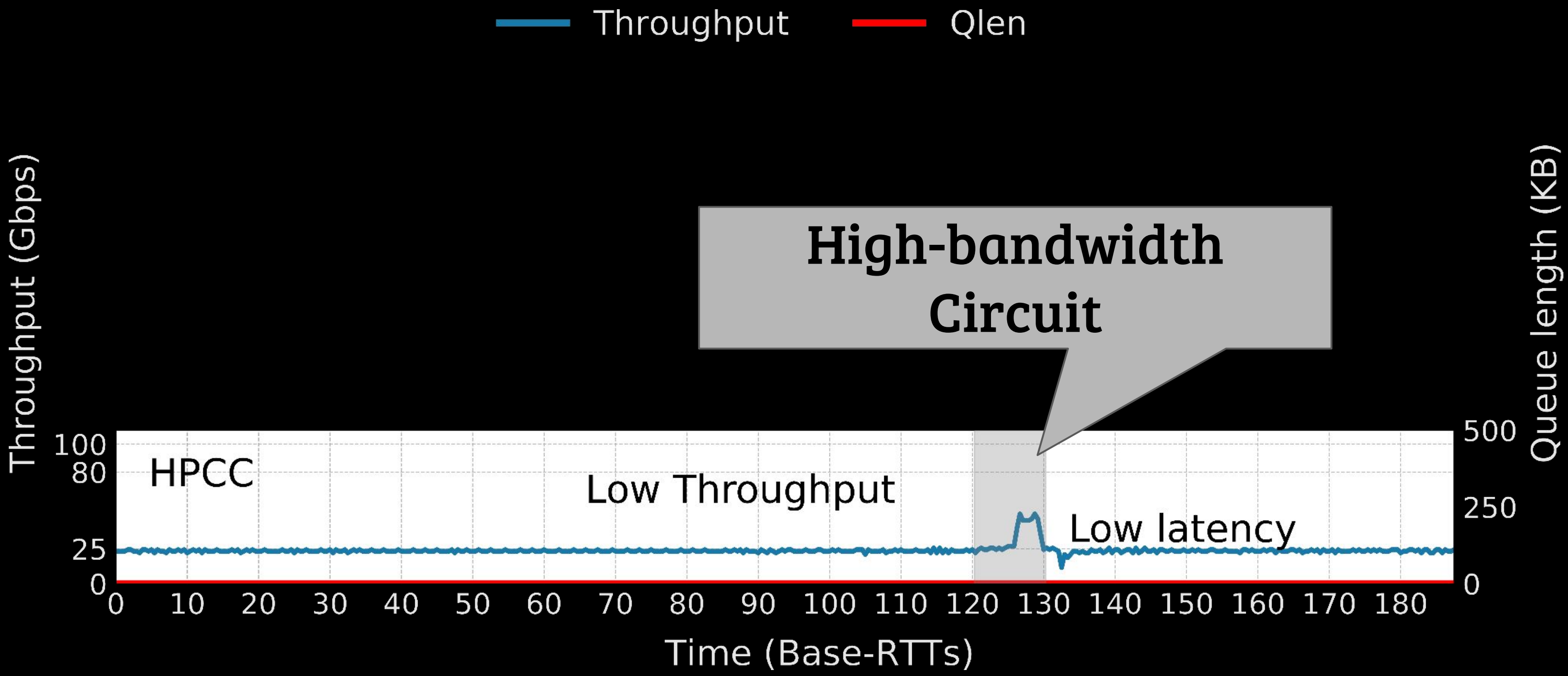
Evaluation - Workload



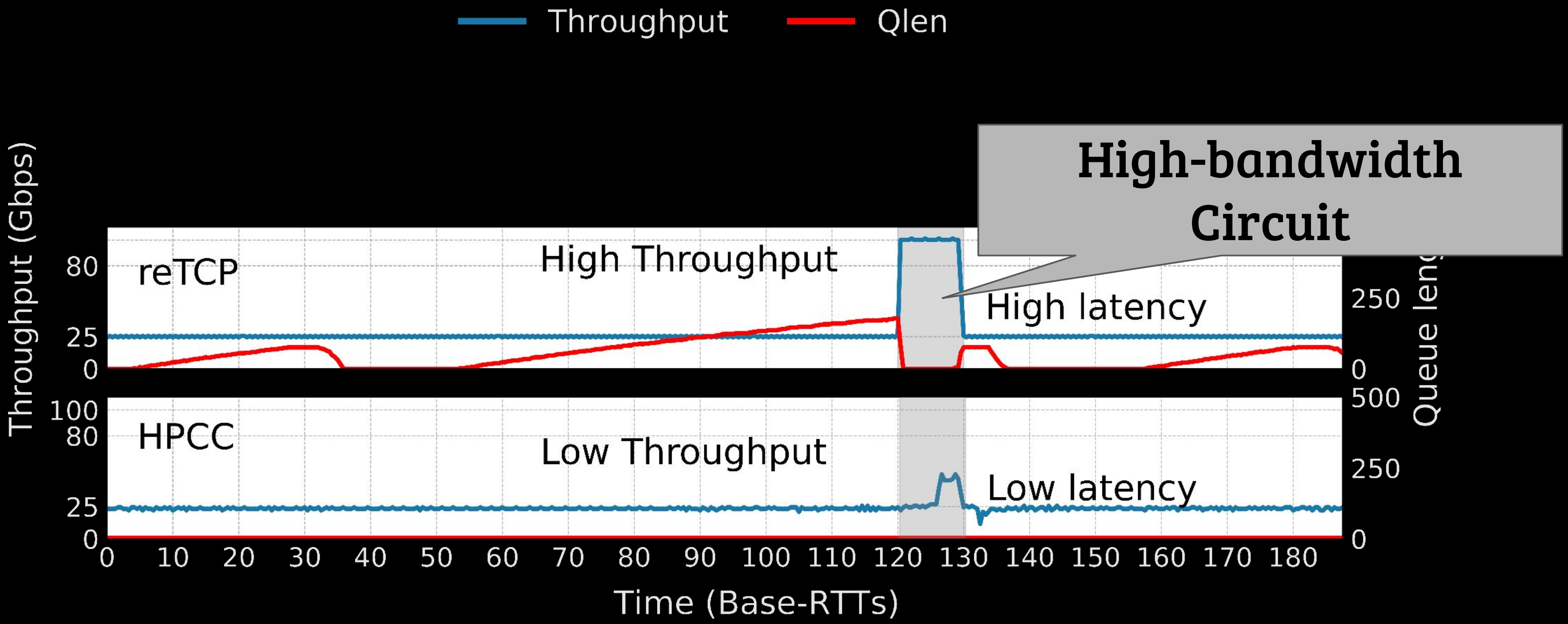
Evaluation - Workload



Evaluation - Reconfigurable Networks

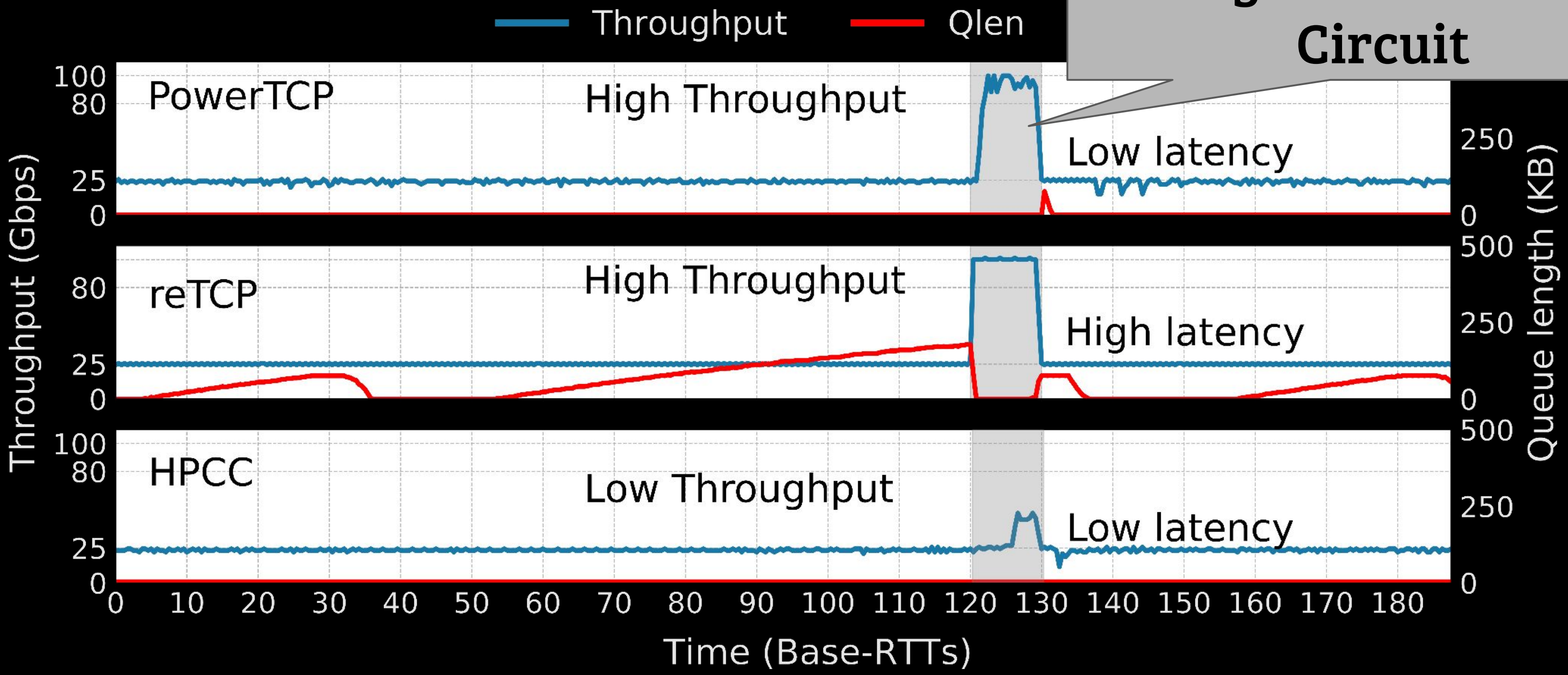


Evaluation - Reconfigurable Networks



High-bandwidth Circuit

Evaluation - Reconfigurable Networks



High-bandwidth Circuit

Conclusion

- Existing CC are fundamentally limited to a single dimension
- Power is an interesting and provably good measure for CC
- PowerTCP: a novel control law based on Power
- Improves FCTs for short flows and even for long flows

Thank you

References

[1] Alizadeh M, Greenberg A, Maltz DA, Padhye J, Patel P, Prabhakar B, Sengupta S, Sridharan M. Data center tcp (dctcp). InProceedings of the ACM SIGCOMM 2010 Conference 2010 Aug 30 (pp. 63-74).

[2] Alizadeh M, Yang S, Sharif M, Katti S, McKeown N, Prabhakar B, Shenker S. pfabric: Minimal near-optimal datacenter transport. ACM SIGCOMM Computer Communication Review. 2013 Aug 27;43(4):435-46.

[3] Bai W, Hu S, Chen K, Tan K, Xiong Y. One more config is enough: Saving (DC) TCP for high-speed extremely shallow-buffered datacenters. IEEE/ACM Transactions on Networking. 2020 Dec 9;29(2):489-502.

[4] Foerster KT, Schmid S. Survey of reconfigurable data center networks: Enablers, algorithms, complexity. ACM SIGACT News. 2019 Jul 24;50(2):62-79.

[5] Kumar G, Dukkipati N, Jang K, Wassel HM, Wu X, Montazeri B, Wang Y, Springborn K, Alfeld C, Ryan M, Wetherall D. Swift: Delay is simple and effective for congestion control in the datacenter. InProceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication 2020 Jul 30 (pp. 514-528).

[6] Li Y, Miao R, Liu HH, Zhuang Y, Feng F, Tang L, Cao Z, Zhang M, Kelly F, Alizadeh M, Yu M. HPCC: High precision congestion control. InProceedings of the ACM Special Interest Group on Data Communication 2019 Aug 19 (pp. 44-58).

References

[7] Mittal R, Lam VT, Dukkipati N, Blem E, Wassel H, Ghobadi M, Vahdat A, Wang Y, Wetherall D, Zats D. TIMELY: RTT-based congestion control for the datacenter. ACM SIGCOMM Computer Communication Review. 2015 Aug 17;45(4):537-50.

[8] Montazeri B, Li Y, Alizadeh M, Ousterhout J. Homa: A receiver-driven low-latency transport protocol using network priorities. InProceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication 2018 Aug 7 (pp. 221-235).

[9] Mukerjee MK, Canel C, Wang W, Kim D, Seshan S, Snoeren AC. Adapting TCP for reconfigurable datacenter networks. In17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20) 2020 (pp. 651-666).

[10] www.google.com/about/datacenters/gallery/