A Network Calculus Model for Congestion Control in Data Center Network

Natchanon Luangsomboon

The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

This talk

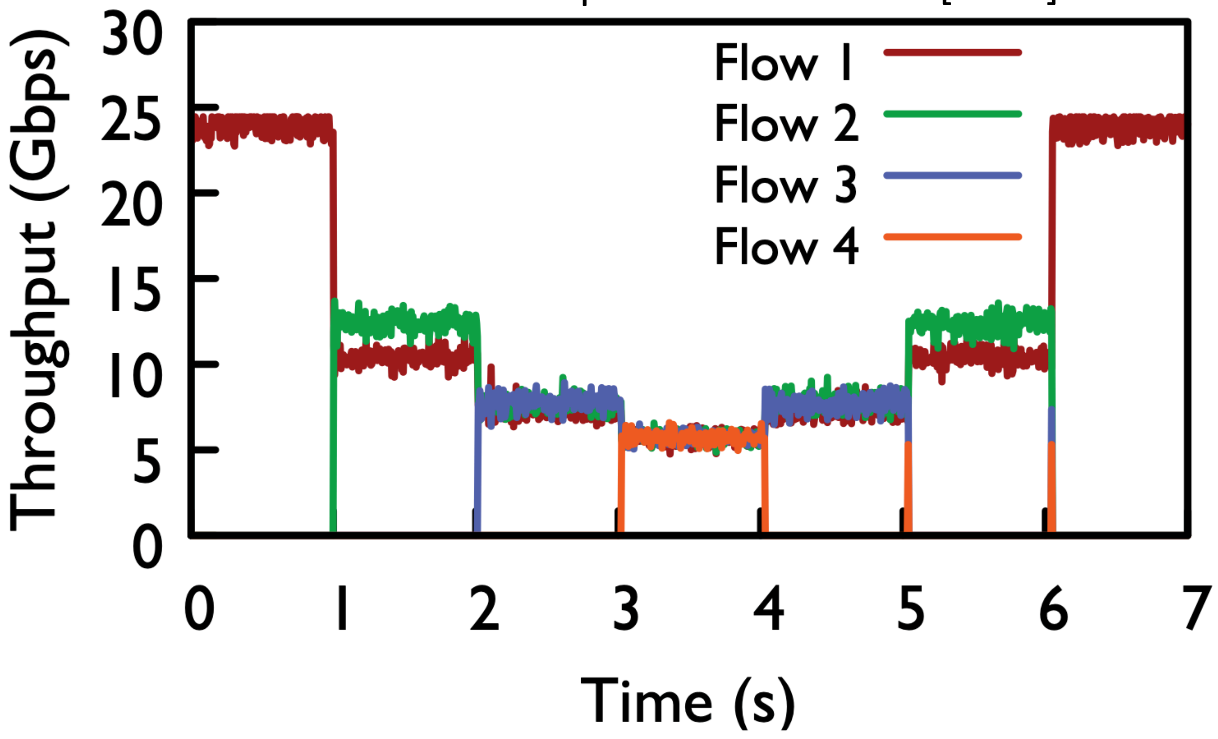
- Use network calculus to model data center network as multi-flow window flow control
- Analyze its bounds on transmission rate and fairness given the time-variable congestion windows

- Goal: a congestion control algorithm (CCA) that is fair and fully utilizes the link bandwidth based on the multi-flow model

Fairness is an important aspect of CCAs

- The fainess experiment is a common test for CCAs
- BBRv1 is criticized for being unfair when used with other CCAs, e.g., CUBIC, Reno
    - Its inter-protocol fairness has been improved in subsequent versions

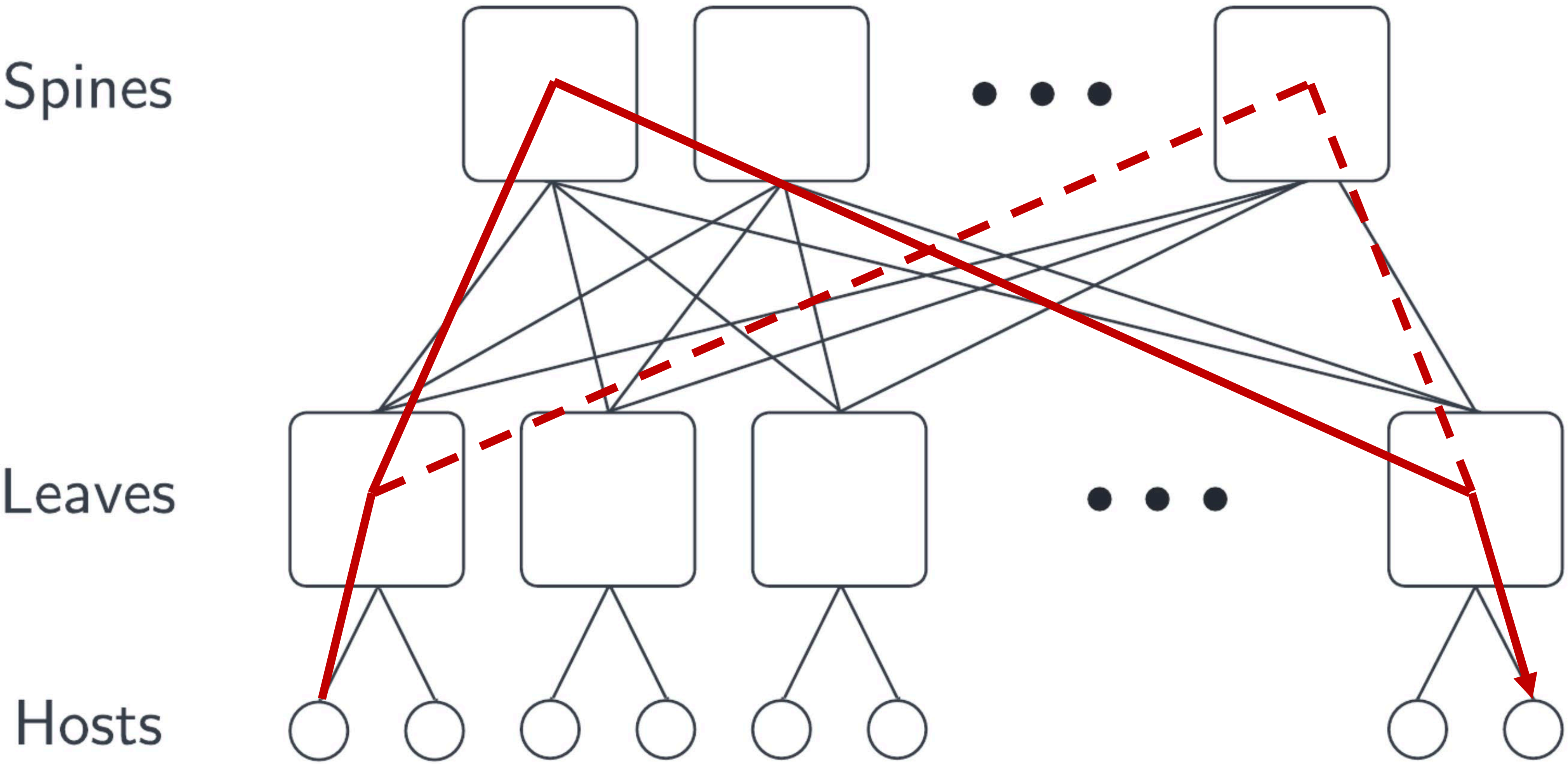
Fairness experiment of HPCC [Li'19]



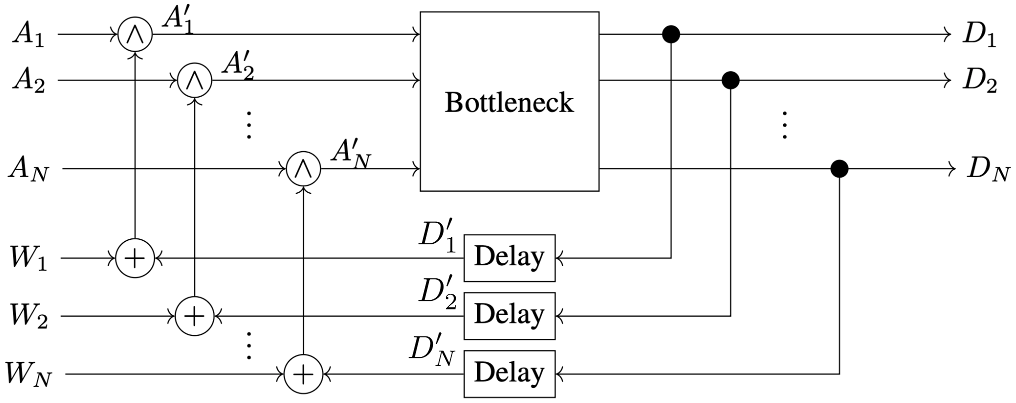
Related works

- Fairness and Stability of End-to-End Congestion Control [Kelly’03]
- Congestion Control for Large-Scale RDMA Deployments (DCQCN) [Zhu’15]
- Model-based Insights on the Performance, Fairness, and Stability of BBR [Scherrer’22]
- Toward Stability Analysis of Data Transport Mechanisms: A Fluid Model and Application (Reno & CUBIC) [Vardoyan’18]
- …

Data center network has homogeneous delay



Network model and notations



$$N_{\text{net}}(t) := \{i \in N \mid B_i(t) = W_i(t)\}$$

For each flow ,

- $A_i(t)$: Available data from the application
- $A'_i(t)$: Data transmitted from the sender
- $D_i(t)$: Data arrived at the receiver
- $D'_i(t)$: Acknowledgement
- $B_i(t)$: Inflight bytes
- $W_i(t)$: Congestion window
- : feedback delay
- : List of all flows
- $N_{\text{net}}(t)$: List of network limited flows

Additional bottleneck switch assumptions

- The bottleneck switch transmits data in a FIFO order, i.e., for all ,

$$A'(s) \le D(t) \iff \forall i, A'_i(s) \le D_i(t)$$

- The switch transmits one packet at a time

$$\forall s \exists t, A'(s) = D(t) \quad \text{and} \quad \forall s \exists t, A'_i(s) = D_i(t)$$

Swith transmits at s

Remains idle for ε

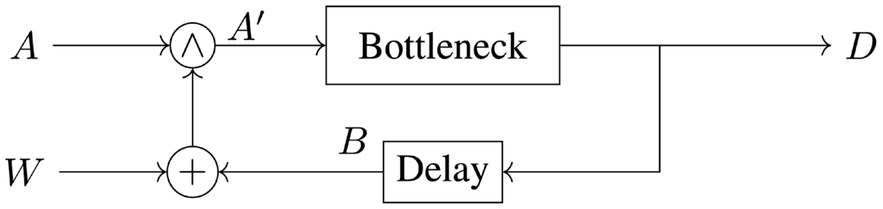
- Each transmission is at least apart

$$\boxed{\forall s, D(s) < \lim_{v \to s^+} D(v)} \implies \boxed{\lim_{v \to s^+} D(v) = D(s + \epsilon)}$$

Minimum W_i to “fill the pipe” (BDP)

Given upper/lower service curves at the switch, what is the minimum W_i s.t. the entire network has the same upper/lower service curves?

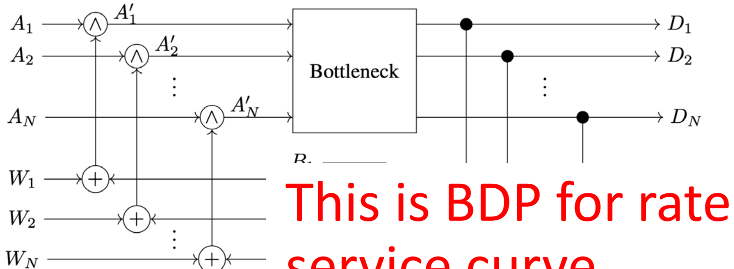
Single-flow case [Agrawal’99]



$$A = \sum_i A_i, D = \sum_i D_i, \ldots$$

$$W = \sum_i (A'_i - B_i)$$

Multi-flow case



This is BDP for rate service curve

$$\forall t, W(t) \geq \sum_{t\in\mathbb{R}} \{\underline{S}(t) - \underline{S} \otimes \underline{S}(t-d)\} \implies \underline{S} = \underline{S}_{\text{net}}$$

$$\forall t, W(t) \geq \sum_{t\in\mathbb{R}} \{\overline{S}(t) - \overline{S} \otimes \overline{S}(t-d)\} \implies \overline{S} = \overline{S}_{\text{net}}$$

Max-min fairness

- Each flow is associated with a weight ϕ_i
- If two flows , are network-limited throughout an interval

$$\left| \frac{D_i(s,t)}{\phi_i} - \frac{D_j(s,t)}{\phi_j} \right| \leq X$$

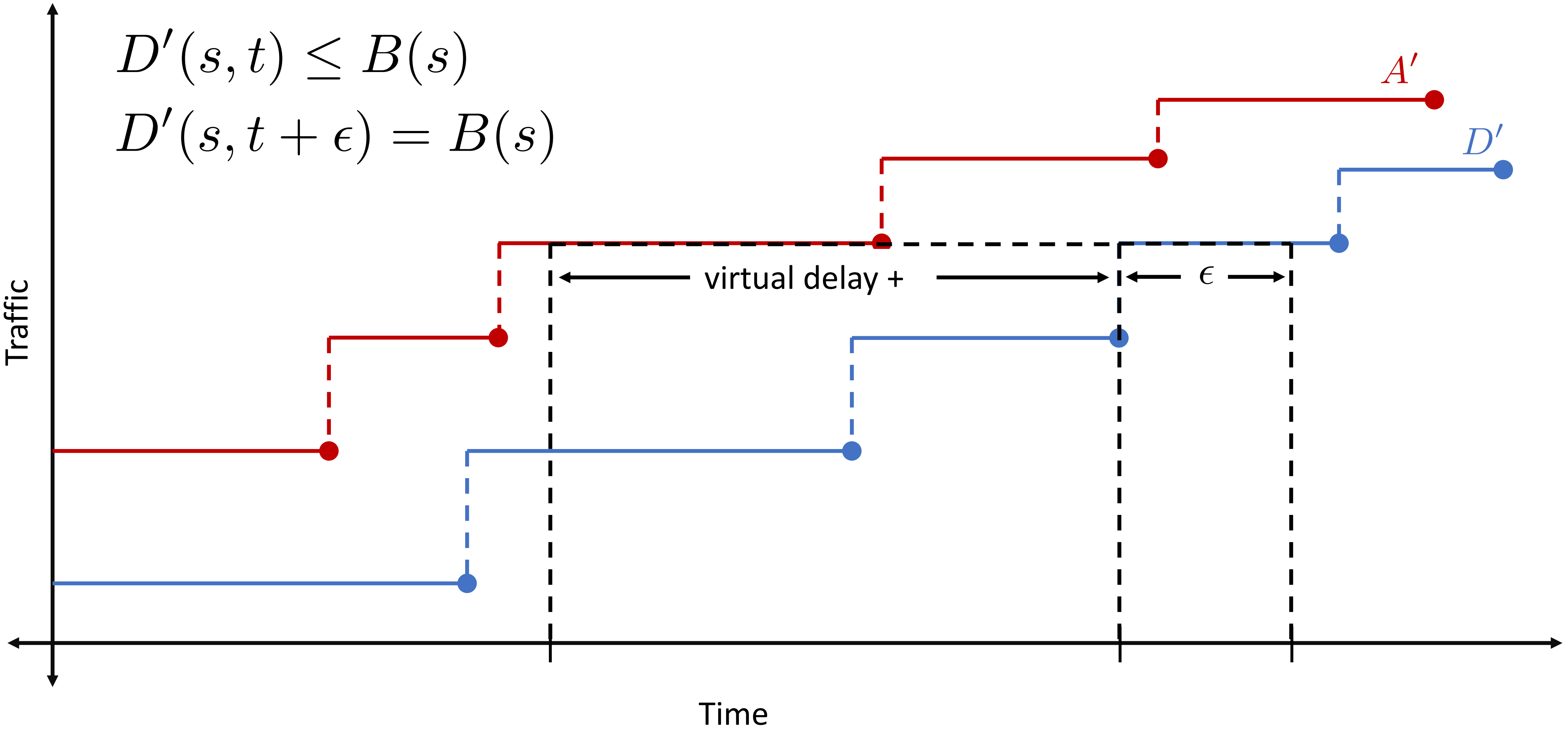
$$\Updownarrow$$

$$\left| \frac{D_i'(s-d,t-d)}{\phi_i} - \frac{D_j'(s-d,t-d)}{\phi_j} \right| \leq X$$

Relationship between backlog and transmission over a virtual delay

$$D'(s,t) \leq B(s)$$

$$D'(s,t+\epsilon) = B(s)$$

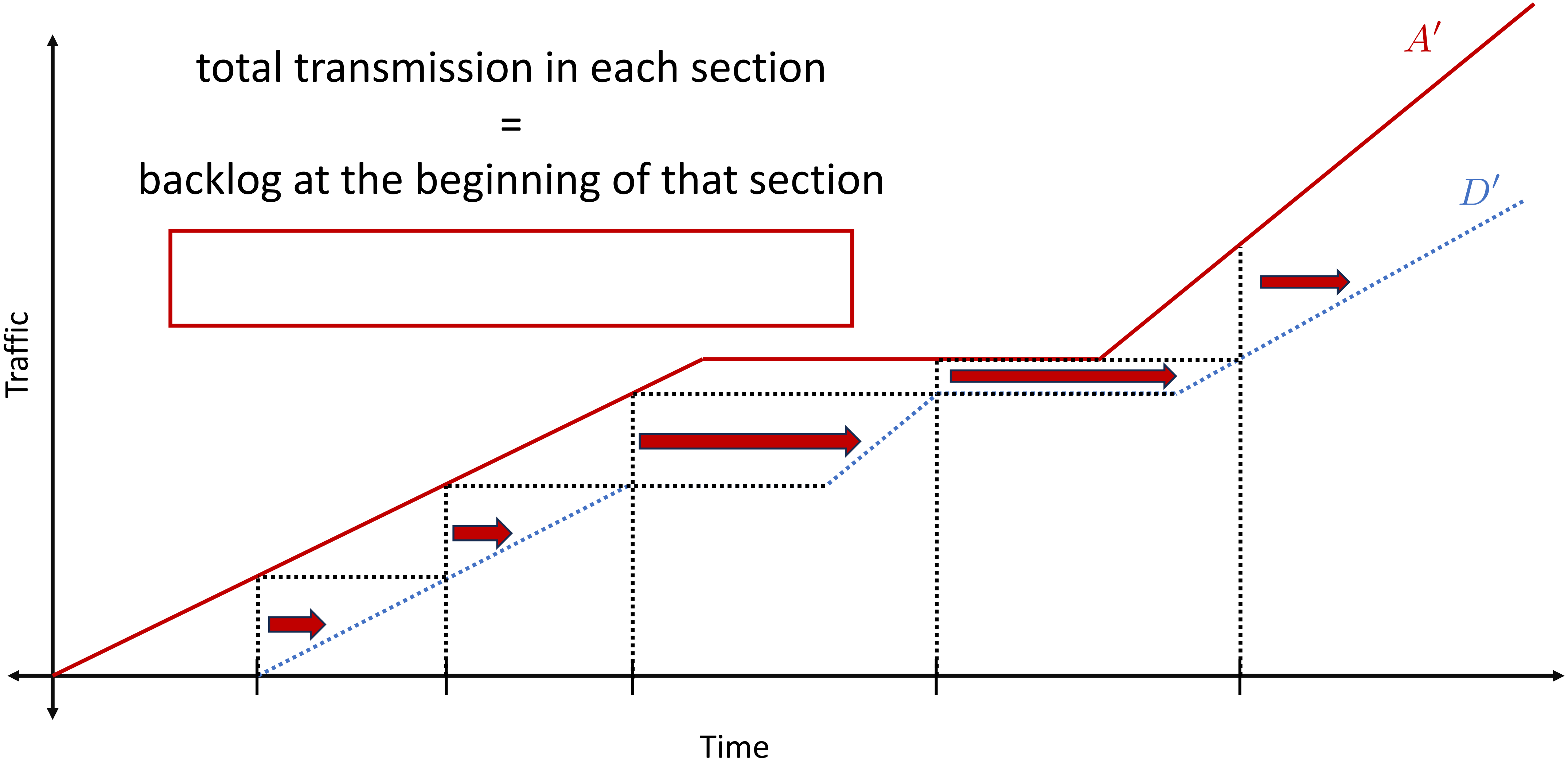


Splitting the timeline into sections

total transmission in each section

=

backlog at the beginning of that section



Pivot definition

- Each flow defines pivots $p_0^i, p_1^i, p_2^i, \ldots$ and $P_i = \{p_k^i\}_{k \in \mathbb{N}_0}$ s.t.

$$A_i(p_0^i) = 0 \qquad p_{k+1}^i = \inf\{\tau \geq p_k^i \mid A'(p_k^i) \leq D(\tau)\} + \epsilon + d$$

p_k^i + virtual delay

- From the definition,

$$\Longrightarrow D_i'(p_{k+1}^i) = A_i'(p_k^i)$$

$$D_i'(p_k^i, p_{k+1}^i) = A_i'(p_k^i) - D_i'(p_k^i) = B_i(p_k^i)$$

$$\boxed{\forall m \leq n,\ D_i'(p_m^i, p_n^i) = \sum_{k=m}^{n-1} D_i'(p_k^i, p_{k+1}^i) = \sum_{k=m}^{n-1} B_i(p_k^i)}$$

Pivots from different flows interleave each other

- For any flows , and indices , where $p^i_k \le p^j_l \le p^i_{k+1}$,

$$D'(p^i_{k+1}) = A'(p^i_k) \le A'(p^j_l) = D'(p^j_{l+1}) \quad \Longrightarrow \quad p^i_{k+1} \le p^j_{l+1}$$

$$\Longrightarrow \quad p^i_k \le p^j_l \le p^i_{k+1} \le p^j_{l+1} \le p^i_{k+2} \le p^j_{l+2} \le p^i_{k+3} \le \cdots$$

$$P_i(s,t) := \{u \in [s,t] \mid u \in P_i\}$$

$$\Downarrow$$

upper- and lower-bounds over an interval

- Given interval , let and
- If contains no pivot, it is between and
- Otherwise, $D_i(p_m, p_n) \leq D_i'(s,t) \leq D_i'(p^i_{m-1}, p^i_{n+1})$ and so

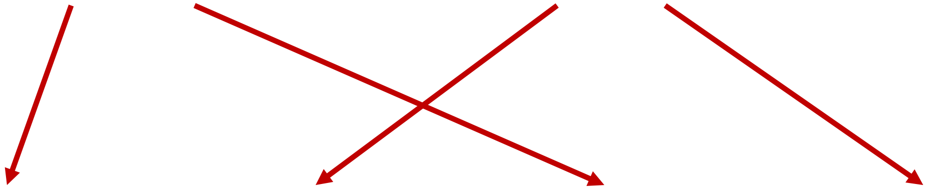
Bound on difference over an interval

- Let and

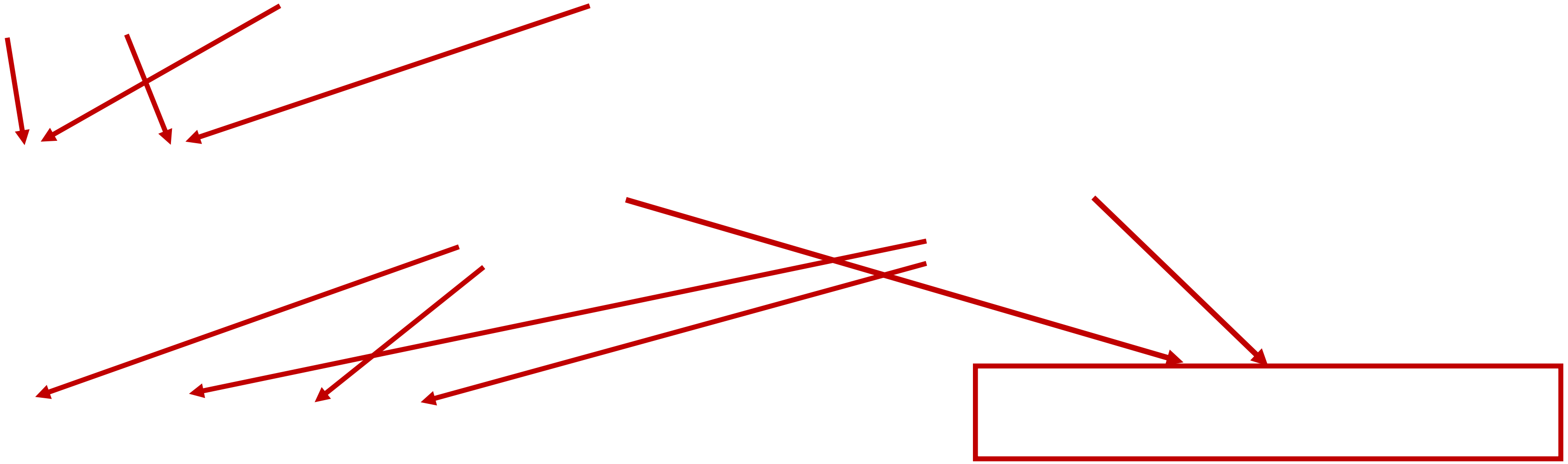
Bound on max-min fairness

- If for some constants

  then the network is max-min fair

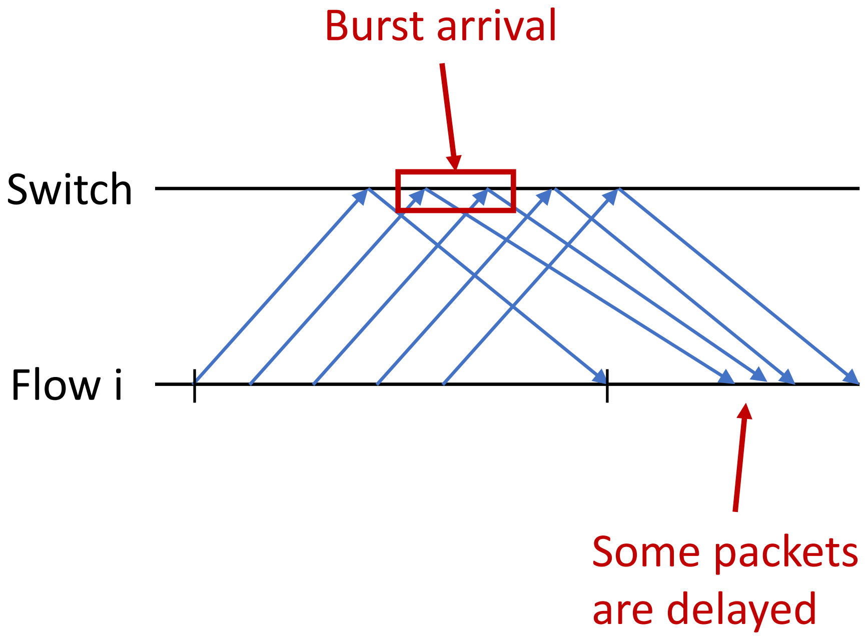


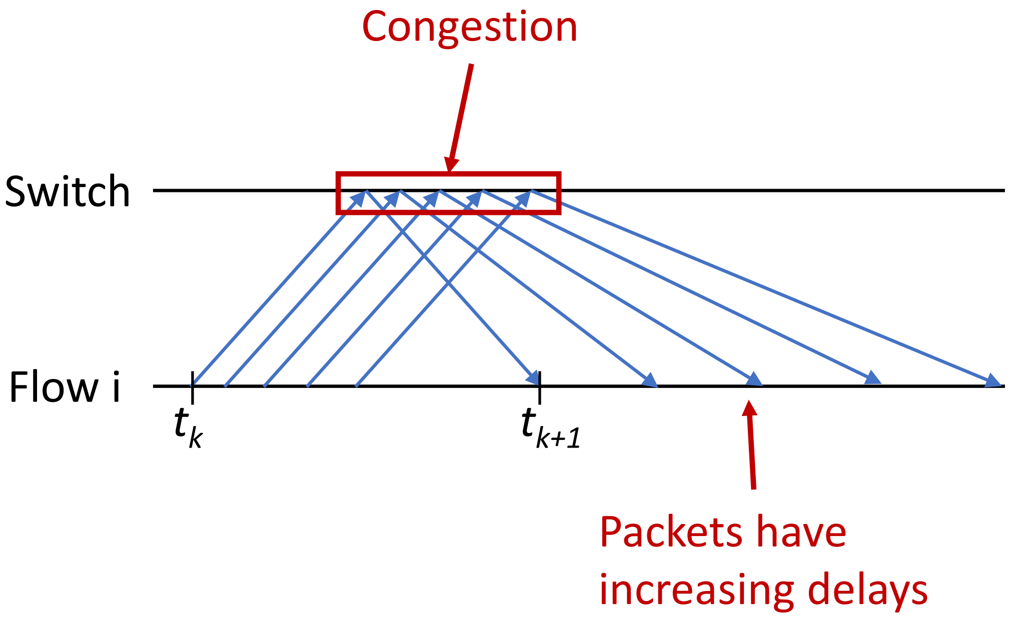
Bound on max-min fairness (Con’t)



Using the ACK rate to differentiate short-term and long-term bursts

- We can utilize that fairness bound only depends on at pivots
  - Flow i transmits data at a fixed rate for
  - Acknowledgement rate can identify the network status





Approximating incast traffic using ACK rate

- For a work-conserving bottleneck switch with rate , If the arrival A_i' is a constant bit rate (CBR), so is the departure D_i' [Yashar'09]

$$A_i'(t) = C_i t \implies D_i'(t) = \begin{cases} C_i t & \text{, if } C \geq \sum_i C_i \\ \frac{C_i}{\sum_i C_i} C t & \text{, if } C < \sum_i C_i \end{cases}$$

- If $A_i'(t) = C_i t$ and $D_i'(t) = R_i[t-d]^+$ for some R$_i$

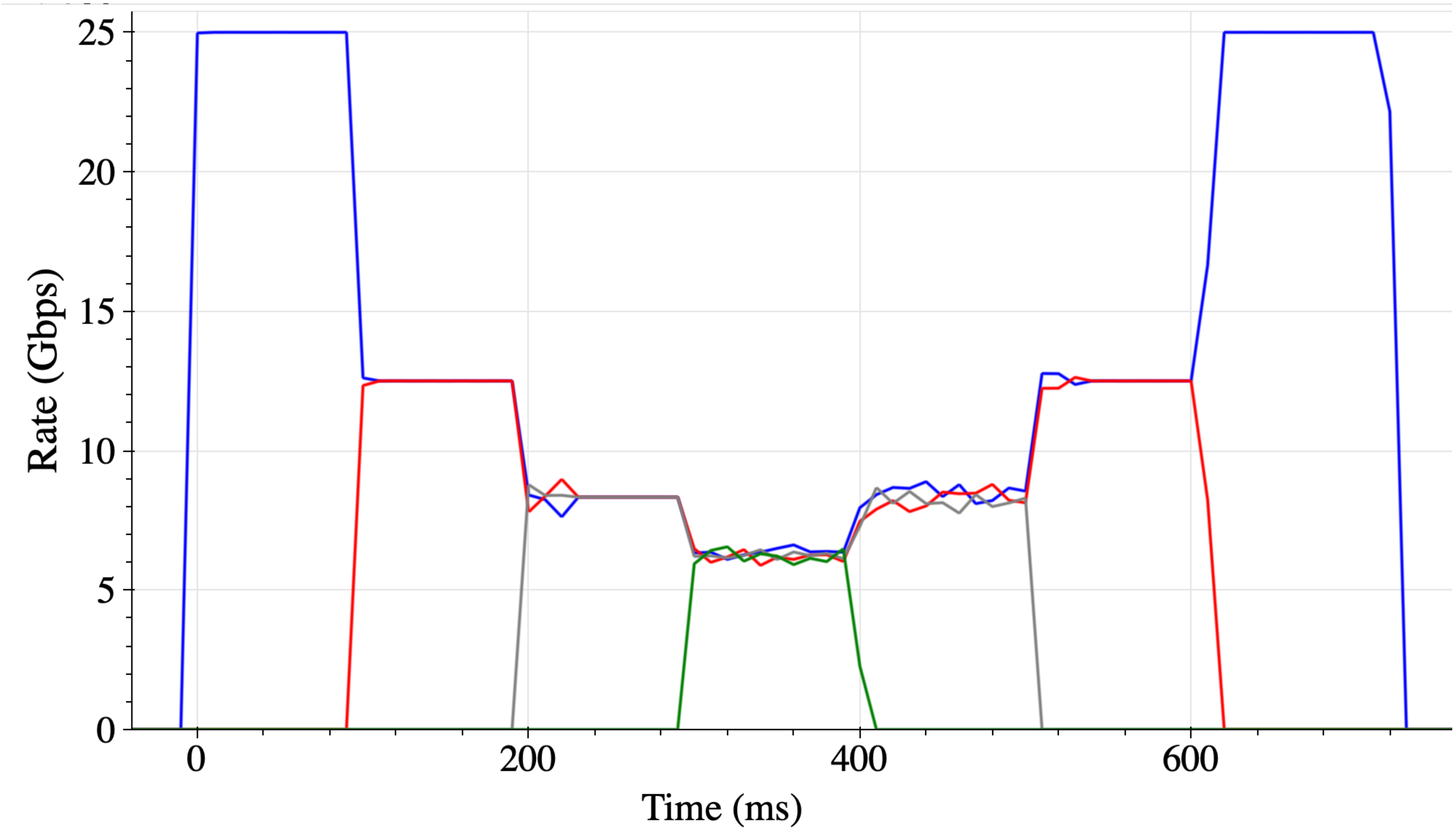
$$R_i = C_i \implies C \geq \sum_i C_i$$

$$R_i < C_i \implies R_i = \frac{C_i}{\sum_i C_i} C \longrightarrow \boxed{\sum_i C_i = \frac{C_i C}{R_i}}$$

Rate-check congestion control

- The sender operates in rounds
- The first packet transmission of each round is a pivot
- The sender transmits at rate C_r in each round r for a duration of d
- Inflight bytes at the beginning of round r is $C_{r-1}d$
- At the beginning of round r, approximate R_{r-2}

Fairness experiment



Conclusion

- We modeled data center network as multi-flow window flow control
- CCAs “fill the pipe” if the total congestion windows over network-limited flows exceed BDP
- CCAs is max-min fair if the congestion windows <u>at pivots</u> converge toward values proportional to their weights
    - The convergence does not depend on congestion windows outsize of the pivots

Future work

- Allow the feedback delay for each flow to differ from one another
- Relax the absolutely converge requirements

Thank you!