CS 925

Lecture 5

Traffic Management

Tuesday, February 6, 2024

Little's formula

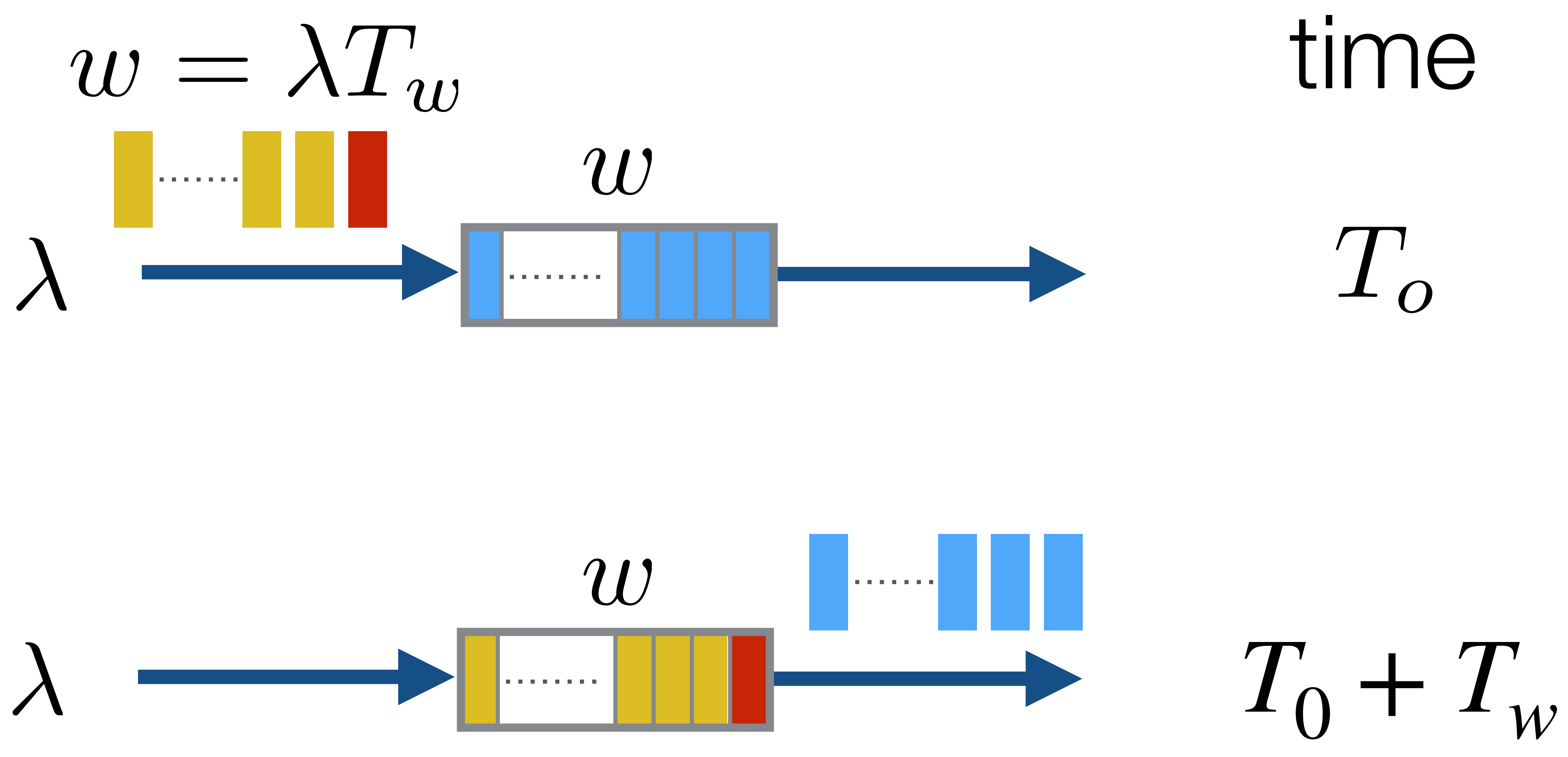
- Warning: lots of handwaving follows

$$w = \lambda T_w$$

- Assume the system is stationary (can be modeled by a stationary process)
  - observing an item: the number of items in the system is, on average, the same when it enters as it is when it leaves
  - it takes T_w from entry to departure, during that period λT_w items arrive (and depart) to maintain the constant number of items in the system
  - assuming FIFO, all items that were in the system prior to the arrival of the observed one must be gone, so the system contains λT_w of the items

Little's formula

- The question remains: Is the system stationary?
  - under what conditions?
  - does it converge to that state?



Networks of Queues

- Traffic partitioning and merging, queues in tandem,…
- Jackson's Theorem (1963):
  - Assuming:
    - nodes provide independent service
    - Poisson arrivals from outside
    - fixed partitioning probability
    - no transport delay
  - then, mean delays can be added together
- … not really: the theorem does not hold, an error in the proof was found in 2003!

Network Performance

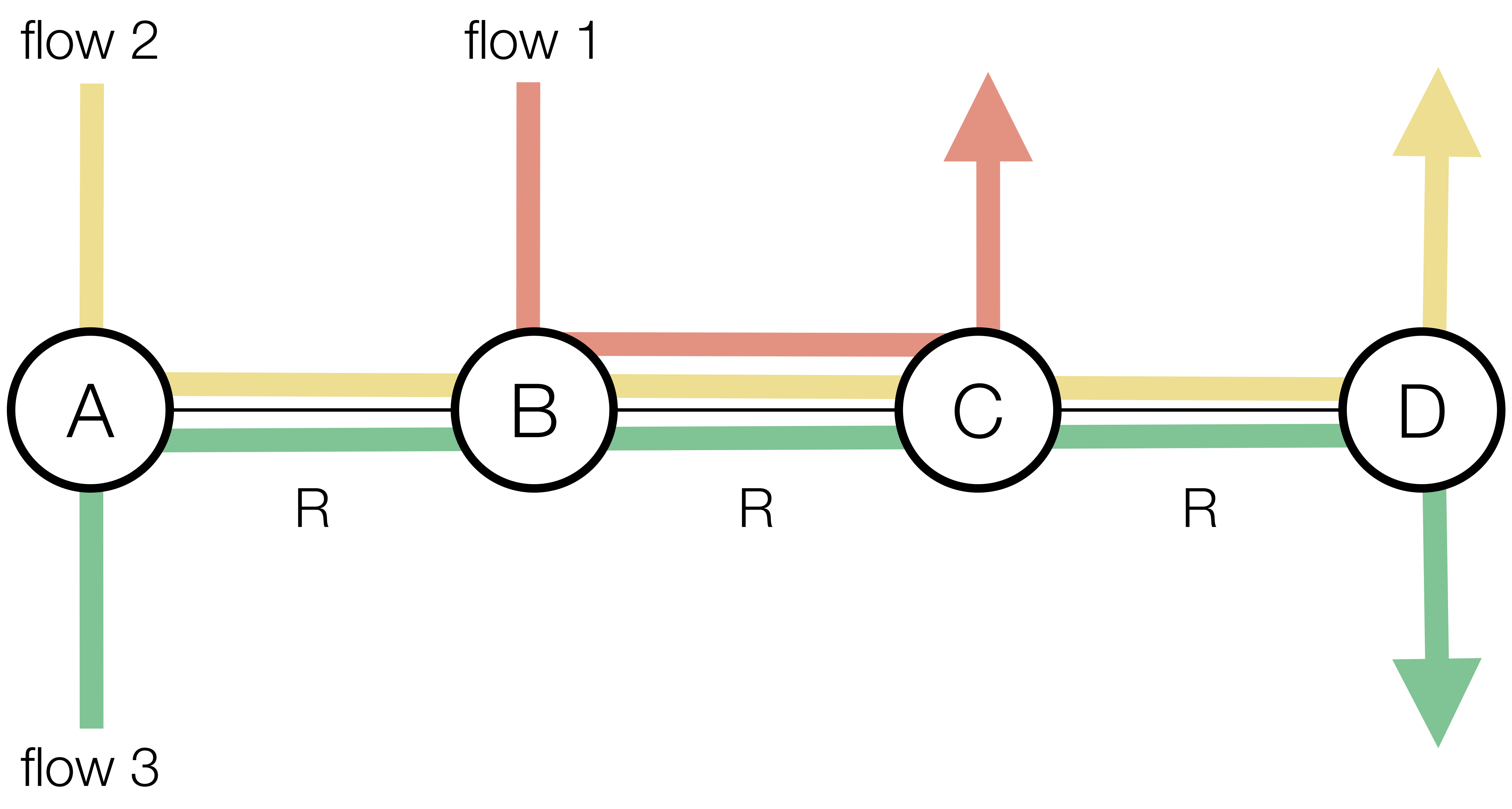
- Load vs Latency diagram
  - impact of load on the delay in delivery
- Offered vs Carried Load diagram
  - impact of load on effective throughput
- Loss vs Throughput diagram
  - impact of packet loss on throughput

Traffic Management

Traffic Management

- A process to:
  - maximize utility
  - ensure fairness
  - deliver “quality” service (QoS, QoE, …)
- Where?
  - transport (e.g., TCP window)
  - network (e.g., obsolete ICMP Source Quench)
  - link (e.g., Data Center Bridging)
  - application (e.g., HTTP/2, HTTP/3)

Utilization vs fairness



	Max utilization	Max fairness
flow 1	0	R/3
flow 2	R/2	R/3
flow 3	R/2	R/3

R - link rate

Layers of Traffic Management

- Within a device (router/switch/host)
  - what to do to deliver desired results?
- Within a protocol (protocol layer)
  - how instruct individual devices what they are supposed to do?
- Within a network
  - how to ensure that appropriate level of service is delivered?

Considerations

Obvious but worth reminding ourselves:

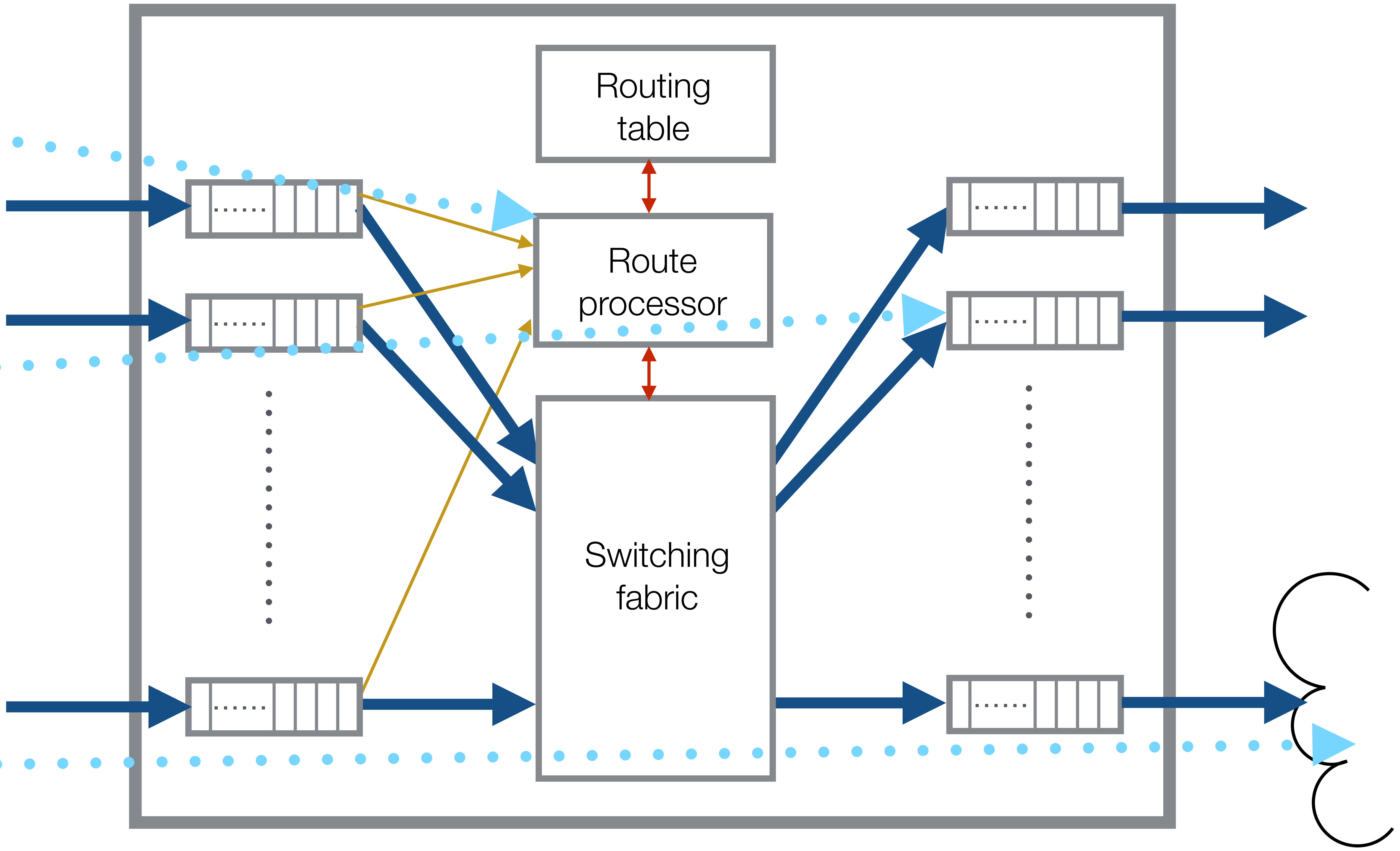
- In low load situations everyone gets the best service possible
  - unless we don't want to create unrealistic performance expectations
- In high load situations, better service for some means degraded service for others
  - how to determine who "deserves" better service?
  - greater good or more profit?

Questions

- How do we know what to do?
  - methods and techniques that translate user/application demands in traffic management objectives
- How do we instruct the network elements?
  - protocols to facilitate network management information exchange
- What do the network elements need to do?
  - methods and techniques through which the traffic management is implemented in the network

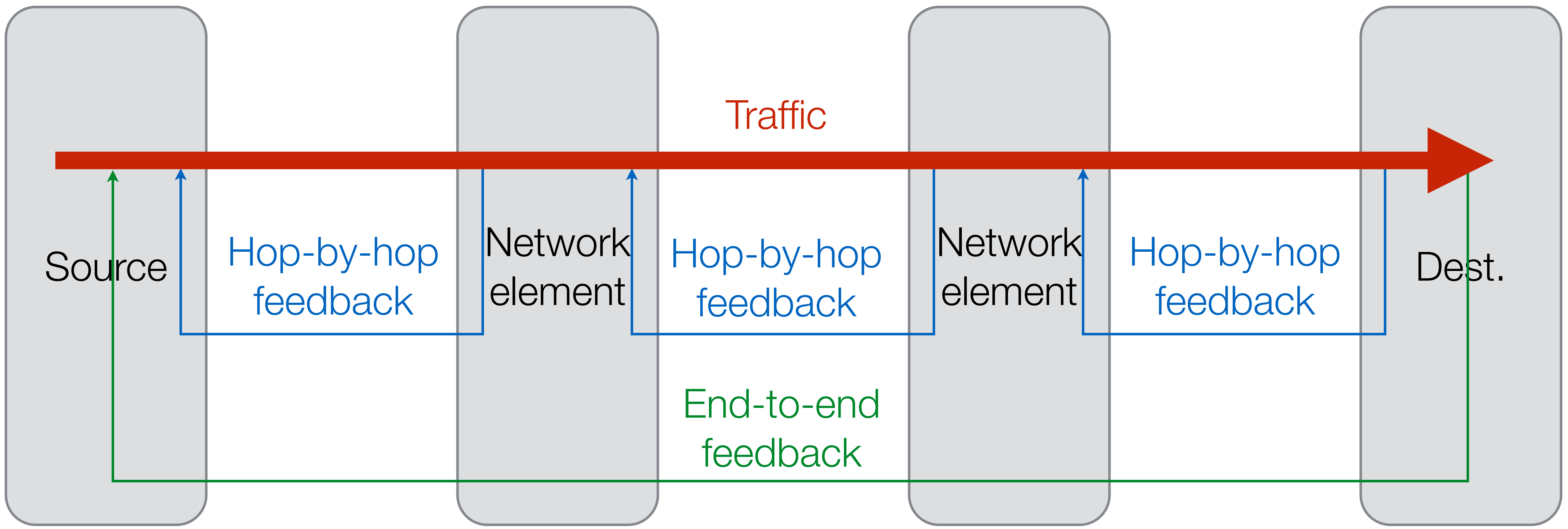
Router/Switch Actions

- Route selection
- Queueing policy
- Subnetwork service request

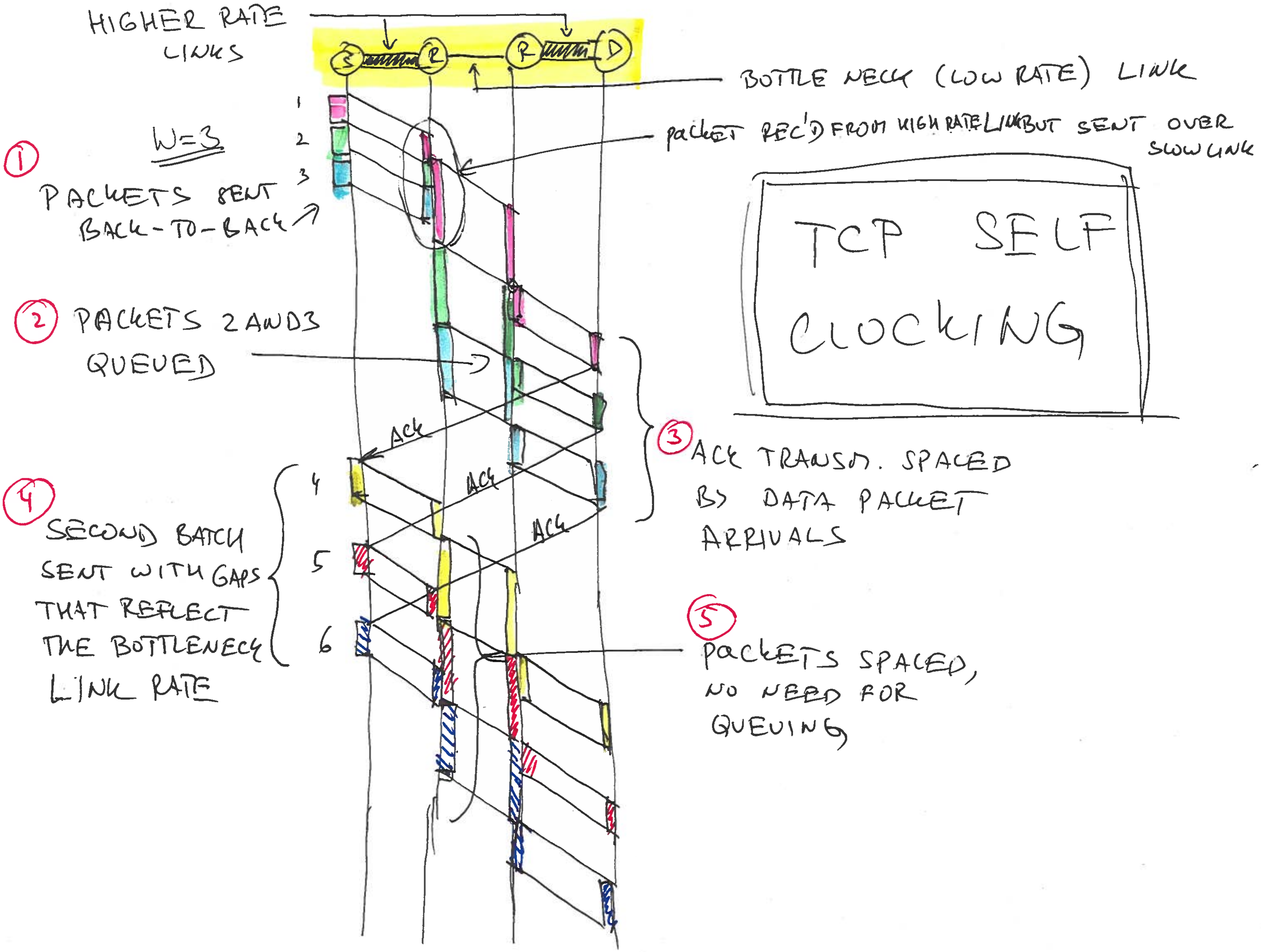


Node to Node Actions

- Back-pressure / feedback / …



- Speed of reaction - feedback loop latency
- Stability


TCP SELF CLOCKING
HIGHER RATE LINKS
S
R
R
D
BOTTLE NECK (LOW RATE) LINK
W=3
1
2
3
① PACKETS SENT BACK-TO-BACK
packet REC'D FROM HIGH RATE LINK BUT SENT OVER SLOW LINK
② PACKETS 2 AND 3 QUEUED
ACK
ACK
ACK
③ ACK TRANSM. SPACED BY DATA PACKET ARRIVALS
4
5
6
④ SECOND BATCH SENT WITH GAPS THAT REFLECT THE BOTTLENECK LINK RATE
⑤ PACKETS SPACED, NO NEED FOR QUEUING