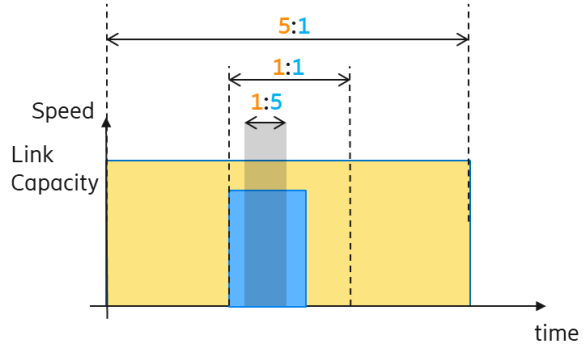
Multi timescale bandwidth profile and its application for burst-aware fairness

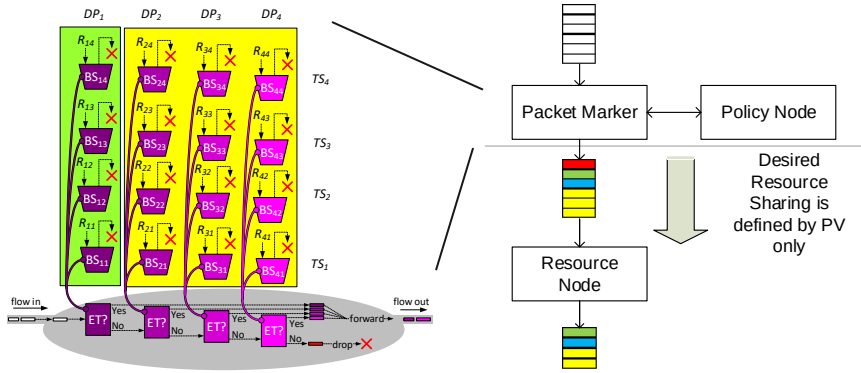
Szilveszter Nádas, Balázs Varga
Ericsson Research

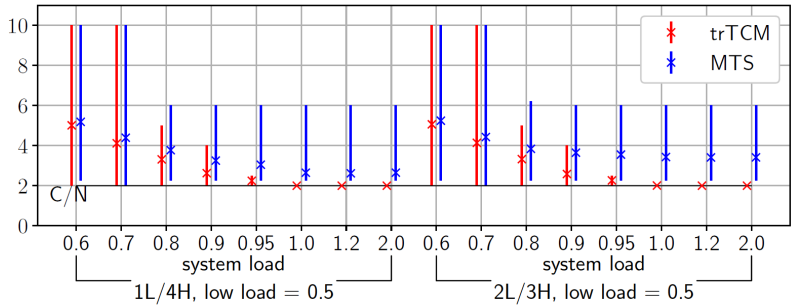
Illés Horváth, András Mészáros, Miklós Telek
Budapest University of Technology and Economics

Overview

- We give a definition of fairness on multiple-time scales
  - based on bitrate measurement on multiple time scales
- We propose an implementation
  - we build on Core Stateless Resource sharing and
  - we only update the edge marking to reflect the time-scales
- We show potential advantages and characteristics
  - fluid simulation assuming ideal Congestion Control
  - Two-Rate, Three-Color Marker (trTCM) is used as a reference





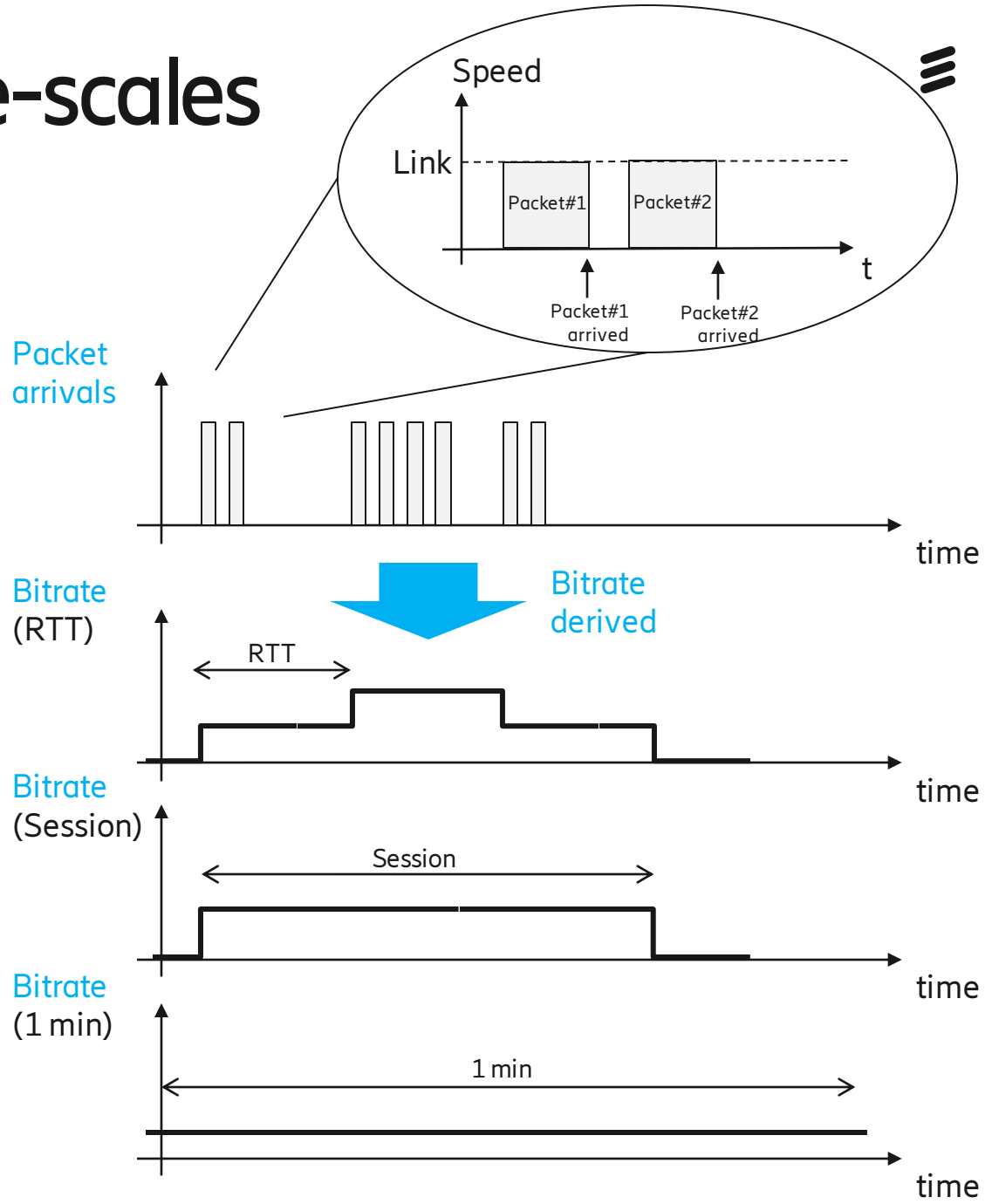


Bitrate measurement and time-scales

- Bitrate is a derived measure
  - Discrete packet arrivals are translated to bitrate
  - Bitrate always has a time-scale associated

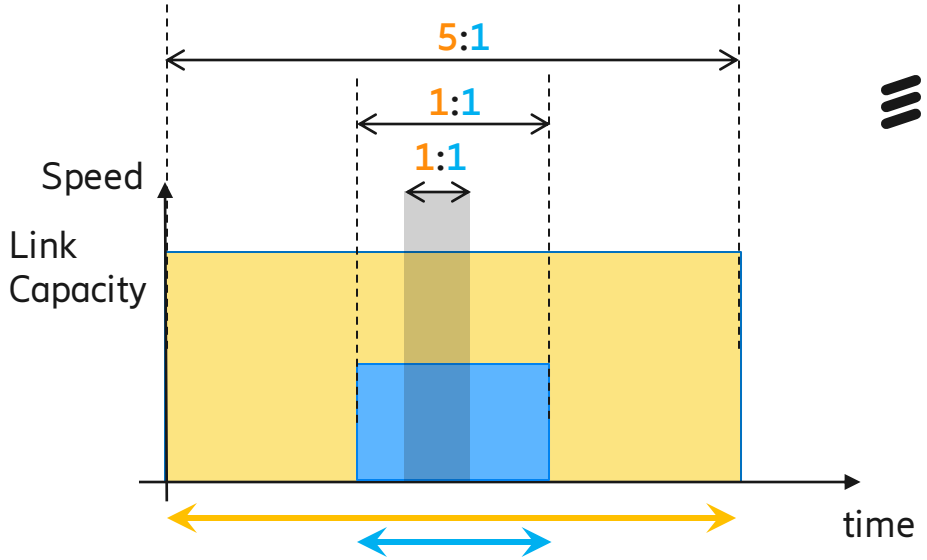
$$Bitrate = \frac{Volume\ (bits)}{Time\ (sec)}$$

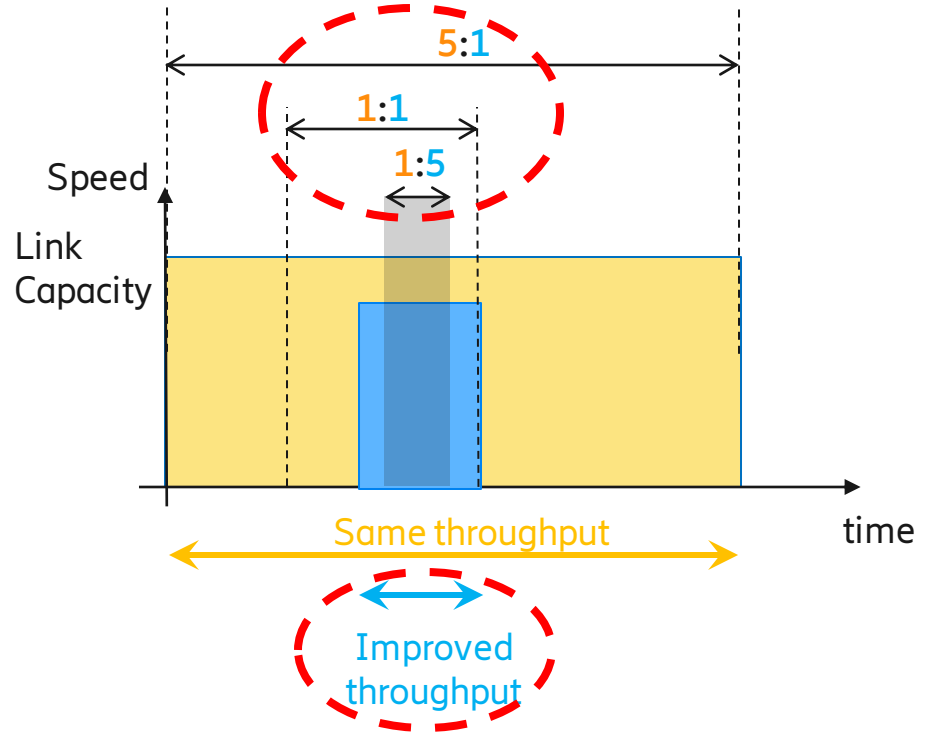
- Natural time-scales:
  - ~ RTT
  - ~ 1s – speed shown in apps
  - ~ Session duration (target)
  - ~ 1 minute: short term history and activity
  - ~ 10 minutes: longer term activity
  - ~ Month: monthly cap



Fairness on multiple time-scales

- When to measure bitrate
  - When source is active – to describe performance
  - During both active and inactive periods – to judge the fairness of resource sharing
- Fairness goal on multiple time-scales
  - Balanced fairness: multiple time scales are considered
  - Allow higher share on shorter time-scales for flows below their fair share in longer time-scales

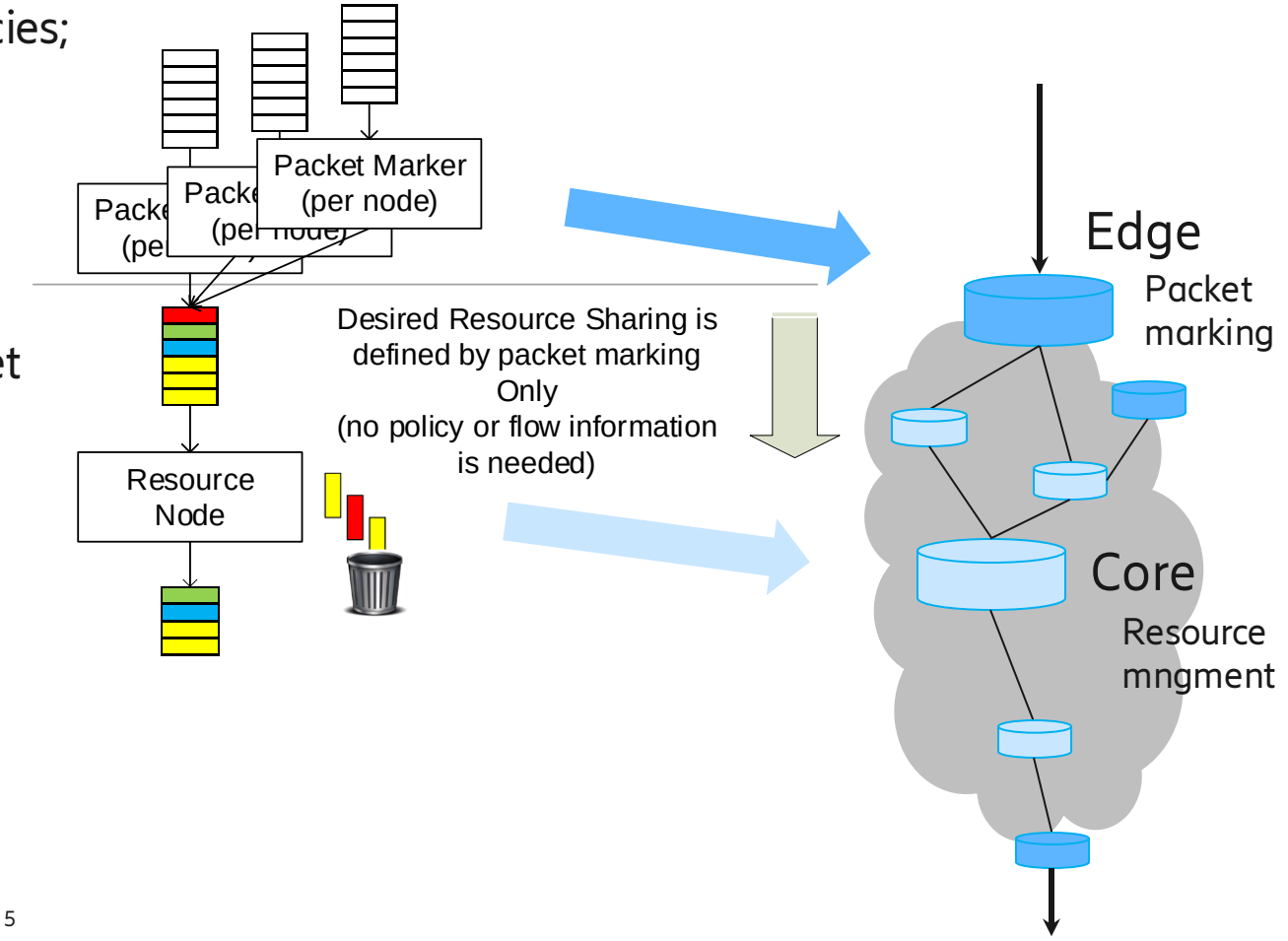




Overview of Core-Stateless Resource Sharing Example: Per Packet Value based CS RS

- PPV is a Core-Stateless Resource Sharing framework, which
  - allows a wide variety of detailed and flexible policies;
  - enforces those policies for all traffic mixes; and
  - scales well with the number of flows

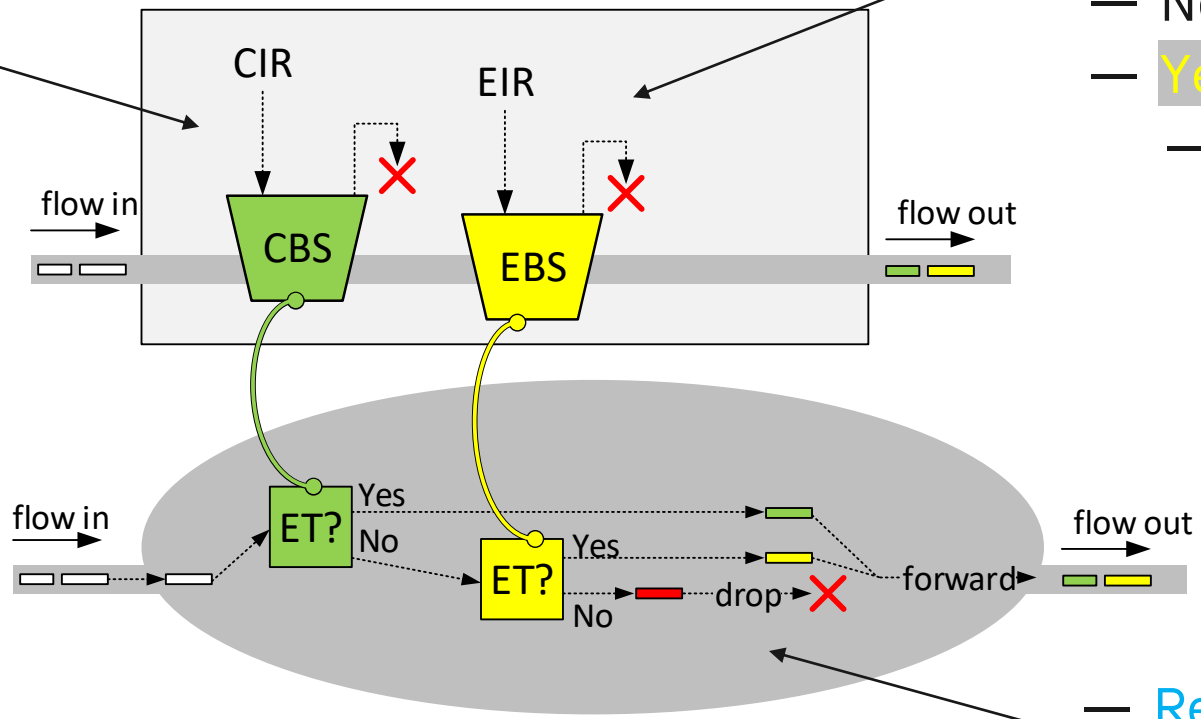
- Packet Marking at the edge
  - encodes policy into a value marked on each packet
- Resource Node – AQM and Scheduling
  - behavior based on packet marking only
  - no need for
    - policy information
    - flow identification
    - separate queues
  - very fast and simple implementations exist



Two-Rate, Three-Color Marker (trTCM)
A simple Core-Stateless Marker

- Committed Information Rate
  - Guaranteed
  - Green
    - DE = False (Drop Eligibility)

- Excess information Rate
  - Allowed into the network
  - No guarantee at all
  - Yellow
    - DE = True

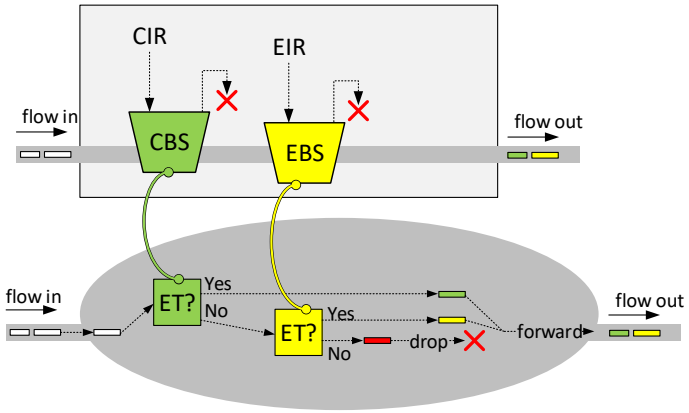


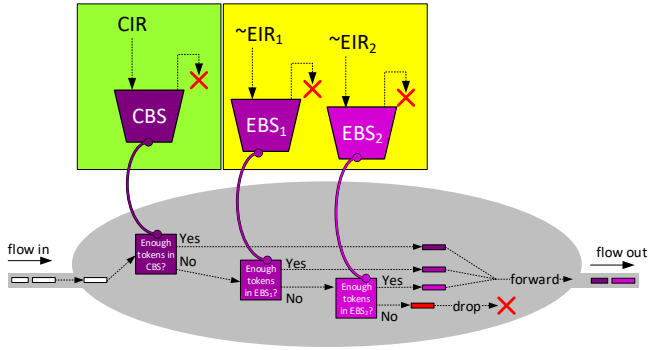
- Rest traffic
  - Red
    - dropped

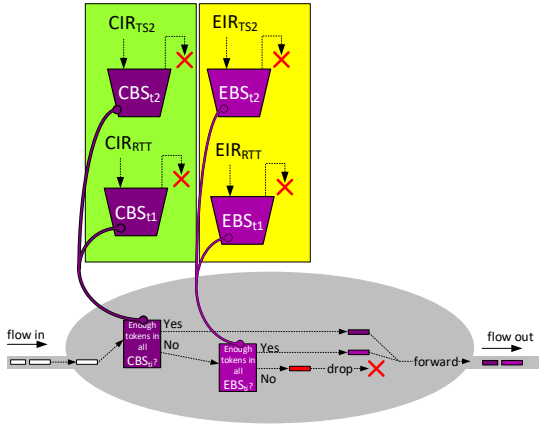
Extending trTCM

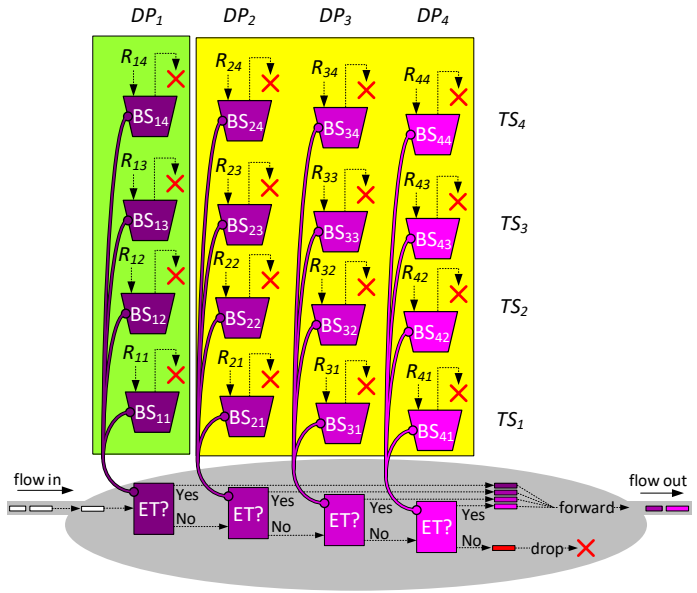
More drop precedences (DPs)

More time-scales (TSs)



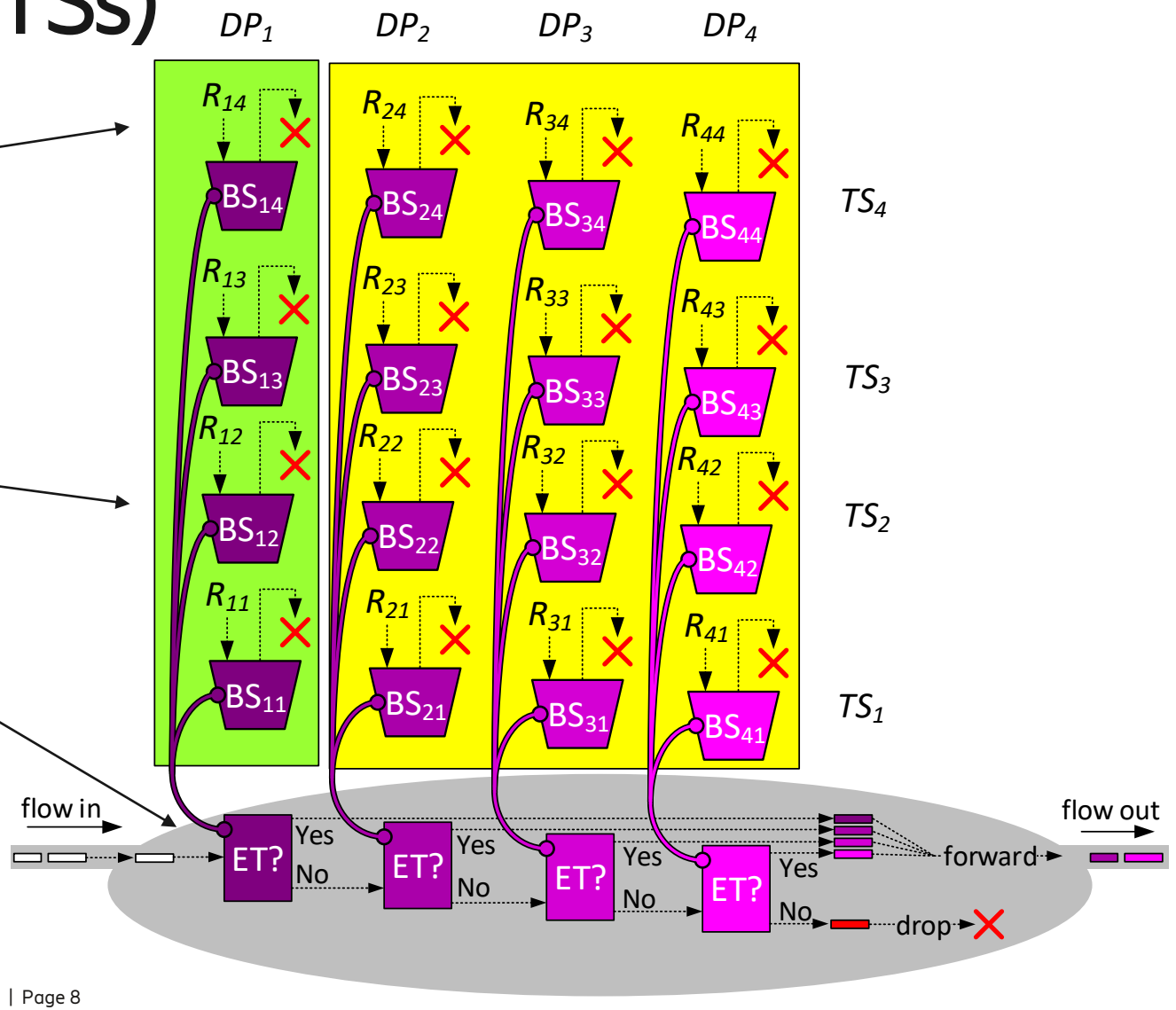






Multi-Timescale Bandwidth profile (MTS-BWP) 4x4 example (i.e., 4DPs, 4TSs)

- Input:
  - $R_{dp,ts}$: token bucket rates
  - $MBS_{dp,ts}$: maximum bucket sizes
- MBS is calculated from R and the Time Scales
  - $MBS_{dp,ts} \approx R_{dp,ts} * TS_{ts}$
- ET?: "Enough Tokens?"
  - Checks whether there are "packet size" worth of tokens in all buckets of that Drop Precedence
  - E.g. R_{14} limits the bitrate of DP_1 packets on TS_4

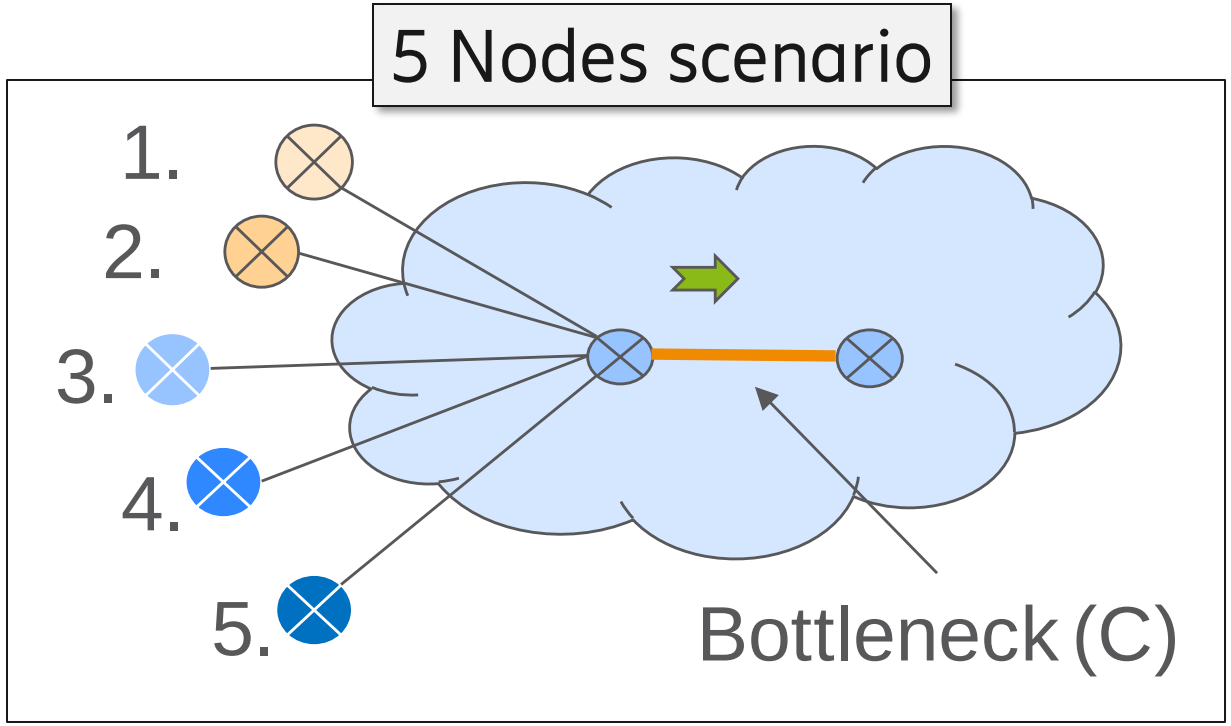


MTS-BWP example Scenario Access/Aggregation Network

- Few nodes sharing a common bottleneck
  - Several flows/users in one node
  - One MTS-BWP per node

The advantage we are looking for

- Nodes with good history can temporality access high portion of bottleneck capacity
  - I.e. high peak rates achieved for small bursts (feels like an underloaded system)
  - At the same time multi-timescale fairness is maintained



- C = 10 Gbps
- N = 5 Nodes
- C/N = 2 Gbps (fair share of a node)

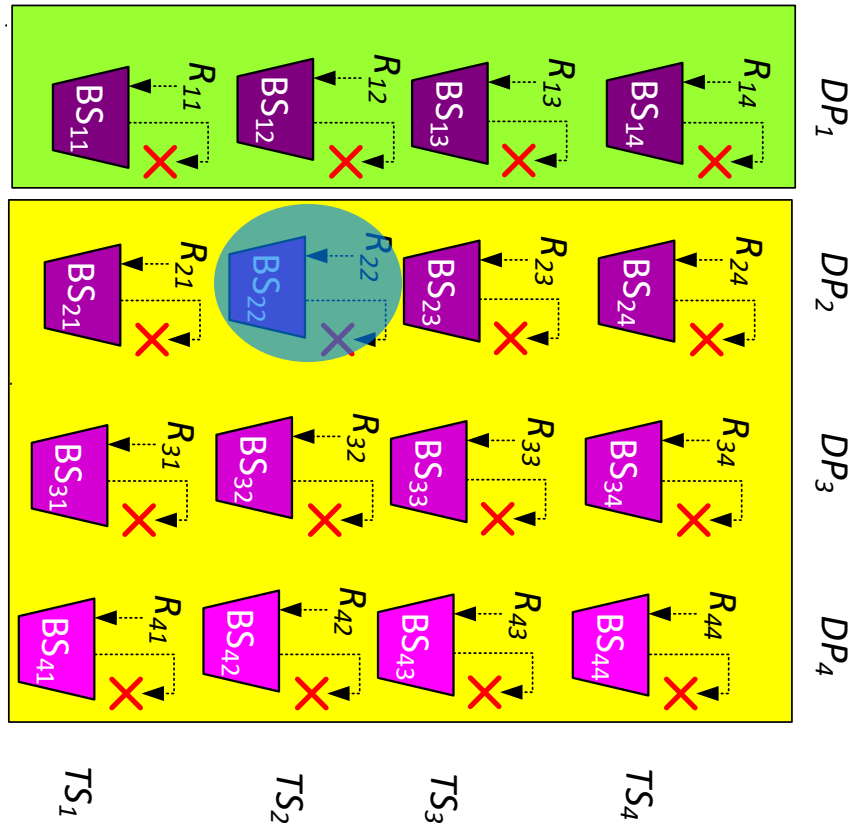
MTS-BWP R matrix Design highlights

$$\boldsymbol{R} = \begin{bmatrix} R_{11} & R_{12} & R_{13} & R_{14} \\ R_{21} & R_{22} & R_{23} & R_{24} \\ R_{31} & R_{32} & R_{33} & R_{34} \\ R_{41} & R_{42} & R_{43} & R_{44} \end{bmatrix}$$

$$TS = [0.01,\ 0.133,\ 2,\ 30] \text{ (in sec)}$$

$$\boldsymbol{R} = \begin{array}{c} \begin{matrix} TS_1 & TS_2 & TS_3 & TS_4 \end{matrix} \\ \begin{bmatrix} 2 & 2 & 2 & 0.75 \\ 4 & 2 & 1 & 0.25 \\ 10 & 10 & 1 & 1 \\ 10 & 10 & 10 & 10 \end{bmatrix} \end{array} \begin{matrix} DP_1 \\ DP_2 \\ DP_3 \\ DP_4 \end{matrix}$$

(Rates in [Gbps])



MTS-BWP R matrix Design highlights

$$TS = [0.01,\ 0.133,\ 2,\ 30]\ (\text{in sec})$$

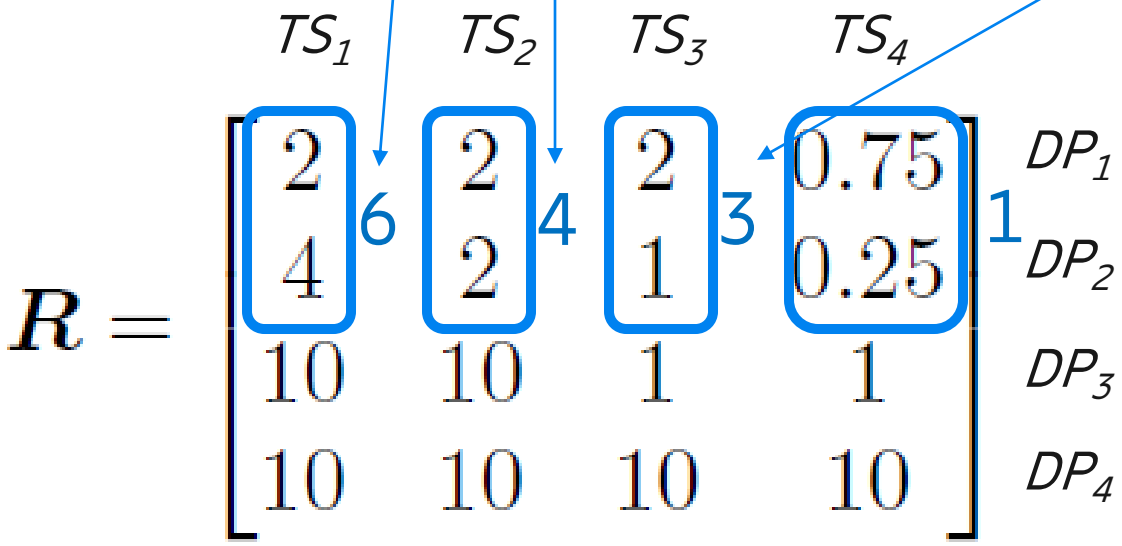
$$\boldsymbol{R} = \begin{array}{c} \begin{matrix} TS_1 & TS_2 & TS_3 & TS_4 \end{matrix} \\ \begin{bmatrix} 2 & 2 & 2 & 0.75 \\ 4 & 2 & 1 & 0.25 \\ 10 & 10 & 1 & 1 \\ 10 & 10 & 10 & 10 \end{bmatrix} \end{array} \begin{matrix} DP_1 \\ DP_2 \\ DP_3 \\ DP_4 \end{matrix}$$

Guaranteed bitrate on different Time-Scales (DP_1 is dropped last)

MTS-BWP R matrix Design highlights

Throughput for small and medium files in nodes with good history

Throughput for nodes with "still" good history



Target throughput when
-All DP_2 packet go through
-For different time scales

DP_2 is designed to go through, when all but one nodes are having "bad history"

MTS-BWP R matrix Design highlights

$$\boldsymbol{R} = \begin{array}{c} \begin{matrix} TS_1 & TS_2 & TS_3 & TS_4 \end{matrix} \\ \begin{bmatrix} 2 & 2 & 2 & 0.75 \\ 4 & 2 & 1 & 0.25 \\ 10 & 10 & 1 & 1 \\ 10 & 10 & 10 & 10 \end{bmatrix} \end{array} \begin{matrix} DP_1 \\ 2\ DP_2 \\ DP_3 \\ DP_4 \end{matrix}$$

C/N = 2

— It is the fair share of two nodes on the same Time-Scale

— It is also the long time fair share

MTS-BWP
Design highlights

$$TS = [0.01,\ 0.133,\ 2,\ 30,\ \infty]\ \text{[sec]}$$

TS_1 TS_2 TS_3 TS_4

2. Consequently, TS_i is also how long the rates in column *i-1* can be maintained

1. TS_i Determines bucket sizes on the same column (*i*)

$$\boldsymbol{R} = \begin{bmatrix} 2 & 2 & 2 & 0.75 \\ 4 & 2 & 1 & 0.25 \\ 10 & 10 & 1 & 1 \\ 10 & 10 & 10 & 10 \end{bmatrix}$$

DP_1
DP_2
DP_3
DP_4

MTS-BWP
Design highlights

30 sec active period, before the node is considered high load

TS_1 TS_2 TS_3 TS_4

$$TS = [0.01,\ 0.133,\ 2,\ 30,\ \infty]\ \text{[sec]}$$

~RTT

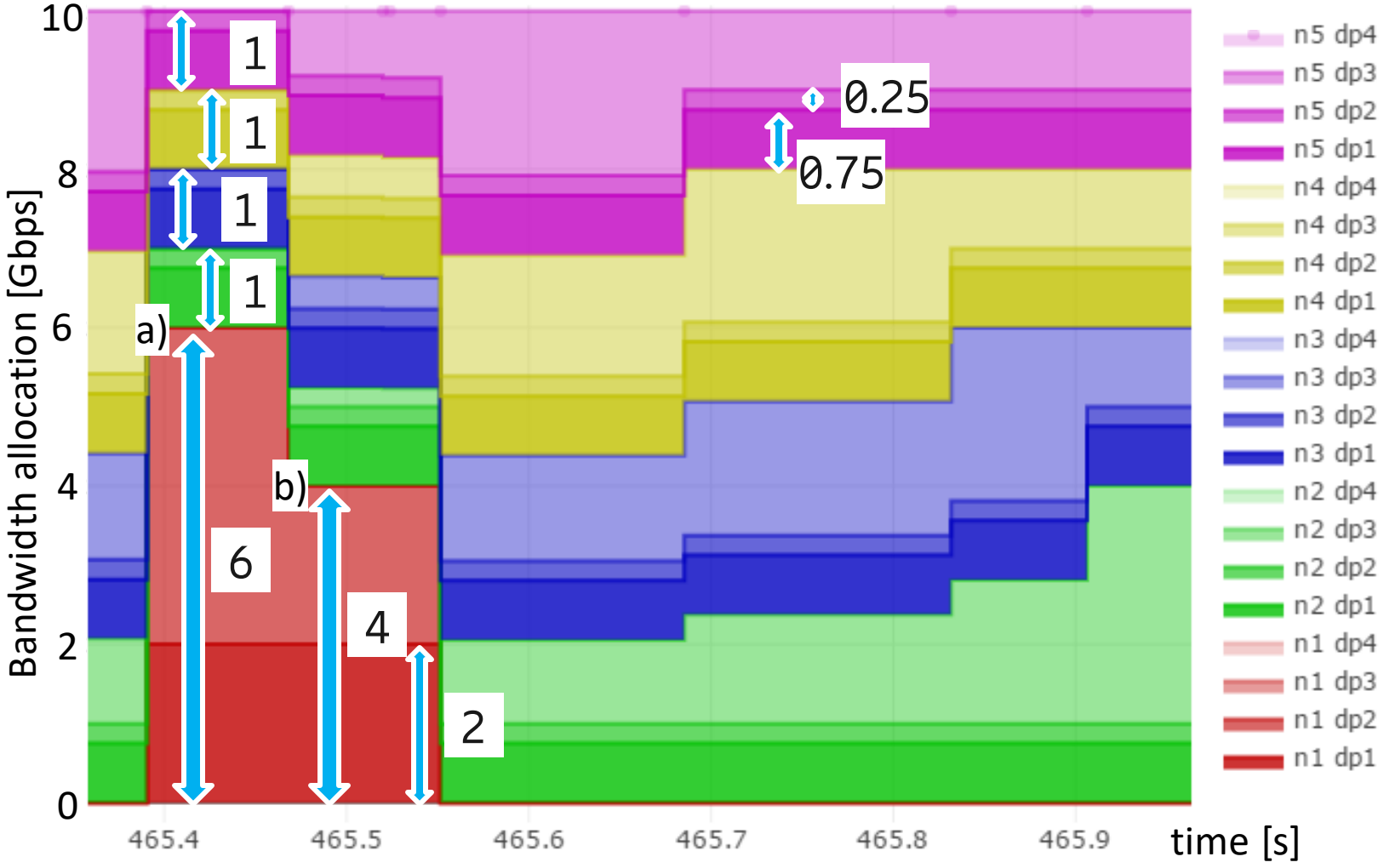
File size = [0.1, 1] (Gbyte)

$$\boldsymbol{R} = \begin{bmatrix} 2 & 2 & 2 & 0.75 \\ 4 & 2 & 1 & 0.25 \\ 10 & 10 & 1 & 1 \\ 10 & 10 & 10 & 10 \end{bmatrix}$$

6 4

DP_1 DP_2 DP_3 DP_4

Fluid Simulation – Time-Series Example



$$\boldsymbol{R} = \begin{matrix} TS_1 & TS_2 & TS_3 & TS_4 \end{matrix} \begin{bmatrix} 2 & 2 & 2 & 0.75 \\ 4 & 2 & 1 & 0.25 \\ 10 & 10 & 1 & 1 \\ 10 & 10 & 10 & 10 \end{bmatrix} \begin{matrix} 6 \\ 4 \\ 3 \\ 1 \end{matrix} \begin{matrix} DP_1 \\ DP_2 \\ DP_3 \\ DP_4 \end{matrix}$$

- Nodes 2-5 – "bad history"
  - On TS_4 (DP_1-DP_2)
    - 1 (.75+.25) Gbps
- Node 1 – "good history"
  - a) Starts on TS_1
    - 6 (2+4) Gbps
  - b) Changes to TS_2
    - 4 (2+2) Gbps
- Extra capacity
  - DP_3

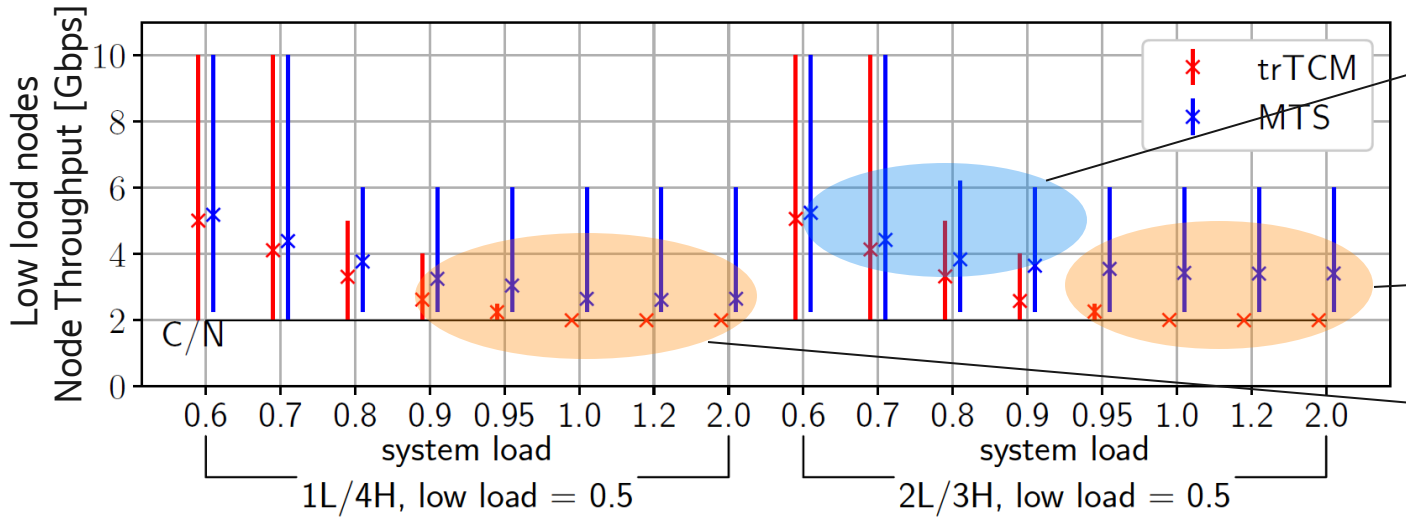
Simulation of Advantages
Fluid Simulator

- Traffic Model
  - Poisson arrivals
  - Two file sizes (small, large)
  - Maximum number of flows (per Node)
- Nominal load (of a Node):
  - the load of Node line divided by its fair share
  - Low load node: Nominal load<1
  - High load node: Nominal load>1
- System load

- Scenario naming: 1L/4H – meaning 1 low load node, 4 high load nodes
  - The load of low load nodes and the system load is varied
  - (The load of high load node is calculated from the above two)

Selected simulation results MTS-BWP vs. trTCM

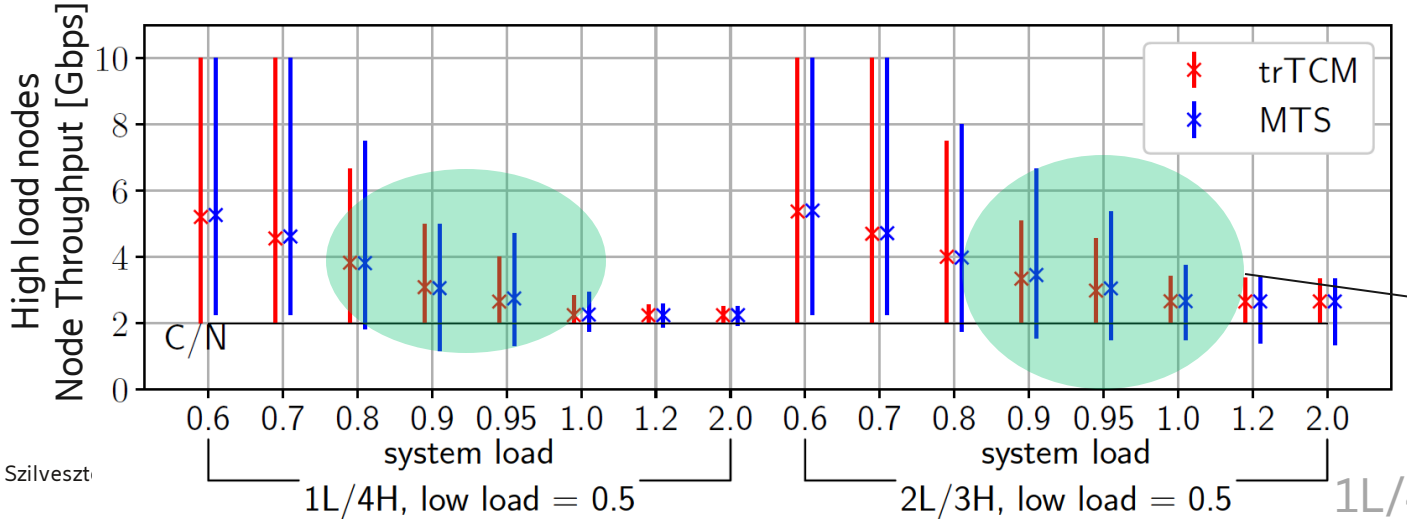
Node throughput: measured when the node is ACTIVE (sending data)



— Small gain
— But good performance anyway

— Increased average and
— 10% best

— Similar, but smaller gains



Experience:
- Increased for low load nodes
- Almost no change for high load ones

— Average remains the same
— Slightly increased variance

1L/4H – meaning 1 low load node, 4 high load nodes

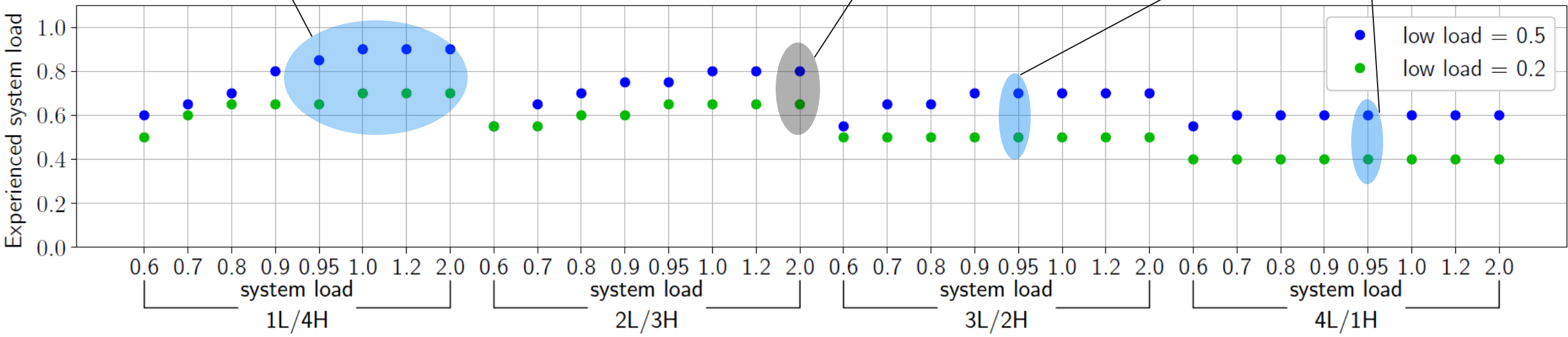
Experienced system load for low load nodes for MTS-BWP

- Defined as the system load in the equivalent trTCM scenario
  - where the average node bandwidth for low load nodes is the same

High load/overloaded system
Good experience for low load node

Experience in an overloaded system is like trTCM with load 0.8/0.65
(For low load nodes with load 0.5/0.2)

Smaller load for low load nodes and higher Nr of low load nodes result in better performance



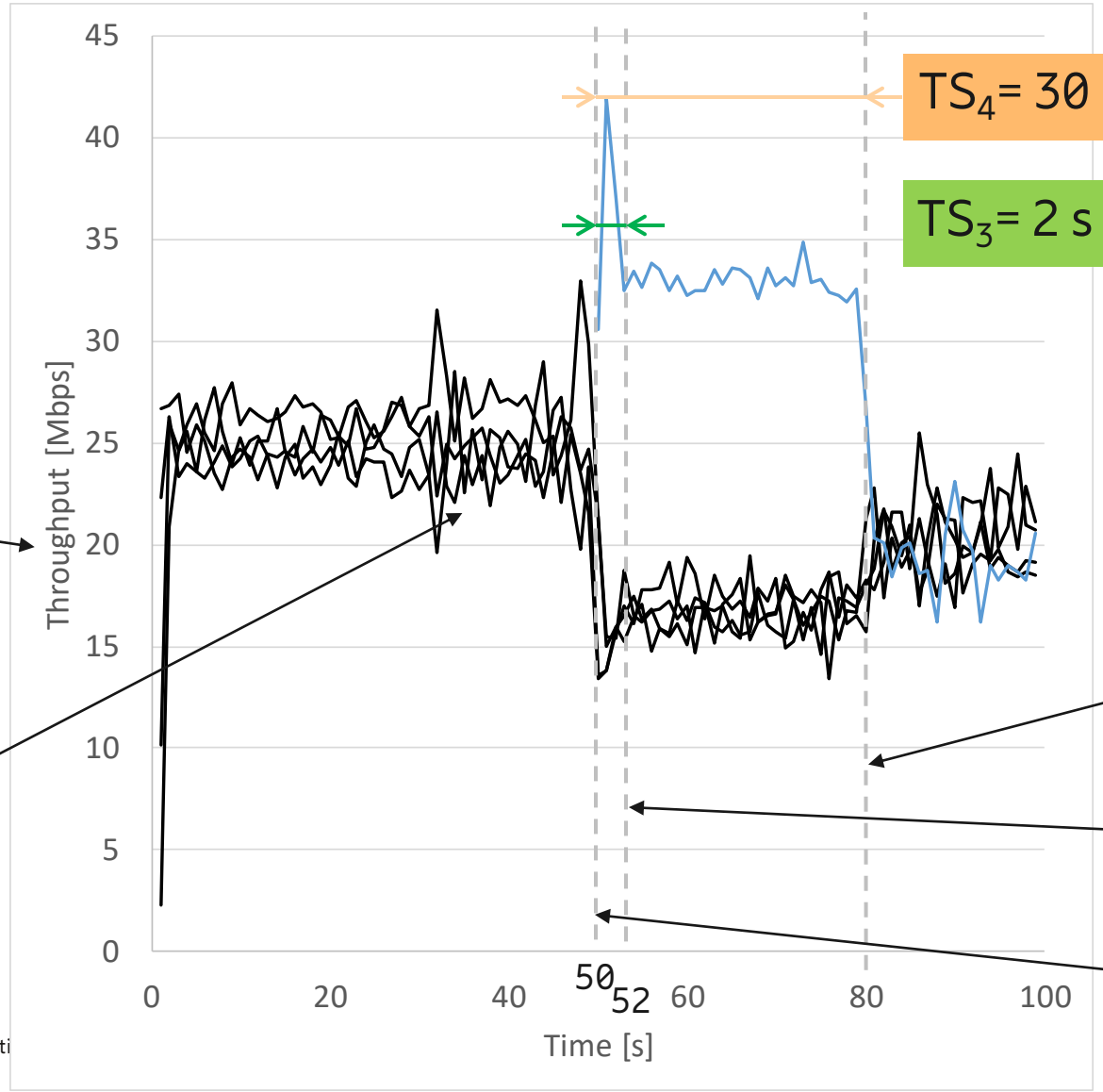
1L/4H – meaning 1 low load node, 4 high load nodes

Preliminary Packet level Simulation Results (ns3-dce) Validate the ideal fluid model

5 Cubic TCPs per Node

1s sliding window avg

4 Nodes, greedy Bad history (t>30s)



"perfect history"

"good history"

"bad history"

$$\boldsymbol{R} = \begin{bmatrix} 20 & 20 & 20 & 7.5 \\ 40 & 20 & 10 & 2.5 \\ 100 & 100 & 10 & 10 \\ 100 & 100 & 100 & 100 \end{bmatrix} \text{ [Mbps]}$$

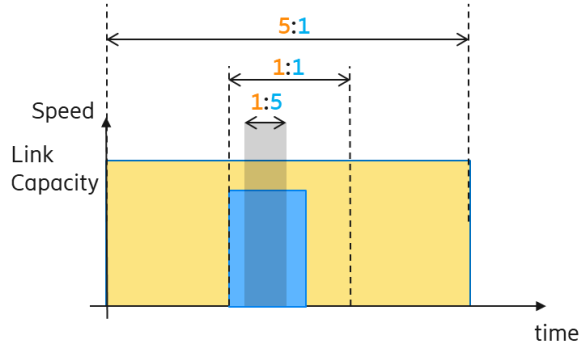
new Node's history becomes equally "bad"
Equal sharing – fair share (20 Mbps)

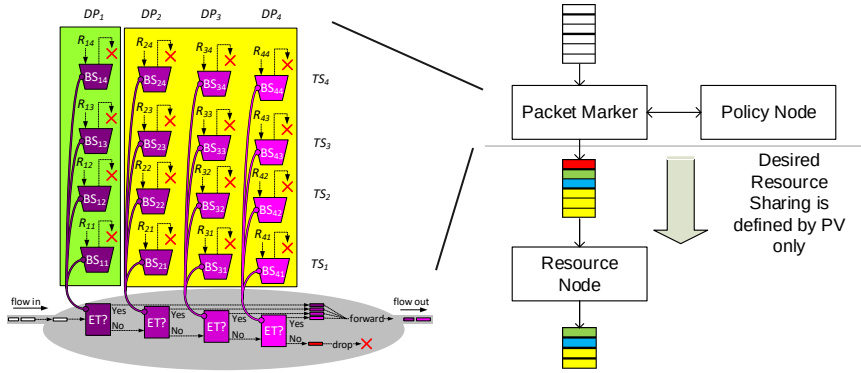
TS changes for the new Node (good history)

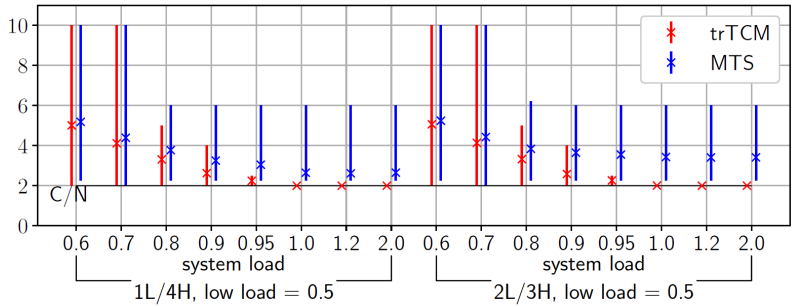
1 new Node arrives (perfect history)

Summary

- We give a definition of fairness on multiple-time scales
  - based on bitrate measurement on multiple time scales
- We propose an implementation
  - we build on Core Stateless Resource sharing and
  - we only update the edge marking to reflect the time-scales
- We show potential advantages and characteristics
  - fluid simulation assuming ideal Congestion Control
  - Two-Rate, Three-Color Marker (trTCM) is used as a reference







References

— Szilveszter Nádas, Balázs Varga, Illés Horváth, András Mészáros, Miklós Telek, "**Multi timescale bandwidth profile and its application for burst-aware fairness**", preprint at https://arxiv.org/abs/1903.08075

— http://ppv.elte.hu – our articles and videos about core-stateless resource sharing

— Michael Menth, Nikolas Zeitler: "**Fair Resource Sharing for Stateless-Core Packet-Switched Networks with Prioritization**", in IEEE Access, vol. 6, 2018, IEEE – another core-stateless resource sharing solution