Enhanced DSCP by in-band signaling for latency guaranteed service

Lin Han (lin.han@futurewei.com)

Yingzhen Qu (yingzhen.qu@futurewei.com)

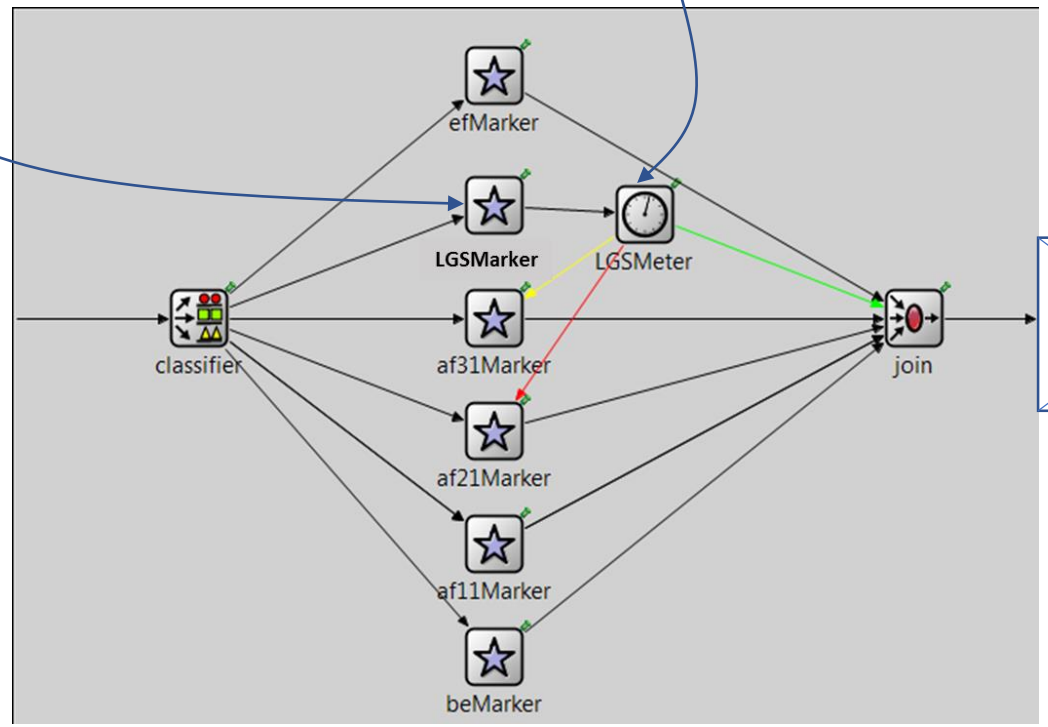
Introduction

- What -
  - Achieve the E2E guaranteed service (Latency, Bandwidth) for IP
  - The Guaranteed latency service is like Detnet (but without using MPLS and RSVP-TE)
- How -
  - Use DiffServ integrated with in-band Signaling
  - In-band signaling is for admission control
- Why -
  - Many existing hardware supports DiffServ already
  - Class Based Queuing (Strict Priority Queuing + Weight Fair Queueing) are simplest implementation for hardware to achieve different class of service
  - Strict Priority Queuing is most effective for bounded latency flows, but the admission control is critical to protect other classes

Network modeling

New DSCP value is used to identify the new class for LGS flows

Rate is checked for LGS flows, colored packets are remarked to lower class or discarded



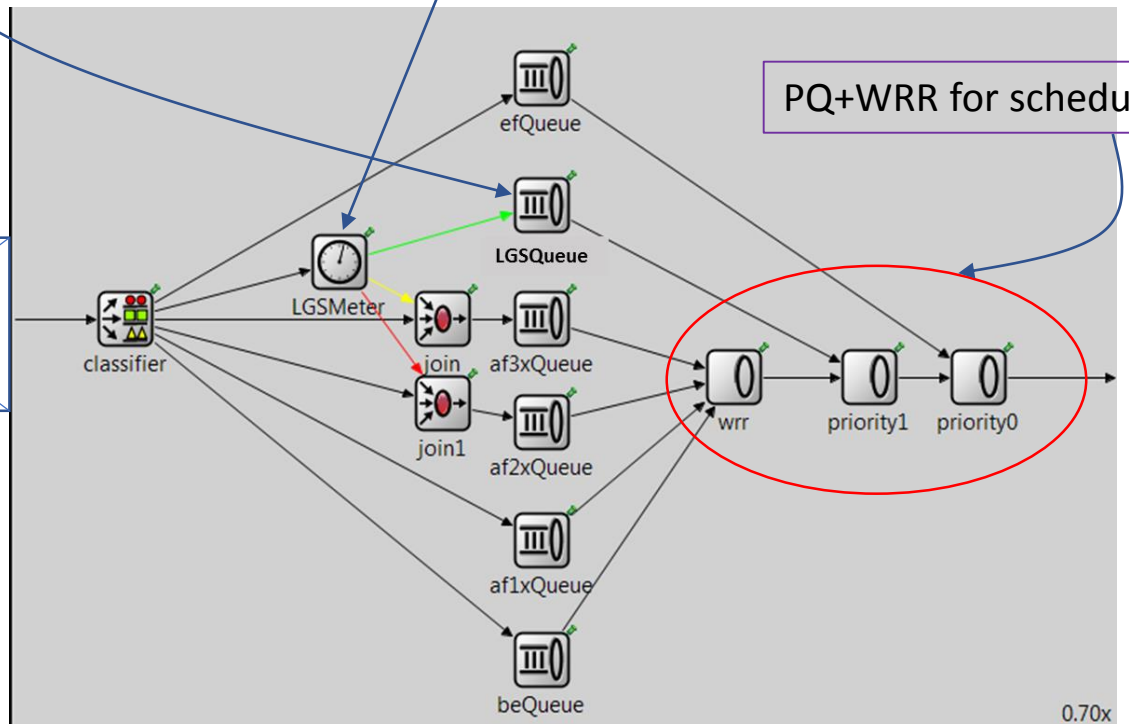
Ingress Traffic Classification Module at All Ingress Routers' ingress interface

Switch Fabric

LGS flows are queued into the 2nd priority queue

Rate is checked for LGS flows, colored packets are remarked to lower class or discarded

PQ+WRR for scheduling



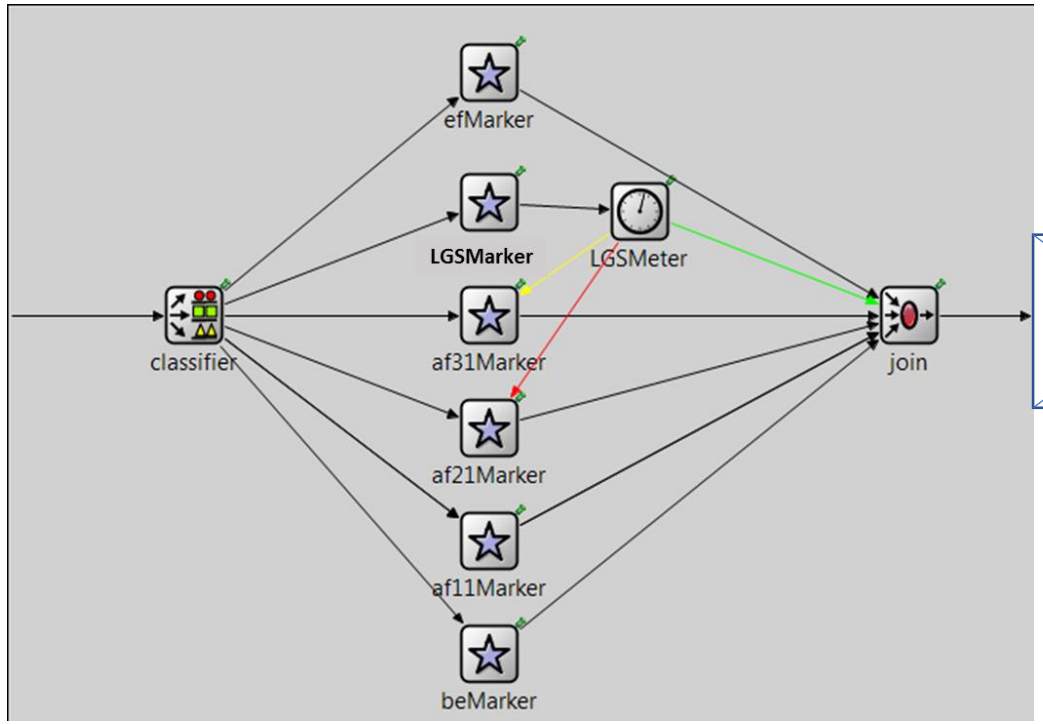
Egress Queuing and Scheduling Module at All Routers' egress interface

Admission Control by In-band Signaling

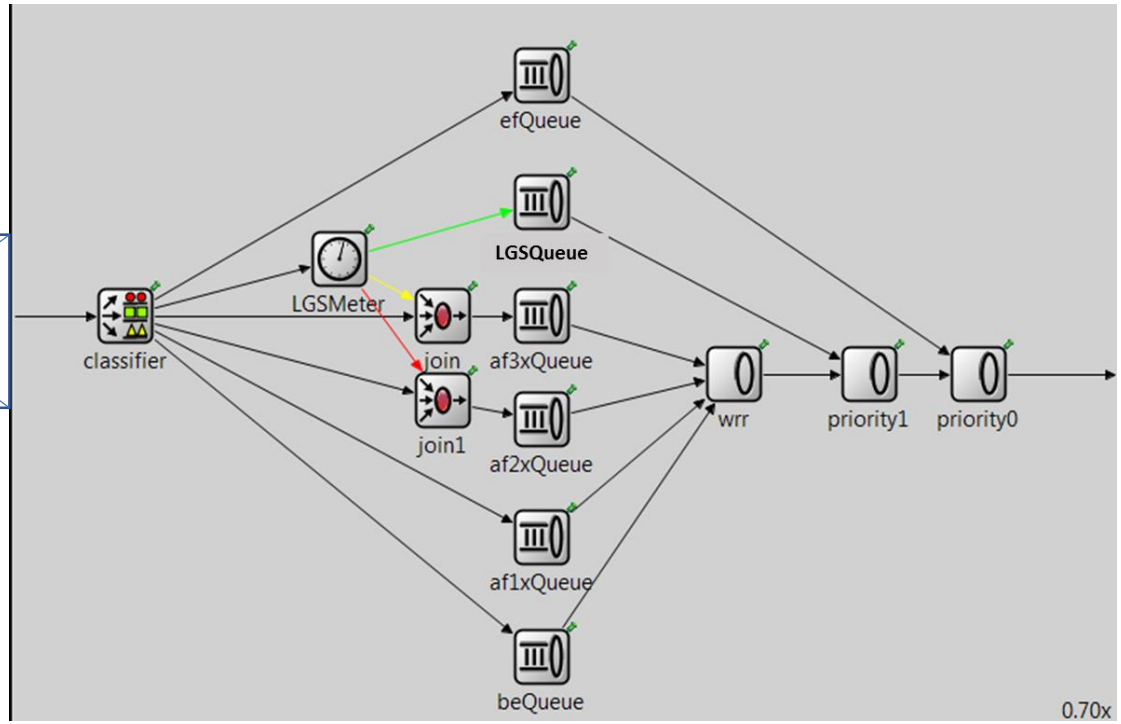
Admission Control: In-band signaling will notify the host about the resource reservation status on path, host then knows the status of connection

If (total ingress flow's CIR < ingress interface rate)
Allowed, classify as LGS flow, update meter DB, flow DB and in-band signaling
else
not allowed, classified as BE, update in-band signaling

If (total egress flow's CIR < egress interface rate)
allowed, update meter DB, flow DB and in-band signaling
else
not allowed, update in-band signaling



Switch Fabric

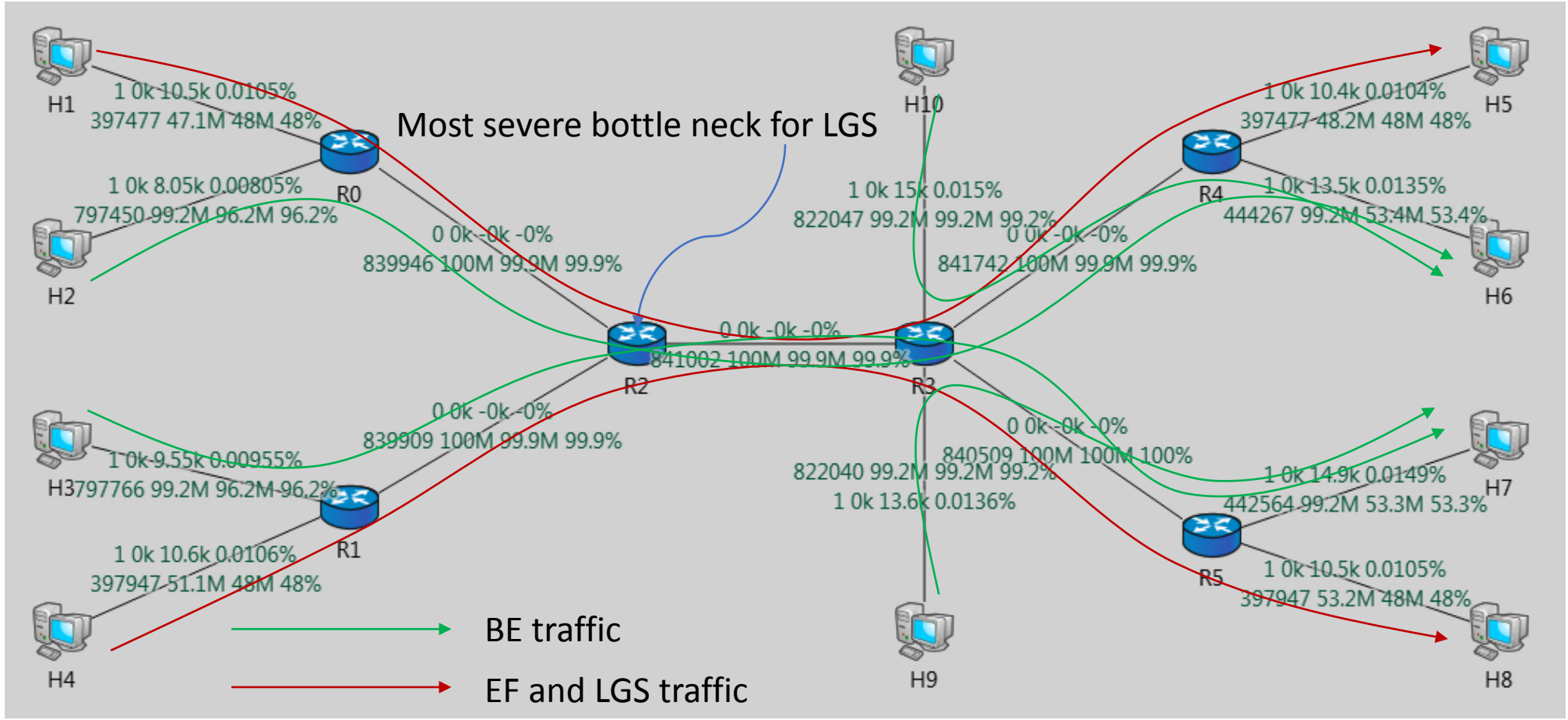


Ingress Traffic Classification Module at All Ingress Routers' ingress interface

Egress Queuing and Scheduling Module at All Routers' egress interface

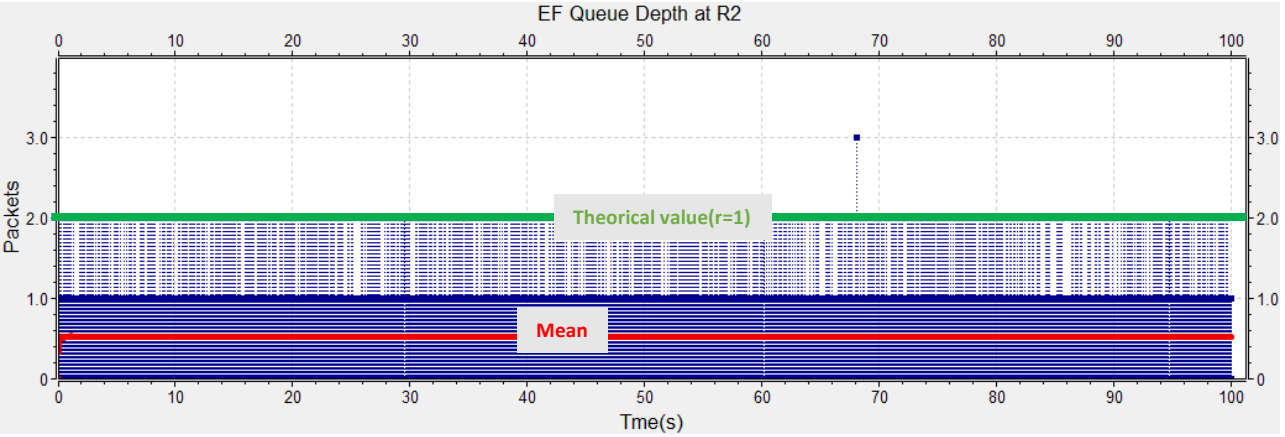
Experiment results and analysis

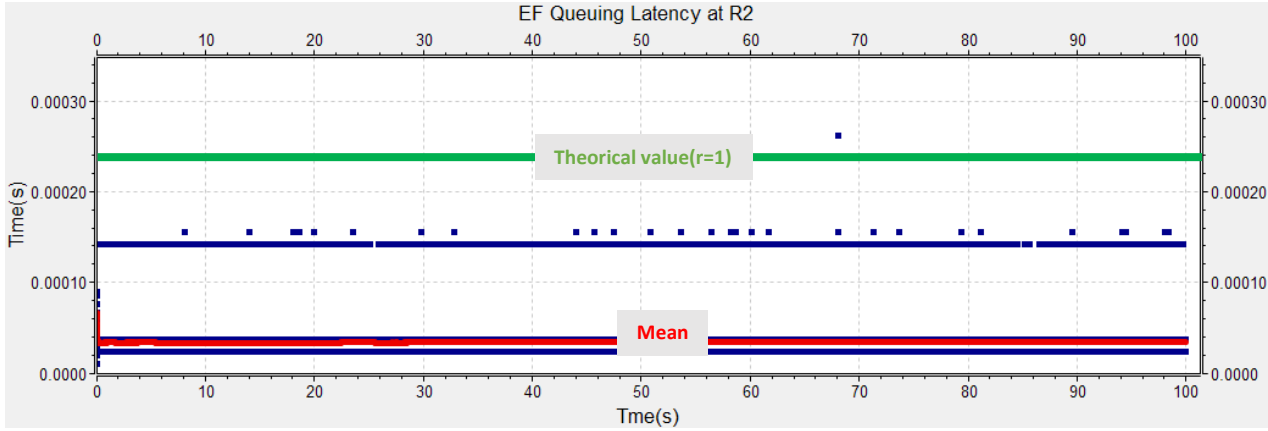
- Severely congested at all router on path by uncontrolled BE traffic

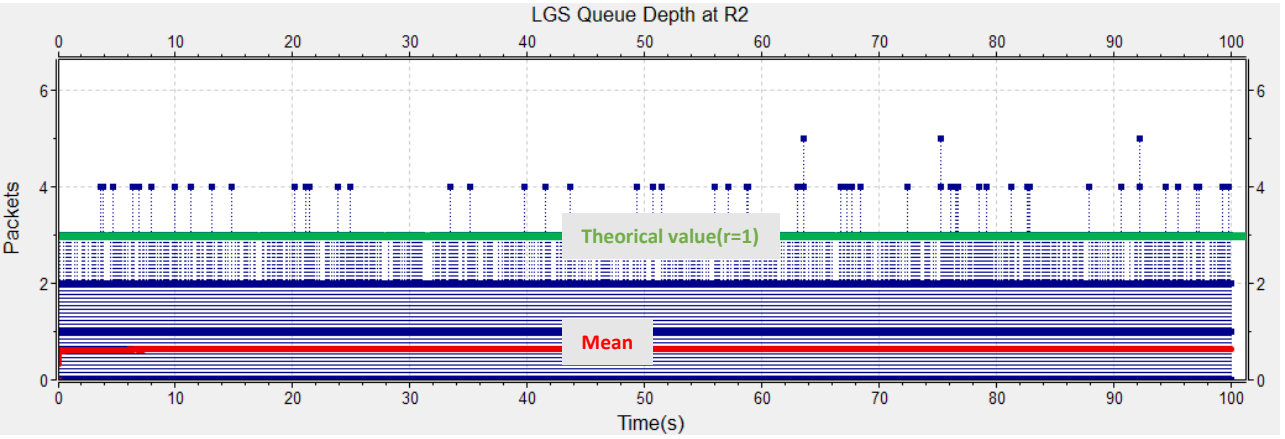


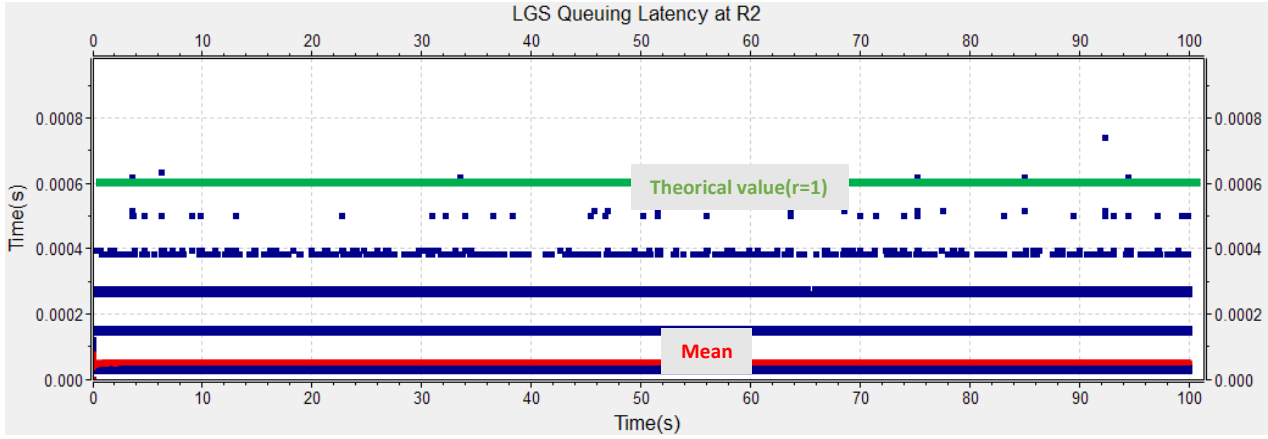
Continuous and burst UDP traffic are used

Case 1: Buffer Depth and Packet Latency: Measurement Vs Estimation (EF+LGS < 50% link rate)









EF flows =4, LGS flows = 16; R_{EF}=8.14M, R_{LGS}=26.67M; Burst traffic, Burst duration = Exp(0.02s), Burst sleep = Exp(0.001s).
Fully congested at all routers by BE, All Link Util =~100

Conclusion

- In-band signaling
  - Configuration automation for DiffServ
  - Can support the guaranteed service for E2E flows in a managed domain
  - New service can coexist with the current DiffServ
  - The current DiffServ capable hardware can be reused with minimum changes
- Design targets can be satisfied
- Only flow state is kept, No per flow scheduling. It is not heavy considering the new service is only a small portion of the whole network.
- Further research
  - The TCP CC needs to be changed to get the benefits of the work. The BGS and LGS may have different CC.
  - New meter to smooth the burst of traffic to achieve better latency and jitter
  - Other technologies such as DeltaQ to smooth the burst in network
  - In-band signaling work with other queuing algo to achieve the LGS
  - Interworking with other protocols such as TSN and Detnet

Q & A

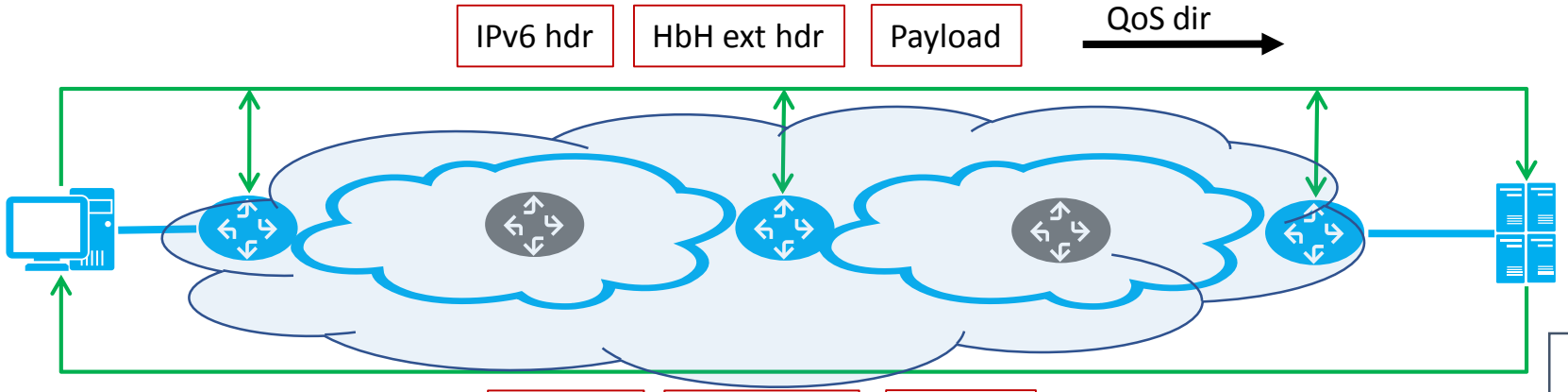
Introduction

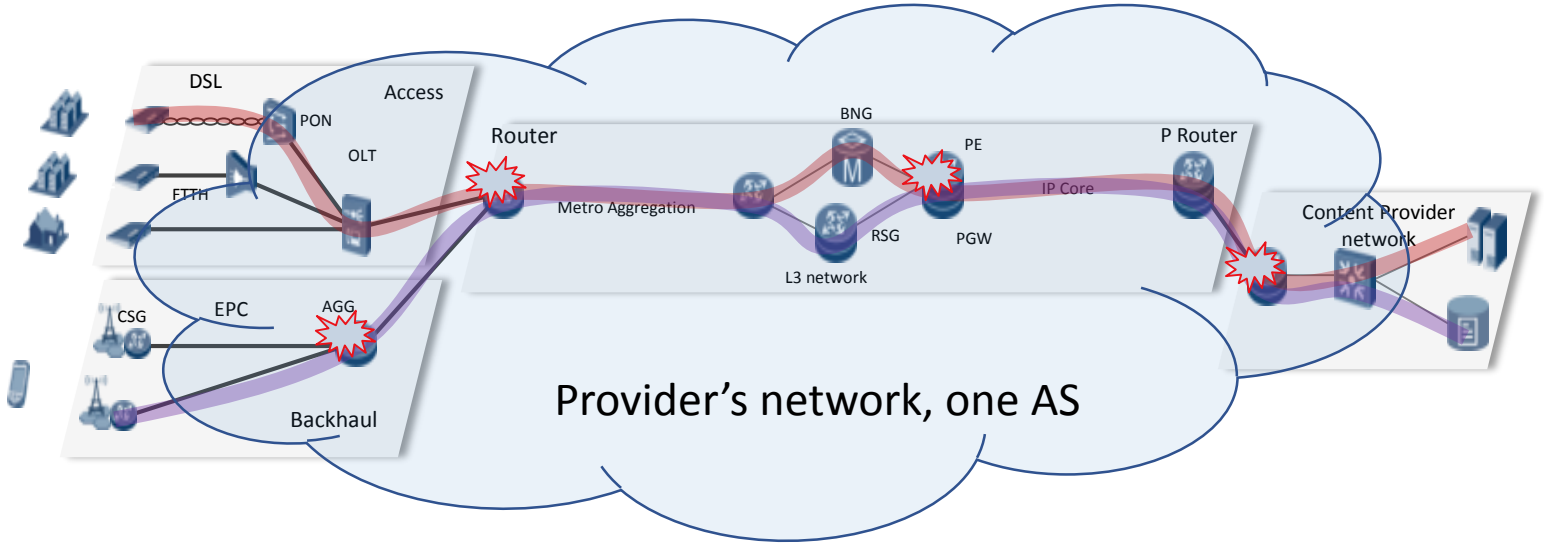
- In-band signaling was introduced in TSVWG.
- The presentation will use it to work with DSCP idea to achieve the bounded latency service.
  - Only focus on
    - the architecture, math and experiments on bounded latency service
  - No talks about
    - In-band signaling details, TCP/UDP changes, congestion control, etc
- Objective
  - Provide more scalable control and data plane for guaranteed service, latency and bandwidth
  - Provide the E2E Bandwidth Guaranteed Service (BGS) and Latency Guaranteed Service (LGS)
- Solution:
  - In-band signaling carries the user's service expectation (Bandwidth and/or Latency)
  - Network device on the path check if the resource is enough and update signaling accordingly.
  - Flows will be classified according to flow's expectation and resource reservation status.
    - New DSCP values are needed to represent new classes, the number is TBD
  - Class Based queuing and scheduling applied to the class of traffic
    - For LGS flow, either the 1st or 2nd PQ are mapped.
    - For BGS flow, the WRR shared Q are mapped.

Design targets

- Service Guarantee
  - Provide the guaranteed and minimized (queuing) latency for LGS (Latency Guaranteed Service) flows,
    - The maximum latency is guaranteed, minimized and predictable at each hop, if LGS flow rate is confirmed with their pre-claimed parameters (CIR,PIR,CBS,EBS)
  - Provide the guaranteed bandwidth for BGS (Bandwidth Guaranteed Service) flows
    - The bandwidth of CIR is guaranteed at each hop, if BGS flow rate is confirmed with their pre-claimed parameters (CIR,PIR,CBS,EBS)
- No starvation
  - LGS flows will never starve other type low priority flows (BGS and BES)
  - BGS flows will never starve BE flows
- No sacrifices of link utilization
  - When the total rate of LGS flow is less than the committed rate (sum of CIR of all LGS flows), other class flows (BGS and BES) can use the remained bandwidth
  - When the total rate of LGS flow is less than the committed rate (sum of CIR of all LGS flows), and, the total rate of BGS flow is less than the committed rate (sum of CIR of all BGS flows), other class flows (BES) can use the remained bandwidth
- Fairness within the same class
  - All LGS flows will share the bandwidth within its class
  - All BGS flows will share the bandwidth within its class
  - All BE flows will share the bandwidth for within BE class
- Does not impact the current DiffServ, and can coexist.

Review of in-band signaling





- Configure to process hop-by-hop IPv6 ext hdr
- Only needed for throttle devices
- Signal processing distributed to NPU on line card
- Per flow state kept in line card
- Per flow state self-maintained by data flow, deleted if aged
- Basic security for signaling and data forwarding are provided

Queuing Latency estimation per hop

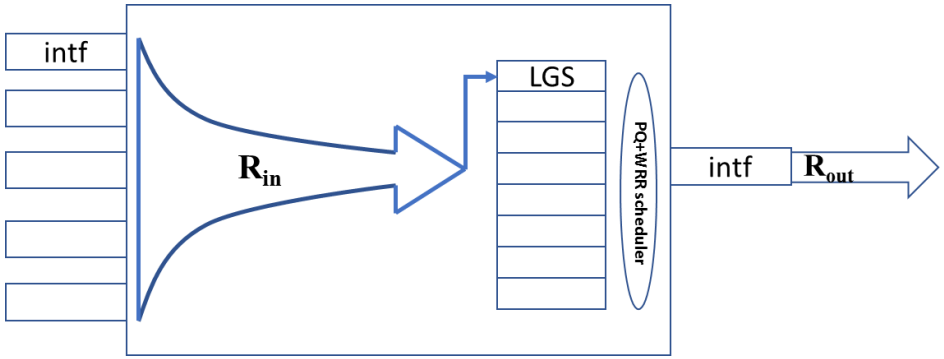
- Estimate at Worst scenario conservatively per hop, thus E2E latency can be guaranteed to be below max value statistically
- Only queuing delay estimated, other delays (switching, look up, propagation) are relatively small and fixed.
- Assume all packet size are the longest packet size
- If LGS flows queued in the highest priority queue
  - The maximum queued packet in the queue is $N_{max}^{LGS} = \lceil 2(R_{in}^{LGS}/R_{out}) + 1 \rceil$
  - The maximum delay of queuing at the hop is $D_{max}^{LGS} = N_{max}^{LGS} * L_{max} * 8/R_{out}$

$R_{in}^{LGS} = r * CIR_{LGS} = r \sum_{i=1}^{n} cir_i^{LGS}$,

cir_i^{LGS} is the cir for the ith LGS flow, r is the ratio of peak cir to avg cir

R_{out} is the link rate;

$L_{max}$is the maximum packet size



Queuing Latency estimation per hop

- If LGS flows queued into the 2nd highest priority queue, assume the 1st queue is EF:
  - The maximum queued packet in the queue is

$N_{max}^{LGS} = \lceil N_{max}^{EF} * (R_{in}^{LGS}/R_{out}) \rceil + N_{max1}^{LGS}$

$N_{max1}^{LGS} = \lceil 2(R_{in}^{LGS}/R_{out}) + 1 \rceil$

$N_{max}^{EF} = \lceil 2(R_{in}^{EF}/R_{out}) + 1 \rceil$

  - The maximum delay of queuing at the hop is

$D2_{max}^{LGS} = (N_{max}^{EF} + N_{max}^{LGS}) * L_{max} * 8/R_{out}$

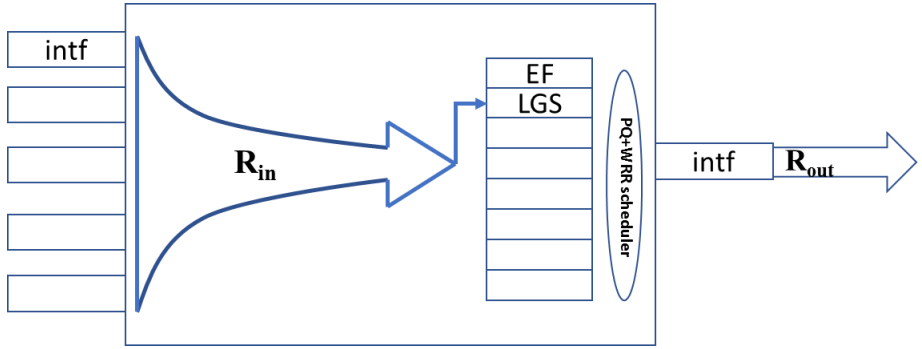
$R_{in}^{LGS} = r * CIR_{LGS} = r \sum_{i=1}^{n} cir_i^{LGS}$,

cir_i^{LGS} is the cir for the ith LGS flow, r is the ratio of peak cir to avg cir

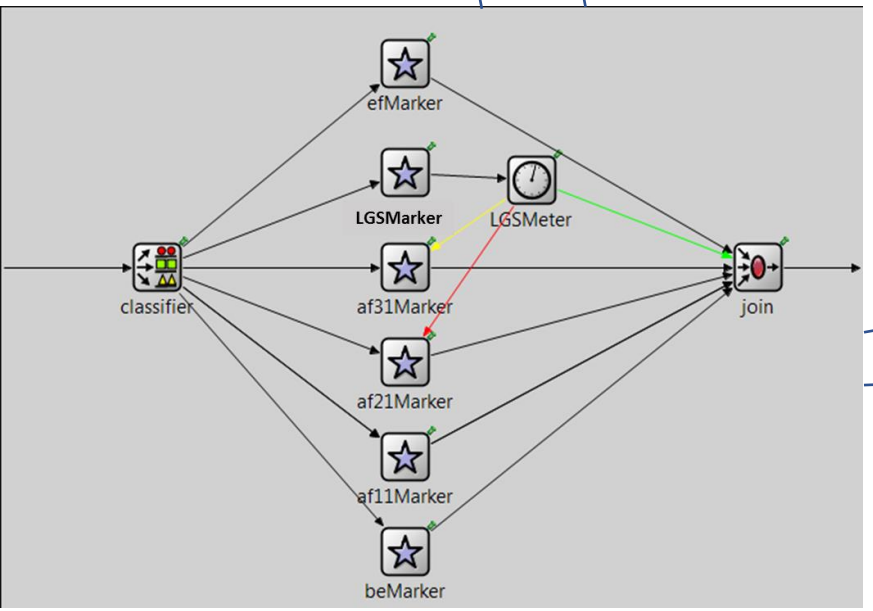
R_{in}^{EF} is the peak rate for EF flows from management knowledges

R_{out} is the link rate;

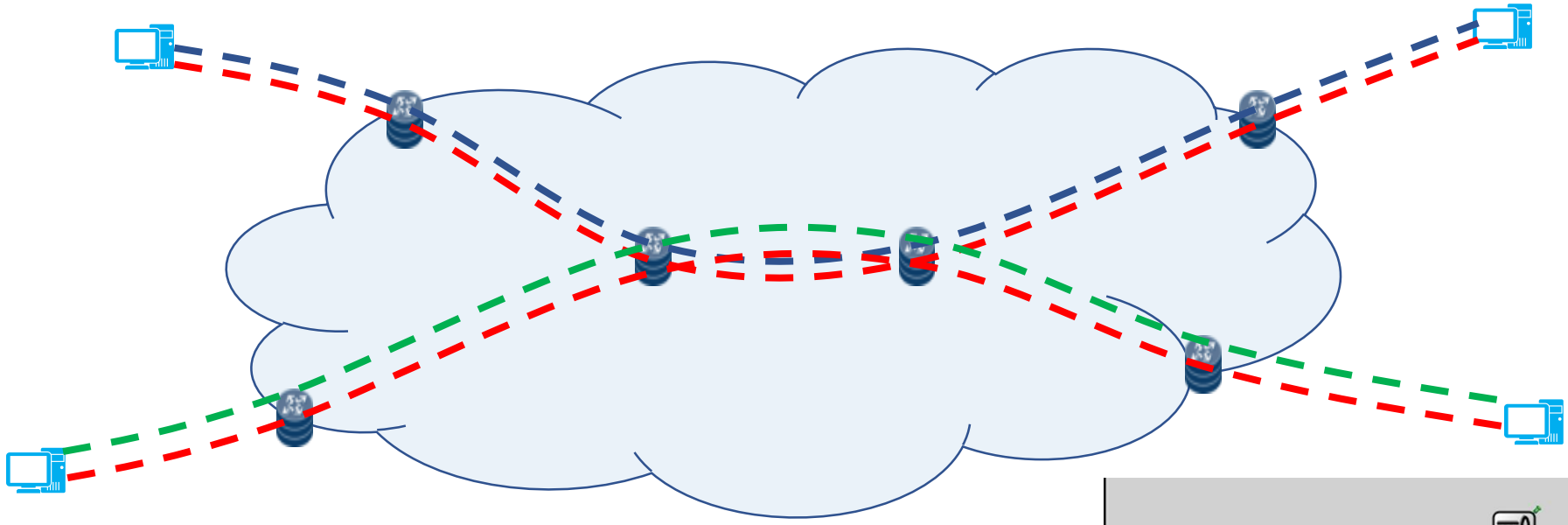
$L_{max}$is the maximum packet size

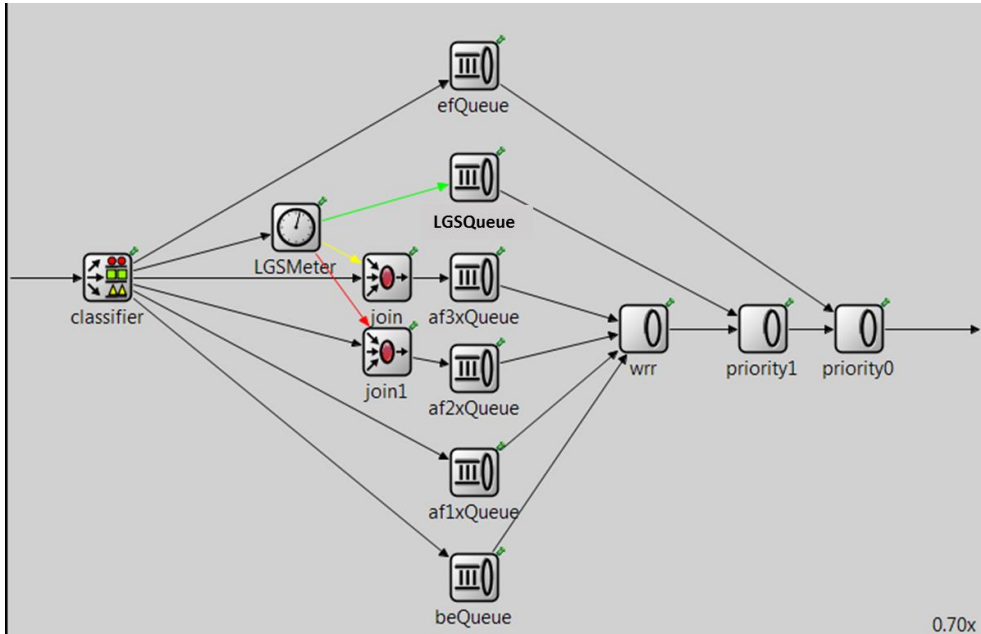


Configuration: Ingress Shaper at all PE Ingress Interface

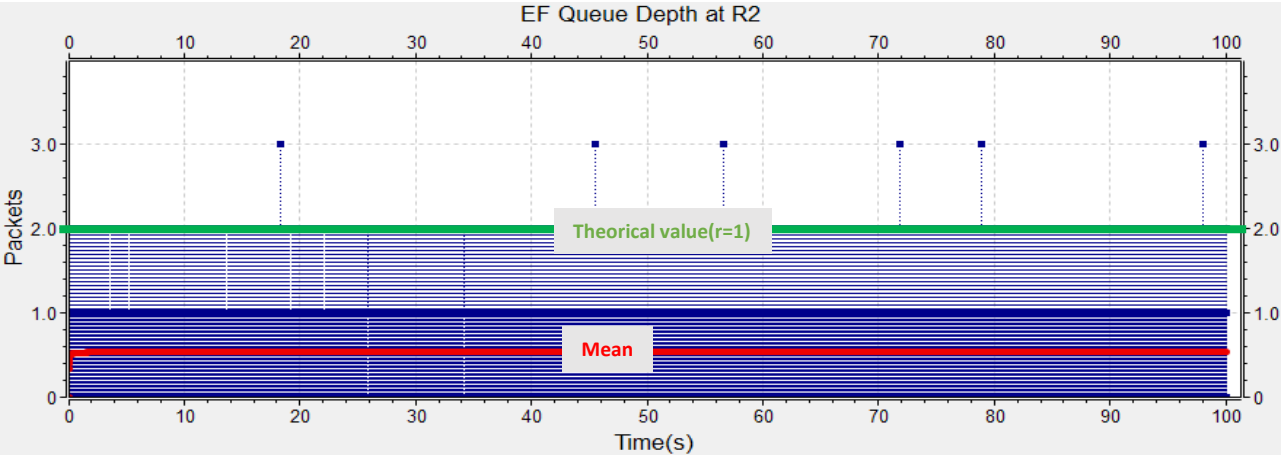


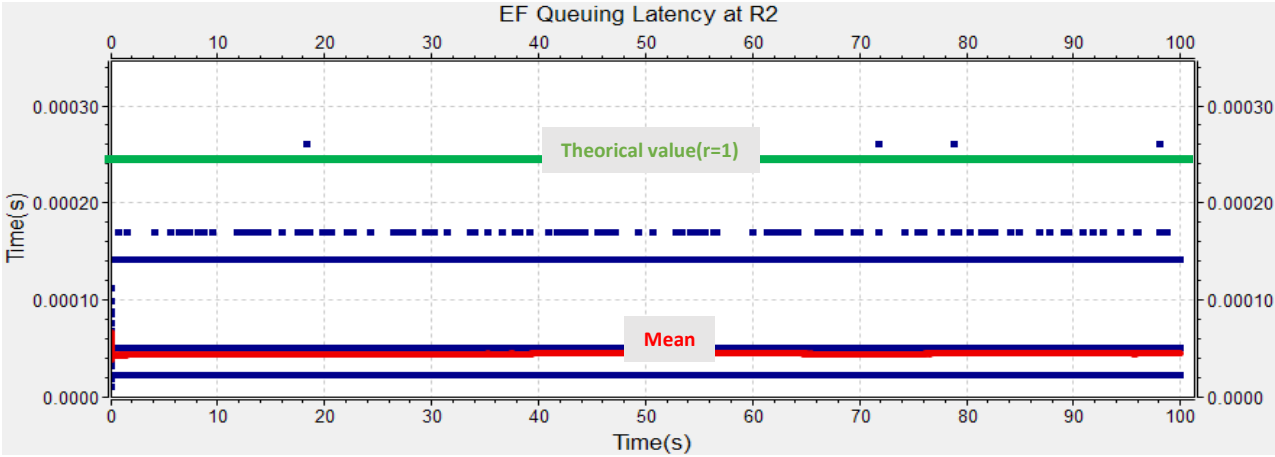
Configuration:
Egress Shaper and Queuing and Scheduler at all Router Egress Interface

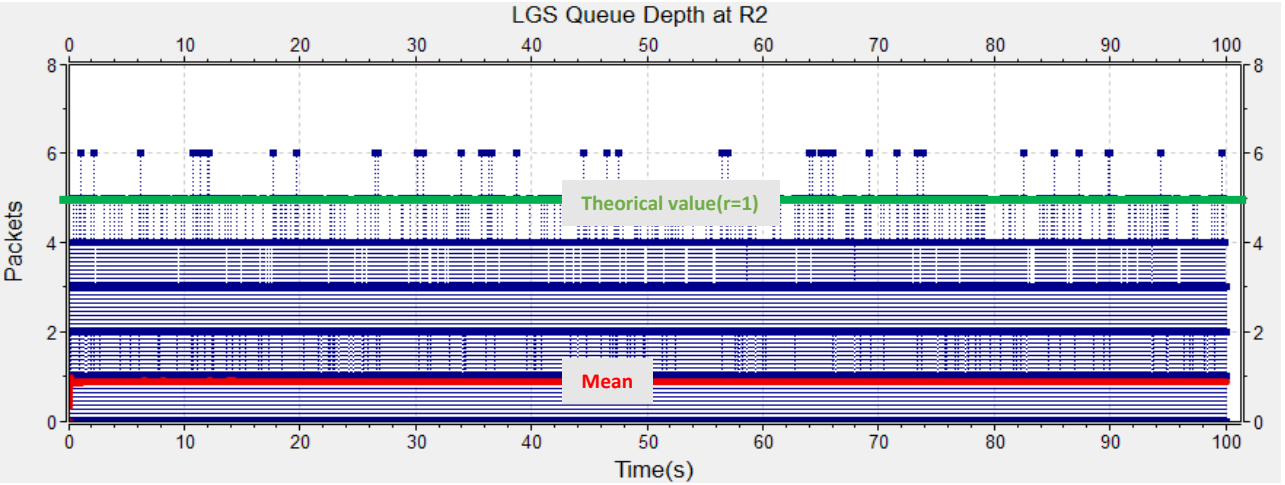


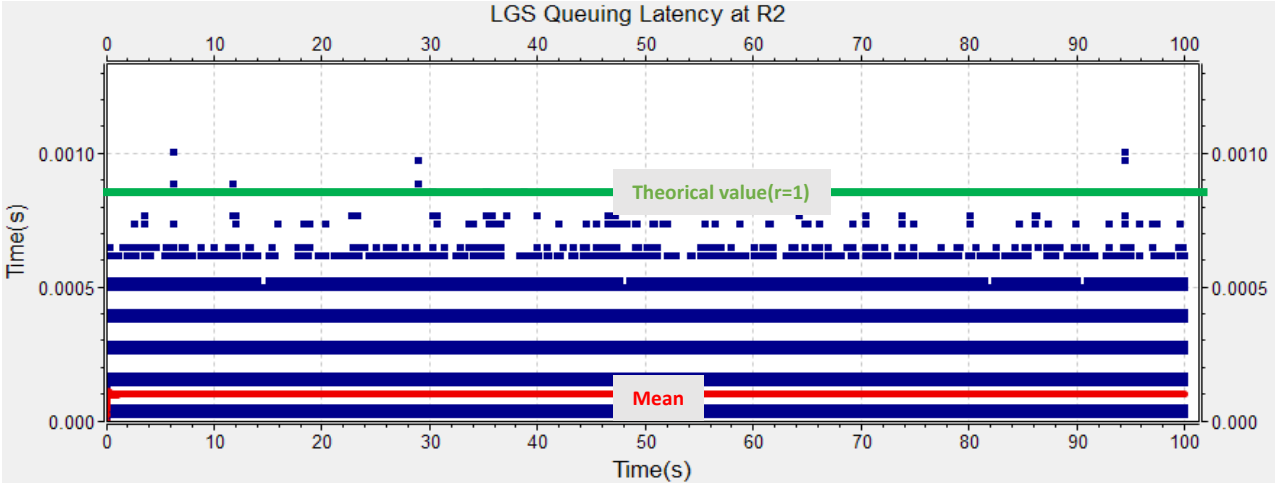


Case 2: Buffer Depth and Packet Latency: Measurement Vs Estimation (EF+LGS ≅ 65% link rate)



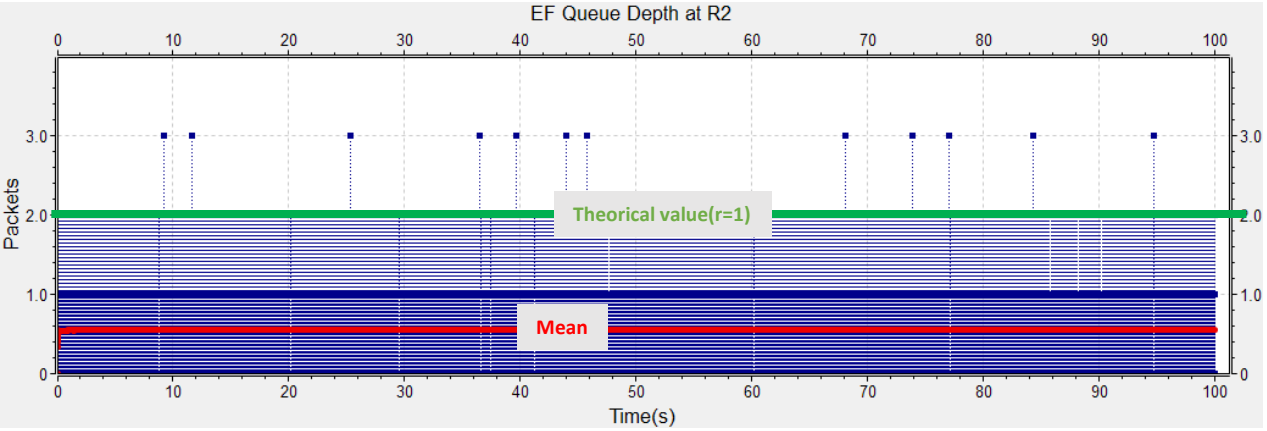


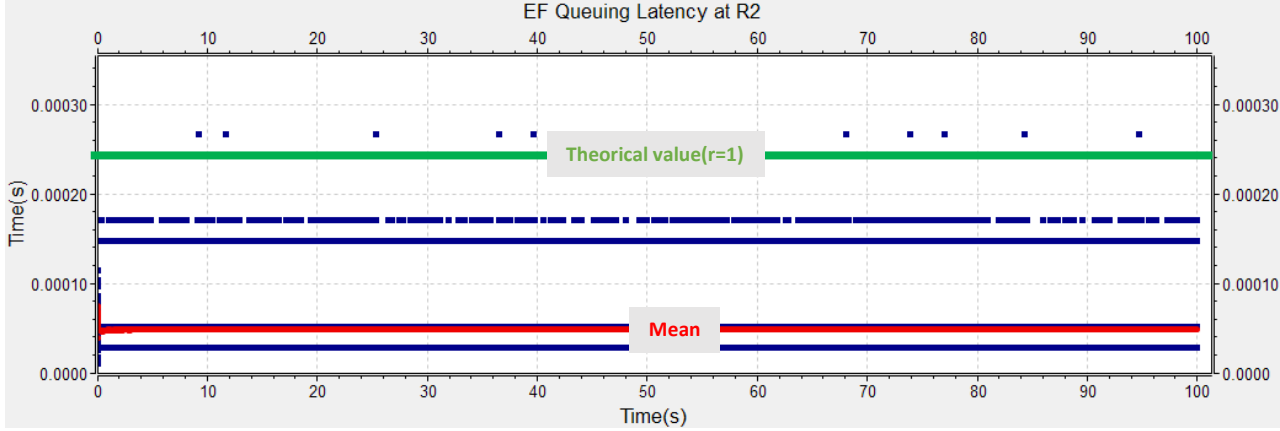


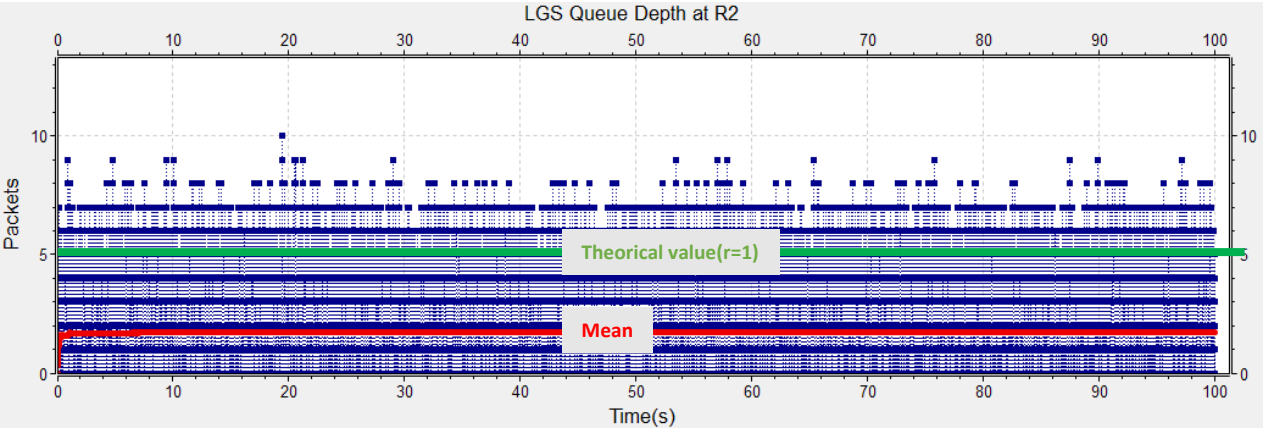


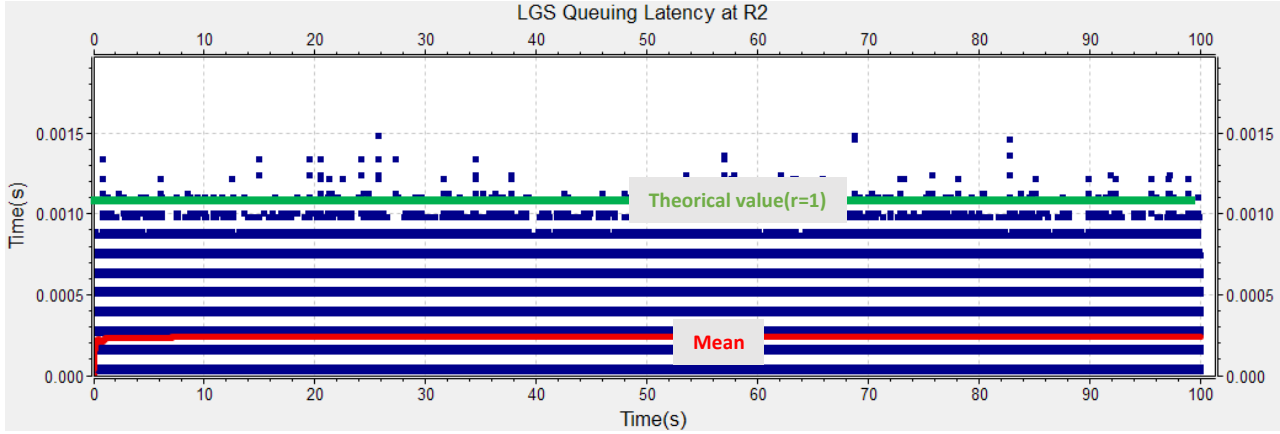
EF flows =4, LGS flows = 16; R_{EF}=14.385M, R_{LGS}=50.11M; Burst traffic, Burst duration = Exp(0.02s), Burst sleep = Exp(0.001s).
Fully congested at all routers by BE, All Link Util =~100

Case 3: Buffer Depth and Packet Latency: Measurement Vs Estimation (EF+LGS ≅ link rate)









EF flows =4, LGS flows = 16; R_{EF}=21.84M, R_{LGS}=69.66M; Burst traffic, Burst duration = Exp(0.02s), Burst sleep = Exp(0.001s). Fully congested at all routers by BE, All Link Util =~100

Experiment results

- Design targets can be satisfied by the Class Based Queuing + In-band Signaling
- The queuing latency of PQ is very short, it is not impacted by the congestion of WRR queues. The latency is related to link rate: 100M ~hundreds us; 100G~ hundreds ns.
- The queuing latency can be estimated with acceptable accuracy when EF+LGS flow rate is not close to the egress link rate.
- The estimation accuracy may decrease when the EF+LGS flow rate is close to the egress link rate. This can be explained by the impact of the jitter of the ingress traffic rate deviated from the average rate. When the jitter of ingress rate cause the rate exceeding the link rate, packet cannot be sent out in time and will build up.