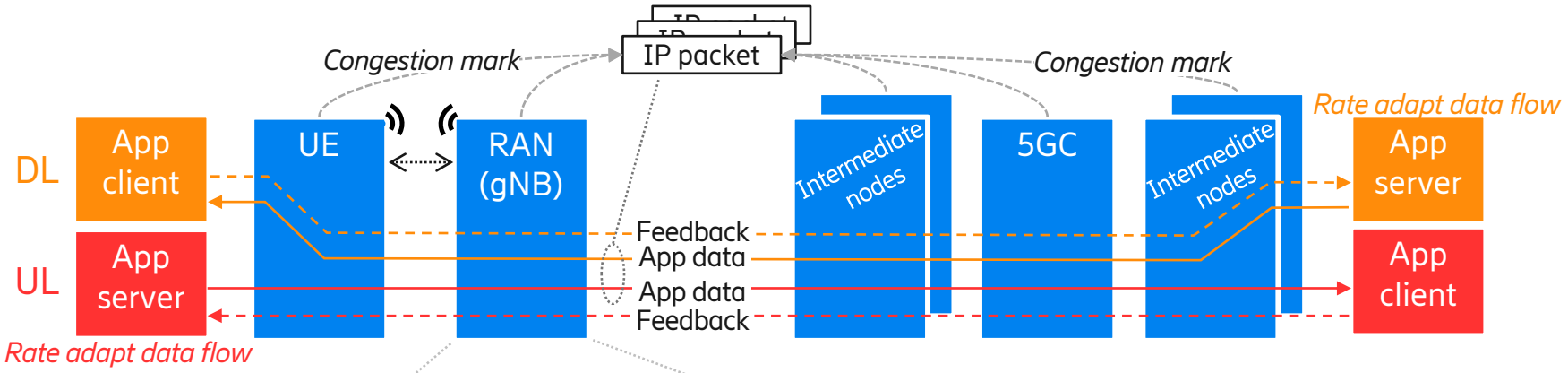
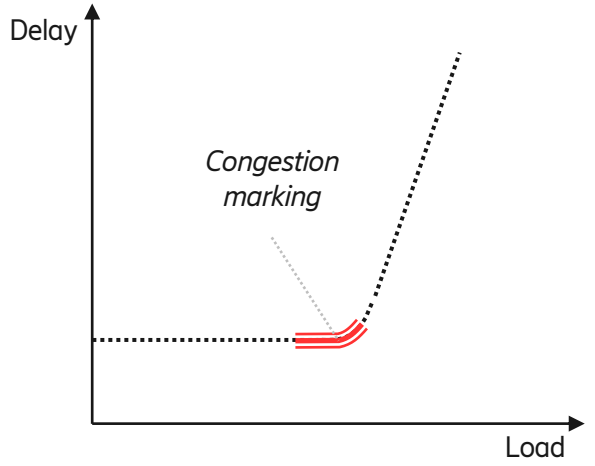
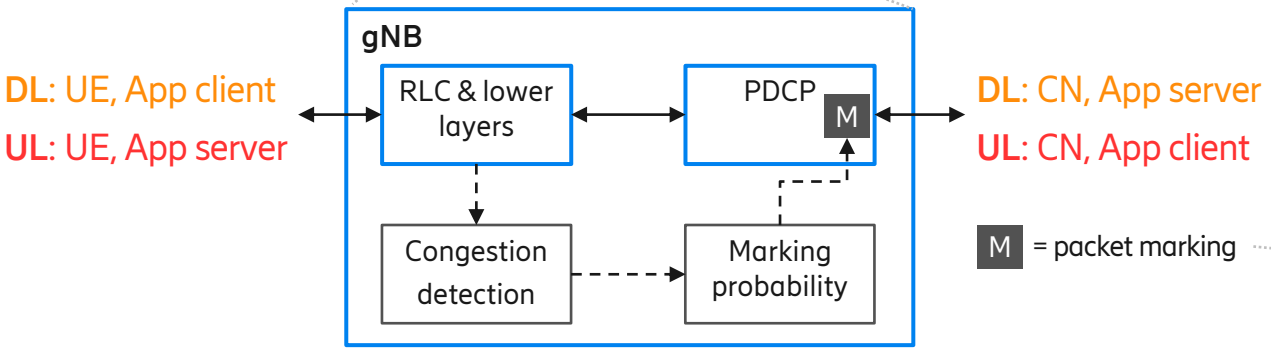
5G and L4S congestion control considerations

Ingemar Johansson, Ericsson AB
ingemar.s.johansson@ericsson.com

RAN supported L4S in 5G, overview







IP header, ECN bits

Non-ECT, 00	Not ECN-capable
ECT-1, 01	L4S-capable transport
ECT-0, 10	ECN-capable transport
CE, 11	Congestion experienced

Same RAN function blocks and information flow for DL and UL L4S

- **Congestion detection**
  - Estimates (queue) delay based on different metrics
- **Marking probability, PMark**
  - Calculates the fraction of packets to mark

Dedicated bearer for Latency-critical traffic

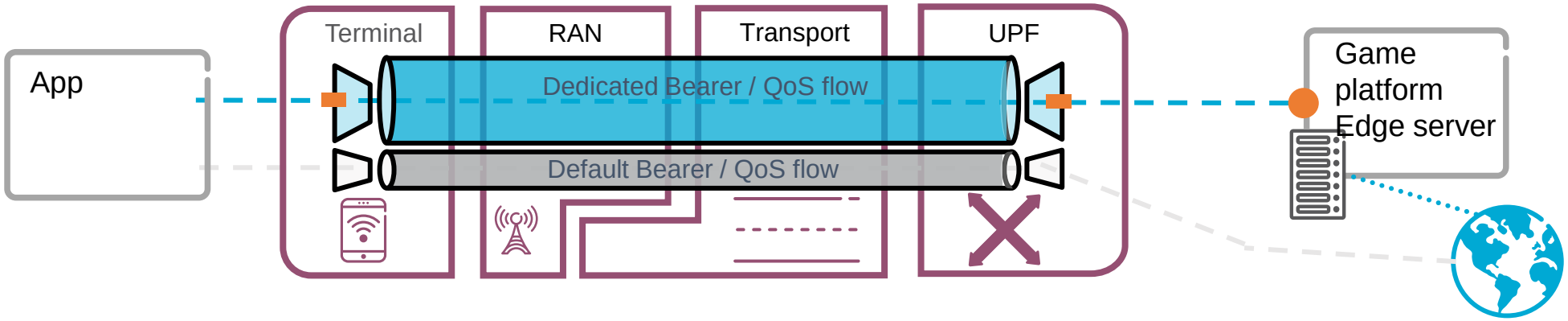
Reusing QoS/bearer features

Why a dedicated bearer/QoS Flow?

- Separate queue for latency-critical traffic (avoid disturbance from legacy traffic)
- Provide L4S marking on this traffic
- Optional: provide QoS/priority
- Opportunity to tune other functions for latency

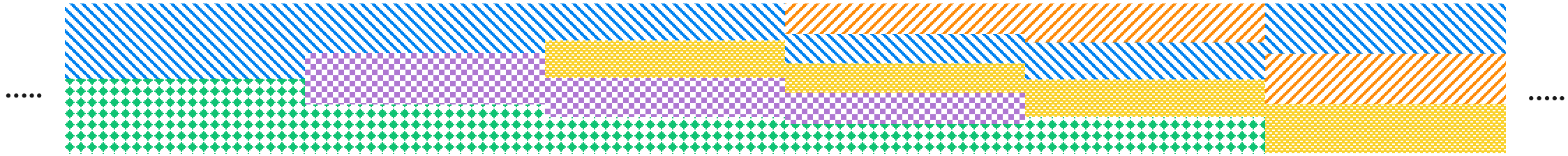
How to use dedicated bearer/QoS Flow?

- Traffic filters to map latency-critical traffic to bearer based on ECN/L4S markings
- Possibly combined with IP of Edge server

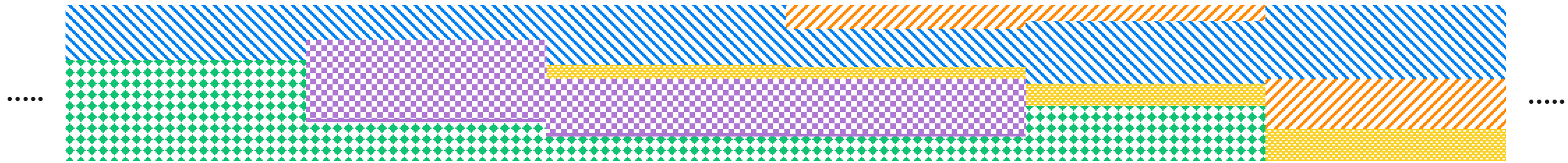


5G resource allocation and throughput

- Resource allocation in frequency and time (average)
  - End user applications may be bitrate limited
  - Resource allocation can drop in a few RTTs or in an instant when users enter (either new or HO from another cell)

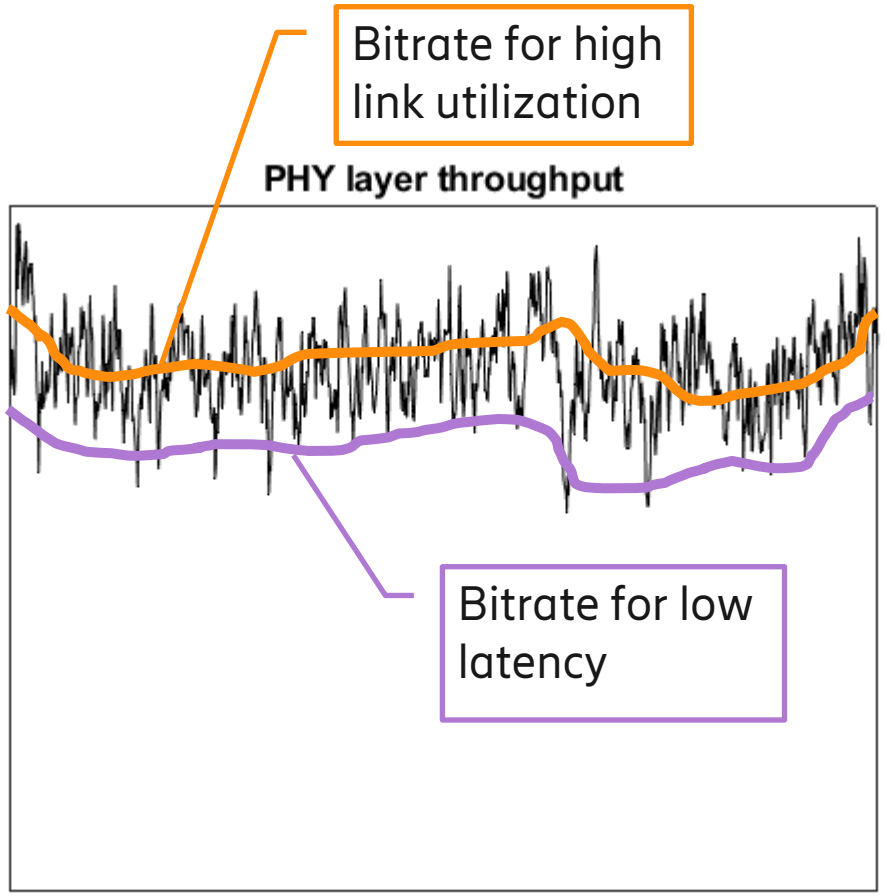


- Actual throughput (average)
  - This is the average..
  - Fast and slow fading gives varying throughput



L4S for all ?, yes and no…

- Cellular transport is subject to fast fading → throughput varies on short time scale
- Trade off between large network buffer, high link utilization **and** small network buffer reduced link utilization
  - It is more efficient to transmit larger chunks of data when a UE is anyway scheduled.
  - Opportunity to schedule a UE when the radio channel is good for that UE, hence increasing overall capacity … if there is a buffer of data
  - Multi-user MIMO will be more efficient if there is some slack to pick the right occasion to schedule the UE
- Possible middleground : BBR{v1,v2} or other rate based CC algos for not so latency sensitive traffic ?



Resource allocation

- A 5G scheduler allocates resource blocks to UEs (User Equipment = Modem or Cellphone)
- Allocated resources depend on activity among other UEs

- Rough overview of the dynamics :
  - New active user in a cell → resource allocation for other users reduce in a few RTTs (think slowstart)
  - New user makes handover from another cell → resource allocation for other users reduce instantly
  - The more users in a cell → smaller change
  - End user applications can be bitrate limited

Actual throughput

- Actual throughput is given by resource allocation and channel quality
- Attenuation
  - Distance to between UE and gNB (base station)
  - Obstruction (walls, trees…)
- Interference

- Slow fading ~100ms to seconds
- Fast fading milliseconds

- Reduced link throughput → reduce application bitrate promptly, or experience delay/loss
  - Very little grace time

- L4S is a congestion signal → does not save applications that react slowly

- QoS can give some slack

Delay jitter

- TDD patterns
- Number of active users
- HARQ reTx (~10% BLER) → ~4ms extra delay
  - Depends on numerology
- Uplink scheduling
  - Scheduling request for initial transmission
  - Pre-scheduling or configured grant scheduling can reduce jitter
- DRX (Discontinuous Reception), battery saving

- Jitter in uplink is typically larger than in downlink
  - SR/BSR process
- Delay jitter makes e.g. packet train measurements to probe for capacity problematic
- Endpoint congestion detection based on delay : queue build-up is hidden in delay jitter
  - L4S congestion signal from gNB can distinguish between delay jitter and actual queue build-up