Admission Control over DiffServ using Pre-Congestion Notification

Philip Eardley, Bob Briscoe, Dave Songhurst - BT Research
Francois Le Faucheur, Anna Charny – Cisco
Kwok-Ho Chan, Joe Babiarz - Nortel

IETF-64 tsvwg Nov 8th 2005

Summary

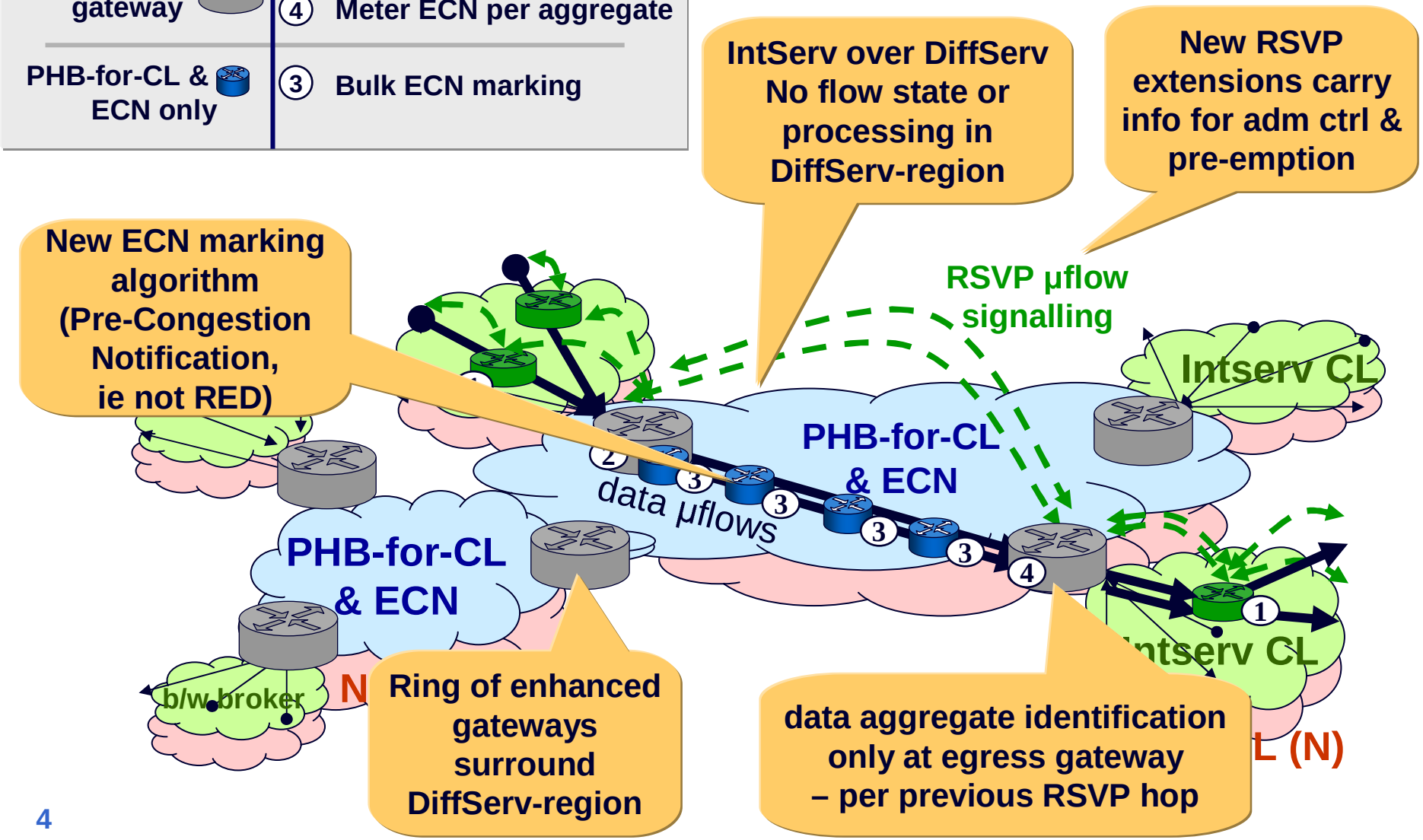
- Aim:
  - End-to-end Controlled Load (CL) service without flow state or signalling in the core / backbone
- Solution:
  - Builds on IntServ over DiffServ
  - new flow admission control mechanism (discover whether DiffServ region support another flow)
  - new flow pre-emption mechanism (if disaster means no longer possible to support all admitted CL flows, discover how many to pre-empt)
- drafts
  1. framework (architecture & use-case)
     - draft-briscoe-tsvwg-cl-architecture-01.txt
     - intention: informational
  2. Router marking behaviour definition
     - Coming soon…
     - intention: standards track
  3. RSVP extensions
     - draft-lefaucheur-rsvp-ecn-00.txt
     - intention: standards track

Summary [2]

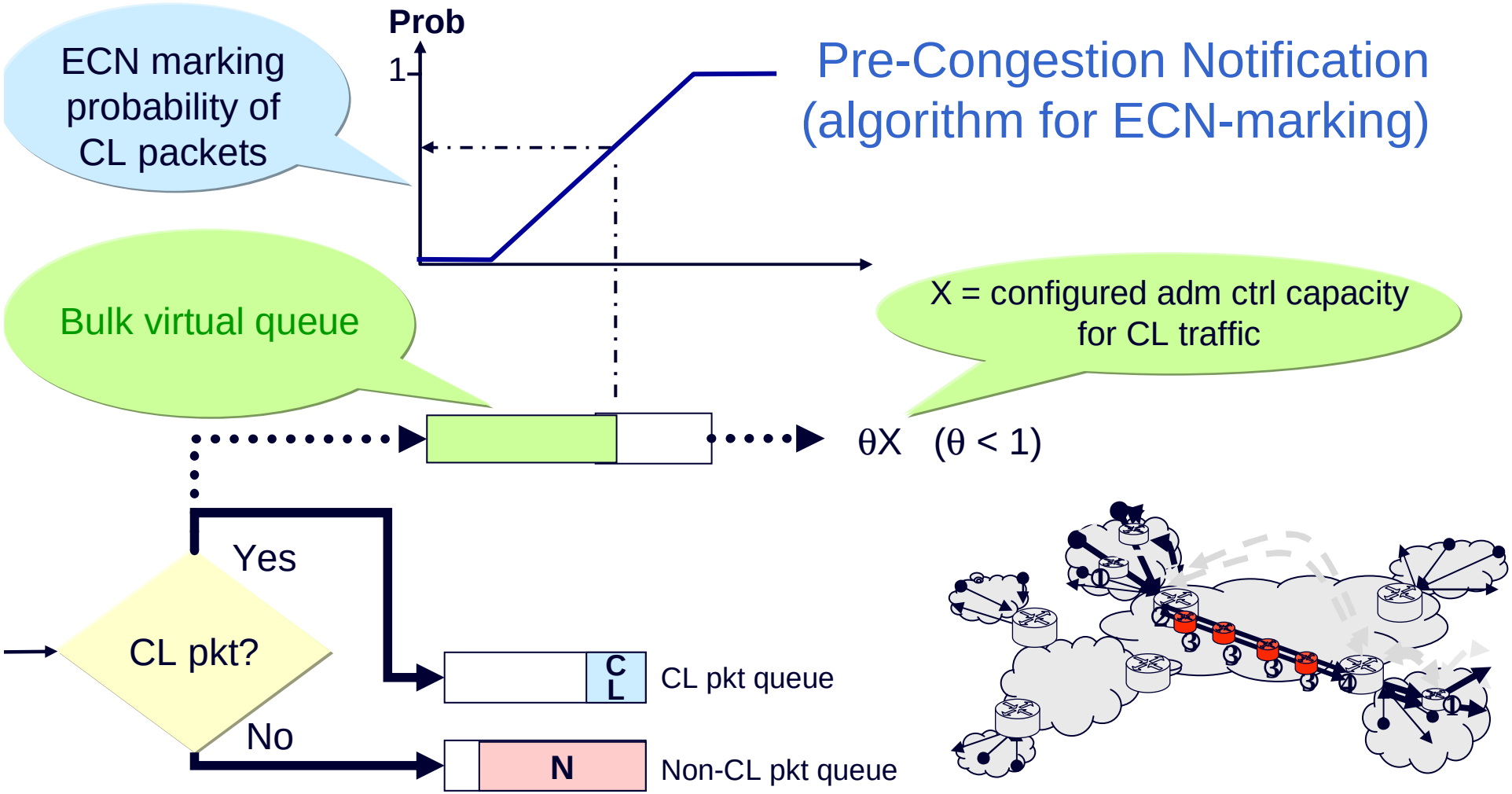
- History & changes
  - Previous draft, draft-briscoe-tsvwg-cl-architecture-00.txt, from BT only.
  - BT, Cisco & Nortel have been working together intensively
  - Admission control:
    - New consistent terminology: Pre-Congestion Notification, a new algorithm for ECN-marking CL-packets (as allowed by RFC3168 [ECN])
    - Intent is to fully aligned with RFC3168 (same ECN codepoints)
  - Flow pre-emption mechanism added
  - RSVP extensions done (could also use other signalling protocols, eg NSIS)
- Assumptions:
  - Edge-to-edge Aggregation: many flows over DiffServ region
  - Trust: all nodes in DiffServ region trust each other (but doesn't have to be any trust relationship with end-hosts)
  - Separation: all nodes in DiffServ region upgraded with Pre-Congestion Notification (ie satisfies draft-floyd-ecn-alternates-03.txt)

IP routers	Data path processing
Reservation enabled	① Reserved flow processing
RSVP/ECN gateway	② Policing flow entry to CL ④ Meter ECN per aggregate
PHB-for-CL & ECN only	③ Bulk ECN marking

end to end controlled load (CL) service using new edge-to-edge adm ctrl mechanism



Pre-Congestion Notification (algorithm for ECN-marking)



- Bulk virtual queue (a conceptual queue, used for measurement):
  - drained somewhat slower than the rate configured for adm ctrl of CL traffic
  - therefore build up of virtual queue is ‘early warning’ that the amount of CL traffic is getting close to the configured capacity
  - NB mean number of pkts in real CL-queue is still very small

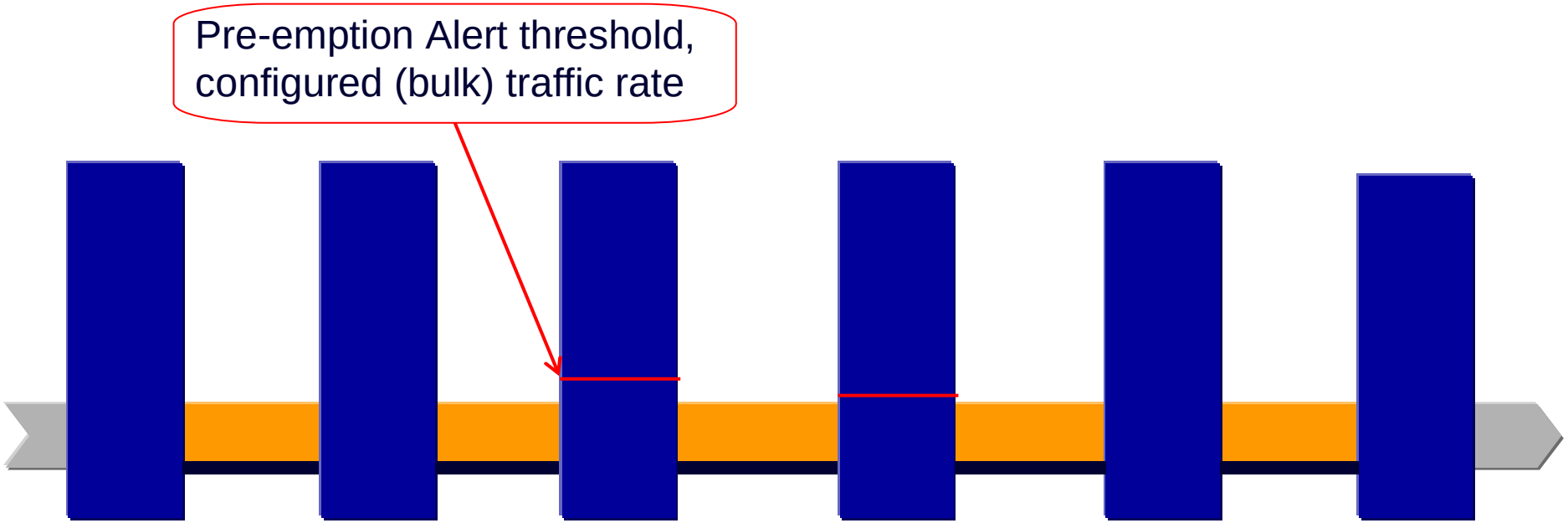
- Solution principles:
  - All routers in the DiffServ region can ECN-mark CL-pkts as ‘early warning’ of congestion, using the new algorithm
    - NB Bulk marking (not per flow)
  - Egress gateway meters ECN marks (moving average) (*congestion-level-estimate*)
    - NB Aggregate metering, ie per ingress (not per flow)
  - Ingress gateway admits new flow if *congestion-level-estimate* < threshold
    - *congestion-level-estimate* piggybacked on RSVP RESV (egress to ingress)

flow pre-emption

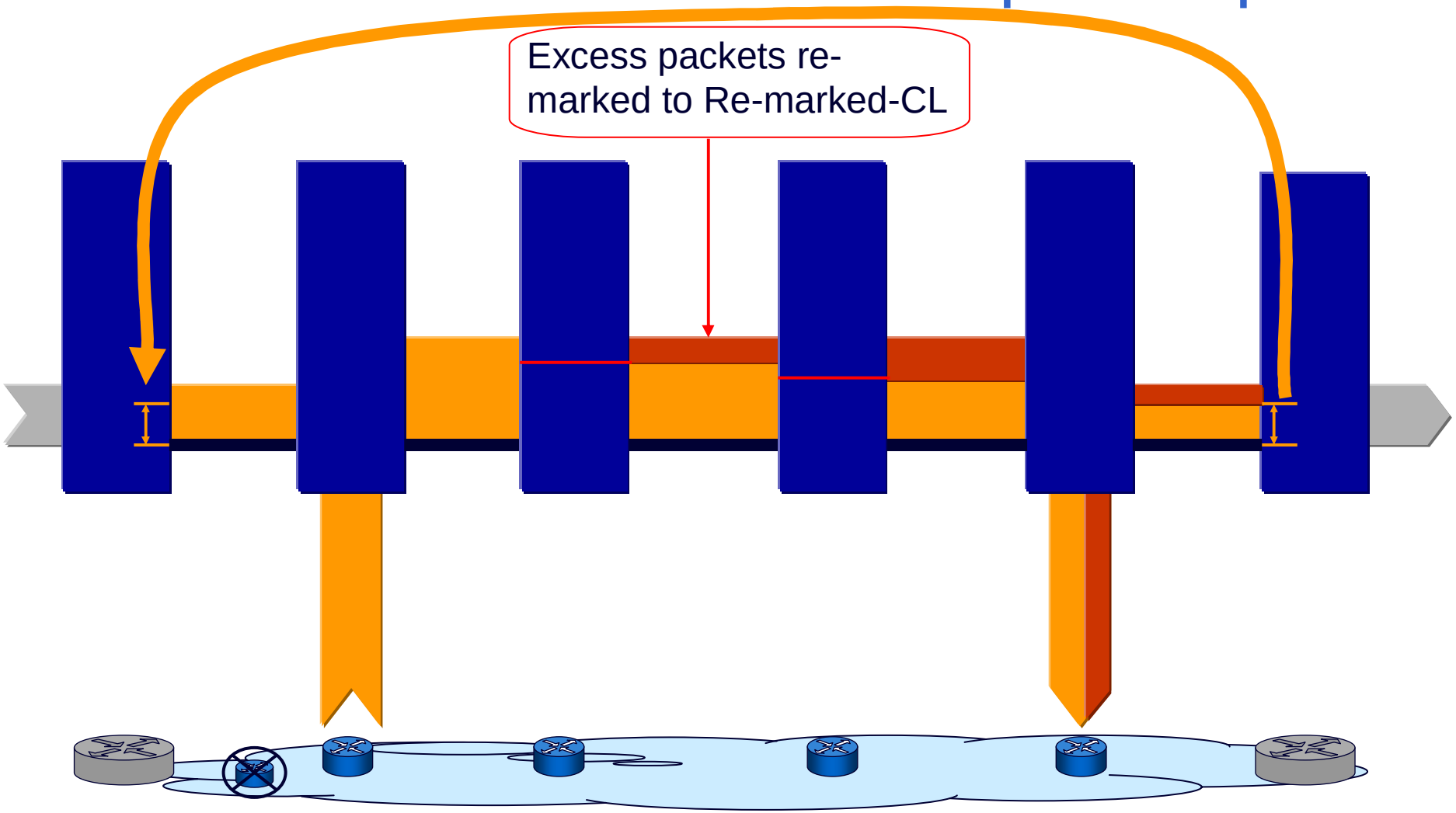
- the need for flow pre-emption
  - Coping with node/link failures (including multiple failures) in core networks is essential QoS issue
  - Consequent re-routing can cause severe congestion on some links and hence degrade the QoS
  - Need to support emergency/military calls (MLPP), especially in disaster scenarios
- rate-based pre-emption mechanism
  - Drop sufficient of the previously admitted CL microflows that the remaining ones again receive QoS commensurate with the CL service
  - Thus quickly restores acceptable QoS to lower priority classes
  - Better than just waiting for CL-sessions to end (which would eventually restore QoS)
- Solution is two-step process:

1. Alert the ingress that pre-emption *may* be needed
2. Ingress determines the right amount of CL-traffic to drop (if any)

flow pre-emption

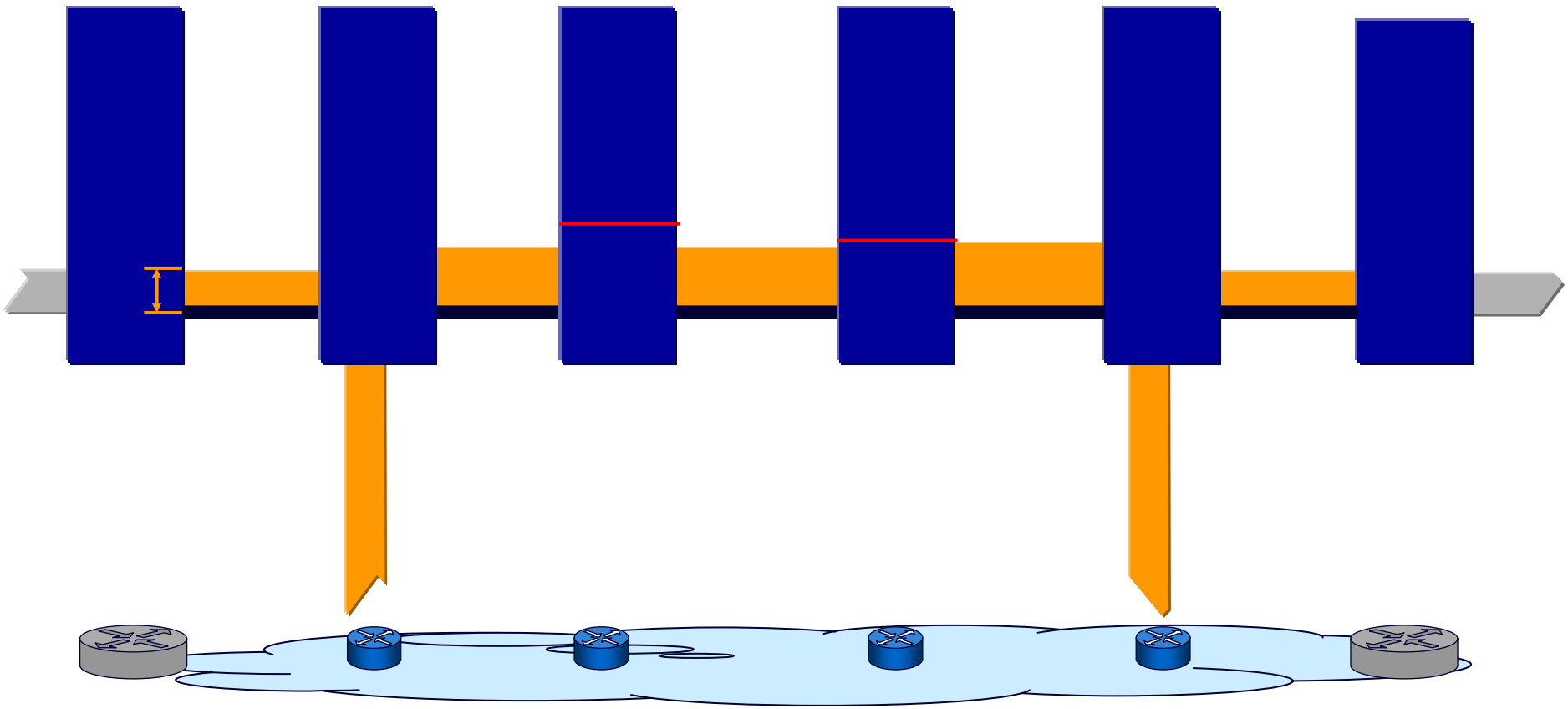


flow pre-emption



- Re-marked-CL triggers egress to measure *sustainable-aggregate-rate* ie how much CL traffic fits across the DiffServ region

After flow pre-emption



summary

- controlled load (CL) service
  - Builds on IntServ over DiffServ
- New mechanisms for DiffServ region
  - Distributed-measurement based Adm Ctrl
  - Rate-based flow Pre-emption
  - Based on bulk pre-congestion marking across the edge-to-edge region
- Standardisation required:
  - New router behaviour for Pre-Congestion Notification (ECN field) and Pre-emption Alert
  - RSVP extension – opaque object to carry congestion-level-estimate & sustainable-aggregate-rate
- We are working to finalise router behaviour draft

benefits…

- Statistical QoS guarantee
  - IntServ over DiffServ end-to-end, and new adm ctrl mechanism over edge-to-edge DiffServ region
  - Preserve QoS to as many flows as possible if heavy congestion, through new pre-emption mechanism
- Support of emergency & military MLPP
  - By flow pre-emption if heavy congestion
- Scales well & resilient
  - No signal processing or path state held on interior routers
- Control load dynamically
  - Avoid potential catastrophic failure problem for big networks with DiffServ architecture & statically provisioned capacity
- Minimal new standardisation
- Incremental deployment
- Deployment path for ECN
  - Operators can gain experience of ECN before end terminals are ECN capable

We would like to get your feedback & further build consensus on the drafts, aiming to move to WG item at next ietf

Extensions (in progress / potential)

(Section 5 of framework draft)

- Inter-operator (DiffServ region spans multiple, non-trusting domains)
  - ECN-based anti-cheating mechanism, same as in draft-briscoe-tsvwg-re-ecn-tcp-00
  - passive inter-domain policing (bulk metering only – nothing per flow)
  - Status: work done, draft soon (BT)
- Adaptive bandwidth for CL service
  - CL & non-CL share BW, based on relative demands, aims for economic efficiency whatever the traffic load matrix
  - Status: work done, on hold?
- MPLS-TE
  - Extend framework for adm ctrl into a set of MPLS-TE aggregates
  - need MPLS header to include the ECN field, which is not precluded by RFC3270
  - Status: is there community interest in this?
- Non-RSVP signalling
  - Eg NSIS could be used
  - Status: NSIS-community interest / help sought

Relationships to other QOS mechanisms
(Section 6 of framework draft)

- IntServ Controlled Load
  - Somewhat better, as get ‘early warning’ before router queue builds. Also more robust to route changes.
- IntServ over DiffServ
    - Same architecture
    - We have: RSVP-awareness confined to “border nodes” (gateways); “router marking” (by ingress)
- Differentiated Services
  - DiffServ protocol but not (info) DiffServ architecture (that has static provisioning, through traffic conditioning agreements at ingress)
- ECN
  - Comply with IP aspects of RFC3168 (ECN), but new feedback mechanism instead of TCP aspects of RFC3168
- RTECN
  - Very similar approach, but RTECN is host-to-host rather than edge-to-edge as here
- RMD
  - Broadly similar, especially RMD’s measurement-based adm ctrl mode
  - But RMD does hop-by-hop adm ctrl (all interior nodes in DiffServ region are QoS-NSLP aware & process RESERVE msg to compare the requested resources with {capacity minus current load})
  - Includes Severe Congestion handling – our Pre-emption has same aim but different method
- RSVP Aggregation over MPLS-TE
  - possible to extend our framework for adm ctrl of microflows into a set of MPLS-TE aggregates
  - would require MPLS header to include the ECN field (not precluded by RFC3270)