congestion marking for low delay (& admission control)

Bob Briscoe
BT Research
Mar 2005

RFC3168 (ECN in TCP/IP)

- For a router, the CE codepoint of an ECN-Capable packet SHOULD only be set if the router would otherwise have dropped the packet as an indication of congestion to the end nodes. When the router's buffer is not yet full and the router is prepared to drop a packet to inform end nodes of incipient congestion, the router should first check to see if the ECT codepoint is set in that packet's IP header. If so, then instead of dropping the packet, the router MAY instead set the CE codepoint in the IP header.[…]
- The above discussion of when CE may be set instead of dropping a packet applies by default to all Differentiated Services Per-Hop Behaviors (PHBs) [RFC 2475]. Specifications for PHBs MAY provide more specifics on how a compliant implementation is to choose between setting CE and dropping a packet, but this is NOT REQUIRED. A router MUST NOT set CE instead of dropping a packet when the drop that would occur is caused by reasons other than congestion or the desire to indicate incipient congestion to end nodes (…)

wider agenda I

- Meeting purpose: To find concensus combining the best parts of each proposal.

- Meeting approach:
  - Expose differences in requirements spaces between the main proposals
  - Elicit feedback on these differences in requirements from operators
  - Expose differences in technical approach between the main proposals
  - Identify which technical differences are essential for which requirements
  - Establish which technical differences need to be preserved (config options) and which can be discarded in favour of features of the other approaches
  - Decide which standards this will require and agree who will do what

  By the end, we will have succeeded if we are no longer talking about each approach as an integrated whole, but instead extracting the separate functional parts from each.

- Proposed agenda items

- 1. Reminder of proposals
  - Session admission control (Nortel)
  - Guaranteed QoS Synthesis (BT)
  - Flow-state aware (BT/Angram)
  - [Resource Management for DiffServ (RMD) (Ericsson)?]

  [I could also give an informal overview of the research literature on this field (Distributed Measurement-based Admission Control) if that is of general interest]

wider agenda II

- 2. Differences in emphasis of requirements
  - CAC: how strong? pre or post data?
  - pre-emption (emergency services etc)
  - Partitioned bandwidth or not?
  - interconnect: only bulk data, flow-aware?
  - applications: CBR, VBR? more specific (telephony, video)?
  - trust, proxies, authentication etc.
  - end-to-end? edge-edge?
  - business models (e.g. co-ordination of CAC response across domains, to synch charging triggers) simplex/duplex, etc?

  3. Wire protocols
  - 'signalling' from network elements to CAC decision points
    - remaining capacity, congestion?
    - marking algorithm(s)
  - 'signalling' from end-points (or their proxies) to network elements
    - bandwidth requests? nothing?

  4. Probing
  - at flow-start, at aggregate start, after idle/silence
  - passive (as data) or active (router alert)

  5. Deployment/interworking routemaps
  - differences in emphasis on what a good deployment strategy is
  - target architecture (not just constrained by incremental deployment, but where are we trying to get to)

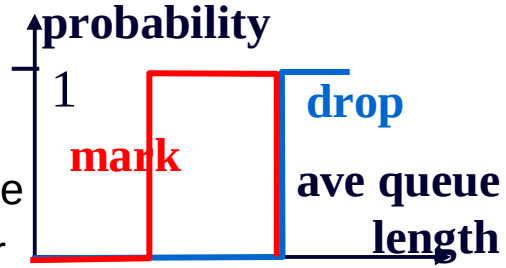
  6. Standards requirements
  - agree necessary standards actions
  - other claims on the protocol fields we're wanting to use

congestion of a class or a resource?

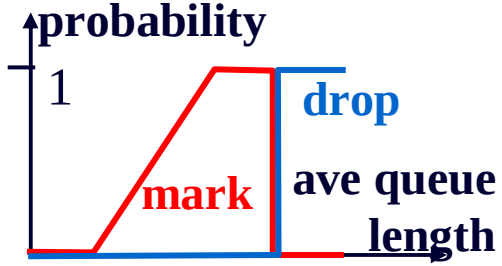
- operator should be able to configure either
- if traffic can borrow from other classes
  - ‘congestion’ should mean of the resource the traffic could use
- if traffic is confined to the resources of one class
  - ‘congestion’ should mean of the class

BT

congestion marking for admission ctrl on/off vs. gradual rise



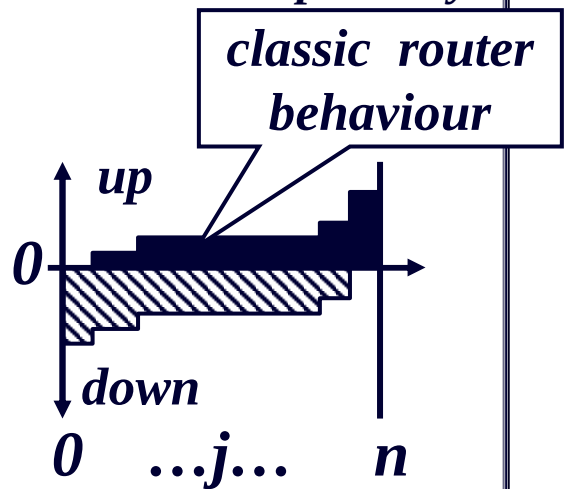
- on/off
  - disadvantage: requires smoothing, delaying response
    - could smooth at queue, or at admission controller
  - advantage: when (delayed) response triggered, one probe sufficient
- gradual rise: use virtual queue (cf. bulk reverse token bucket)
  - disadvantage: requires multiple probes to discover level
  - advantages:
    - can keep buffers empty – low delay
    - less sensitive to
    - can find congestion due to multiple bottlenecks
    - direct economic interpretation
- to be explored:
  - on/off based on virtual queue
  - might have all the advantages



re-ECN

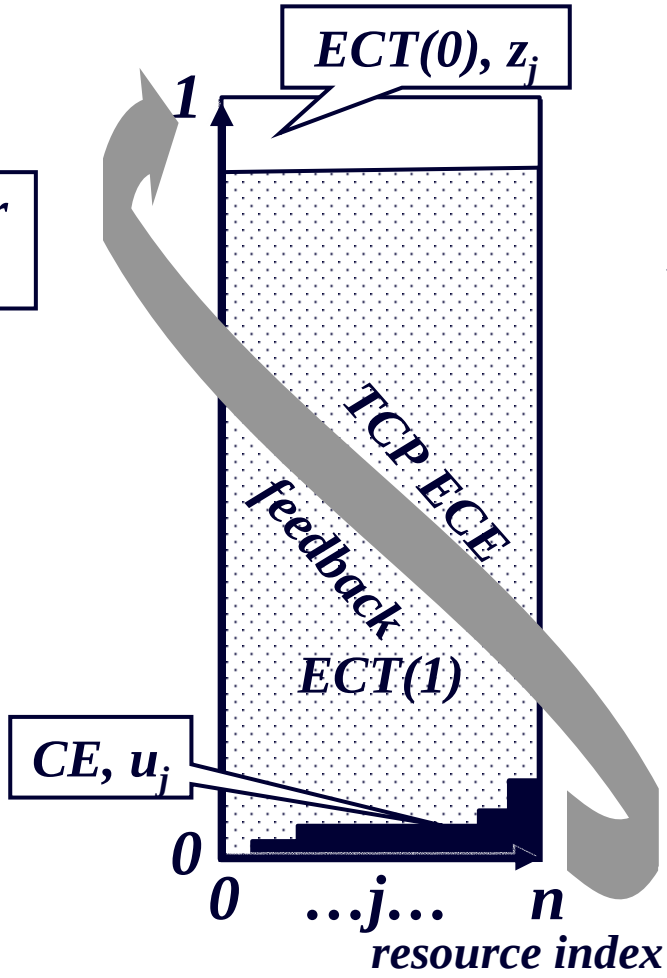
goals

- *backward compatibility*

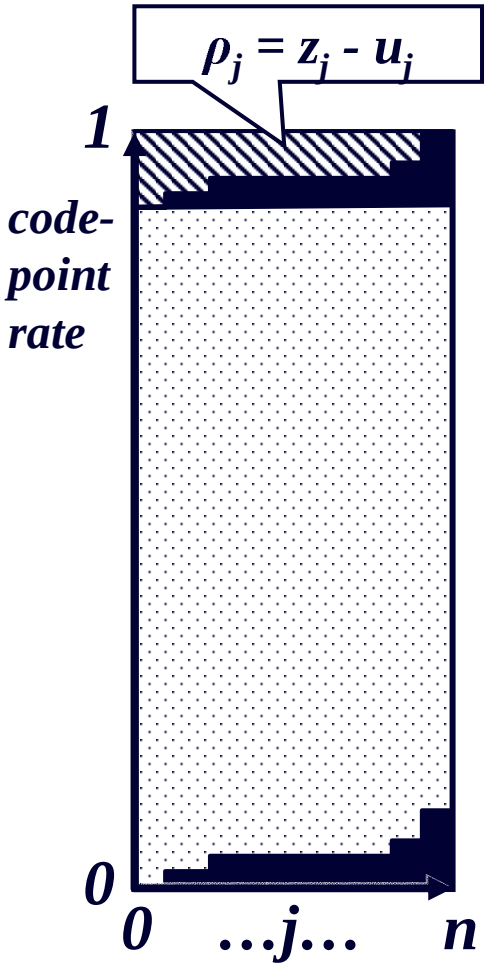


- *packets should carry downstream path congestion metric, ρ_j*

mechanism (approx)



'coding' (approx)



partitioning capacity (scheduler config)
Diffserv SLA provisioning v differential congestion marking

- if sharing capacity between classes
- higher priority traffic should be marked to reflect congestion it causes to any other traffic:
  - its own class and lower classes, but not higher
  - e.g. priority queue marked on length of itself plus lower priority queue

- don’t have to do this cross-class marking
  - but should not standardise anything that precludes it
  - lose ability to optimise network’s economics

BT

definition

- The congestion caused by a packet at single resource is the probability that the event X_i will occur if the packet in question is added to the load, given any pre-existing differential treatment of packets.
- Where X_i is the event that another selected packet will not be served to its requirements by the resource during its current busy period.

- This definition maps directly to economic cost
  - also usefully approximated by algorithms like RED

BT

spare slides

identifier-free differential treatment

- ✓ congestion status of network element
  - signalled irrespective of identifiers (bulk)
  - identifier in packet merely carries signal to dest (& source)
  - no policing, authorisation or authentication on network elements
  - policing can be done only at first ingress
  - border policing can be emulated by passive metering
- ✗ available capacity of network element
  - must be allocated per identifier
  - open to abuse:
    - split identity
    - arbitrage
  - requires policing ⇒ poor scaling at trust boundaries

BT