Re-ECN:
Adding Accountability for Causing Congestion to TCP/IP

Bob Briscoe, BT & UCL
Arnaud Jacquet, BT
Alessandro Salvatori, BT

IETF-64 tsvwg Nov 2005

initial draft

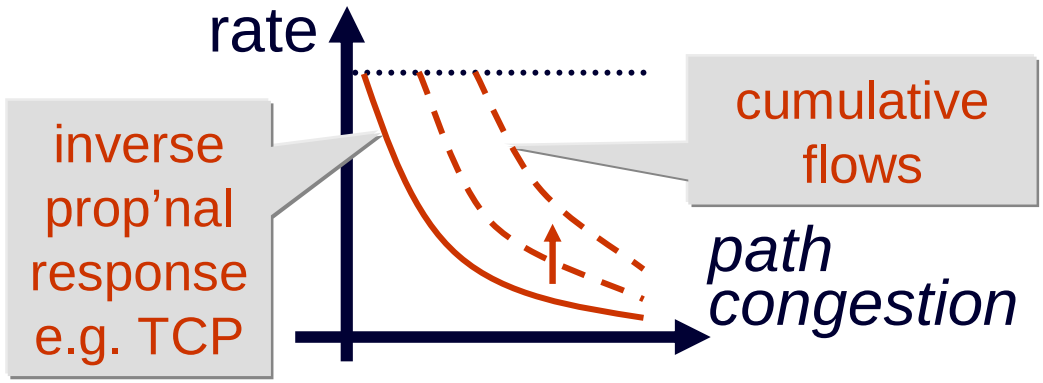
- IETF-63 Paris July 05
  - new research results (SIGCOMM'05) using ECN nonce codepoints
  - TSVWG chair asked for our proposal by IETF-64
  - hold ECN nonce (RFC3540) at experimental status
- re-ECN: adding accountability for causing congestion to TCP/IP
  - **initial draft:** draft-briscoe-tsvwg-re-ecn-tcp-00.txt *
  - **other formats:** www.cs.ucl.ac.uk/staff/B.Briscoe/pubs.html#retcp
  - **ultimate intent:** standards track (hope for working group draft soon)
  - **intent today:** get you excited enough to read it, and break it
  - **status:** haven't simulated this 2-bit IPv4/v6 proposal yet
    - our simulations based on a multibit ECN IPv6 extension header

* changed 2 field names since draft-00 – new terminology in this presentation

the problem: accountability for causing congestion

- main concern
  - non-compliance with e2e congestion control (e.g. TCP-friendly)?
  - how can ingress netwk detect whole path congestion? police cc?
- not just per-flow congestion response
  - **smaller:** per-packet
    - single datagram ‘flows’
  - **bigger:** per-user
    - a congestion metric so users can be held accountable
    - 24x7 heavy sources of congestion, DDoS from zombie hosts
  - **even bigger:** per-network
    - a metric for holding upstream networks accountable if they allow their users to congest downstream networks

previous work



- detect high *absolute* rate [commercial boxes]
  - but nothing wrong with high rate at low congestion
- sampled rate response to *local* congestion [RED + sin bin]
  - but congestion typical at both ends (access networks)
- transport control *embedded in* networks [ATM]
  - but limits behaviours to those standardised by network operators
- *honest* senders police feedback from rcvrs [ECN nonce]
  - but not all senders are community spirited (VoIP, video, p2p?, DoS)
- per-packet, per-user & per-network congestion policing
  - minimal previous work

basic idea (IP layer)

code-point	standard designation
00	not-ECT
10	ECT(0)
01	ECT(1)
11	CE

- sender re-inserts congestion feedback into forward data: “re-feedback”

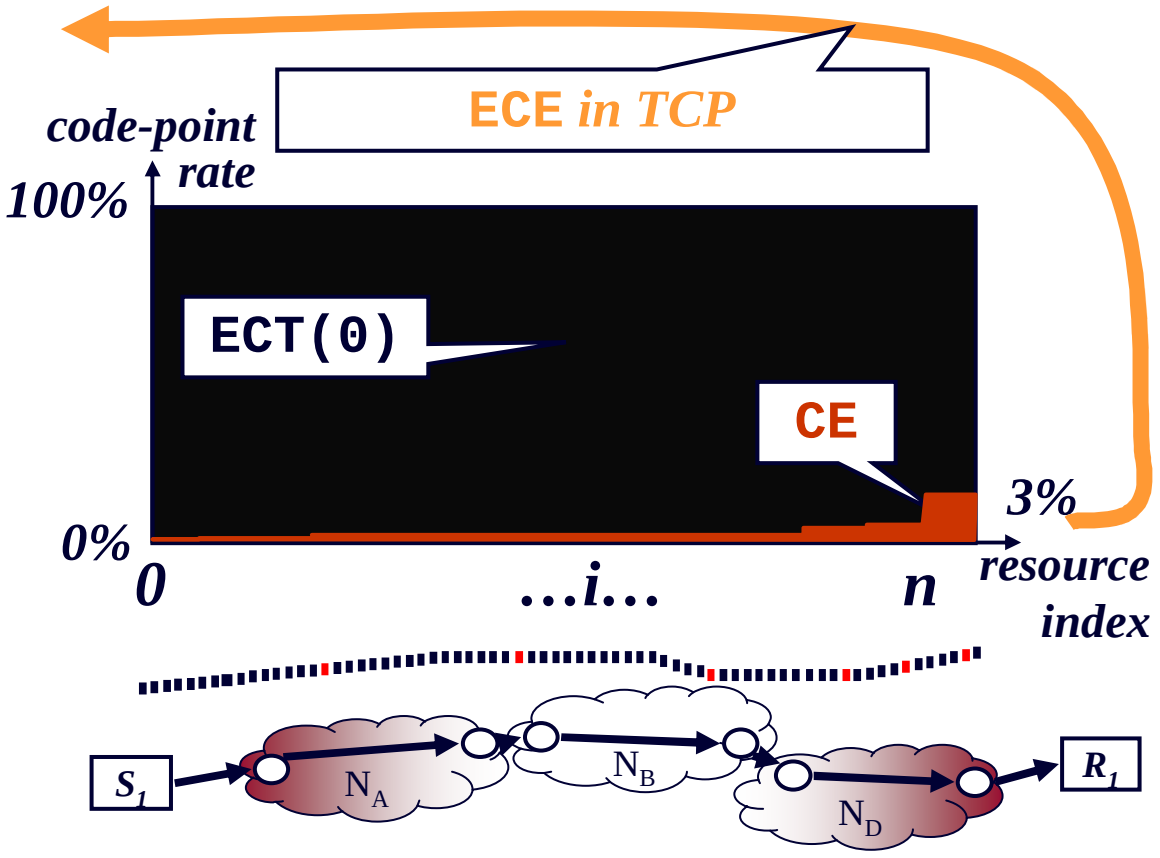
on every **Echo-CE** from transport (e.g. TCP)

sender sets **ECT(0)**
else sets ECT(1)

- and new Feedback-Established (FE) flag

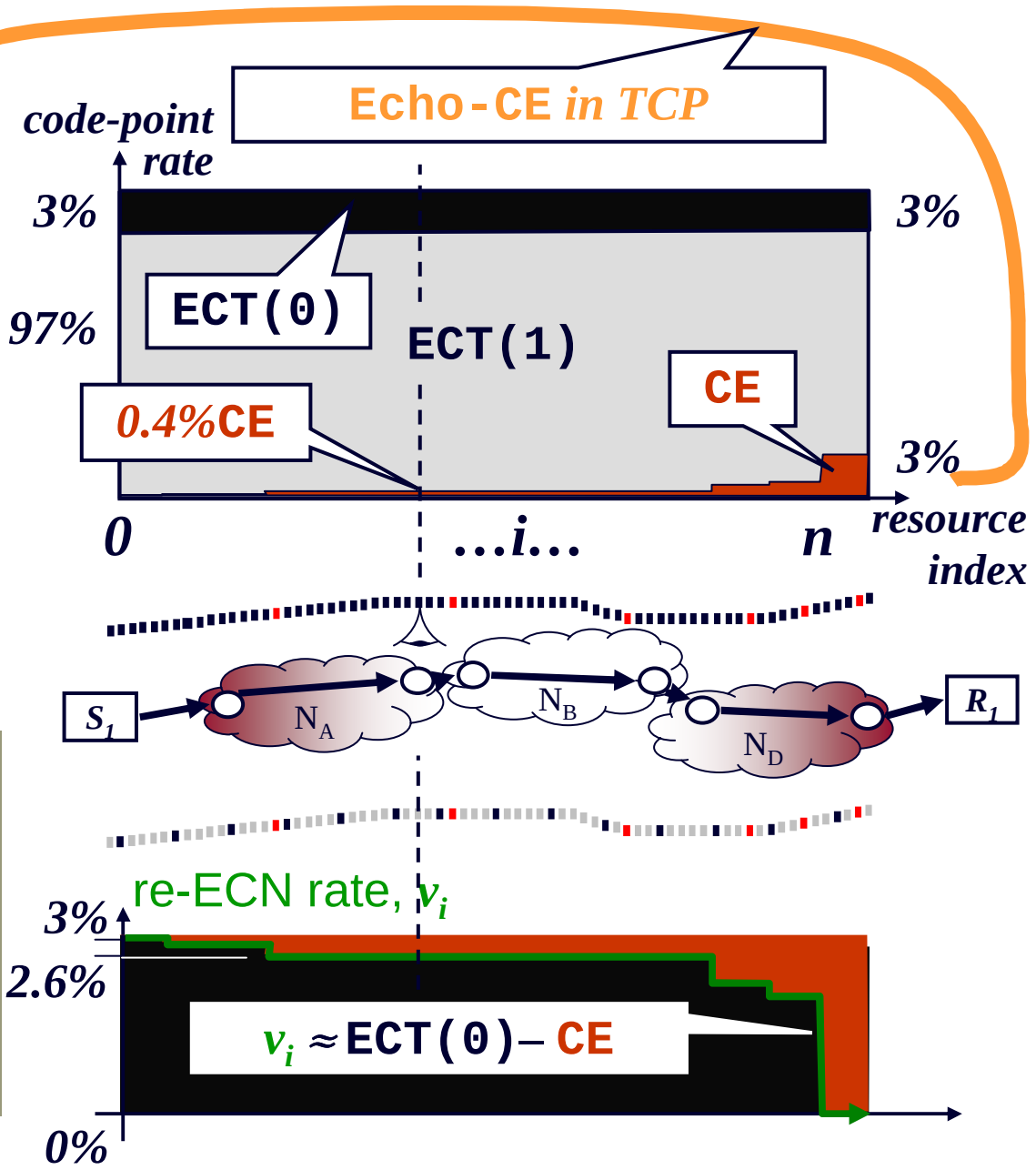
ECN
(recap)

code-point	standard designation
00	not-ECT
10	ECT(0)
01	ECT(1)
11	CE



re-ECN
(sketch)

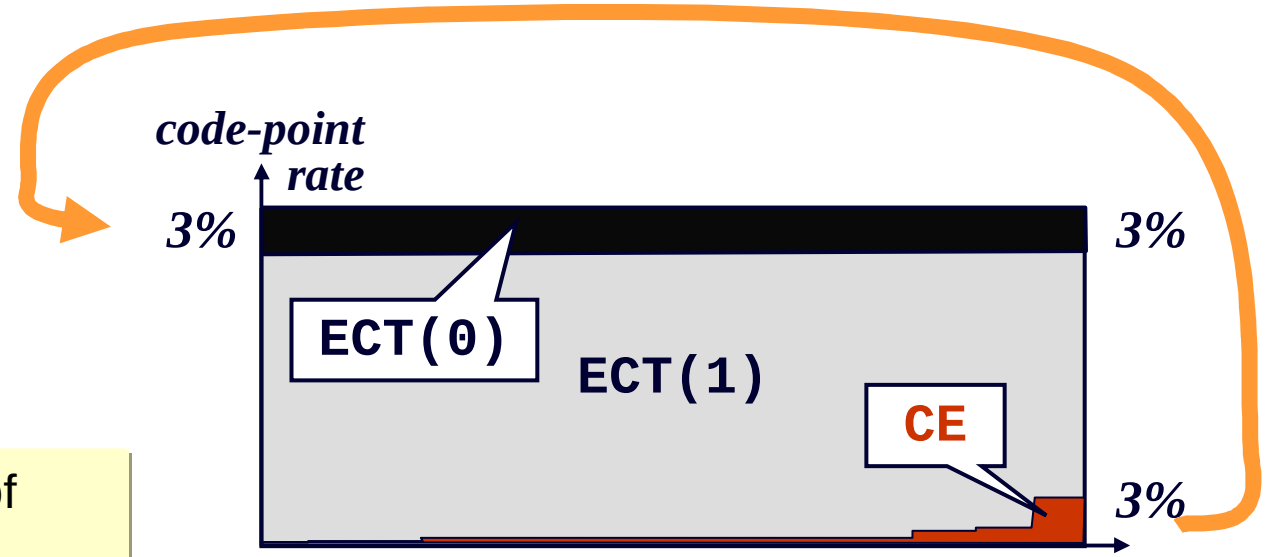
code-point	standard designation
00	not-ECT
10	ECT(0)
01	ECT(1)
11	CE

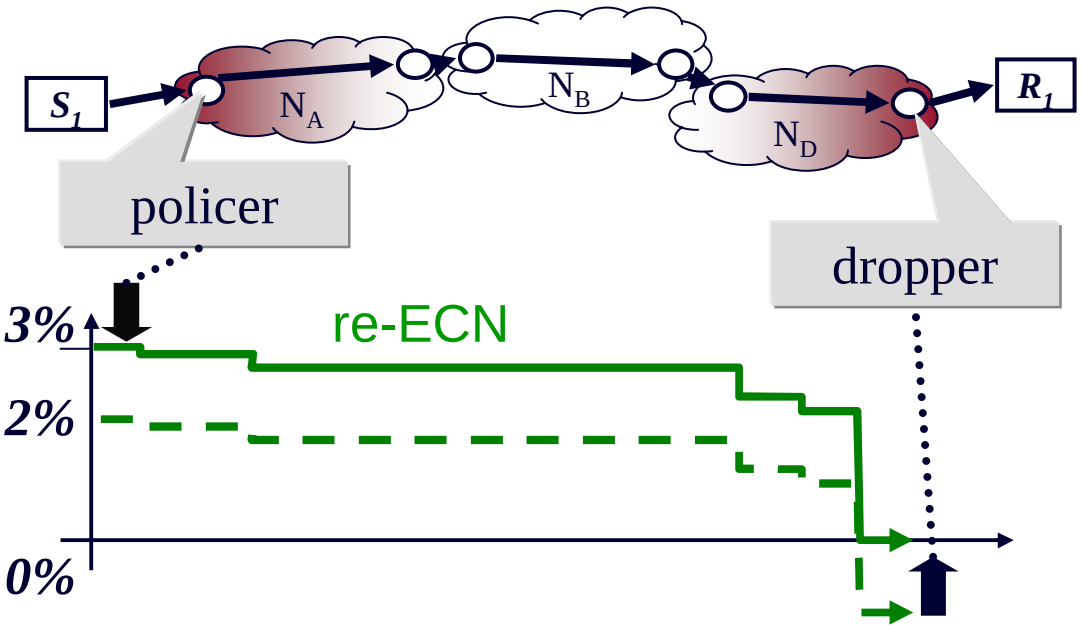


- on every **Echo-CE** from TCP, sender sets **ECT(0)**, else sets ECT(1)
- at any point on path, diff betw rates of **ECT(0)** & **CE** is downstream congestion
- routers unchanged

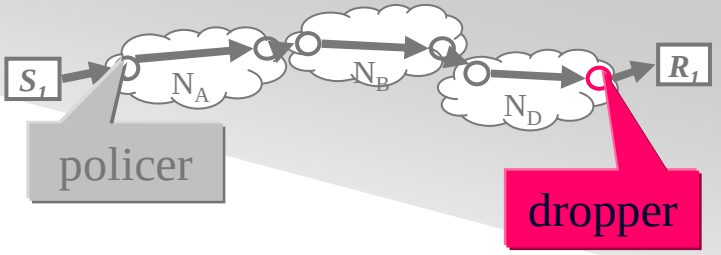
incentive framework
(user-network)

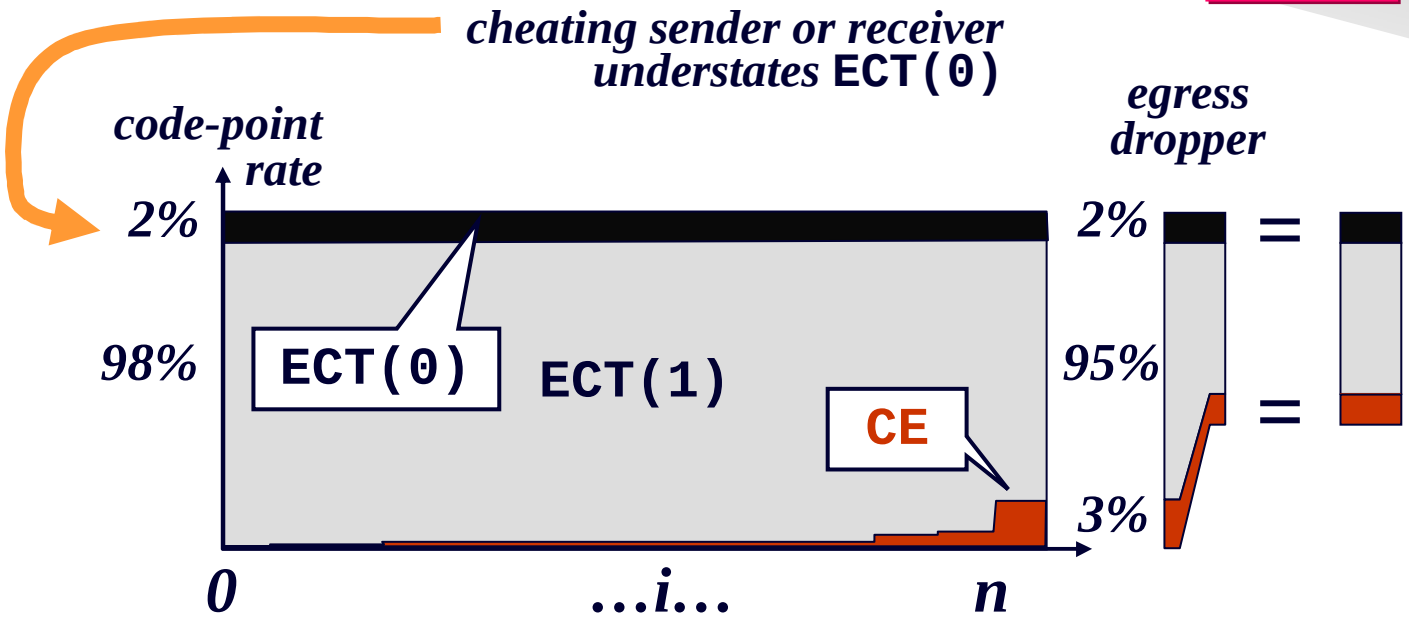
- packets carry view of downstream path congestion to each router
- so ingress can police rate response
  - using path congestion declared by sender
- won't snd or rcv just understate congestion?
- no – egress drops negative balance





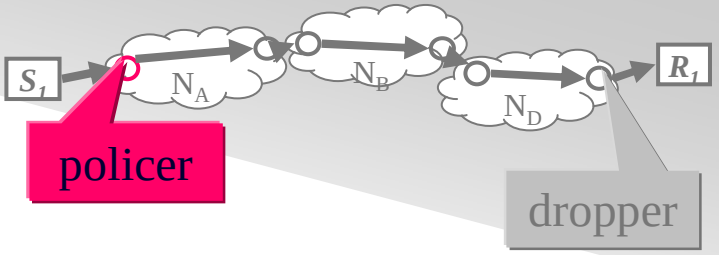
egress dropper (sketch)





- drop enough traffic to make rate of **CE** = **ECT(0)**
- goodput best if rcv & snd honest about feedback & re-feedback

- simple per pkt algorithm
  - max 5 cmp's, 5 adds, 1 shift
- dropper treats traffic in bulk
- can spawn focused droppers
  - misbehaving aggregates/flows prevalent in drop history



ingress policer (sketch)

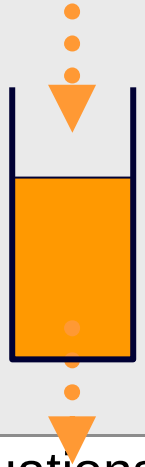
- packets arrive carrying view of downstream path congestion
- can police to any desired rate equation, eg TCP
- token bucket implementation: drop whenever empties
  - bounded flow-state using sampling

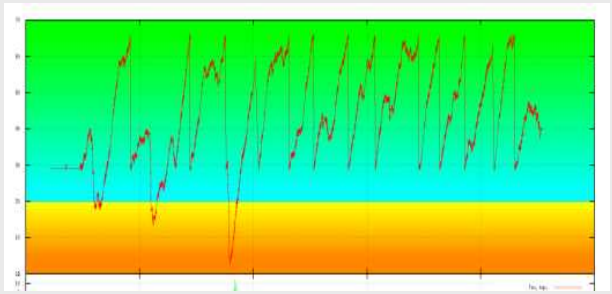
compliant rate

$$x_{TCP} \approx \frac{ks}{T\sqrt{p}}$$

actual rate

$$x = s/\Delta t$$





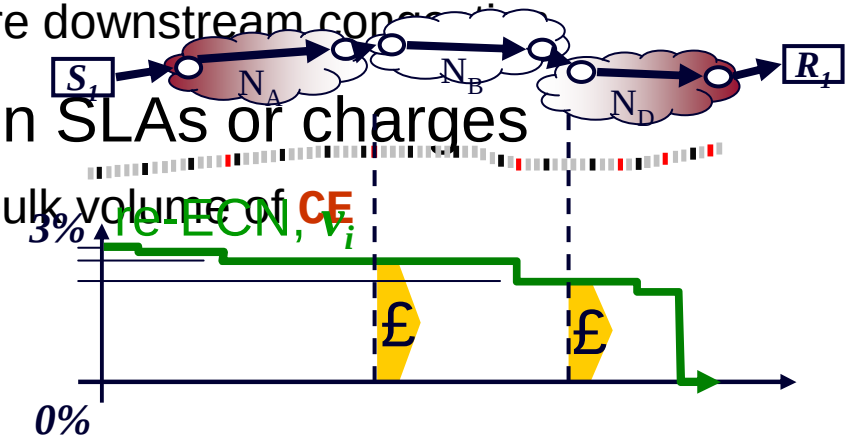
k	$\sqrt{(3/2)}$
s	packet size
T	RTT
p	marking rate
Δt	inter-arrival time

- above equations are conceptual, in practice can re-arrange
  - you get ***1/p*** by counting bytes between **`ECT(0)`** marks
  - high perf. root extraction per **`ECT(0)`** mark challenging (like pulling teeth)
- for RTT need sister proposal for ‘re-TTL’ (tba)

accountability for congestion

other applications

- congestion-history-based policer (congestion cap)
  - throttles causes of past heavy congestion (zombies, 24x7 p2p)
- DDoS mitigation
- QoS & DCCP profile flexibility
  - ingress can unilaterally allow different rate responses to congestion
- load sharing, traffic engineering
  - multipath routers can compare downstream conge
- bulk metric for inter-domain SLAs or charges
  - bulk volume of **ECT(0)** less bulk volume of **CE**
  - upstream networks that do nothing about policing, DoS, zombies etc will break SLA or



flow start

- re-ECN TCP capability handshake in draft
- feedback established (FE) flag in IPv4 header or IPv6 extension
  - future-proofing if short flows or single datagrams dominate traffic mix
  - FE flag only set by sender, only read by re-ECN security apps
  - leave FE=0 at flow start
  - if packet has FE=0, don't include its ECN marking in bulk averages
  - sender incentive to be truthful about FE flag
  - bit 48 (Currently Unused) flag in IPv4 header?
- TCP flow start specifics in draft
- guidelines for adding re-ECN to other transports in draft

re-ECN incremental deployment

- only REQUIRED change is TCP sender behaviour
- precision only if receiver is re-ECN capable too
- optional compatibility mode for ‘legacy’ ECN rcvrs
  - inclined to leave it out (so few Legacy-ECN hosts out there)
- no change from ECN behaviour for
  - routers
  - tunnels
  - IPsec
  - middleboxes etc
- add egress droppers and ingress policers over time
  - policers not necessary in front of trusted senders

re-ECN deployment transition

- if legacy firewalls block FE=1, sender falls back to FE=0
  - FE=0 on first packets anyway, so see connectivity before setting FE=1
  - if sender has to wrongly clear FE=0, makes dropper over-strict for all
- sender (and receiver): re-ECN transport (from legacy ECN)
  - ingress policer (deliberately) thinks legacy ECN is highly congested
    - 50% for nonce senders, 100% for legacy ECN
  - policers should initially be configured permissively
  - over time, making them stricter encourages upgrade from ECN to re-ECN

re-ECN deployment incentives

- access network operators
  - revenue defence for their QoS products
  - can require competing streaming services over best efforts to buy the right to be unresponsive to congestion
- egress access operators: dropper
  - can hold upstream neighbour networks accountable for congestion they cause in egress access
  - without egress dropper, border congestion could be understated
- ingress access operators: policer
  - if downstream networks hold upstream accountable (above)
  - ingress will want to police its heavy & malicious users
  - ingress can choose to rate-limit Not-ECT
- backbone networks
  - unless hold upstream accountable will be held accountable by downstream
- vendors of policing equipment
  - network operators invite to tender
- sender (and receiver): re-ECN transport (from Not-ECT)
  - network operator pressure encourages OS vendor upgrades (sweetener below)
  - Not-ECT rate-limits (above) encourage user upgrades
- end device OS vendors
  - network operators hold levers (policers) to encourage customer product upgrades

everyone gains from adding accountability to TCP/IP
except the selfish and malicious

re-ECN limitations

- snd or rcv can turn off ECN altogether to avoid policing
  - example: suffer drops (say 5%) instead of marking
  - but just add 5% FEC to compensate
  - not policed, so can add say 50% FEC to get 145% goodput
  - effectively how VoIP over BE works today
  - (ECN nonce no better in this respect)
  - solution: rate limit Not-ECT traffic in the future???
- dependency on getting re-TTL standardised
- takes a while for dropper & policer to detect malice
  - binary marking inherently slow to signal changes
- flow state at ingress policer & egress dropper
  - initial designs of policer and dropper with bounded state using sampling
  - don’t need port numbers – can just use IP address(es)

summary

- accountability for congestion
  - long-standing weakness of the Internet architecture
  - re-ECN appears to be a simple architectural fix in 1.5 bits
- main weakness with binary marking is attack detection speed
- request that ECN nonce is held as experimental
  - nonce only useful if sender polices receiver on behalf of network
  - re-ECN allows networks to police both sender and receiver and each other
  - re-ECN offers other accountability uses
  - but community needs time to assess
- makes ECN deployment more likely
  - change tied to new capabilities/products
  - not just performance enhancement

plans in IETF

- finish re-ECN draft
  - currently the text runs out after the TCP/IPv4 protocol spec
- re-TTL draft
- informational draft
  - on security applications, incl performance

- we strongly encourage review on the tsvwg list

- we are well aware this will be a long haul

Re-ECN:
Adding Accountability for Causing Congestion to TCP/IP

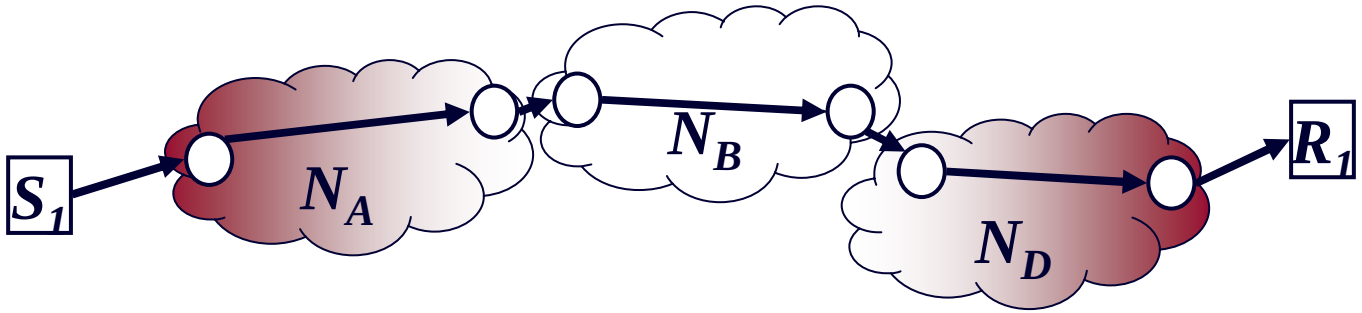
draft-briscoe-tsvwg-re-ecn-tcp-00.txt

Q&A

UCL

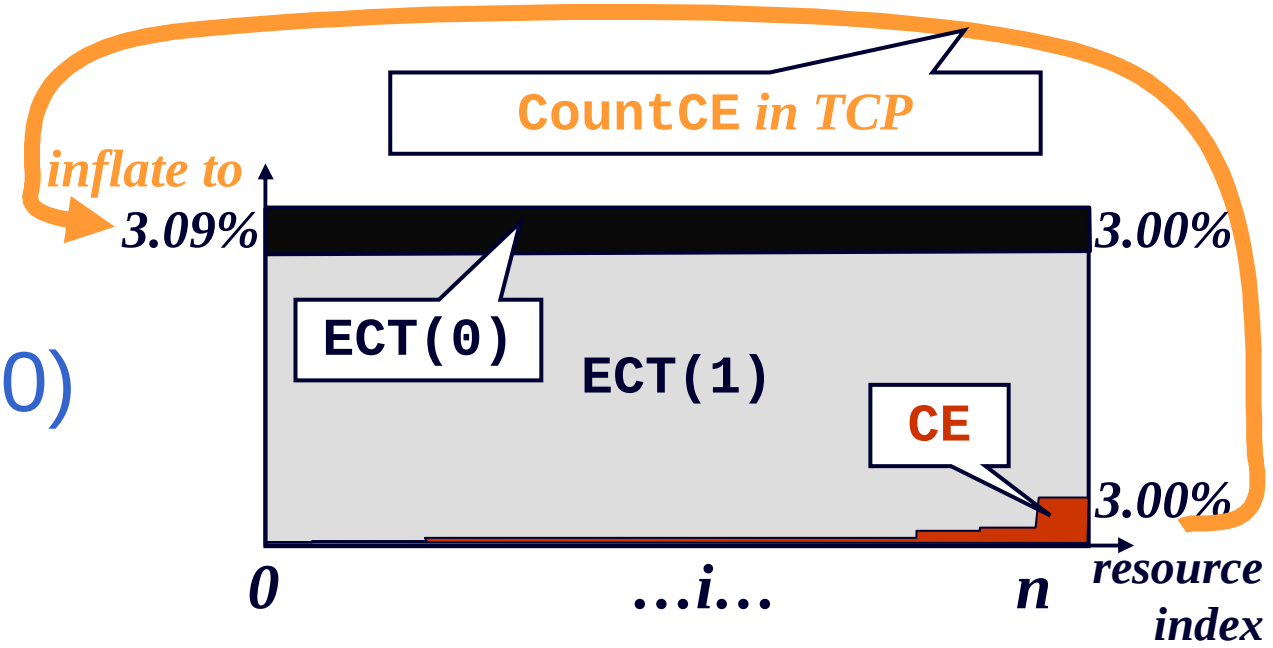
BT

path congestion typically at both edges



- congestion risk highest in access nets
  - cost economics of fan-out
- but small risk in cores/backbones
  - failures, anomalous demand

allowance for losing some ECT(0)



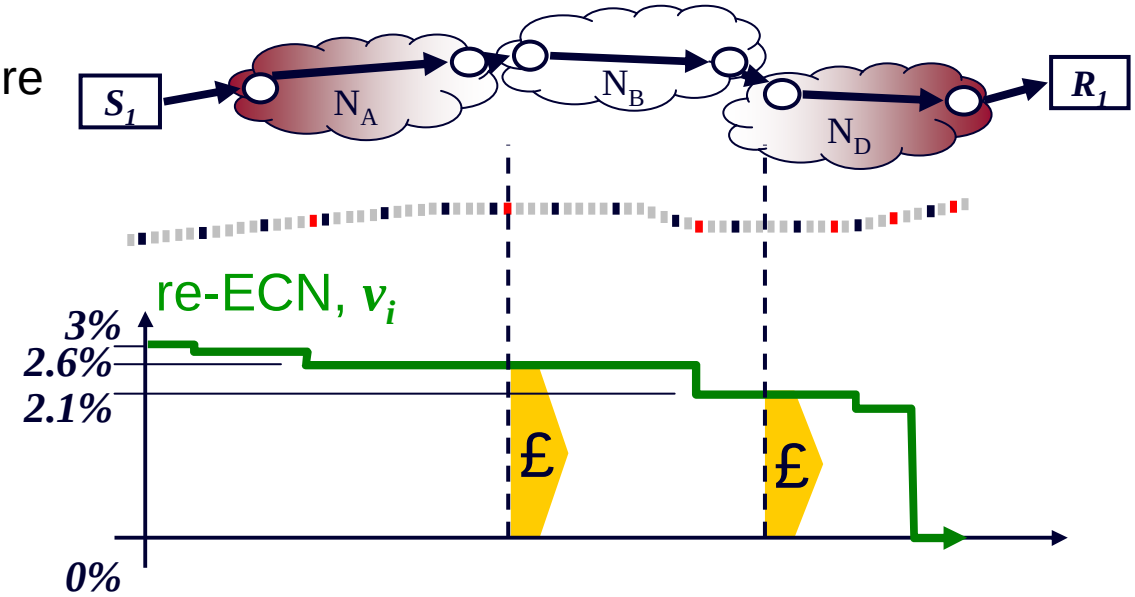
- aim for equal rates of **ECT(0)** and **CE** at egress
  - sender inflates **ECT(0)** to 3/97 = 3.09%
  - allows for 3% of 3.09% = 0.09% **ECT(0)** getting marked **CE**
  - simple packet counting algorithm for sender in draft (self-clocked)
- 'legacy' ECN receiver repeats **ECE** for a round trip until **CWR**
  - hides second and subsequent **CE** per RTT
  - new **CE** counter technique in draft
    - uses three flags in TCP options as a 3-bit **CountCE** counter, modulo 8
    - still safe against pure ACK losses
      if ack'd seqno gap ≥ 8, assume all missed ACKs marked

flow start

- re-ECN capability handshake in draft
- feedback established (FE) flag in IPv4 header or IPv6 extension
  - future-proofing if short flows or single datagrams dominate traffic mix
  - set by sender, used by re-ECN applications
  - leave FE=0 at flow start
  - if packet has FE=0 don't include its ECN marking in bulk averages
  - bit 48 (Currently Unused) flag in IPv4 header?
- getting feedback established, general idea for TCP
  - start with **`ECT(0)`** (be conservative until feedback established)
  - only set FE=1 on packets released by feedback
    - packets 2 and 6, 8, 10 etc during slow-start (assuming init window =4)
    - once in congestion avoidance, set FE=1 on all packets
- guidelines for adding re-ECN to other transports in draft

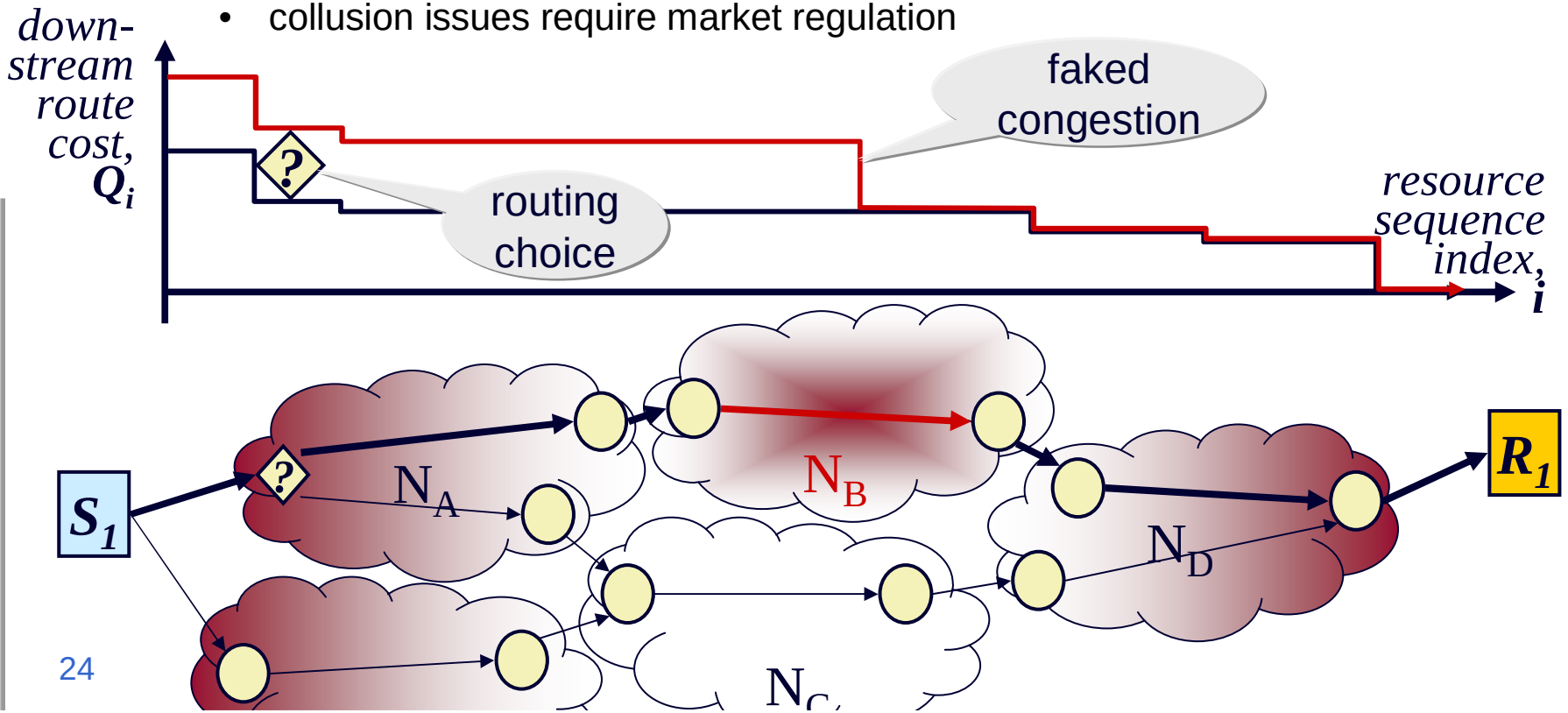
inter-domain accountability for congestion

- metric for inter-domain SLAs or charges
  - bulk volume of ECT(0)less bulk volume of CE
  - measure of downstream congestion allowed by upstream nets
  - volume charging tries to do this, but badly
  - aggregates and deaggregates precisely to responsible networks
  - upstream networks that do nothing about policing, DoS, zombies break SLA or get charged more



congestion competition – inter-domain routing

- if congestion → profit for a network, why not fake it?
  - upstream networks will route round more highly congested paths
  - N_A can see relative costs of paths to R_1 thru N_B & N_C
- the issue of monopoly paths
  - incentivise new provision
  - collusion issues require market regulation



BT IPR related to draft-briscoe-tsvwg-re-ecn-tcp-00.txt

- See IPR declaration at https://datatracker.ietf.org/public/ipr_detail_show.cgi?&ipr_id=651 which overrides this slide if there is any conflict

1)	WO 2005/096566	30 Mar 2004	published
2)	WO 2005/096567	30 Mar 2004	published
3)	PCT/GB 2005/001737	07 May 2004	
4)	GB 0501945.0 (EP 05355137.1)	31 Jan 2005	
5)	GB 0502483.1 (EP 05255164.5)	07 Feb 2005	

- BT hereby grants a royalty-free licence under any patent claims contained in the patent(s) or patent application(s) disclosed above that would necessarily be infringed by implementation of the technology required by the relevant IETF specification ("Necessary Patent Claims") for the purpose of implementing such specification or for making, using, selling, distributing or otherwise lawfully dealing in products or services that include an implementation of such specification provided that any party wishing to be licensed under BT's patent claims grants a licence on reciprocal terms under its own Necessary Patent Claims.