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

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

IETF-65 tsvwg Mar 2006

problem statement (§1)

- previous draft-00 focused on how to do policing
  - problem solved is actually how to allow *some* networks to do policing
- conservative networks
  - might want to throttle if unresponsive to congestion (VoIP, video, DDoS)
- middle ground
  - might want to cap congestion caused per user (e.g. 24x7 heavy sources)
- liberal networks
  - open access, no restrictions
  - evolution of hi-speed/different congestion control,... new worms

- many believe Internet is broken
  - not IETF role to pre-judge which is right answer to these socio-economic issues
  - Internet needs all these answers – balance to be determined by natural selection
  - ‘do-nothing’ doesn’t maintain liberal status quo, we just get more walls
- re-ECN goals
  - just enough support for conservative policies without breaking ‘net neutrality’
  - manage evolution of new congestion control, even for liberal → conservative flows
  - nets that allow their users to cause congestion in other nets, can be held accountable

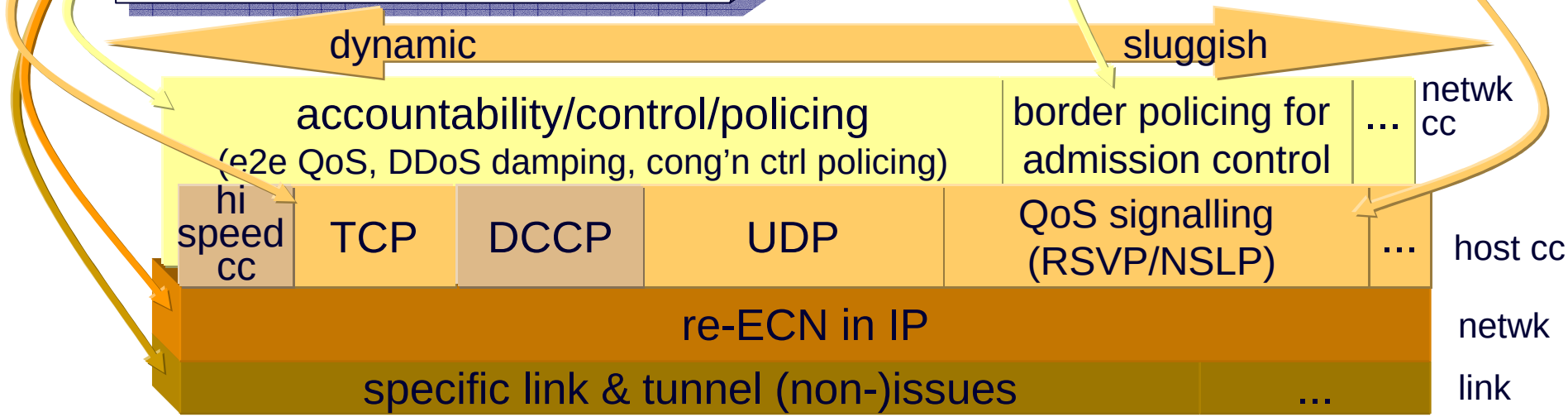
doc roadmap

Re-ECN: Adding Accountability for Causing Congestion to TCP/IP
draft-briscoe-tsvwg-re-ecn-tcp-01
intent

§3: overview in TCP/IP
§4: in TCP & others — *stds*
§5: in IP
§6: accountability apps — *inform'l*

Emulating Border Flow Policing using Re-ECN on Bulk Data
draft-briscoe-tsvwg-re-ecn-border-cheat-00
intent: informational

RSVP Extensions
for Admission Control over Diffserv
using Pre-congestion Notification
draft-lefaucheur-rsvp-ecn-00
intent
adds congestion f/b to RSVP — *stds*

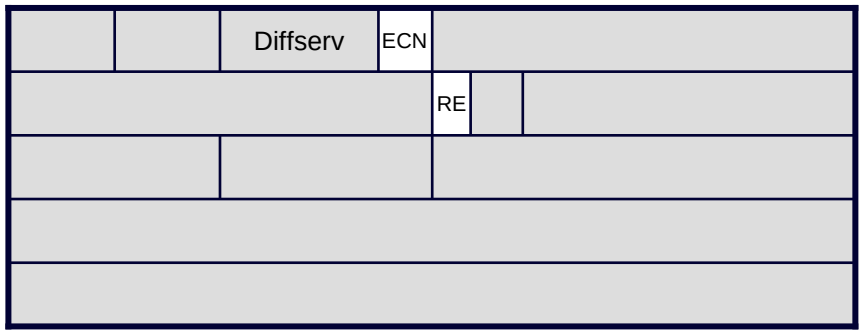


completely updated draft-01

- Re-ECN: Adding Accountability for Causing Congestion to TCP/IP
- IETF-64 Vancouver Nov 05
  - **initial draft, intent then:**
    - hold ECN nonce (RFC3540) at experimental
    - get you excited enough to read it, and break it
  - thanks to reviewers (on and off-list); you broke it (co-author noticed flaw too)
- now
  - **updated draft:** draft-briscoe-tsvwg-re-ecn-tcp-01.txt
  - **ultimate intent:** standards track
  - **immediate intent:** re-ECN worth using last reserved bit in IP v4?

changed re-ECN wire protocol in IPv4 (§3)

- propose Re-ECN Extension (RE) flag
  - for IPv4: propose to use bit 48 (was reserved)
  - set by sender, unchanged e2e



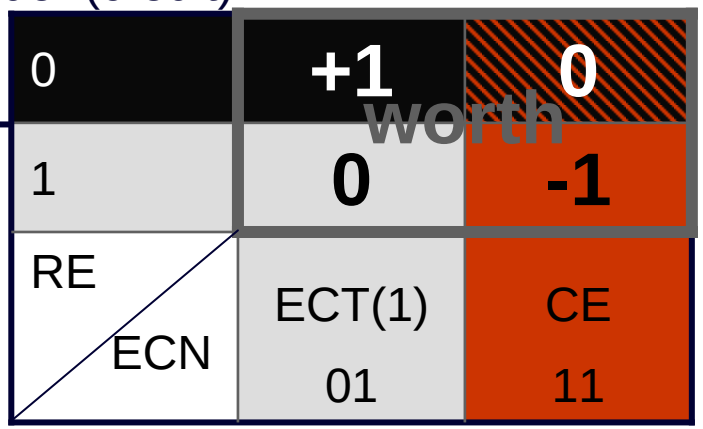
- once flow established
- sender re-inserts ECN feedback into forward data ("re-ECN") as follows
  - re-ECN sender always sets ECT(1)
  - on every **congestion event** from transport (e.g. TCP)

sender blanks **RE**
else sets RE

- conceptually, 'worth' of packet depends on 3 bit `codepoint'
- aim for zero balance of worth in flow

[RFC3168] ECN marking router (debit)

sender (credit)



flow bootstrap

- feedback not established (**FNE**) codepoint; RE=1, ECN=00
  - sent when don't know which way to set RE flag, due to lack of feedback
  - 'worth' **+1**, so builds up credit when sent at flow start
- after idle >1sec
  next packet MUST be **FNE**
  - enables deterministic flow state mgmt (policers, droppers, firewalls, servers)
- **FNE** packets are ECN-capable
  - routers MAY ECN mark, rather than drop
  - strong condition on deployment (see draft)

- **FNE** also serves as state setup bit [Clark, Handley & Greenhalgh]
  - protocol-independent identification of flow state set-up
  - for servers, firewalls, tag switching, etc
  - don't create state if not set
  - may drop packet if not set but matching state not found
  - firewalls can permit protocol evolution without knowing semantics
  - some validation of encrypted traffic, independent of transport
  - can limit outgoing rate of state setup
- considering I-D [Handley & Greenhalgh]
  - state-setup codepoint independent of, but compatible with, re-ECN
- **FNE** is 'soft-state set-up codepoint' (idempotent), to be precise

extended ECN codepoints: summary

- extra semantics backward compatible with previous ECN codepoint semantics

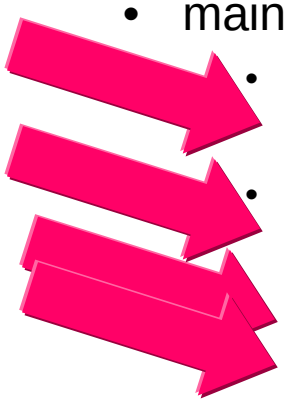
ECN code-point	ECN [RFC3168] codepoint	RE flag	Extended ECN codepoint	re-ECN meaning	`worth'
00	not-ECT	0	Not-RECT	Not re-ECN capable transport	
		1	FNE	Feedback not established	+1
01	ECT(1)	0	Re-Echo	Re-echo congestion event	+1
		1	RECT	Re-ECN capable transport	0
10	ECT(0)	0	---	'Legacy' ECN use	
		1	--CU--	Currently unused	
11	CE	0	CE(0)	Congestion experienced with Re-Echo	0
		1	CE(-1)	Congestion experienced	-1

other changes in draft (27pp → 65pp)

- easter egg added **:)**
- re-ECN in TCP fully spec'd (§4), including ECN-capable SYN
- network layer (§5)
  - OPTIONAL router forwarding changes added
    - preferential drop: improves robustness against DDoS
    - ECN marking not drop of **FNE**
  - control and management section added
- accountability/policing applications described (§6)
  - incentive framework fully described
    - example ingress policers & egress dropper described
    - pseudo-code TBA
  - DDoS mitigation explained
  - why it enables simpler ways to do e2e QoS, traffic engineering, inter-domain SLAs (still ref'd out)
- incremental deployment added (§7) → next slide
- architectural rationale added (§8)
- security considerations added (§10) → next slide but one

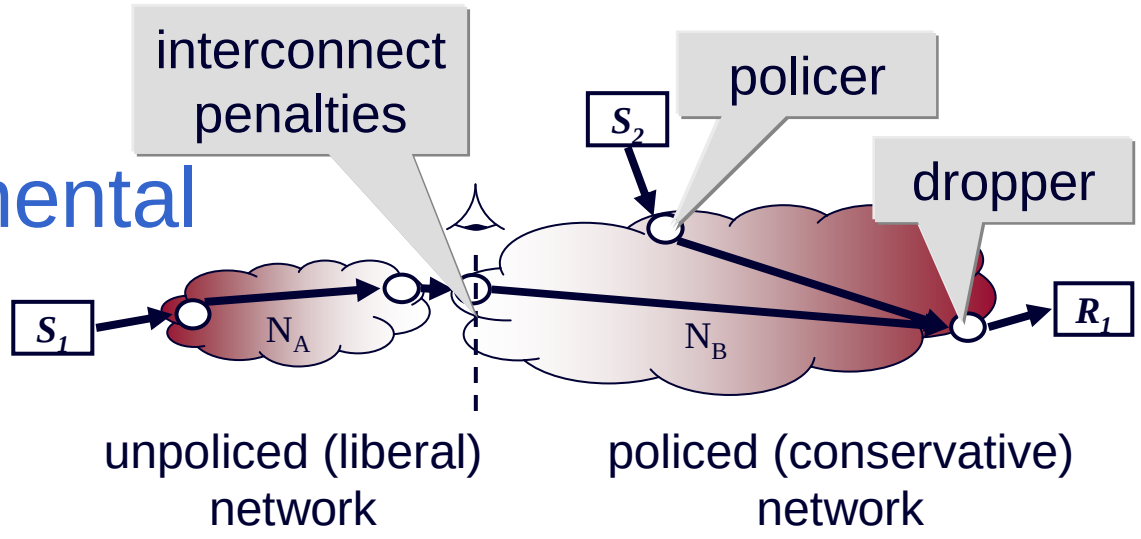
added incremental deployment (§7: 5½pp)

- brings together reasoning for wire protocol choices
- added deployment scenarios & incentives
  - everyone who needs to act, must have strong incentive to act
  - and incentives must arise in the order of required deployment
- main new messages
  - **first step** to break ECN deployment deadlock
    - edge-edge PCN for end-to-end controlled load (CL) QoS
  - **next step:** greed and fear motivators
    - help TCP (naively friendly) against greedy (streaming) apps
    - probably vertically integrated (conservative) operators first
    - 3GPP devices leak deployment to other networks by roaming
  - unilateral deployment per network ...

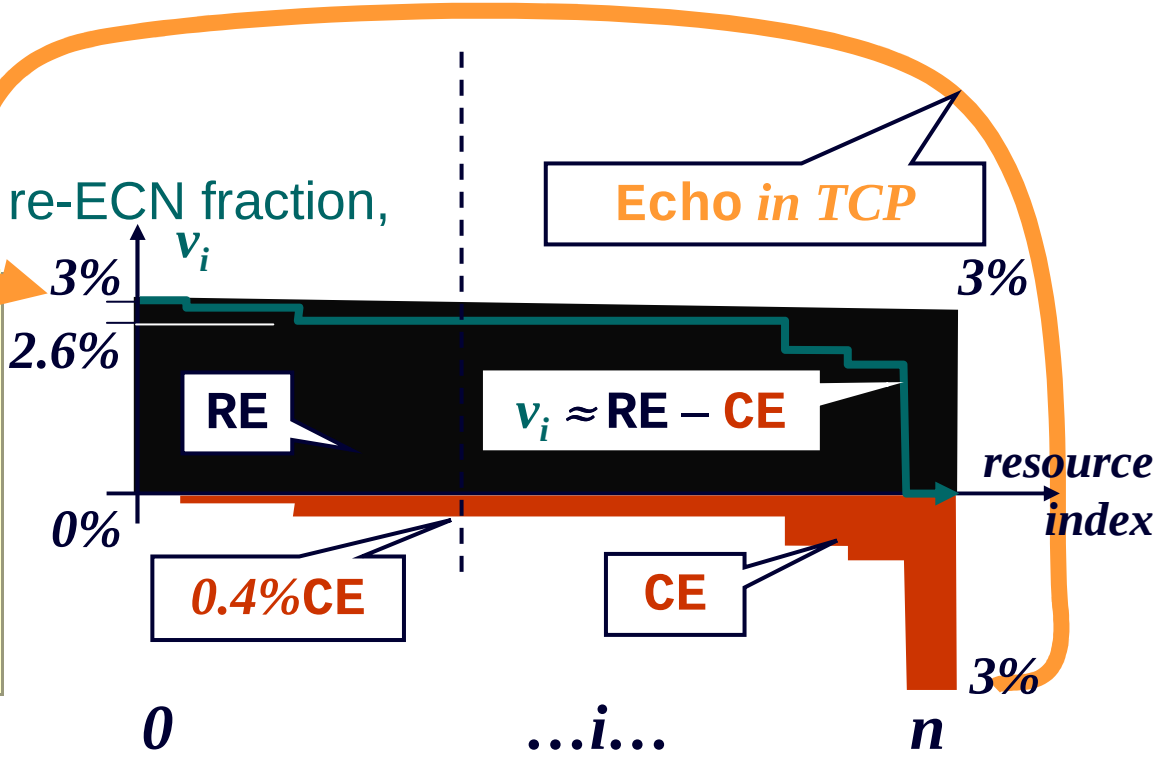


re-ECN incremental deployment

0	+1 worth 0	0
1	0	-1
RE / ECN	ECT(1) 01	CE 11



- on every **congestion event** from TCP, sender blanks **RE**, else sets RE
- at any point on path, diff betw fractions of **RE** & **CE** is downstream congestion
- routers unchanged



added re-ECN security considerations (§10)

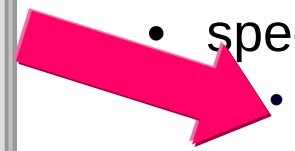
- egress dropper
  - robust against attack that plays-off against ingress policing
  - robust against state exhaustion attacks (by design of **FNE**)
  - write-up of state aggregation implementation TBA
  - believe new protocol allows dropper to be robust against dynamic attacks
- working on preventing collateral damage where malicious source spoofs negative traffic like someone else's flow
- see also
  - limitations text added (§6.3) – presented in Vancouver
  - tsvwg posting "traffic ticketing considered ineffective or harmful" (26 Jan '06)
- security of re-ECN deliberately designed not to rely on crypto
- provoking you to break re-ECN

summary

- enables ‘net neutral’ policing of causes of congestion
  - liberal networks can choose not to police, but still accountable
- simple architectural fix
  - generic accountability hook per datagram
  - requires one bit in IP header
- ECN nonce of limited scope in comparison
- fixed vulnerabilities so far by making it simpler
  - working on robustness to new attacks
- detailed incremental deployment story

plans in IETF

- split draft into two and fill some ‘TBAs’:
  - protocol spec
  - accountability/policing applications
- implementation/simulation next
- re-TTL draft planned (Appendix E gives exec summary)
- independent flow state setup draft (possibly)



- spec detail more than sufficient for intensive review
  - ~20 controversial points highlighted
  - strongly encourage review on the tsvwg list

- changing IPv4 header isn’t a task we’ve taken on lightly

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

draft-briscoe-tsvwg-re-ecn-tcp-01.txt

Q&A

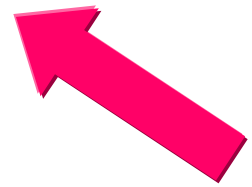
Emulating Border Flow Policing using Re-ECN on Bulk Data

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

IETF-65 tsvwg Mar 2006

simple solution to a hard problem?

- Emulating Border Flow Policing using Re-ECN on Bulk Data
  - **initial draft:** draft-briscoe-tsvwg-re-ecn-border-cheat-00
  - **ultimate intent:** informational
  - **exec summary:** claim we can now scale flow reservations to any size internetwork *and* prevent cheating



doc roadmap

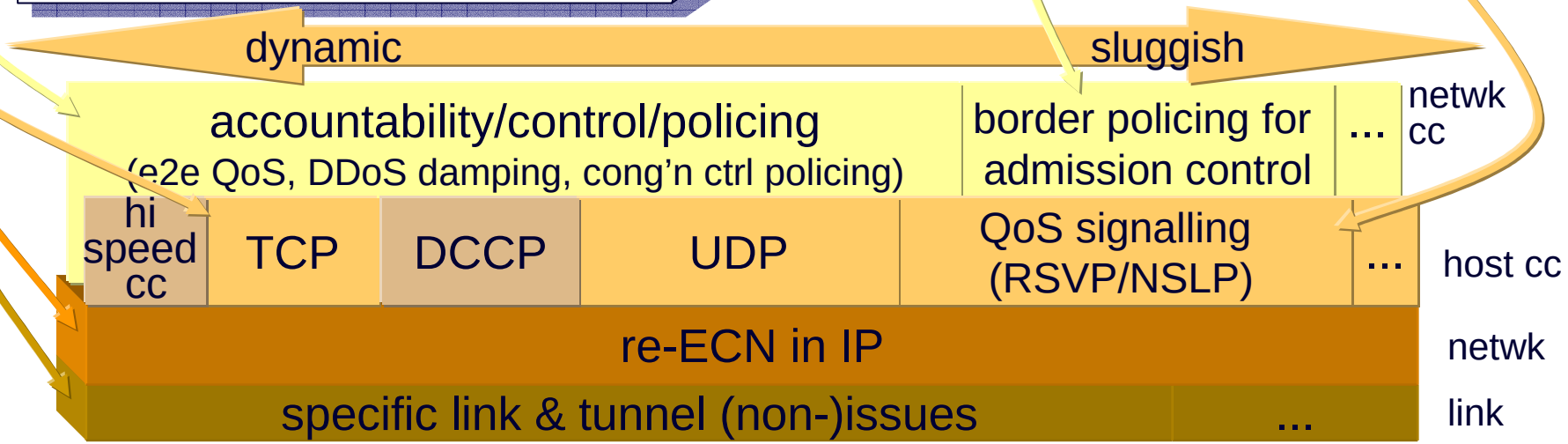
Re-ECN: Adding Accountability for Causing Congestion to TCP/IP
draft-briscoe-tsvwg-re-ecn-tcp-01

intent

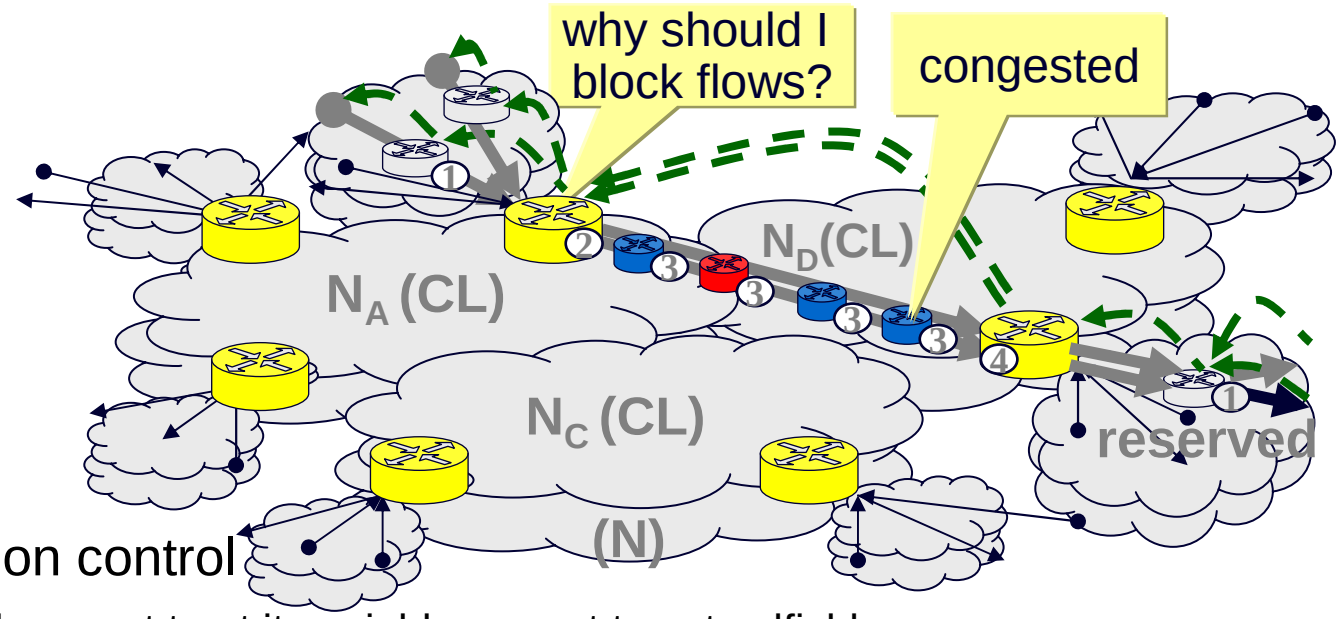
§3: overview in TCP/IP
§4: in TCP & others — *stds*
§5: in IP
§6: accountability apps — *inform'l*

Emulating Border Flow Policing using Re-ECN on Bulk Data
draft-briscoe-tsvwg-re-ecn-border-cheat-00
intent: informational

RSVP Extensions
for Admission Control over Diffserv
using Pre-congestion Notification
draft-lefaucheur-rsvp-ecn-00
intent
adds congestion f/b to RSVP — *stds*



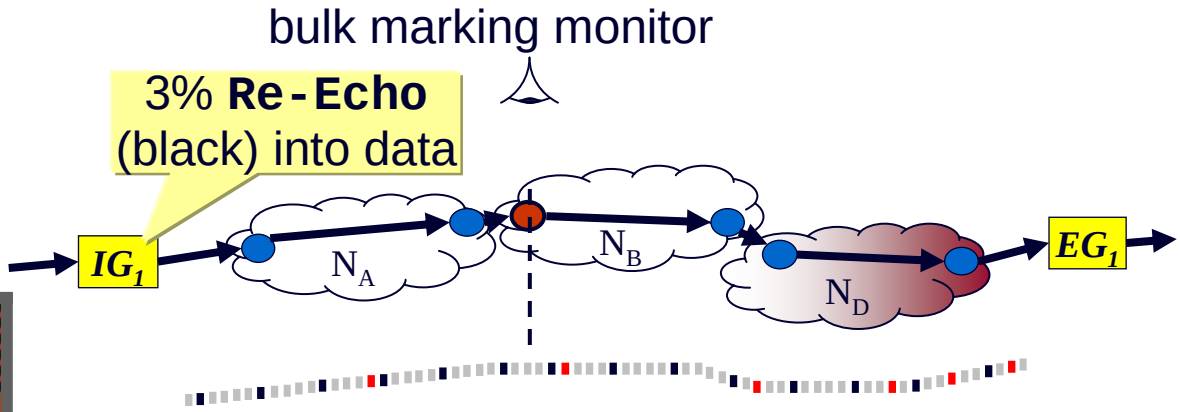
problem statement

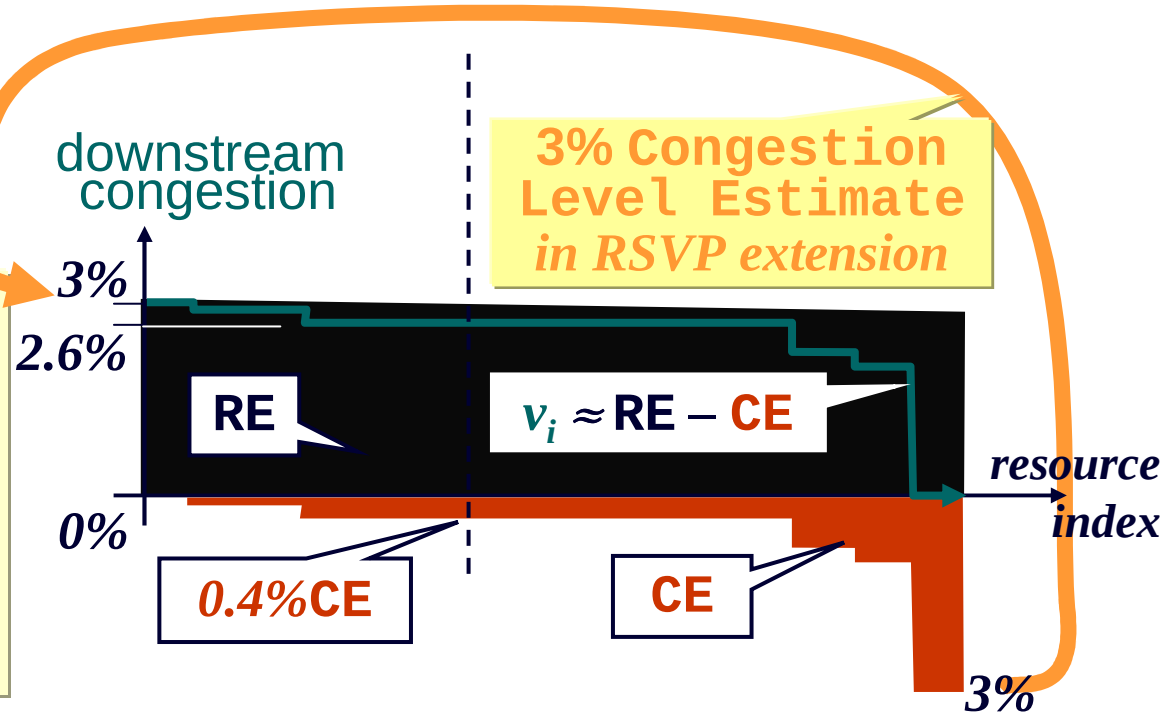


- flow admission control
  - a network cannot trust its neighbours not to act selfishly
  - if it asks them to deny admission to a flow
    - it has to check the neighbour actually has blocked the data
  - if it accepts a reservation
    - it has to check for itself that the data fits within the reservation
- traditional solution
  - flow rate policing at borders
- can pre-congestion-based admission control span the Internet?
  - without per-flow processing at borders?

solution: use re-ECN

0	+1 worth 0	0
1	0	-1
RE / ECN	ECT(1) 01	CE 11

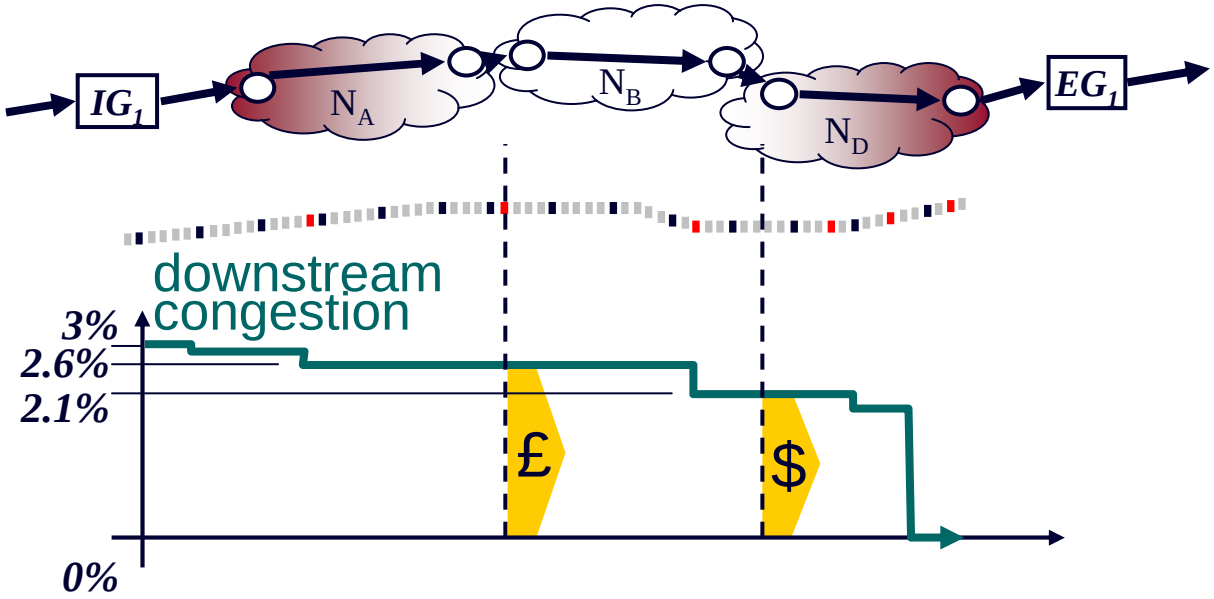




- ingress gateway blanks **RE**, in same proportion as fraction of **CE** arriving at egress
- at any point on path, bulk diff betw fractions of **RE** & **CE** is downstream congestion
- routers unchanged

inter-domain accountability for congestion

- metric for inter-domain SLAs or usage charges
  - N_B applies penalty to N_A in proportion to bulk volume of **RE** less bulk volume of **CE** over, say, a month
  - could be tiered penalties, directly proportionate usage charge, etc.
  - flows and f'back de-aggregate precisely to responsible networks
- see draft for fail-safes against misconfigs etc.

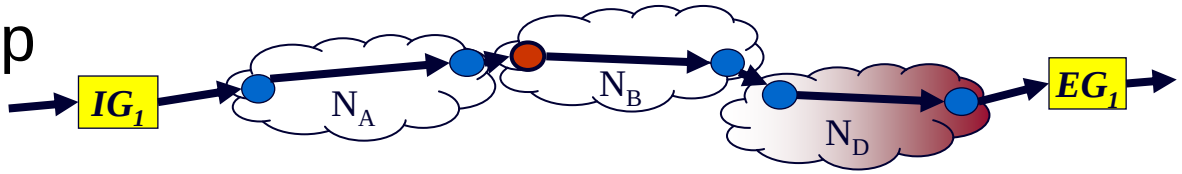


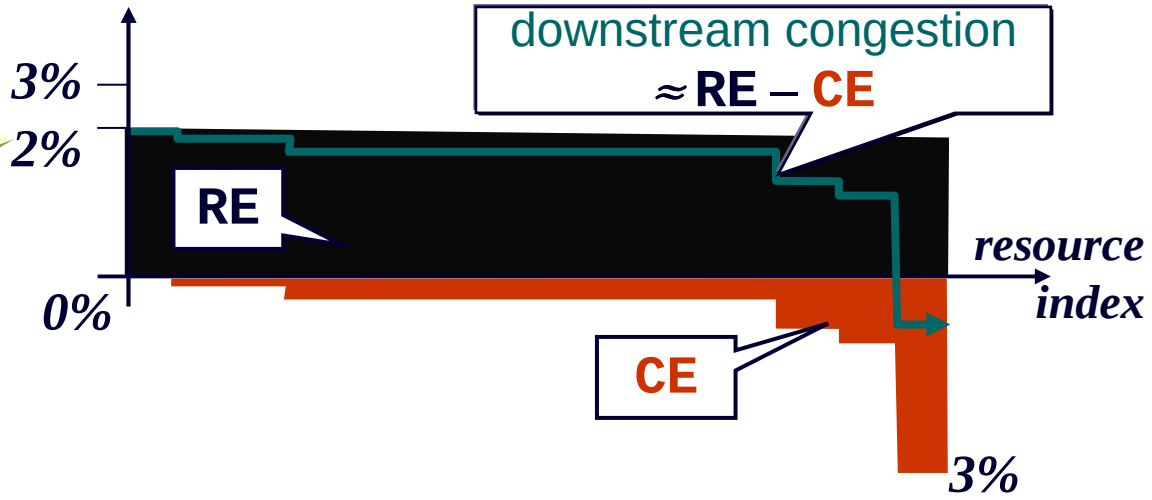
note well: not standardising contracts

- want to avoid protocols that depend on particular business models
  - only standardise the protocol
  - then networks can choose to use the metric in various ways
- the contractual arrangement was an example to prove a solution exists
  - networks can choose other, broadly similar arrangements
  - or choose not to use it, and to do per-flow processing instead
- only concerns interconnection within Diffserv region

why should ingress re-echo honestly?

- if N_D detects persistent imbalance between **RE** and **CE**, triggers sanctions
- probably not drop
  - raise mgmt alarm
  - sanction out of band

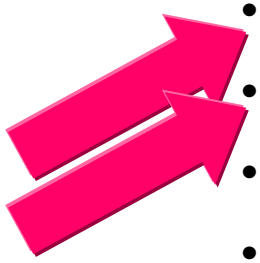




2% **Re-Echo** (black) into data (understatement)

summary

- claim we can now scale flow reservations to any size internetwork *and* prevent cheating
  - without per-flow processing in Internet-wide Diffserv region
  - just bulk passive counting of packet marking over, say, a month
- see draft for
  - why this is a sufficient emulation of per-flow policing
  - results of security analysis, considering collusions etc.
  - protocol details (aggregate & flow bootstrap, etc)
  - border metering algorithms, etc
- comments solicited, now or on list

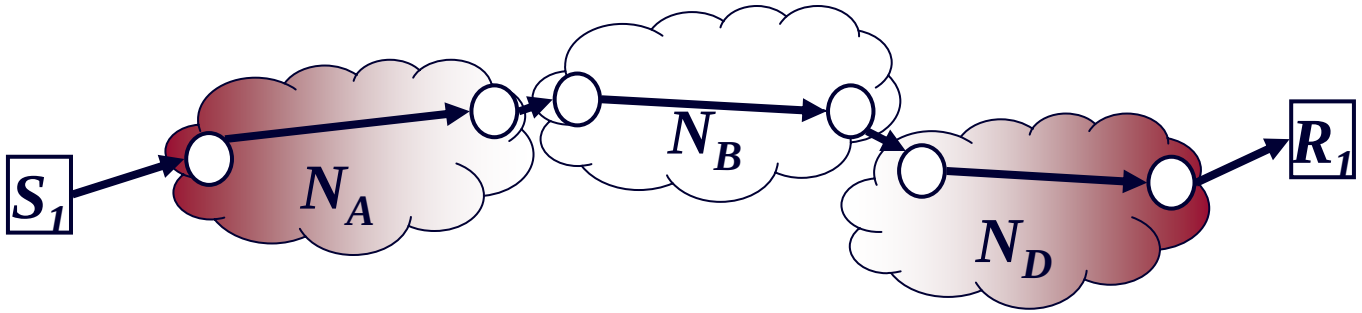


Emulating Border Flow Policing using Re-ECN on Bulk Data

draft-briscoe-tsvwg-re-ecn-border-cheating-00.txt

Q&A

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

you MUST do this
you may not do this

- logically consistent statements
- build-time compliance
  - usual standards compliance language (§2)
- run-time compliance
  - incentives, penalties (§6 throttling, dropping, charging)
- hook in datagram service for incentive mechanisms
    - they can make run-time compliance advantageous to all

previous re-ECN protocol (IP layer)

ECN 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)

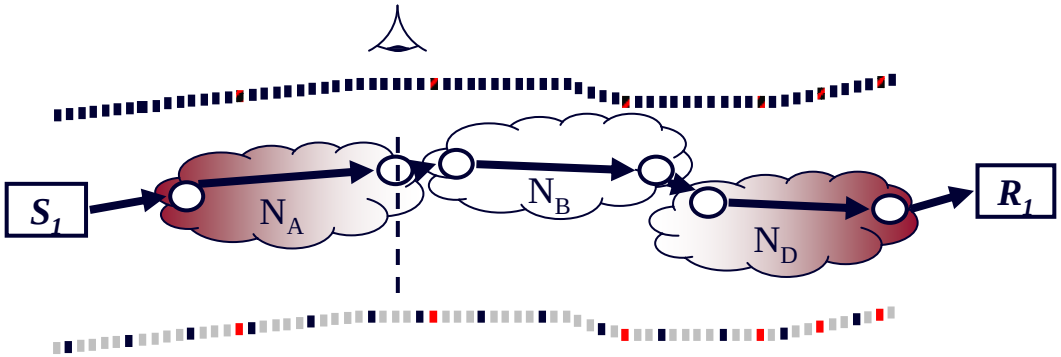
- Feedback-Established (FE) flag

IPv4 control flags		
FE	DF	MF

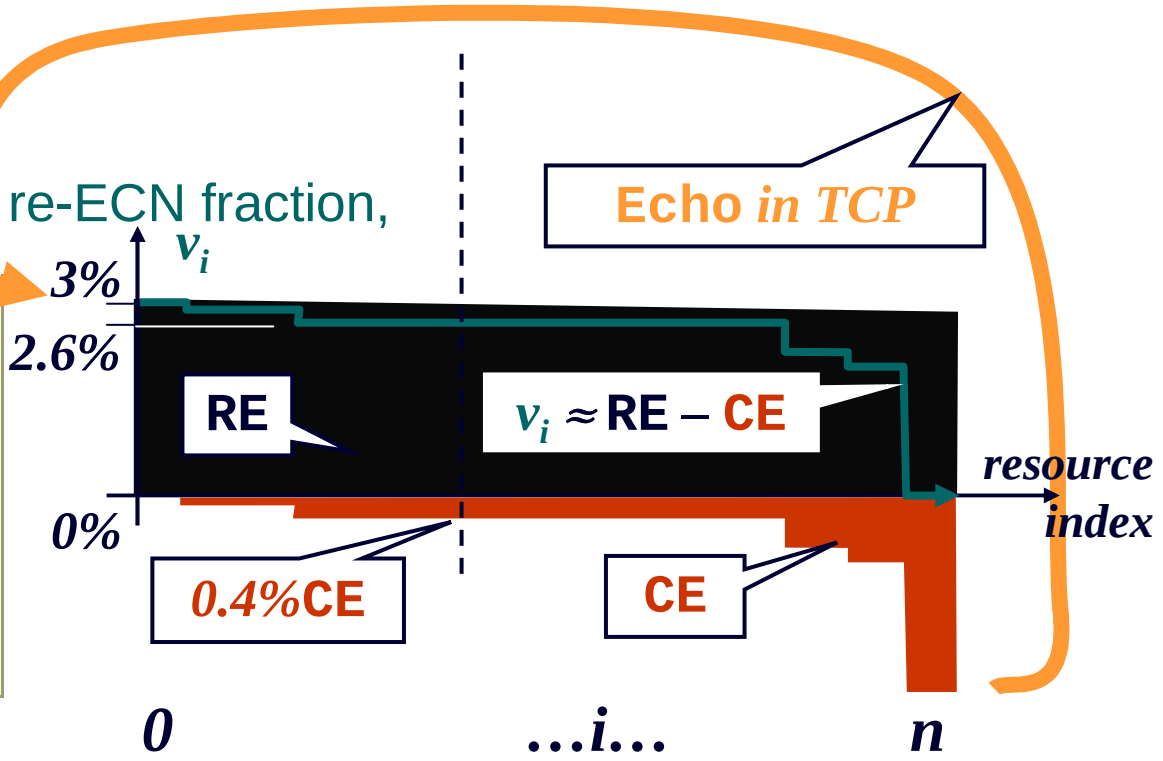
re-ECN
(sketch)

0	+1	0
1	0	-1
RE / ECN	ECT(1) 01	CE 11

worth



- on every **congestion event** from TCP, sender blanks **RE**, else sets RE
- at any point on path, diff betw fractions of **RE** & **CE** is downstream congestion
- routers unchanged



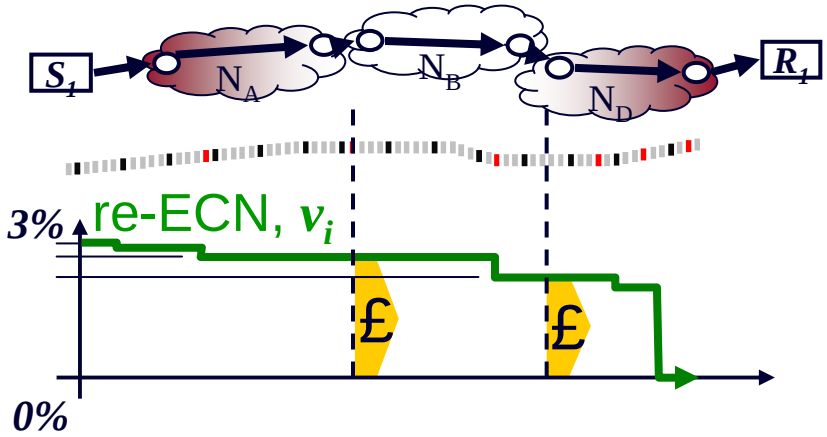
re-ECN in TCP (§4) updated

- flow start now fully spec'd (incl. example session)
- goal: all packets can be ECN capable
  - can now allow ECN capable SYN (and SYN ACK)
    - with a strong deployment condition (see draft)
  - pure ACKs, re-transmissions, window probes: still **`Not-ECT`**
- re-ECN hosts don't need ECN nonce [RFC3540] support

accountability for congestion

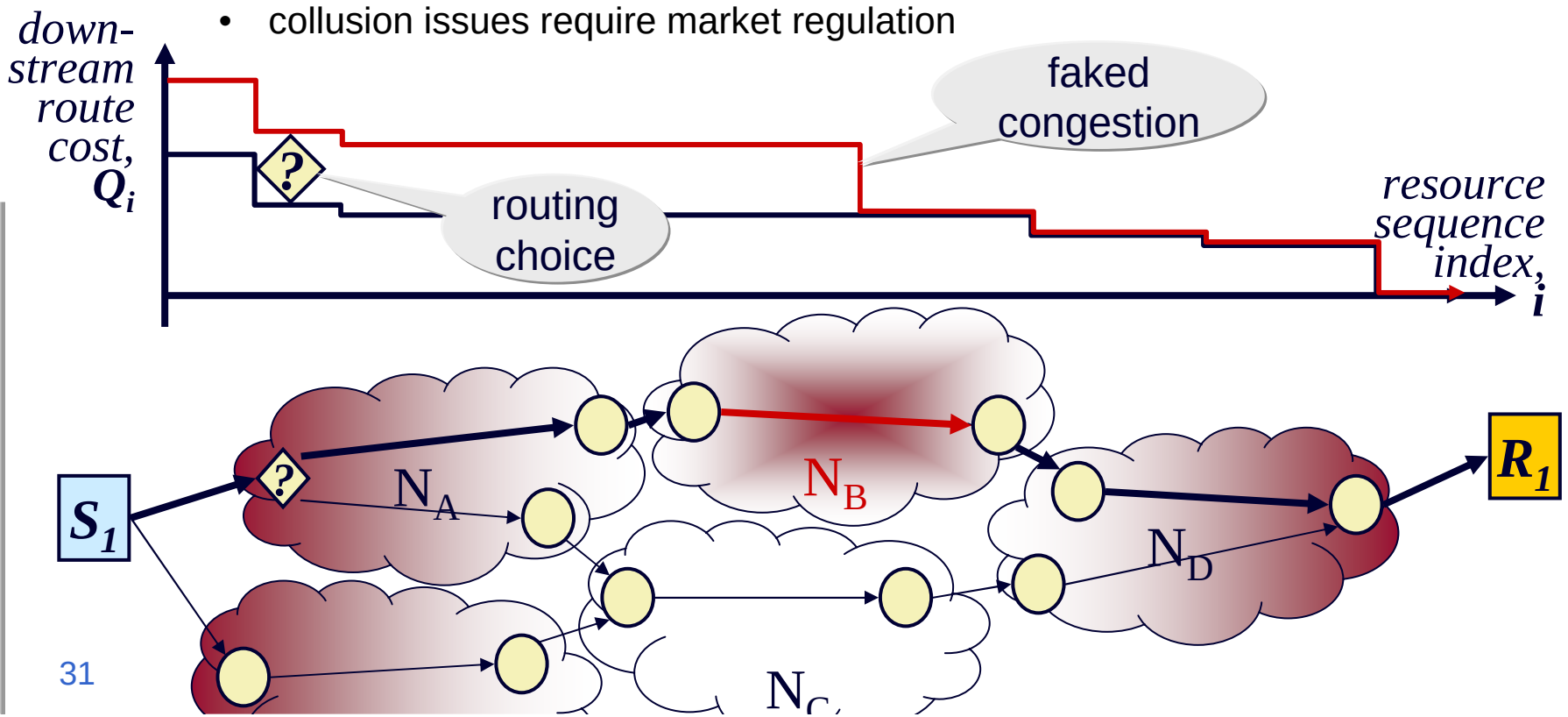
other applications

- 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 congestion
- 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 get charged more

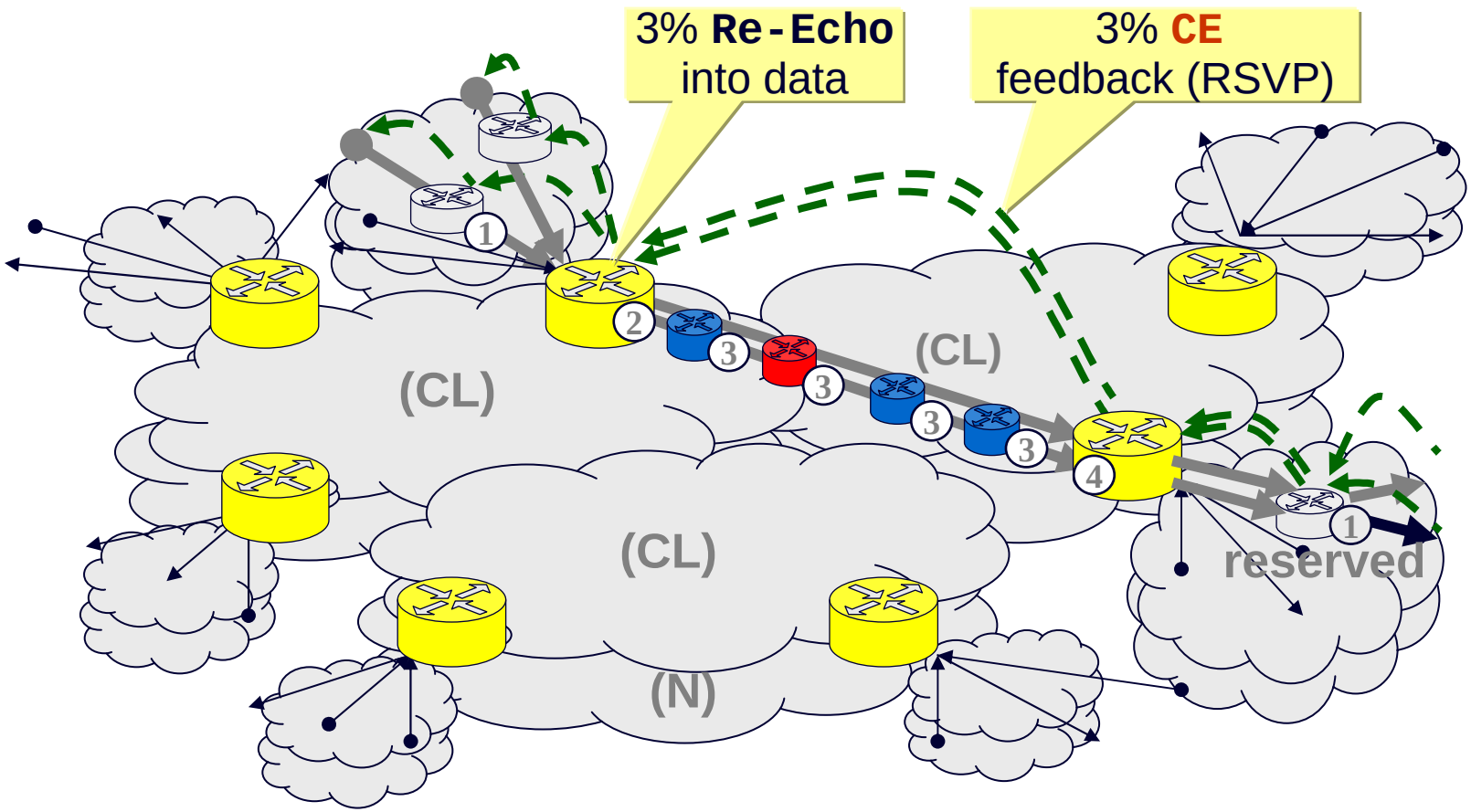


congestion competition – inter-domain routing

- 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



border anti-cheating solution



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.