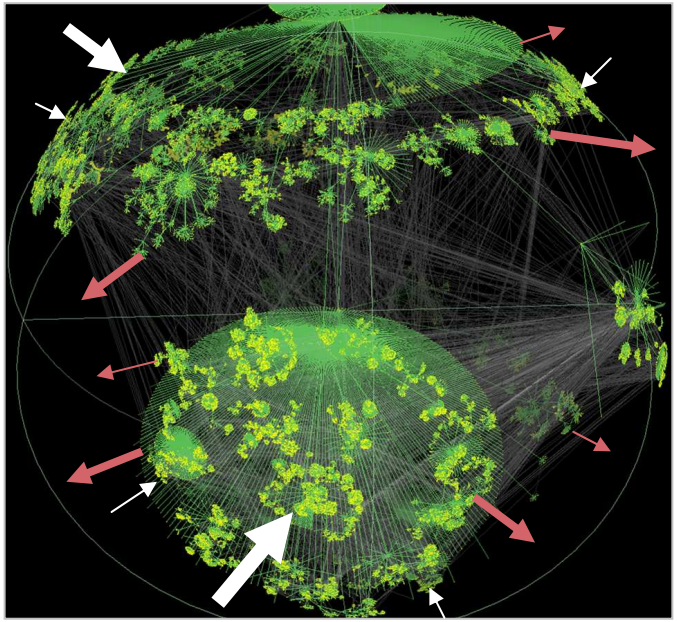
sorting out
Internet resource sharing

Bob Briscoe
Chief Researcher
BT Group

Apr 2008

known problem
since early days

- how to share all the parts of a huge, multi-provider packet multiplexer between competing processes
- keeping one-way datagrams
- allowing for
  - self-interest & malice
    - of users and of providers
  - evolvability
    - of new rate dynamics from apps
    - of new business models
  - viability of supply chain
  - simplicity



Internet topology visualization produced by Walrus (Courtesy of Young Hyun, CAIDA)

- if we do nothing
  - the few are ruining it for the many
  - massive capacity needed to keep interactive apps viable
  - poor incentives to invest in capacity
  - operators are kludging it with DPI
  - solely today's apps frozen into net
  - complex, ugly feature interactions

BT

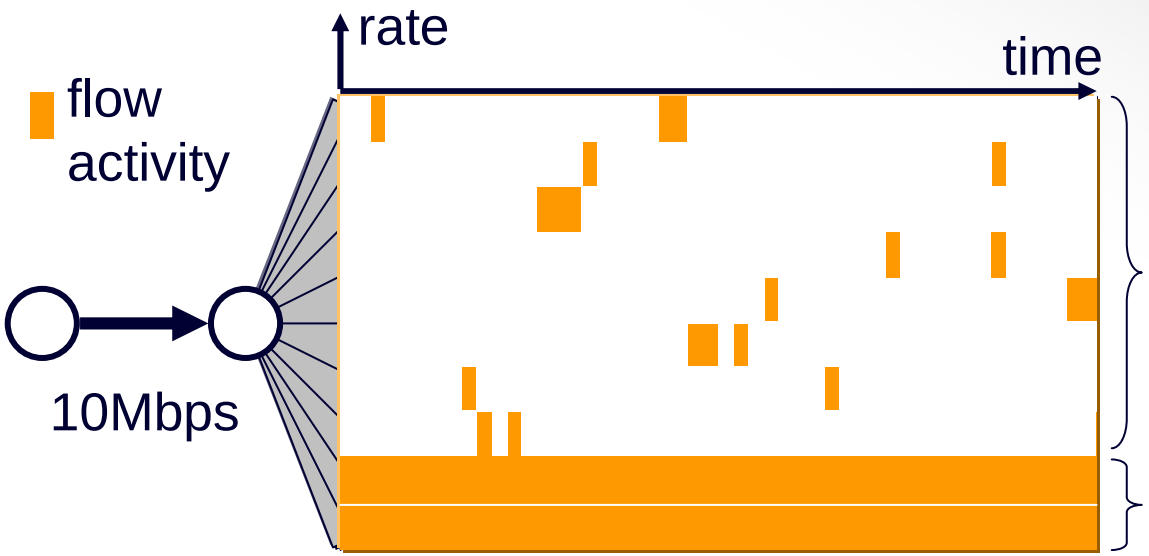
Internet resource sharing

Problems

Bob Briscoe

dismantling the religion of flow rate equality

extra degree of freedom #1: activity factor



2Mbps access each

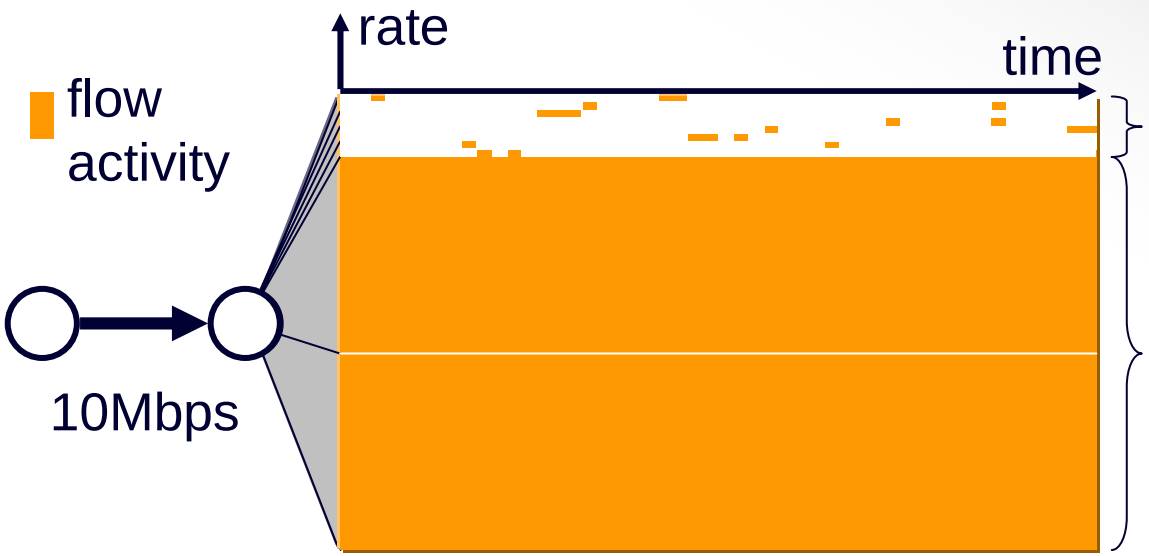
80 users of attended apps

20 users of unattended apps

usage type	no. of users	activity factor	ave.simul flows /user	TCP bit rate /user	vol/day (16hr) /user	traffic intensity /user
attended	80	5%	=	417kbps	150MB	21kbps
unattended	20	100%	=	417kbps	3000MB	417kbps
				x1	x20	x20

dismantling the religion of flow rate equality

degrees of freedom #1&2: activity factor & multiple flows



80 users of attended apps

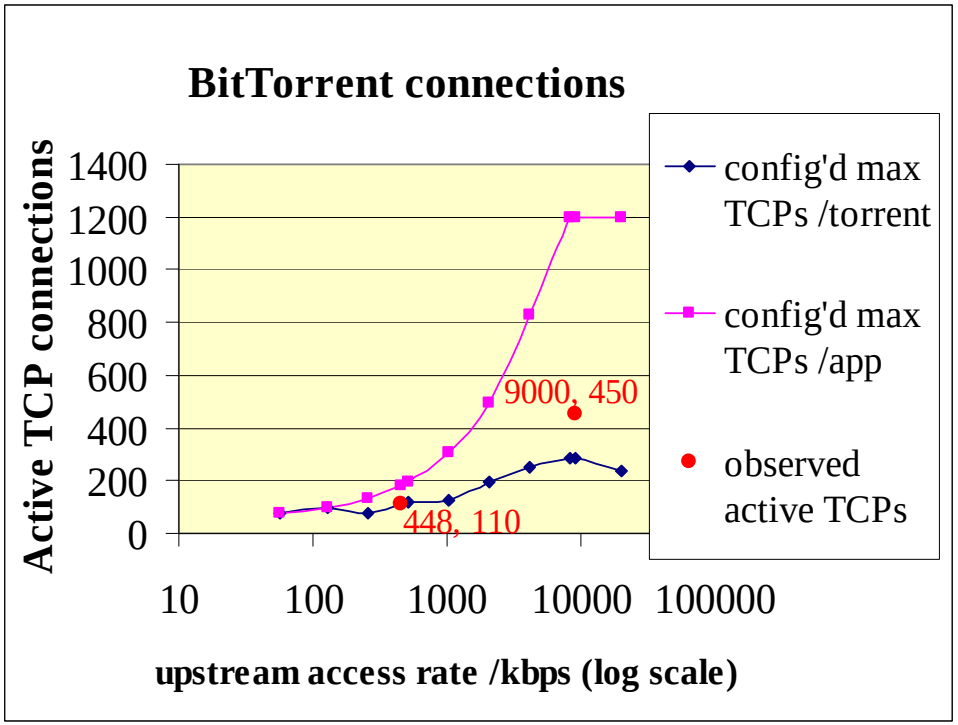
20 users of unattended apps

2Mbps access each

usage type	no. of users	activity factor	ave.simul flows /user	TCP bit rate /user	vol/day (16hr) /user	traffic intensity /user
attended	80	5%	2	20kbps	7.1MB	1kbps
unattended	20	100%	50	500kbps	3.6GB	500kbps
				x25	x500	x500

realistic numbers?

there are elephants in the room

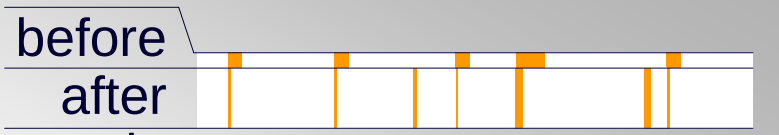


- number of TCP connections
  - Web1.1: **2**
  - BitTorrent: **~100** observed active
    - varies widely depending on
      - no. of torrents per user
      - maturity of swarm
      - config'd parameters

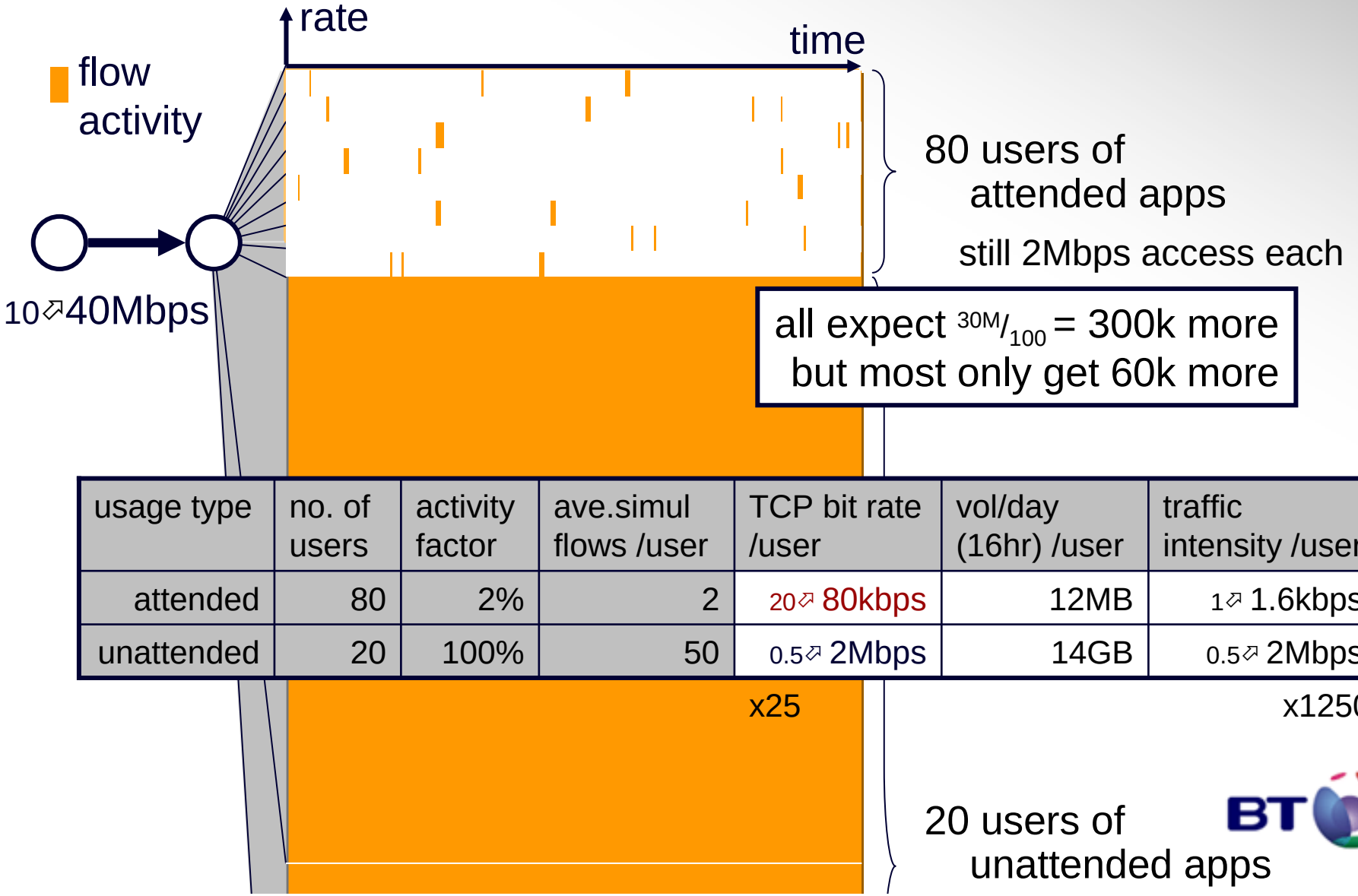
details suppressed:

- utilisation never 100%
  - but near enough during peak period
- on DSL, upstream constrains most p2p apps
  - other access (fixed & wireless) more symmetric

most users hardly benefit from bottleneck upgrade



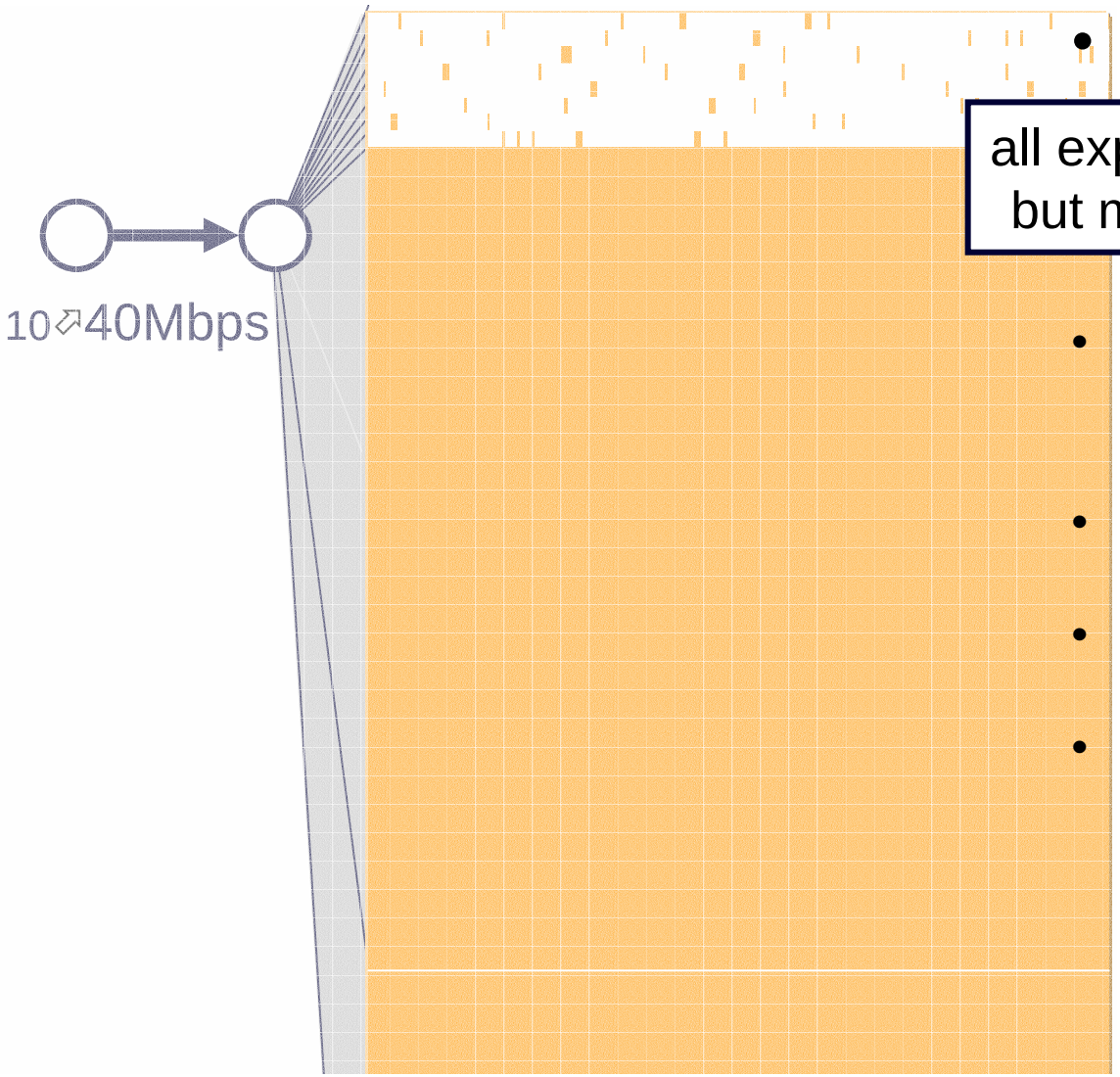
data limited flows
want rate more than volume



usage type	no. of users	activity factor	ave.simul flows /user	TCP bit rate /user	vol/day (16hr) /user	traffic intensity /user
attended	80	2%	2	20⇗ 80kbps	12MB	1⇗ 1.6kbps
unattended	20	100%	50	0.5⇗ 2Mbps	14GB	0.5⇗ 2Mbps

consequence #1

higher investment risk



- recall

all expect $^{30M}/_{100}$ = 300k more
but most only get 60k more

- but ISP needs everyone to pay for 300k more
- if most users unhappy with ISP A's upgrade
- they will drift to ISP B who doesn't invest
- competitive ISPs will stop investing...

BT

consequence #2

trend towards bulk enforcement

- as access rates increase
  - attended apps leave access unused more of the time
  - anyone might as well fill the rest of their own access capacity
- fair queuing eliminates the multiple flow problem
  - but not the activity factor problem
  - and only for congestion where scalability not a problem (backhaul)
- operator choices:
  a) either continue to provision sufficiently excessive shared capacity
  b) or enforce tiered volume limits

consequence #3

networks making choices for users

- characterisation as two user communities over-simplistic
  - heavy *users* mix heavy and light *usage*
- two enforcement choices

  a) bulk: network throttles all a heavy user's traffic indiscriminately
  - encourages the user to self-throttle least valued traffic
  - but many users have neither the software nor the expertise

  b) selective: network *infers* what the user would do
  - using deep packet inspection (DPI) and/or addresses to identify apps
- even if DPI intentions honourable
  - confusable with attempts to discriminate against certain apps
  - user's priorities are task-specific, not app-specific
  - customers understandably get upset when ISP guesses wrongly

BT

consequence #4

future congestion control work-rounds

- trying to satisfy demanding application requirements
  - constrained by staying not ‘much’ faster than TCP
  - resulting ‘over-constrained’ protocols not app-developer’s first choice

hi-speed congestion control >> TCP rate

AND

hi-speed congestion control ≈ TCP rate

- fertile ground for proprietary solutions
  - no peer review of behaviour esp. under anomalous conditions
    - Joost
    - BitTorrent delivery network accelerator (DNA)
    - Mustang TCP
    - etc

BT

distributed denial of service

an extreme of the same problem

- multiple sources able to cause excessive congestion
- defend against what attackers could do
  - not just what they do now
- best time to attack: during a flash crowd
  - maximum impact – when people most want a service
  - least effort to tip network into overload
  - intent is not the only difference
  - lack of response to congestion is generally abnormal
- a congestion control enforcement problem
  - at least as first line of defence

summary so far
no coherent reasoning about resource sharing

- two resource sharing approaches fighting

	the Internet way	operators (& users)
degree of freedom	'flow rate equality'	'volume accounting'
multiple flows	✗	✓
activity factor	✗	✓
congestion variation	✓	✗
application control	✓	✗

- leaving field open for 'hacked' solutions
  - deep packet inspection
  - even 'DoS attacks' by operators on customers (sending TCP resets)
  - self-interested hi-speed transports

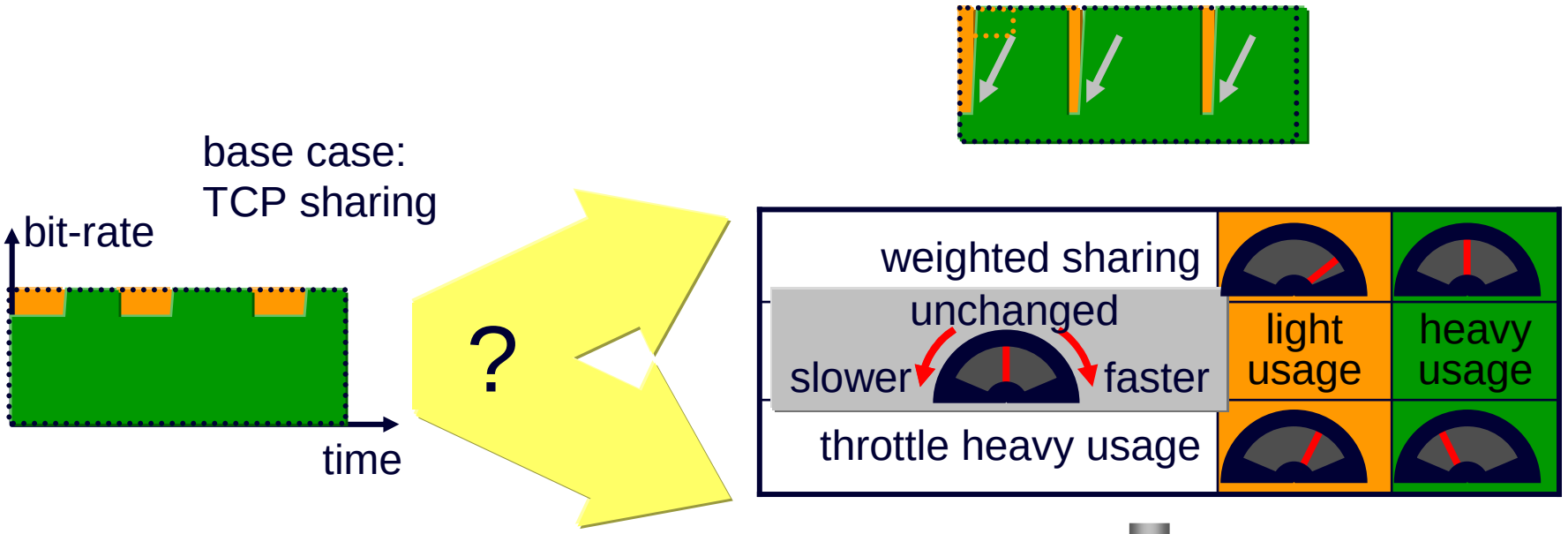
BT

Internet resource sharing

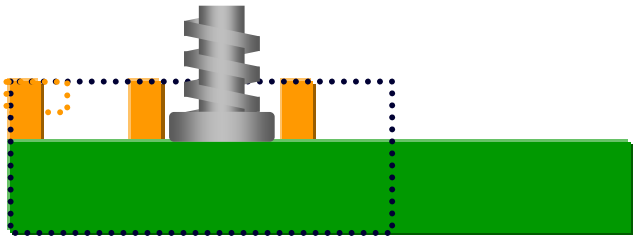
a solution for unicast: re-feedback of ECN (re-ECN)

Bob Briscoe

there must be better solutions than fighting



- light usage can go much faster
- hardly affecting completion times of heavy usage



fairness at run-time, not design time

- protocol designer / IETF doesn't decide fairness
  - whatever protocols *designed* to do, they are being *used* unfairly
- protocol designer / IETF can't decide fairness
  - design-time bodies can't control run-time degrees of freedom
- protocol designer / IETF shouldn't decide fairness
  - shouldn't prejudge fair-use policy agreed between user & ISP
    - whether TCP, max-min, proportional or cost fairness

BT

2-part solution

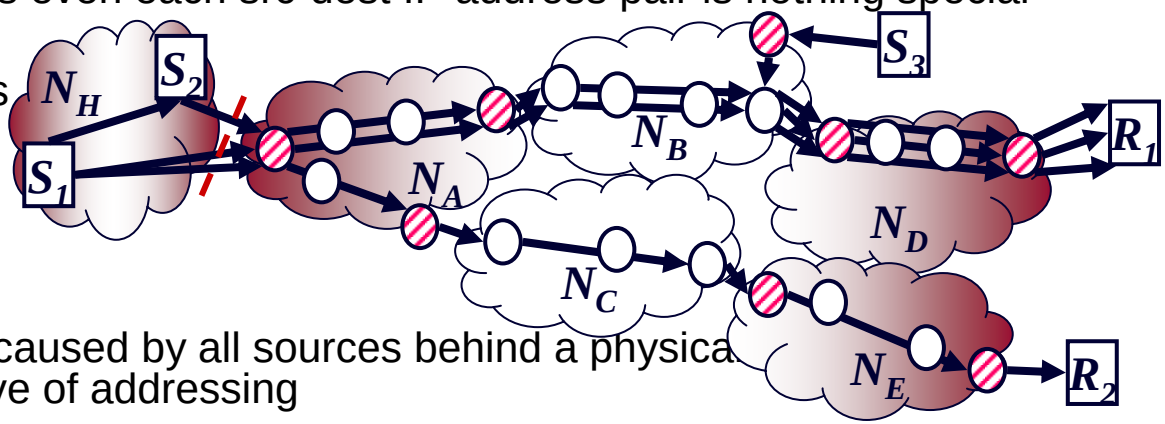
1. transport layer
   deliberately allow very unequal weighted sharing
   - e.g. socket API with weight parameter, w
   - meaning take w shares of a congested resource
2. network layer
   incentive to use the weights sparingly
   - heavy usage w << 1; light usage can set w >> 1
   - flipped round from what we see today (Web w=2, p2p w=50)

- this talk: how to create the network layer incentives
  - we've done #1 but it will be abused without #2

general idea

reveal congestion sender causes throughout rest of the Internet ...at its attachment point

- × bottleneck policers: active research area since 1999
  - detect flows causing unequal share of congestion
  - located at each potentially congested router
  - fooled by splitting flow IDs (or src address spoofing)
  - why equal rights for each flow, when flow IDs can be created arbitrarily?
  - swarmcasting shows even each src-dest IP address pair is nothing special
  - takes no account of how active source is over time



- ✓ re-ECN
  - reveals congestion caused by all sources behind a physical interface, irrespective of addressing
  - no advantage to split IDs
  - accumulates over time
  - like counting volume, but ‘congestion volume’

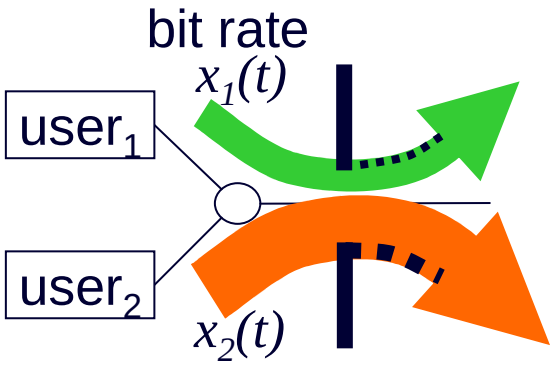
coherent reasoning about resource sharing

degree of freedom	‘flow rate equality’	‘volume accounting’	re-ECN
multiple flows	×	✓	✓
activity factor	×	✓	✓
congestion variation	✓	×	✓
application control	✓	×	✓

BT

core of solution

congestion harm (cost) metric



loss (marking) fraction $p(t)$

- bit rate *weighted* by each flow's congestion, over time

  congestion volume, $v \equiv \int p(t)x_i(t)\,dt$

  summed over all a sender's flows
- result is easy to measure per flow or per user
  - volume of bytes discarded (or ECN marked)
- a precise instantaneous measure of harm, counted over time
  - a measure for fairness over any timescale
  - and a precise measure of harm during dynamics

- intuition: volume is bit rate over time

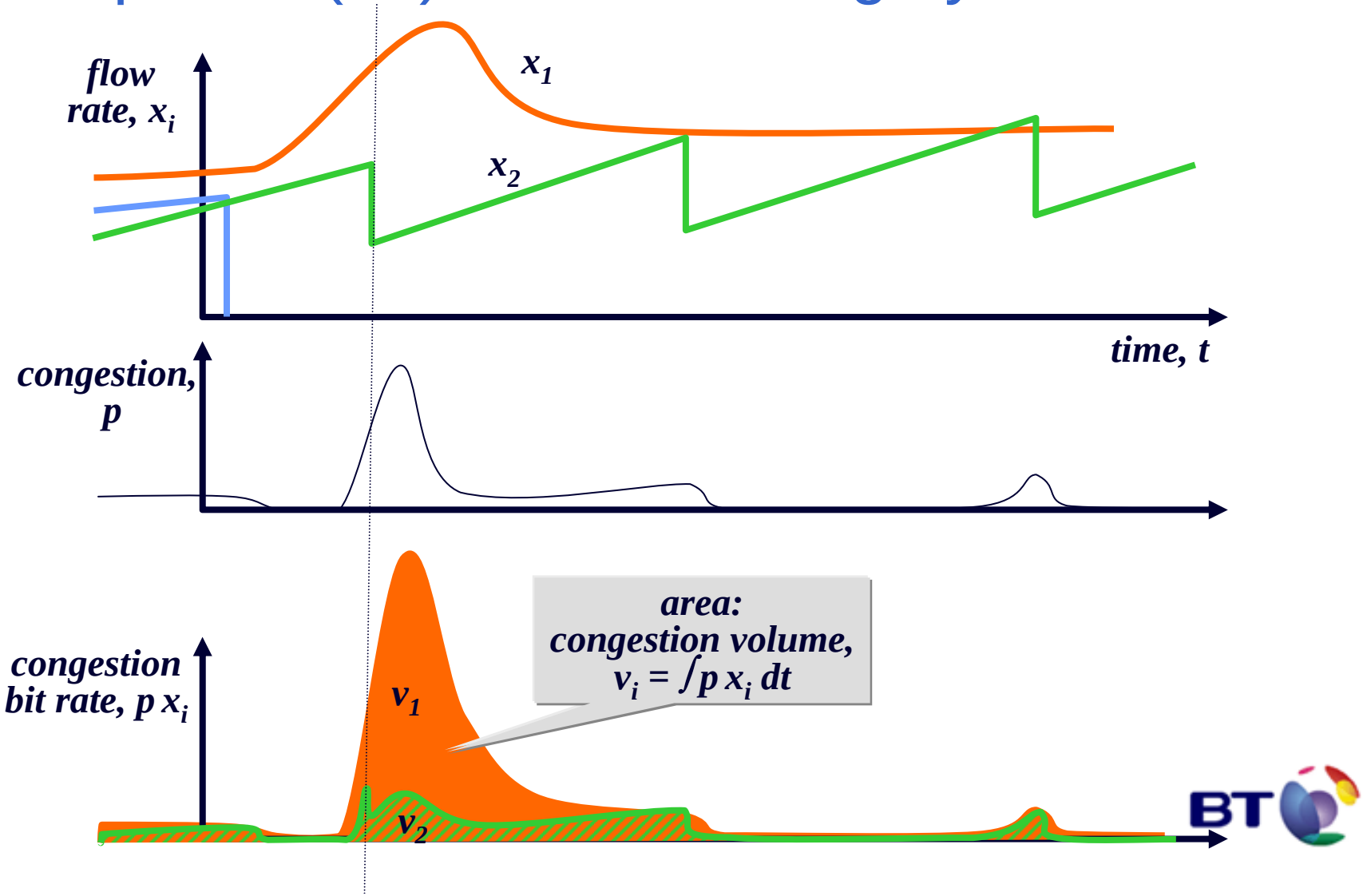
  volume, $V \equiv \int x_i(t)\,dt$

  summed over all a sender's flows
- network operators often count volume only over peak period
  - as if $p(t)=1$ during peak and $p(t)=0$ otherwise

BT

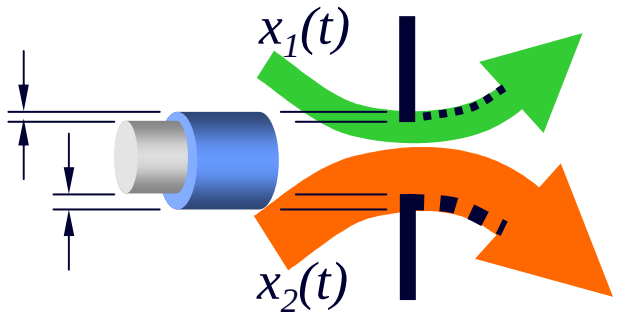
congestion volume

captures (un)fairness during dynamics



calibrating ‘cost to other users’

- a monetary value can be put on ‘what you unsuccessfully tried to get’
  - the marginal cost of upgrading network equipment
    - so it wouldn’t have marked the volume it did
    - so your behaviour wouldn’t have affected others
- competitive market matches...
  - the cost of congestion volume
  - with the cost of alleviating it



note: diagram is conceptual

congestion volume would be accumulated over time

capital cost of equipment would be depreciated over time

- congestion volume is not an extra cost
  - part of the flat charge we already pay
  - but we can’t measure who to blame for what
  - if we could, we *might* see pricing like this...

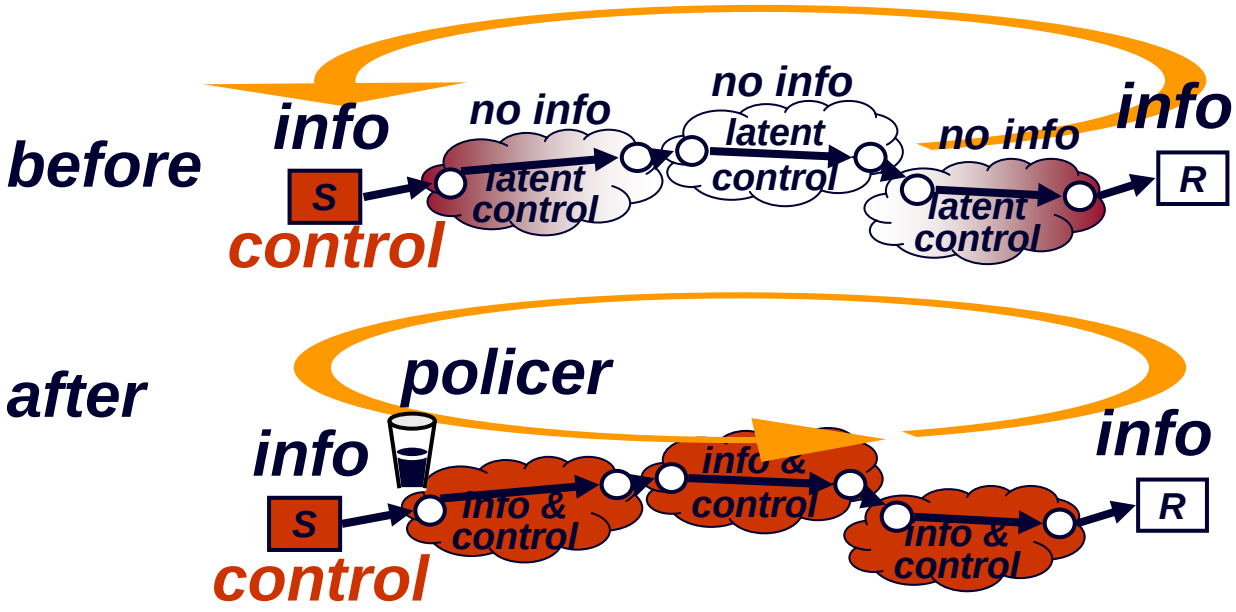
access link	congestion volume allow’ce	charge
100Mbps	50MB/month	€15/month
100Mbps	100MB/month	€20/month

- NOTE WELL
  - IETF provides the metric
  - industry does the business models

BT

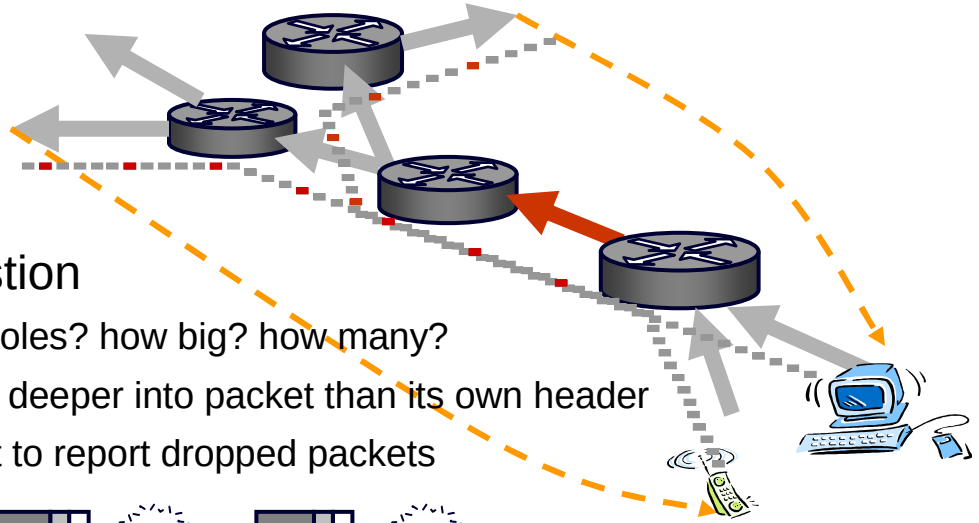
addition of re-feedback – in brief

- *before:* congested nodes mark packets
  receiver feeds back marks to sender
- *after:* sender must pre-load expected congestion
  by re-inserting feedback
- if sender understates expected compared to actual congestion, network discards packets
- result: packets will carry prediction of downstream congestion
- policer can then limit congestion caused (or base penalties on it)



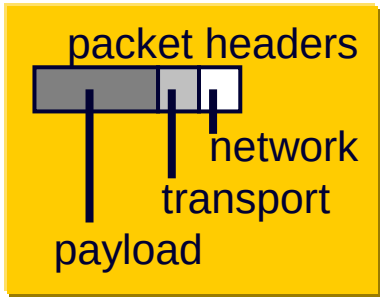
solution step #1: ECN

make congestion visible to network layer



- packet drop rate is a measure of congestion
  - but how does network at receiver measure holes? how big? how many?
  - can't presume network operator allowed any deeper into packet than its own header
  - not in other networks' (or endpoints') interest to report dropped packets

- solution: Explicit Congestion Notification (ECN)
  - mark packets as congestion ***approaches*** - to avoid drop
  - already standardised into IP (RFC3168 – 2001)
  - implemented by most router vendors – very lightweight mechanism
  - but rarely turned on by operators (yet) – mexican stand-off with OS vendors

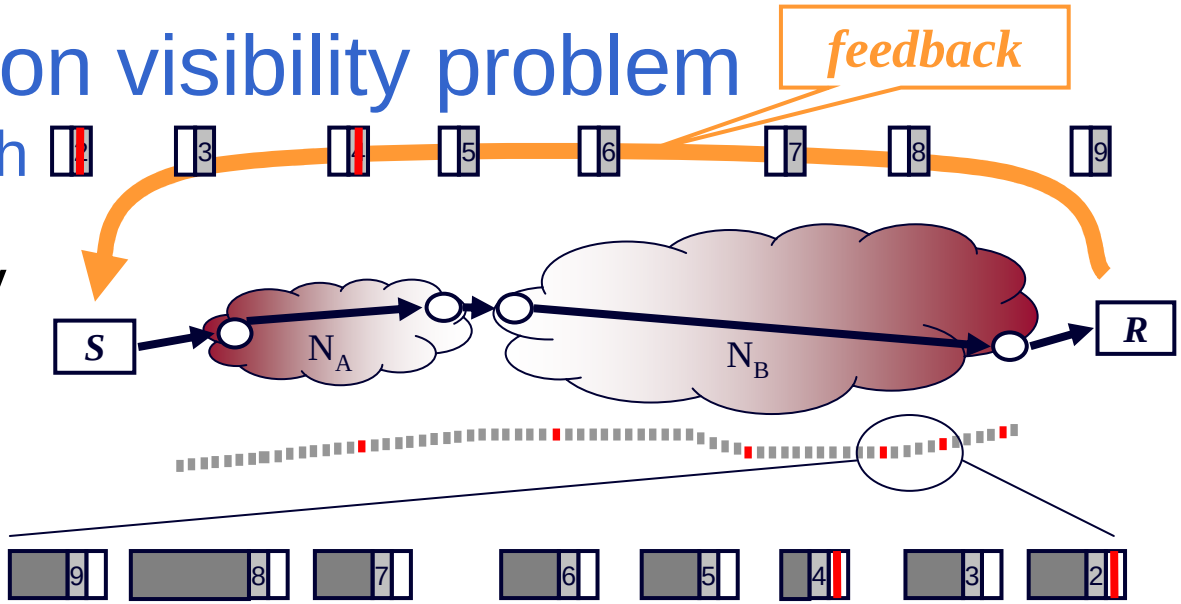


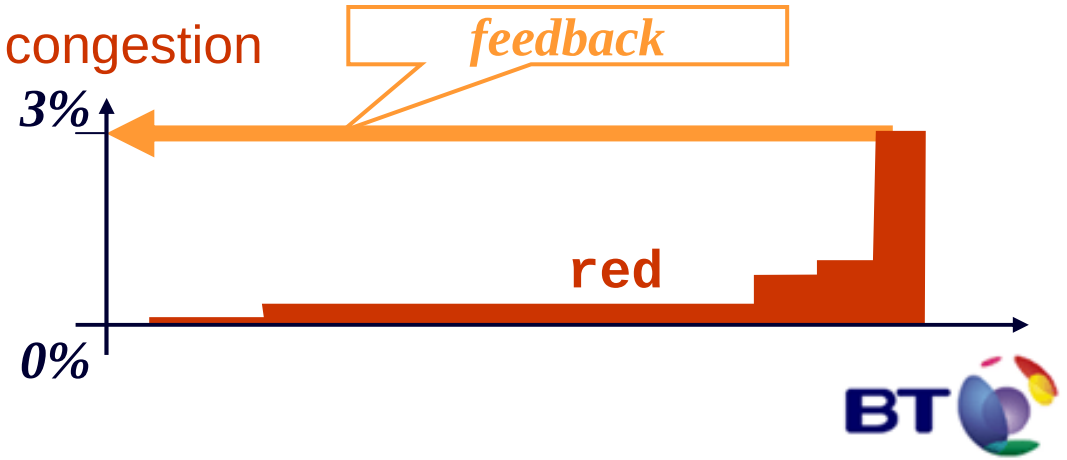
BT

new information visibility problem

ECN is not enough

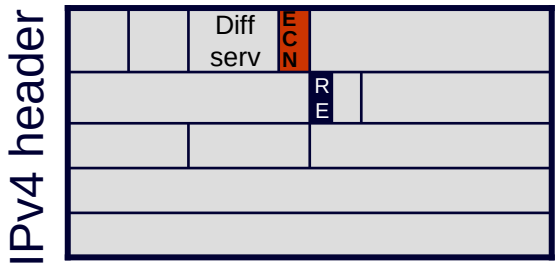
- path congestion only measurable at exit
- can't measure path congestion at entry
  - can't presume allowed deeper into feedback packets

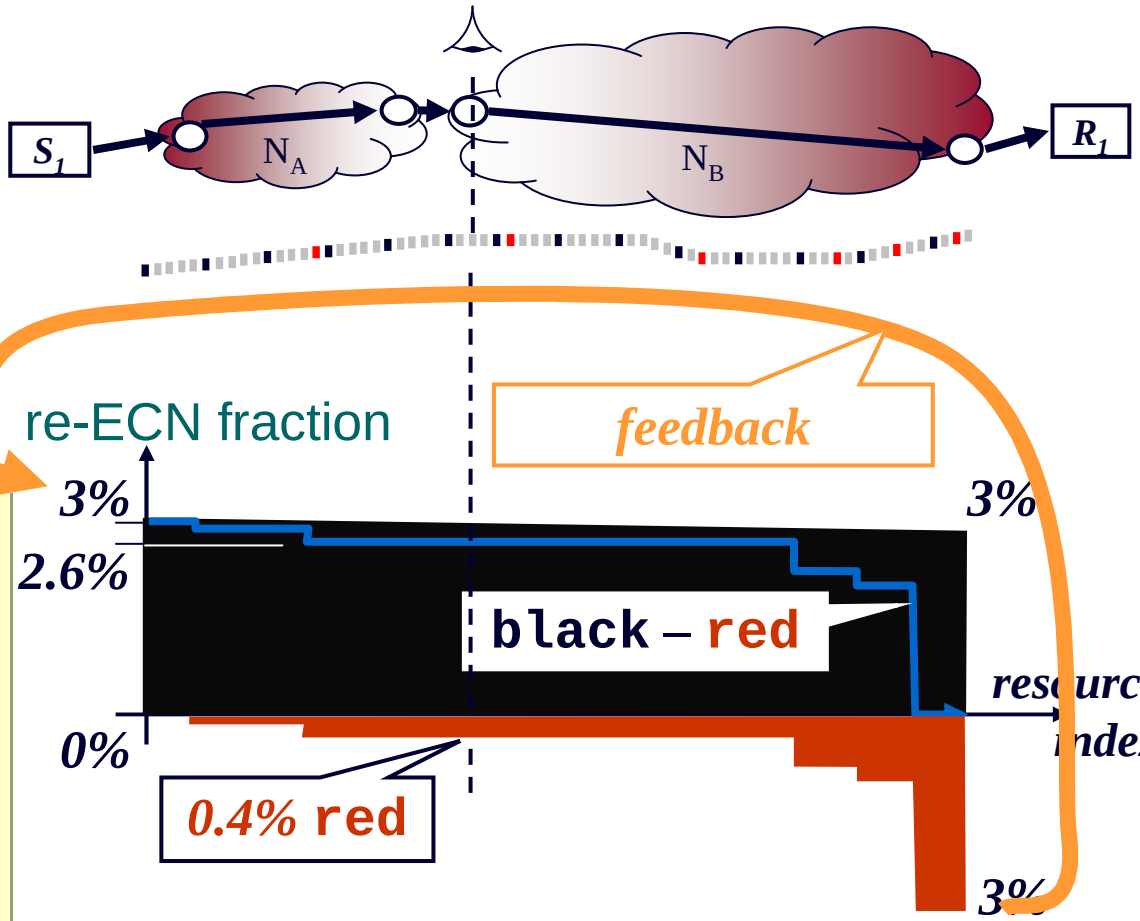




measurable downstream congestion

solution step #2

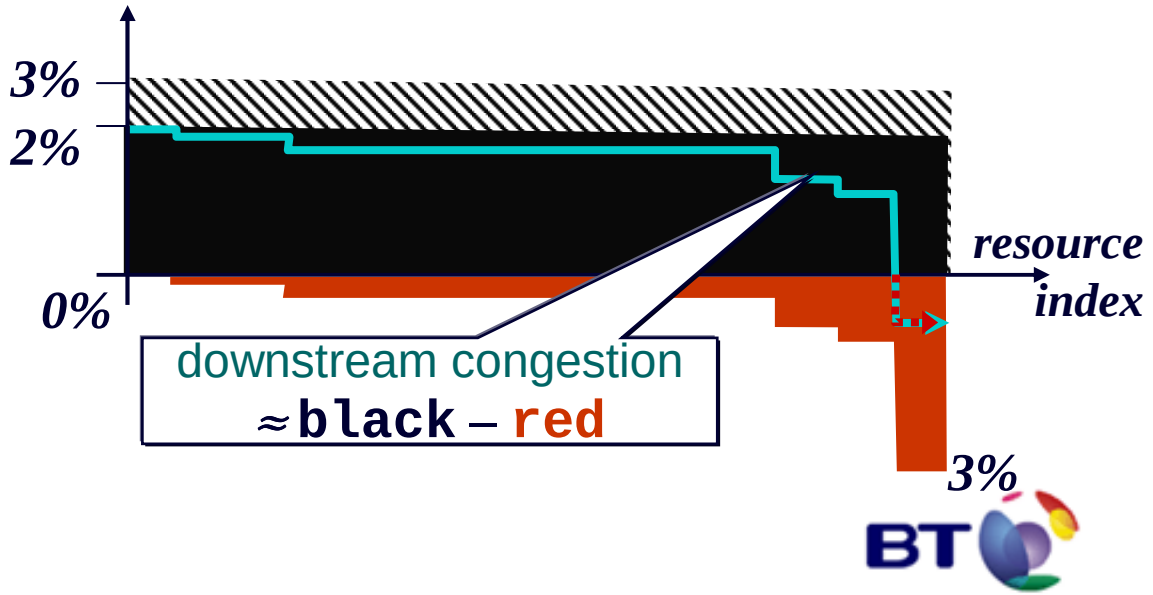




- sender re-inserts feedback by marking packets **black**
- at any point on path,diff betw fractions of **black** & **red** bytes is downstream congestion
- ECN routers unchanged
- **black** marking e2e but visible at net layer for accountability

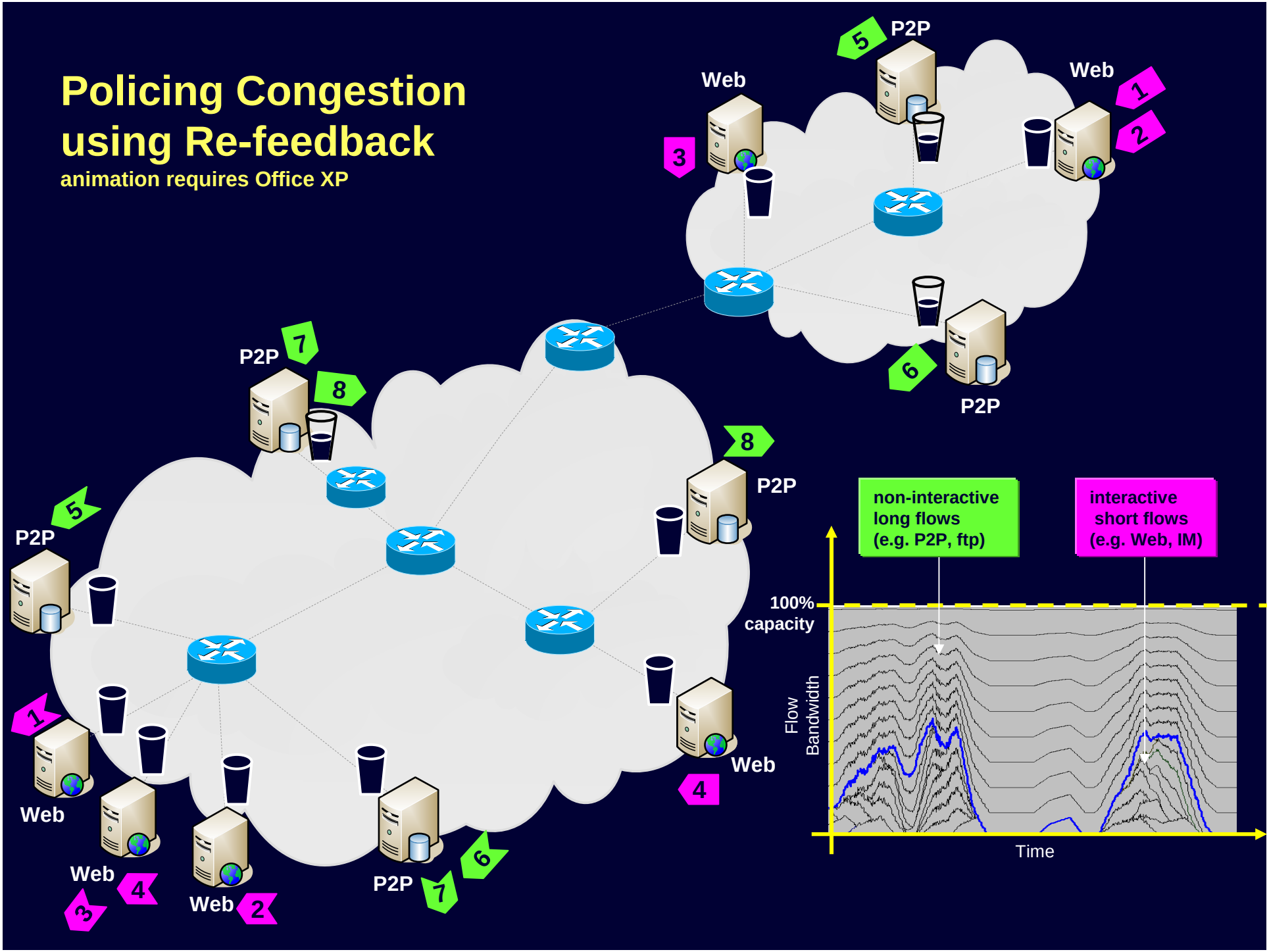
proposed re-ECN service model

- to encourage sender (or proxy) to indicate sufficient expected congestion...
- Internet won't try to deliver packet flows beyond the point where more congestion has been experienced than expected
  - if sender wants to communicate, has to reveal expected congestion
  - even if sender not trying to communicate (e.g. DoS) packets can be dropped rather than enqueued before they add to congestion



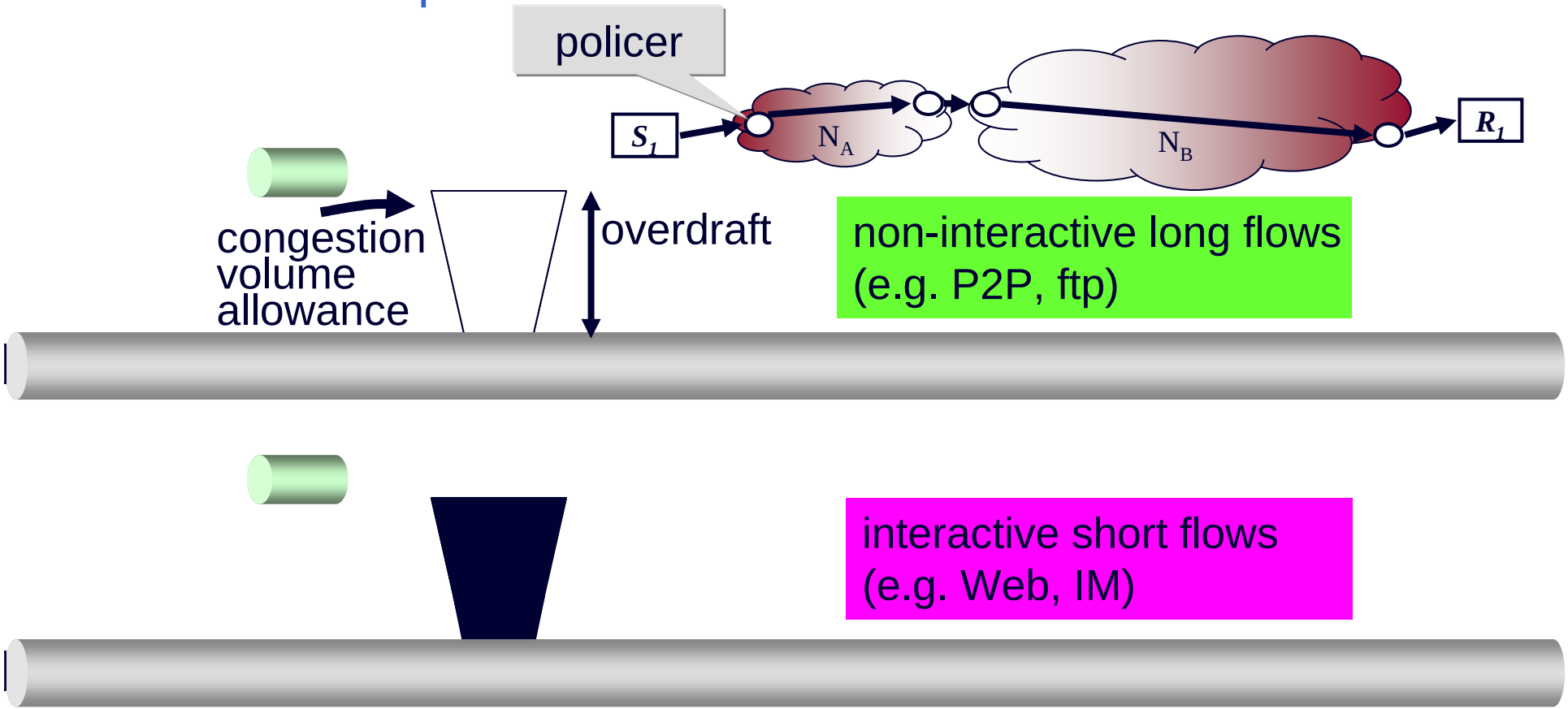
Policing Congestion using Re-feedback

animation requires Office XP



congestion policer – one example: per-user policer differentiator for ISPs

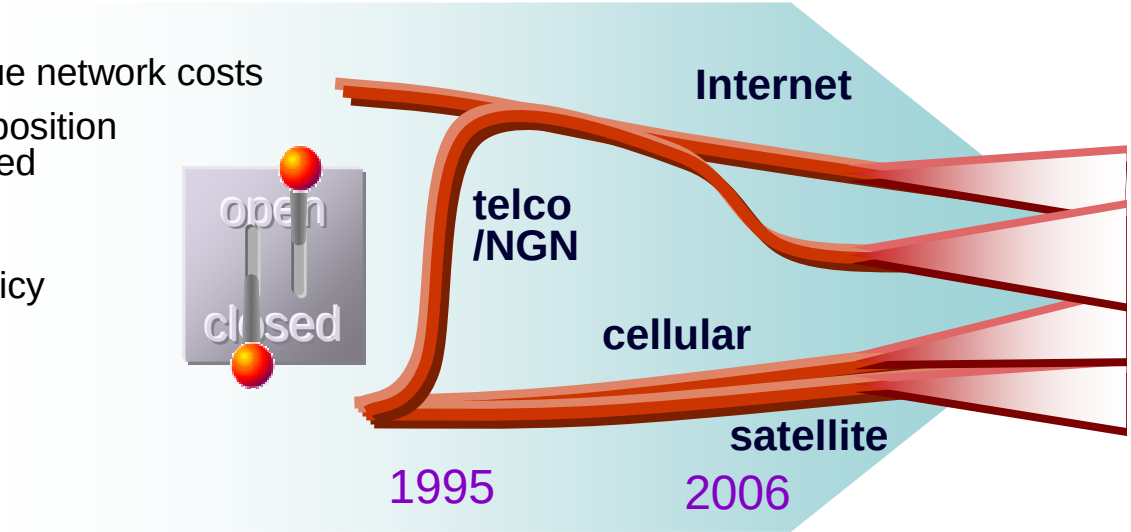
solution step #3



two different customers, same deal

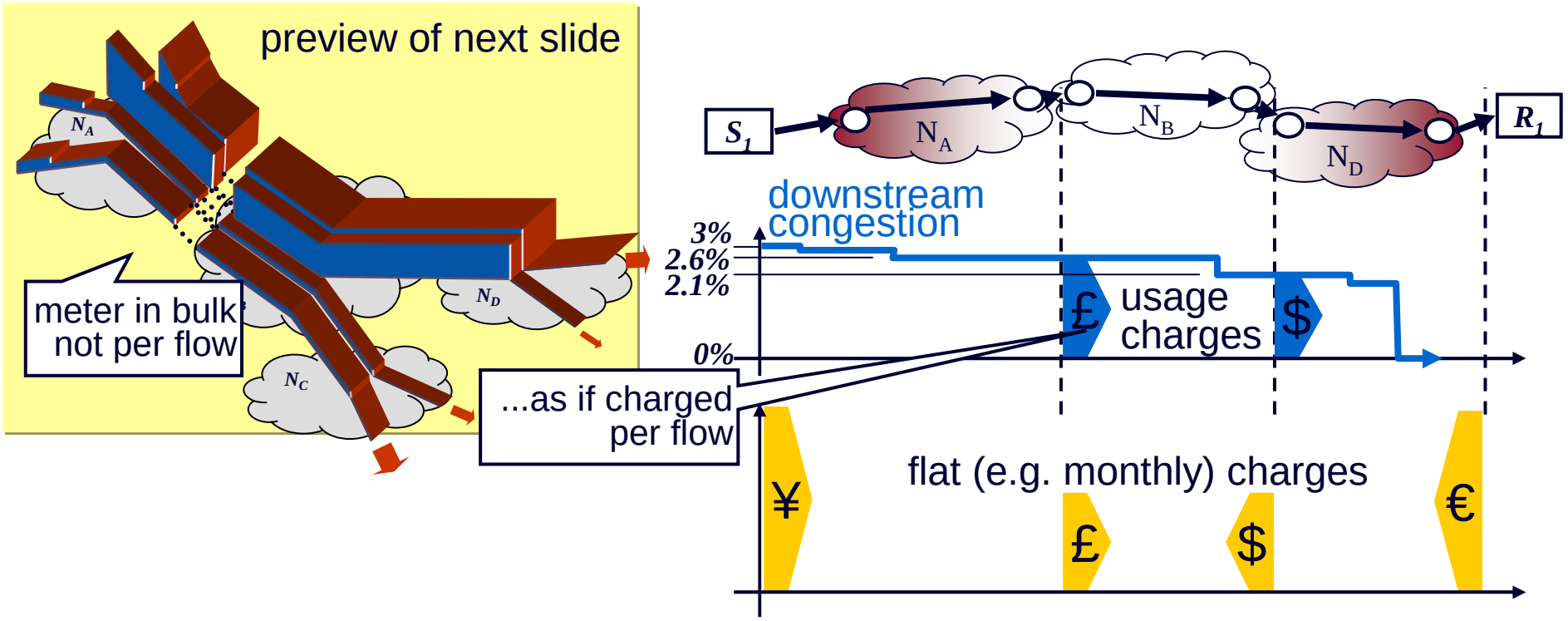
openness can be a tactic not a strategy

- edge congestion policer is the focus all network policy enforcement
  - **open:** per-user policing of bulk congestion volume
    - will allow much more freedom to innovate than current TCP-fairness constraint
    - new behaviours: e.g. very hi-speed, unresponsive, weighted, networked games
    - but all within overall responsibility envelope
  - **closed:** per-flow policing of specific application's congestion response
    - the place where service access is enforced, given IP packets needn't declare service
- Retailers choose
  - how tightly to control true network costs
  - each product's market position between open and closed
- Changing your mind
  - involves changing a policy
  - not new technology
- Wholesaler is agnostic
  - supports all positions
  - simultaneously



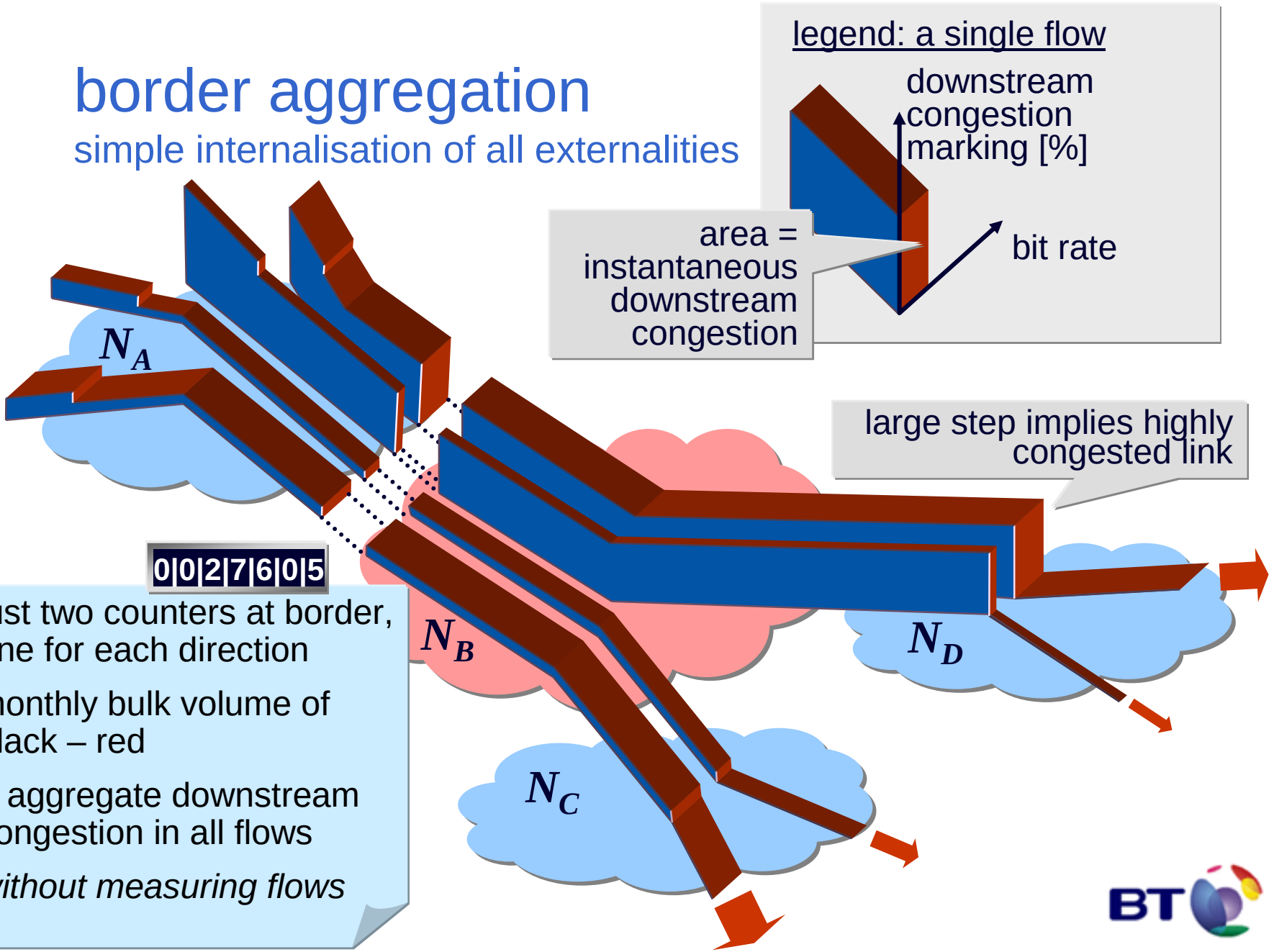
inter-domain accountability for congestion

- metric for inter-domain SLAs or usage charges
  - N_B applies penalty to N_A for bulk volume of congestion per month
  - could be tiered penalties, directly proportionate usage charge, etc.
  - penalties de-aggregate precisely back to responsible networks



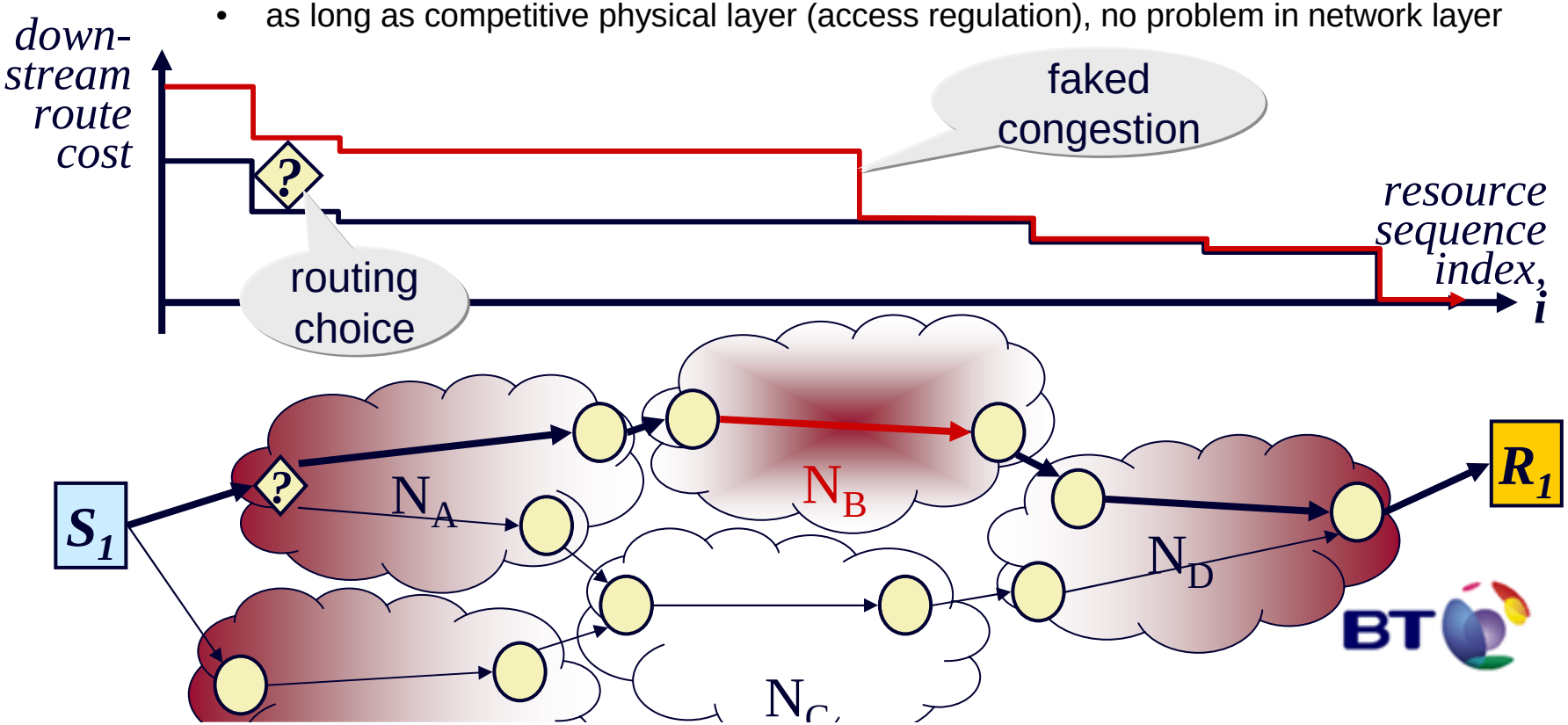
border aggregation

simple internalisation of all externalities



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
  - as long as competitive physical layer (access regulation), no problem in network layer



solution summary
list of problems solved

- sender & forwarder accountability for costs caused by traffic
  - without constraining pricing
- network accountability for insufficient capacity
- enforceable fairness
  - networks can be liberal
  - but conservative networks can protect their interests
  - different fairness regimes possible within groups of endpoints
- incremental deployment without changing forwarding
- first line of defence against DDoS
  - creates strong incentives to deploy DDoS defences
- differentiated bandwidth QoS ‘just happens’
  - bandwidth & jitter guarantees using edge-edge gateways (see PCN)
- all by packets revealing rest-of-path congestion

Internet resource sharing

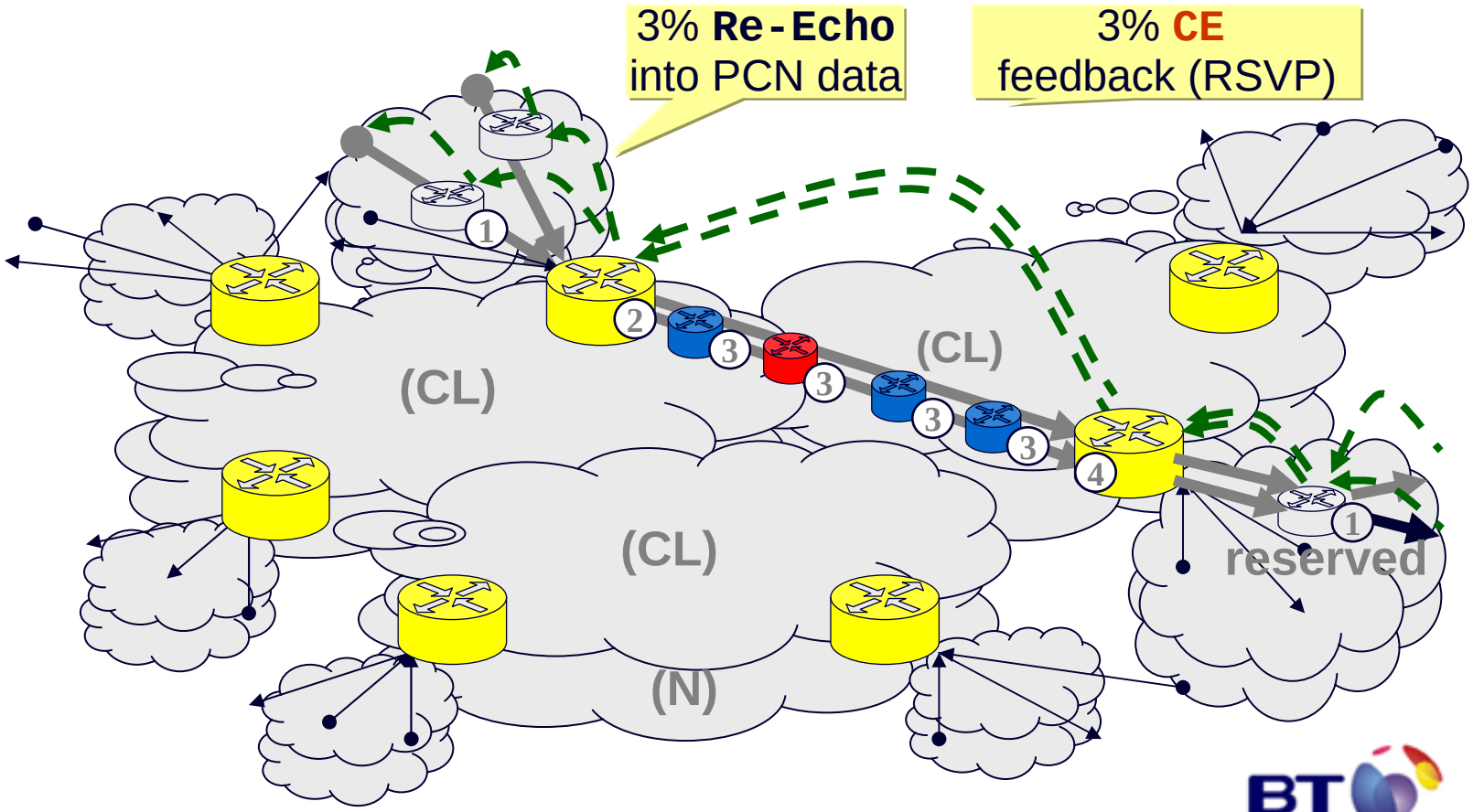
QoS & DoS

Bob Briscoe

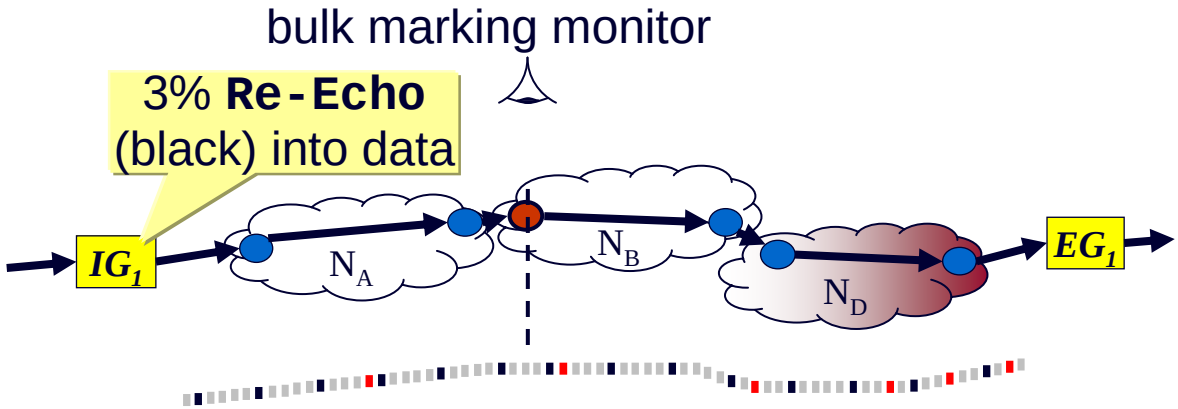
scalable admission control

using pre-congestion notification (PCN)

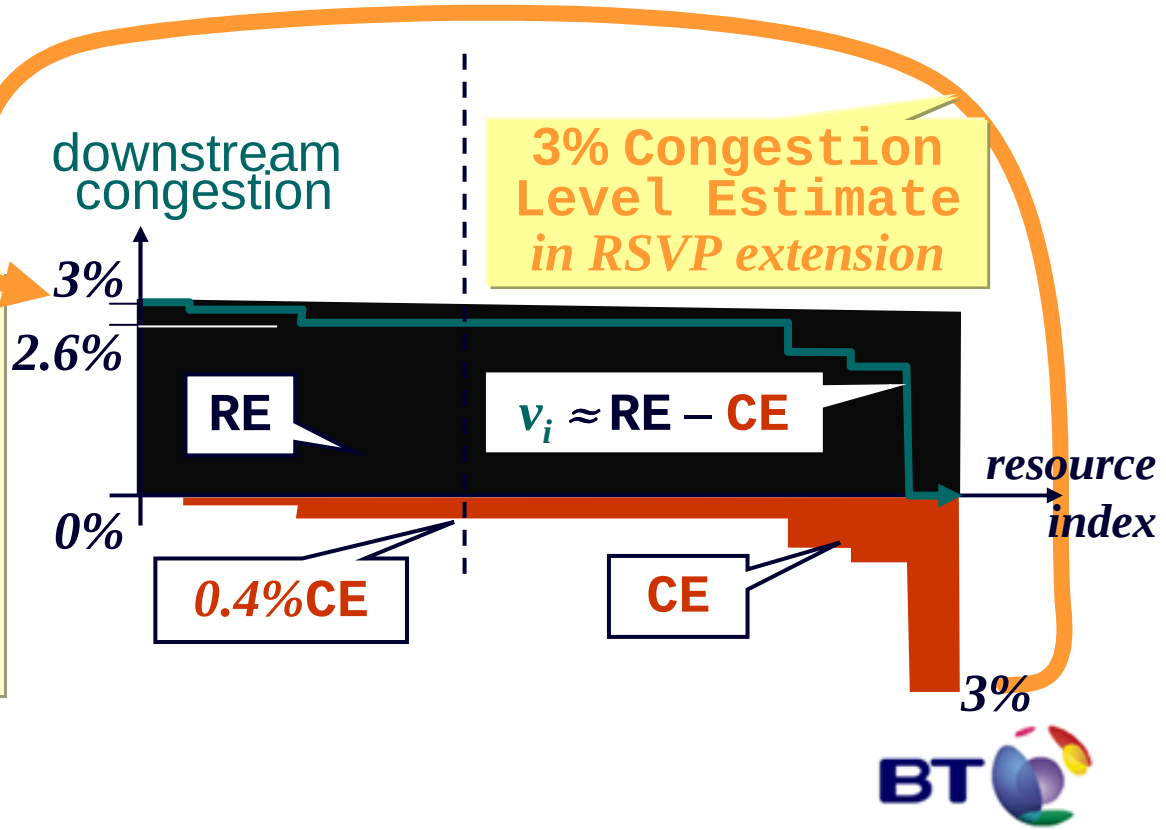
border anti-cheating solution



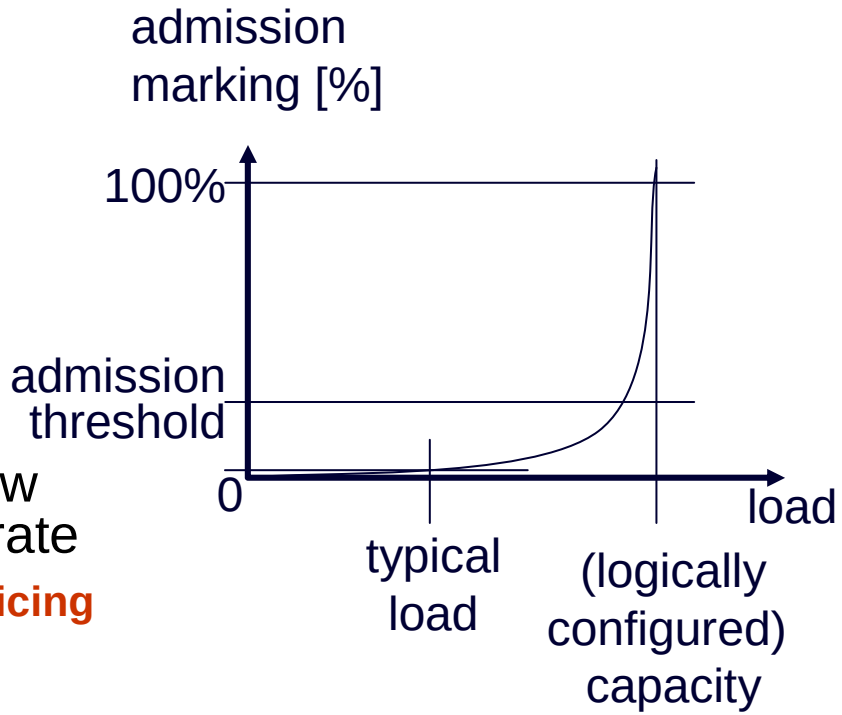
re-PCN



- 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



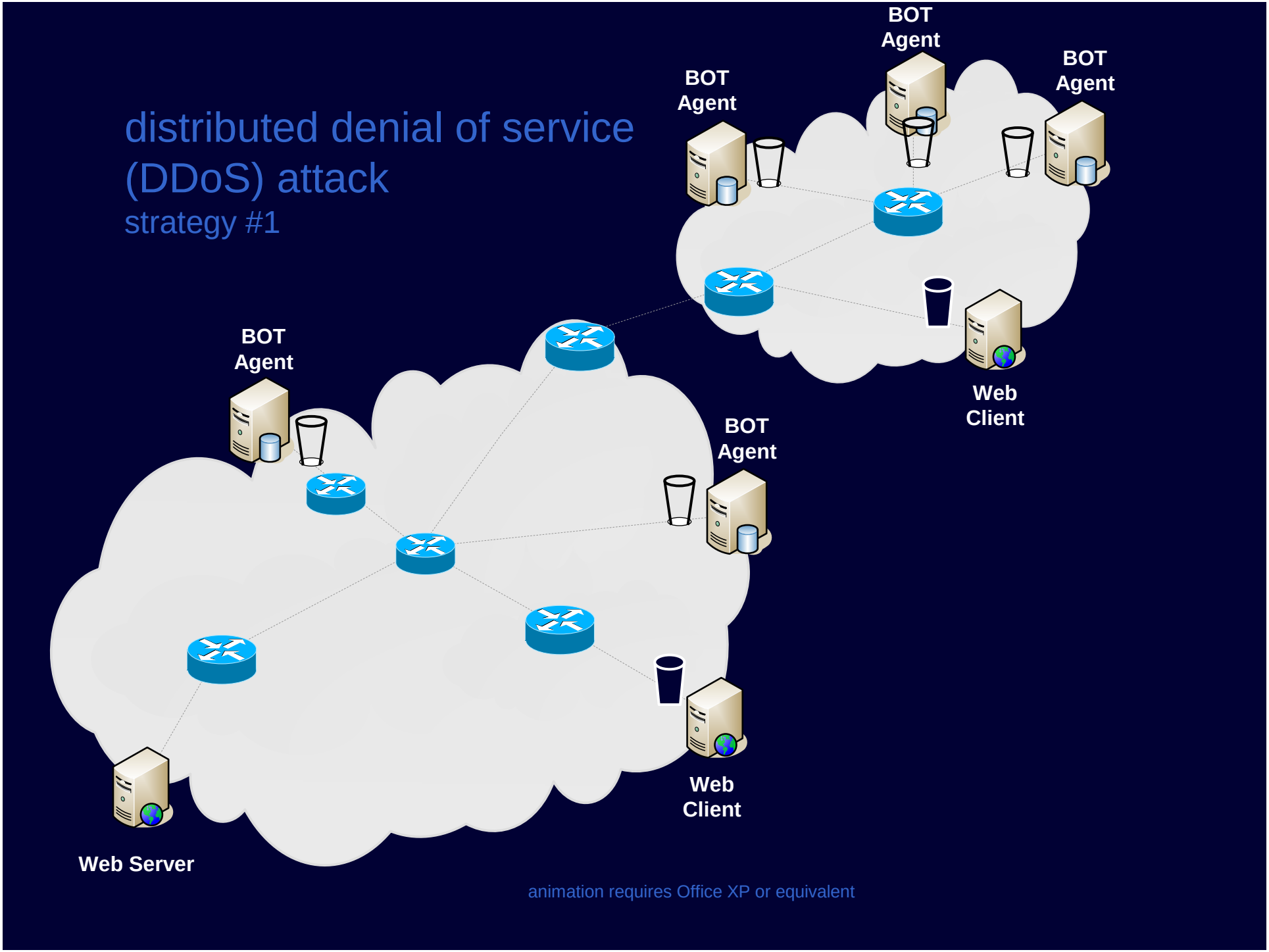
solution rationale



- <0.01% packet marking at typical load
  - addition of any flow makes little difference to marking
- penalties to ingress of each flow appear proportionate to its bit rate
  - **emulates border flow rate policing**
- as load approaches capacity
  - penalties become unbearably high (~1000x typical)
  - insensitive to exact configuration of admission threshold
  - **emulates border admission control**
- neither is a perfect emulation
  - but should lead to the desired behaviour
  - fail-safes if networks behave irrationally (e.g. config errors) – see draft

distributed denial of service (DDoS) attack

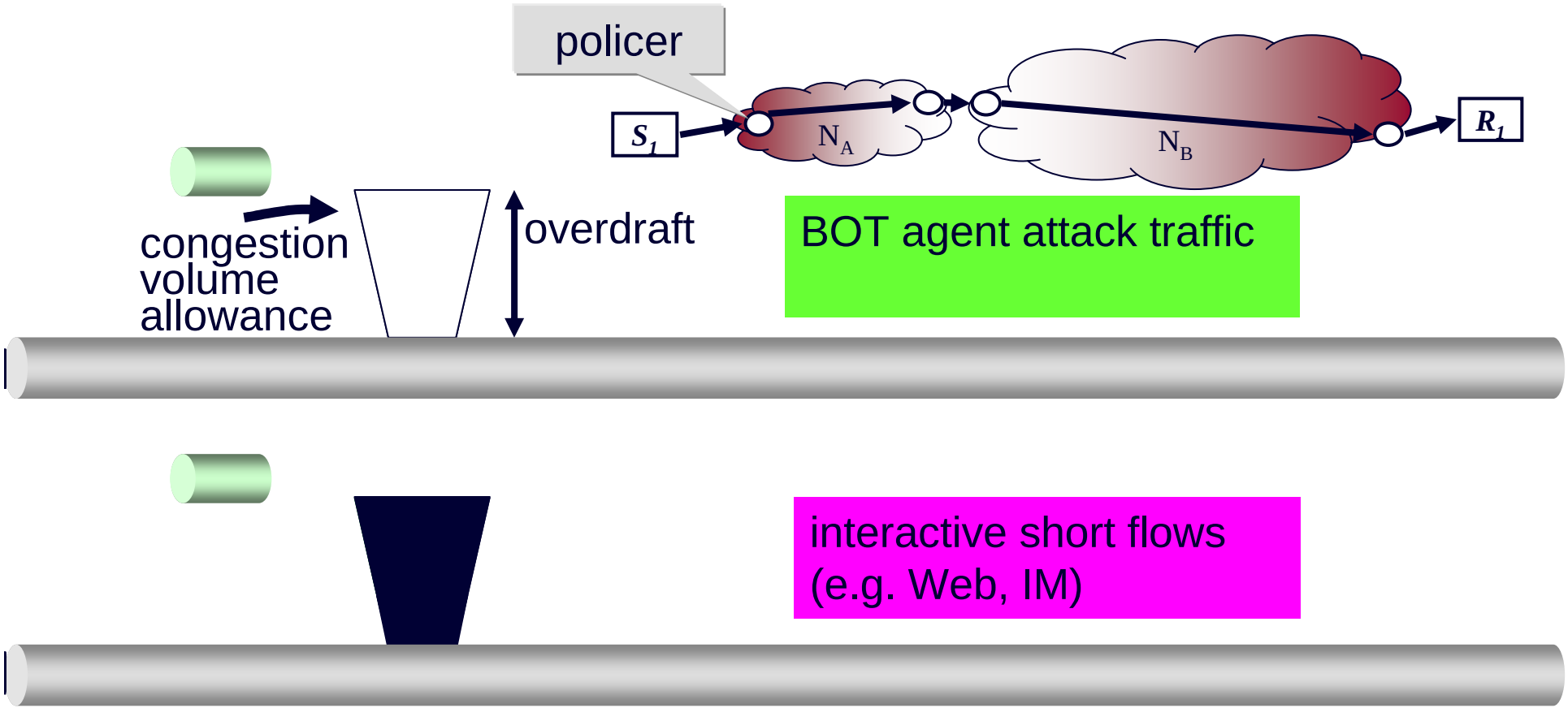
strategy #1



animation requires Office XP or equivalent

per-user congestion policer

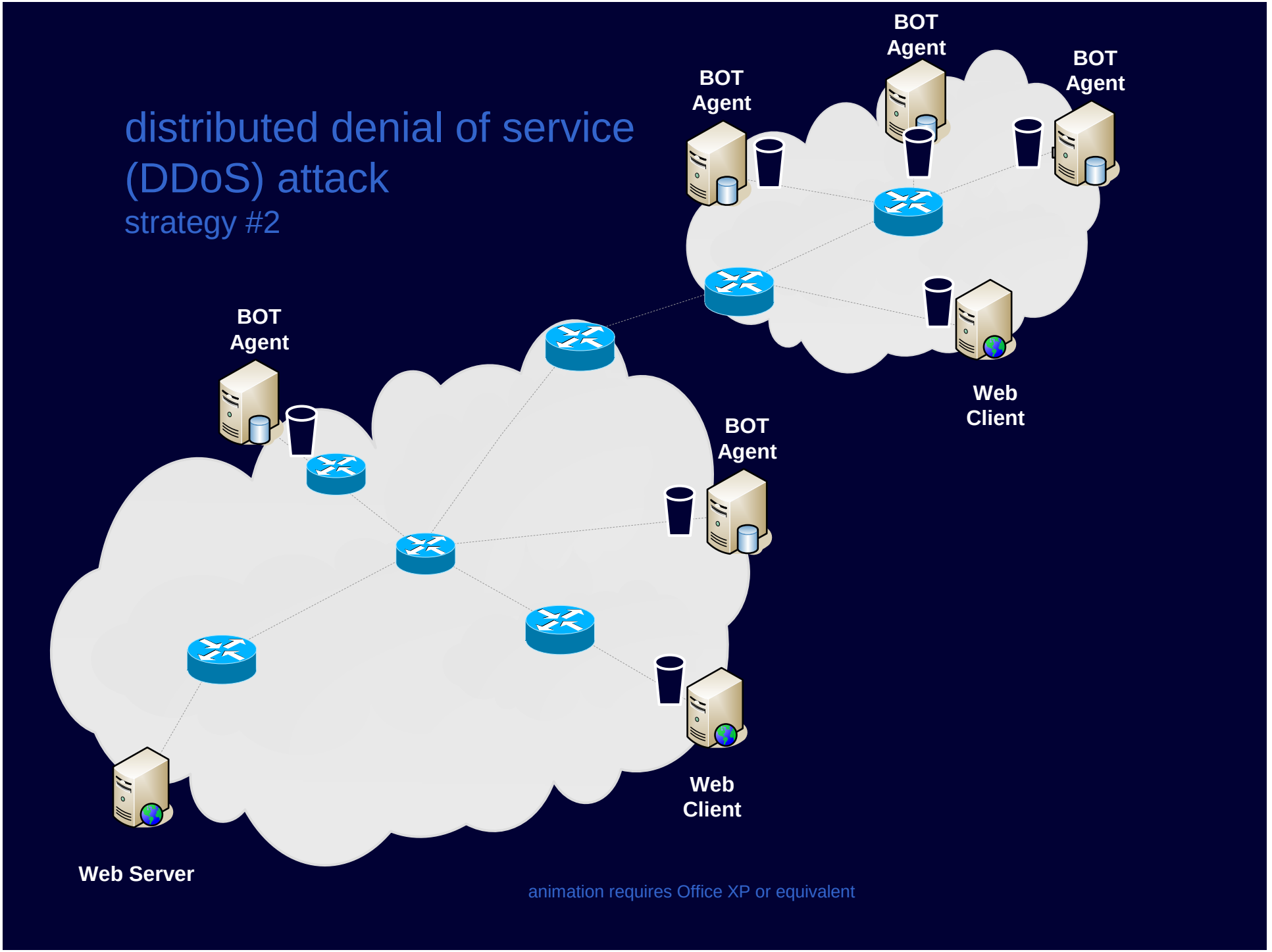
DDoS attack strategy #1



animation requires Office XP or equivalent

distributed denial of service (DDoS) attack

strategy #2



animation requires Office XP or equivalent

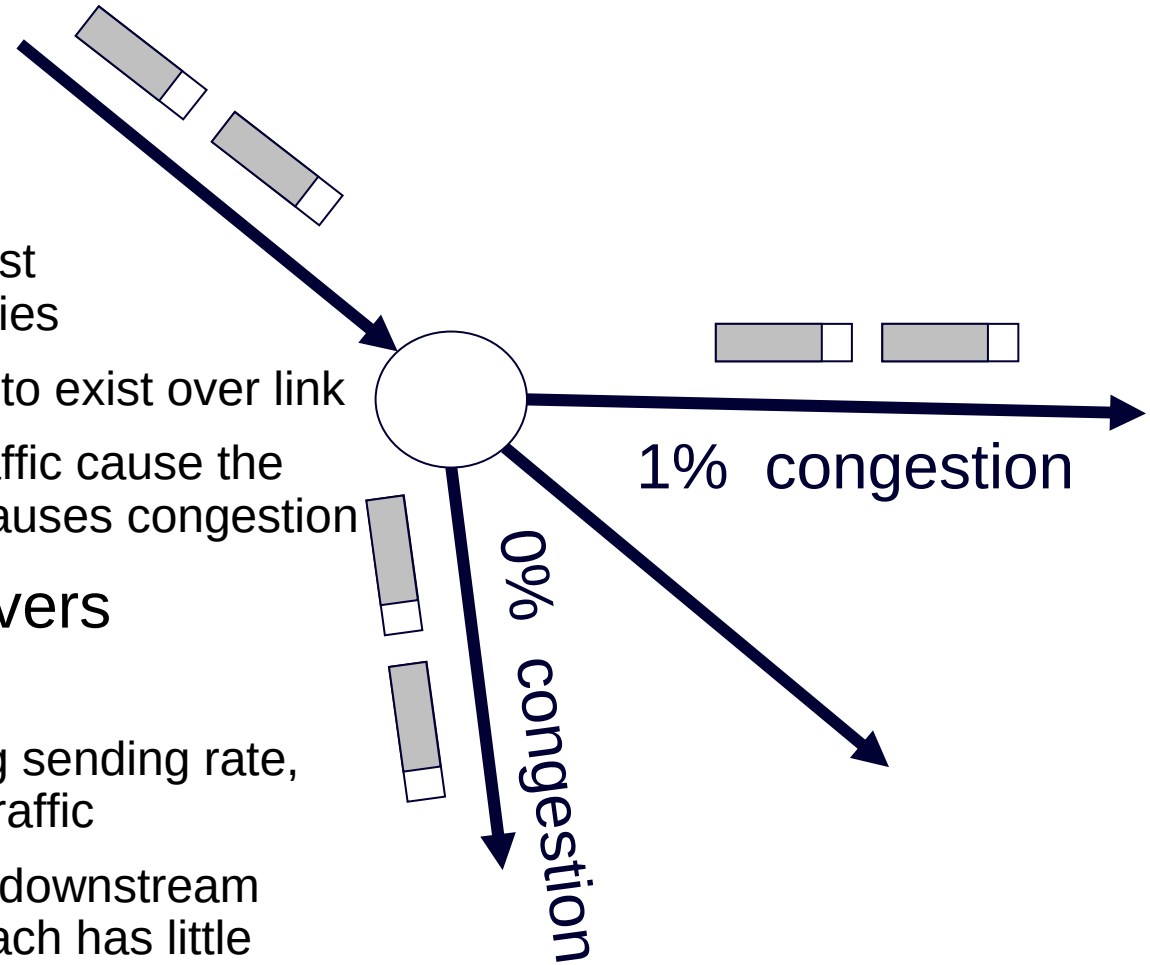
Internet resource sharing

towards a solution for multicast

Bob Briscoe

multicast congestion cost causation?

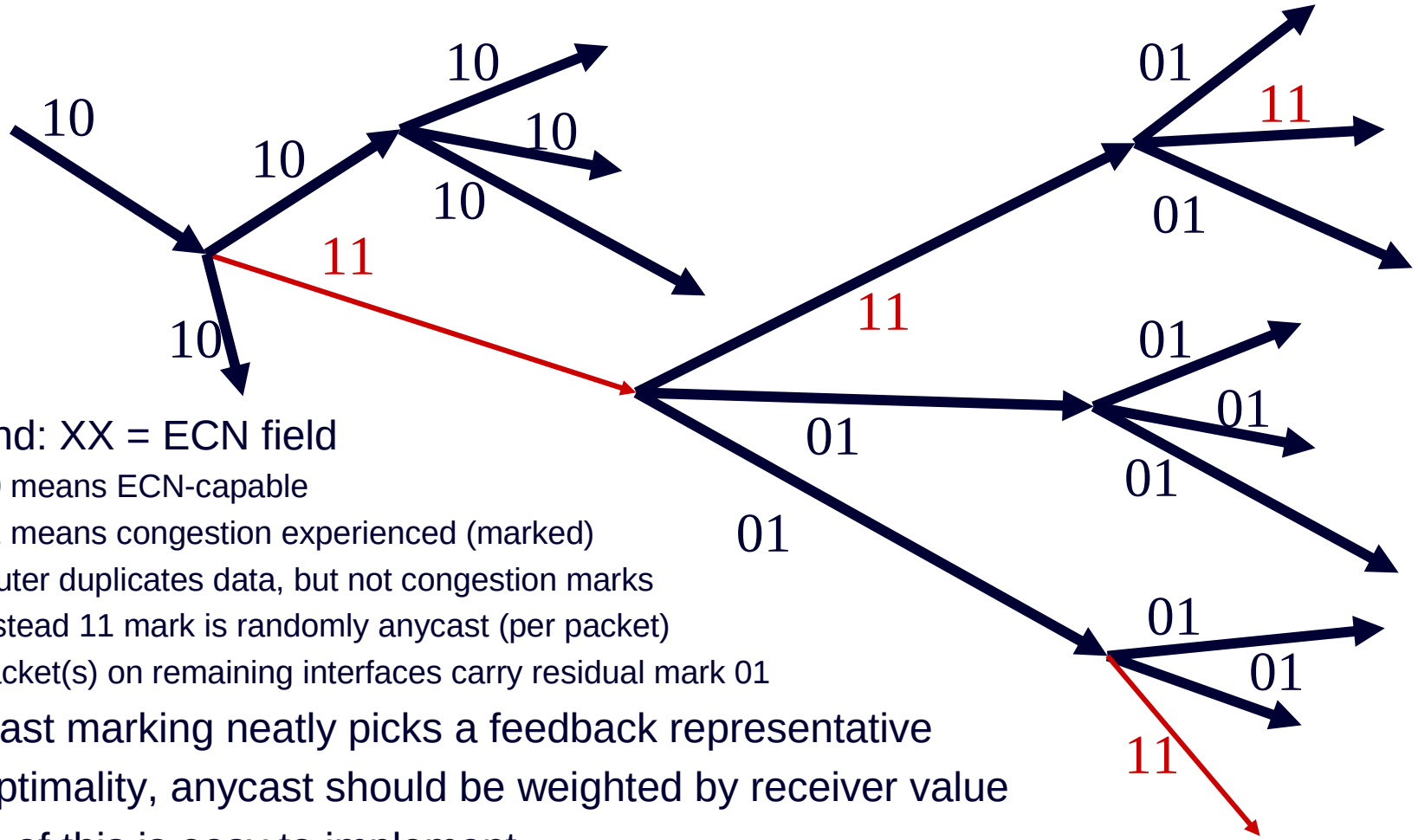
- strictly
  - operator causes packet duplication service to exist and chooses link capacities
  - receivers cause session to exist over link
  - sender & background traffic cause the traffic rate that directly causes congestion
- easier to make receivers responsible for costs
  - but receivers not causing sending rate, only existence of *some* traffic
  - to remove cost, need all downstream receivers to leave, but each has little incentive given cost should be shared



BT

multicast & congestion notification

antidote to arbitrary ‘research’ on fairness between unicast & multicast



- legend: XX = ECN field
  - 10 means ECN-capable
  - 11 means congestion experienced (marked)
  - router duplicates data, but not congestion marks
  - instead 11 mark is randomly anycast (per packet)
  - packet(s) on remaining interfaces carry residual mark 01
- anycast marking neatly picks a feedback representative
- for optimality, anycast should be weighted by receiver value
- none of this is easy to implement
- can’t do any of this with drop

Internet resource sharing

next steps

Bob Briscoe

references

[MacKieVarian95] MacKie-Mason, J. and H. Varian, “Pricing Congestible Network Resources,” IEEE Journal on Selected Areas in Communications, `Advances in the Fundamentals of Networking' 13(7)1141--1149, 1995 http://www.sims.berkeley.edu/~hal/Papers/pricing-congestible.pdf

[Kelly98] Frank P. Kelly, Aman K. Maulloo, and David K. H. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. Journal of the Operational Research Society, 49(3):237–252, 1998

[Gibbens99] Richard J. Gibbens and Frank P. Kelly, Resource pricing and the evolution of congestion control, Automatica 35 (12) pp. 1969—1985, December 1999 (lighter version of [Kelly98])

[Briscoe01] Bob Briscoe and Jon Crowcroft, “An Open ECN service in the IP layer” Internet Engineering Task Force Internet Draft (Expired) http://www.cs.ucl.ac.uk/staff/B.Briscoe/pubs.html#ECN-IP (February 2001)

[Clark05] David D Clark, John Wroclawski, Karen Sollins and Bob Braden, "Tussle in Cyberspace: Defining Tomorrow's Internet," IEEE/ACM Transactions on Networking (ToN) 13(3) 462–475 (June 2005) <portal.acm.org/citation.cfm?id=1074049>

[PCN] Phil Eardley (Ed), “Pre-Congestion Notification Architecture,” IETF Internet Draft draft-ietf-pcn-architecture-02.txt (Nov ’07)

[Siris] Future Wireless Network Architecture <www.ics.forth.gr/netlab/wireless.html>

[M3I] Market Managed Multi-service Internet consortium <www.m3i_project.org/>

re-feedback & re-ECN project page <www.cs.ucl.ac.uk/staff/B.Briscoe/projects/refb/>

[Briscoe05a] Bob Briscoe, Arnaud Jacquet, Carla Di-Cairano Gilfedder, Andrea Soppera and Martin Koyabe, "Policing Congestion Response in an Inter-Network Using Re-Feedback“ In: Proc. ACM SIGCOMM'05, Computer Communication Review 35 (4) (September, 2005)

[Briscoe05b] Commercial Models for IP Quality of Service Interconnect, Bob Briscoe & Steve Rudkin, in BTTJ Special Edition on IP Quality of Service, 23(2) (Apr 2005)

[Briscoe06] Using Self-interest to Prevent Malice; Fixing the Denial of Service Flaw of the Internet, Bob Briscoe, The Workshop on the Economics of Securing the Information Infrastructure (Oct 2006)

[Briscoe07] Flow Rate Fairness: Dismantling a Religion, Bob Briscoe, ACM CCR 37(2) 63--74 (Apr 2007)

[Briscoe08] Re-ECN: Adding Accountability for Causing Congestion to TCP/IP, Bob Briscoe, Arnaud Jacquet, Toby Moncaster and Alan Smith, IETF Internet-Draft <draft-briscoe-tsvwg-re-ecn-tcp-05.txt> (Jan 2008)

Internet resource sharing

Q&A

typical p2p file-sharing apps

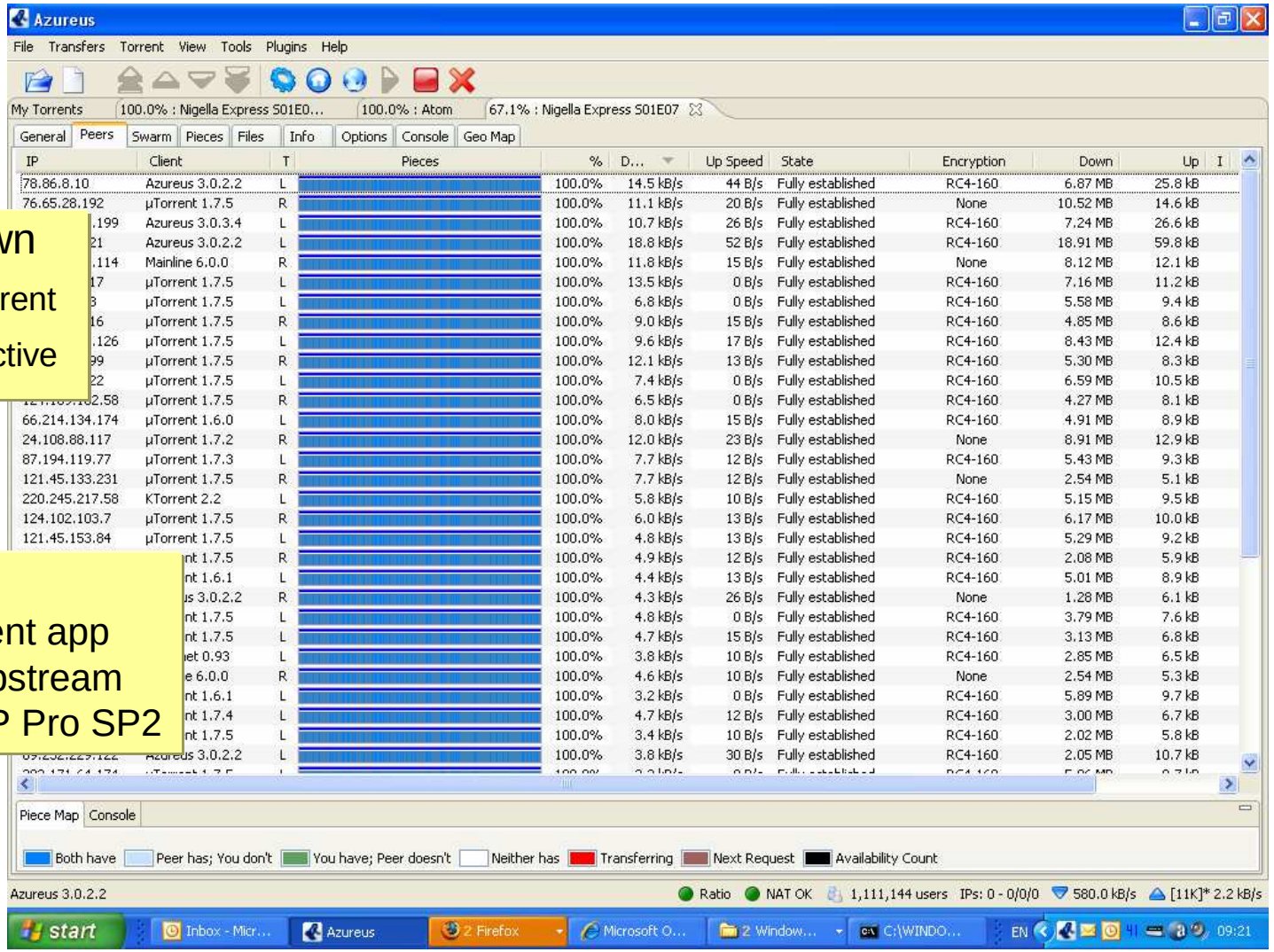
- 105-114 active TCP connections altogether

1 of 3 torrents shown

- ~45 TCPs per torrent
- but ~40/torrent active

environment

Azureus BitTorrent app
ADSL+ 448kb upstream
OS: Windows XP Pro SP2

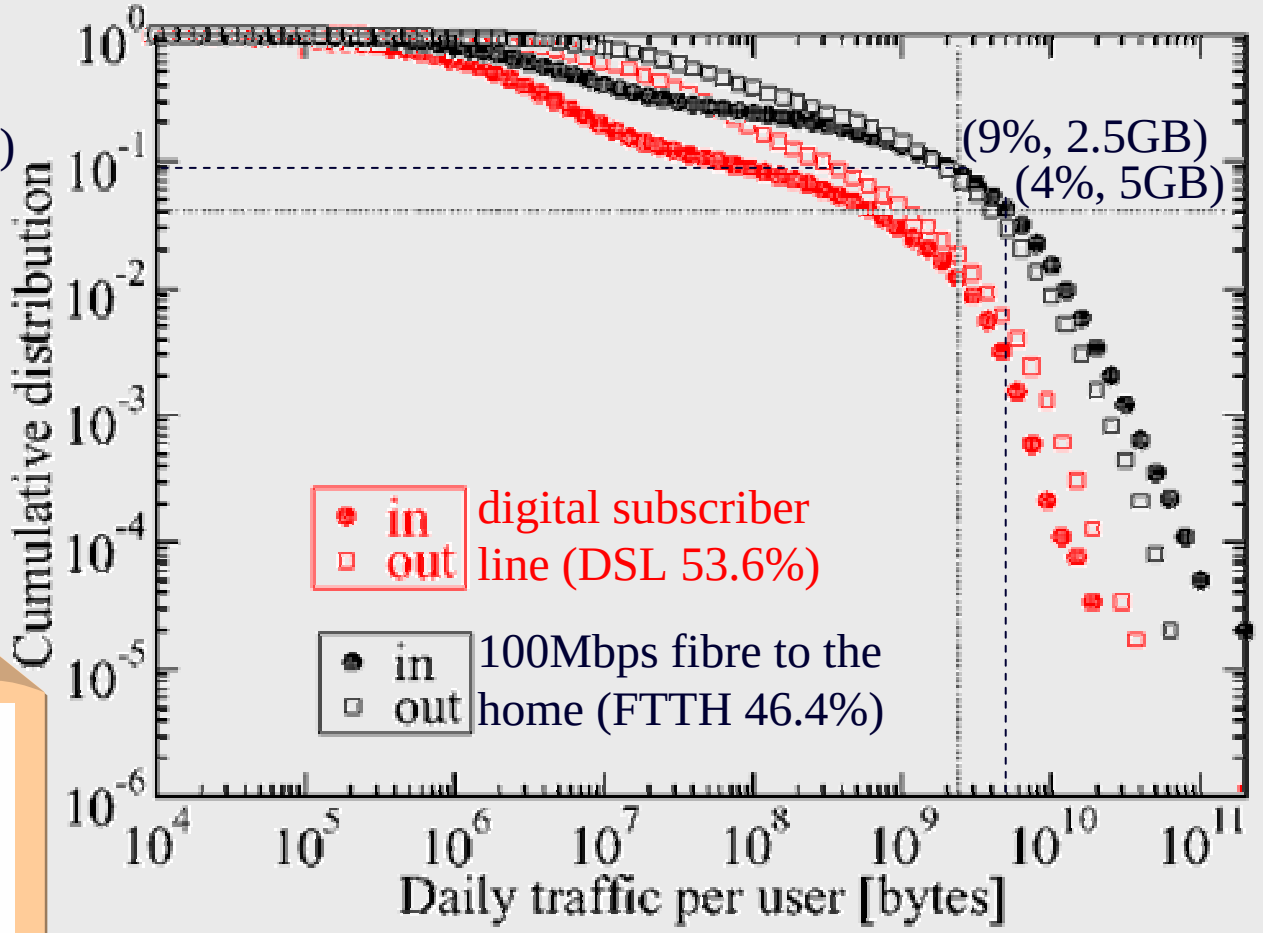


capacity growth will prevent congestion?

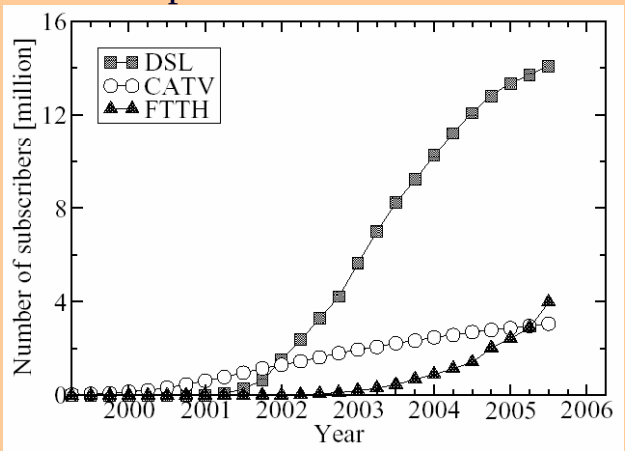
public problem

Distribution of customers' daily traffic into & out of a Japanese ISP (Feb 2005)

(5GB/day equivalent to 0.46Mbps if continuous)



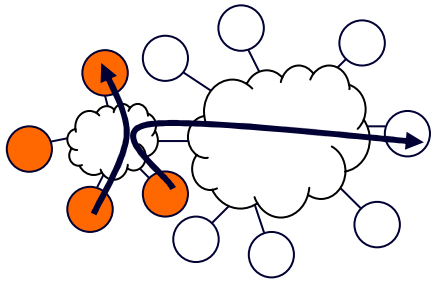
Changing technology shares of Japanese access market



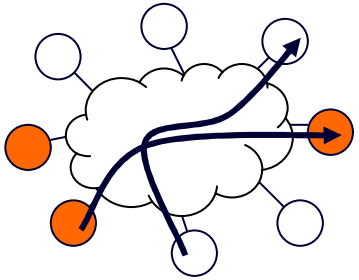
Courtesy of Kenjiro Cho et al
The Impact and Implications of the Growth in Residential User-to-User Traffic, SIGCOMM'06

BT

fairness between fairnesses

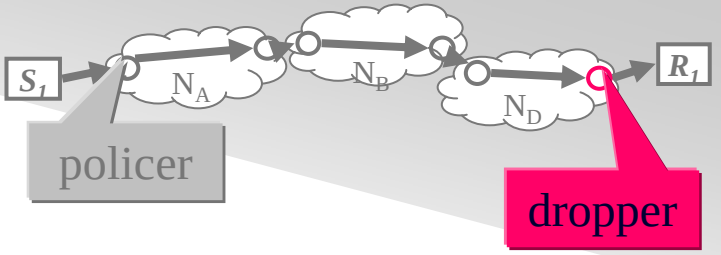


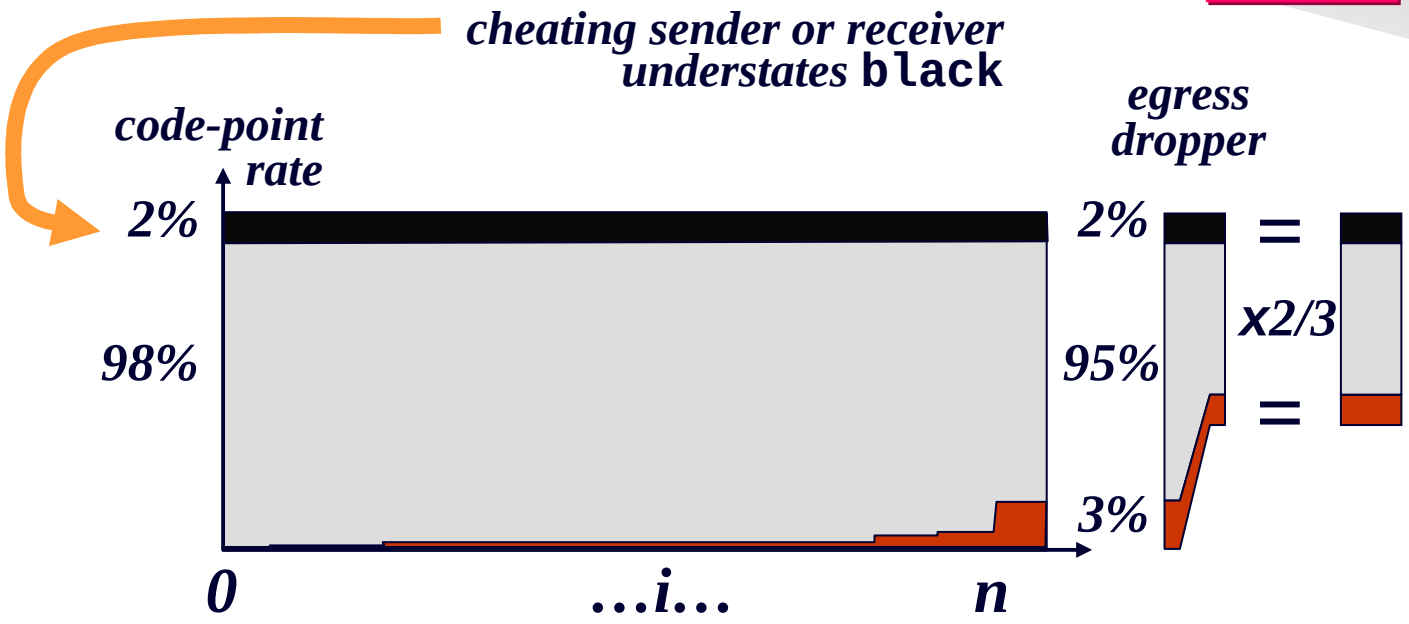
- to isolate a subgroup who want their own fairness regime between them
  - must accept that network between them also carries flows to & from other users
- in life, local fairnesses interact through global trade
  - e.g. University assigns equal shares to each student
    - but whole Universities buy network capacity from the market
  - further examples: governments with social objectives, NATO etc



- cost fairness sufficient to support allocation on global market
  - then subgroups can reallocate the right to cause costs within their subgroup
    - around the edges (higher layer)
  - naturally supports current regime as one (big) subgroup
    - incremental deployment
- different fairness regimes will grow, shrink or die
  - determined by market, governments, regulators, society – around the edges
  - all over congestion marking at the IP layer – neck of the hourglass

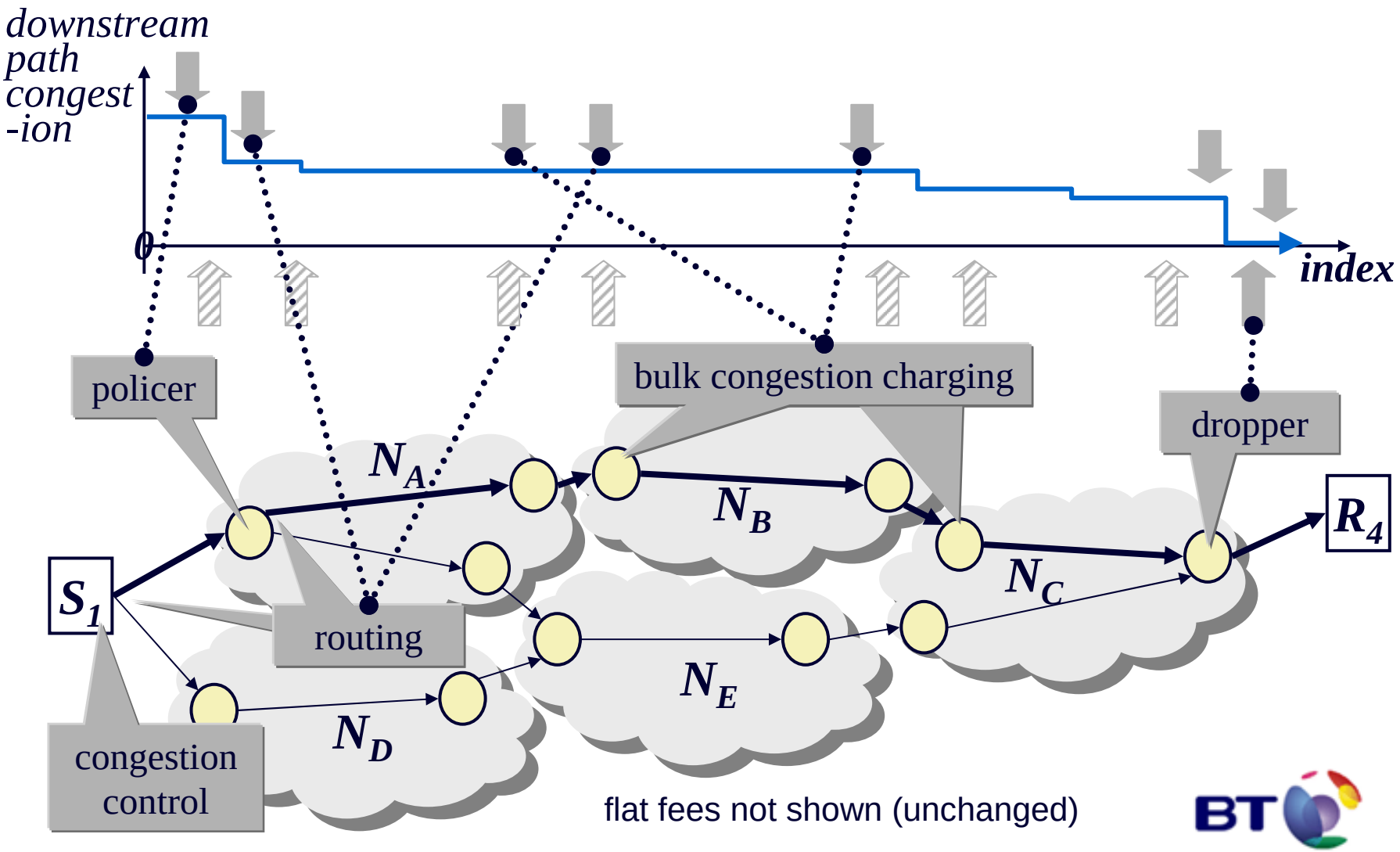
egress dropper (sketch)





- drop enough traffic to make fraction of **red** = **black**
- goodput best if rcvr & sender honest about feedback & re-feedback

incentive framework



re-feedback & routing support

- not done any analysis on this aspect