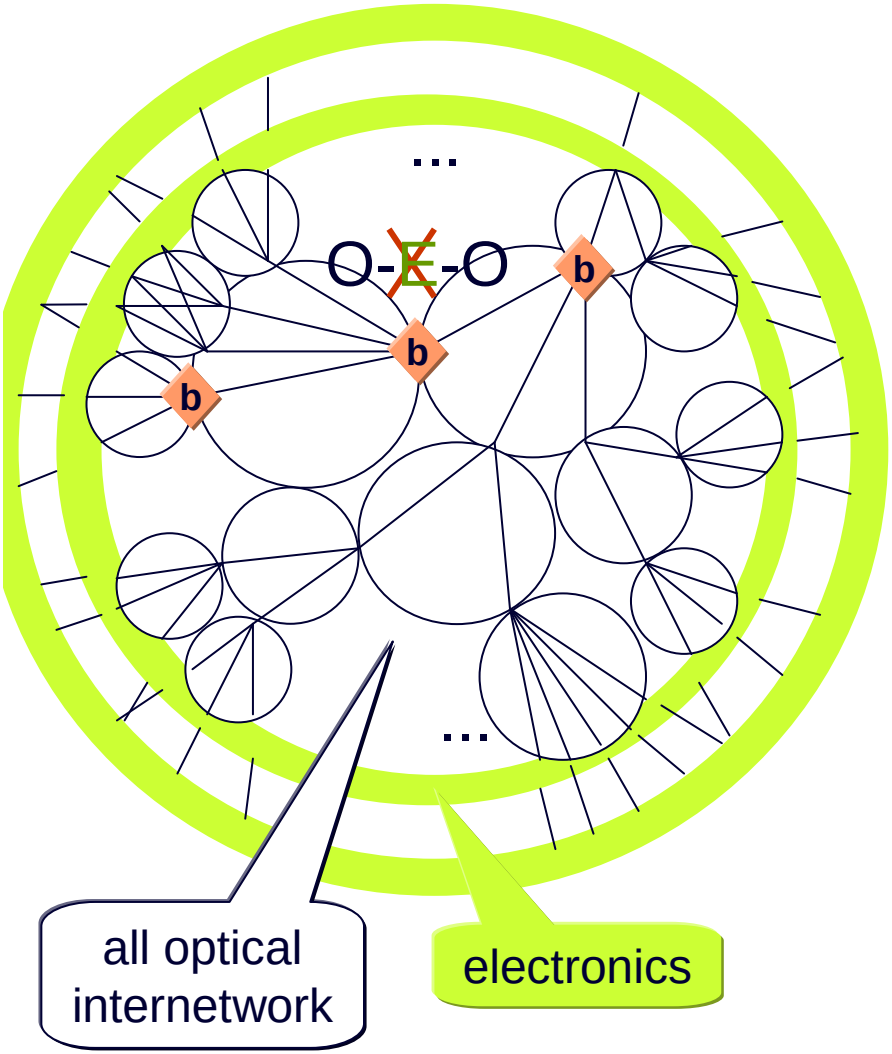
Pushing Packet Processing to the Edge

Scheduling in Optics is the Wrong Answer for Fine-Grained Resource Sharing

Bob Briscoe
Chief Researcher, BT Group

European Conf on Optical Communication (ECOC)
Workshop on Future Internet Design (FID)
Sep 2007

E-O-O-O-O-O-E
joined up thinking?

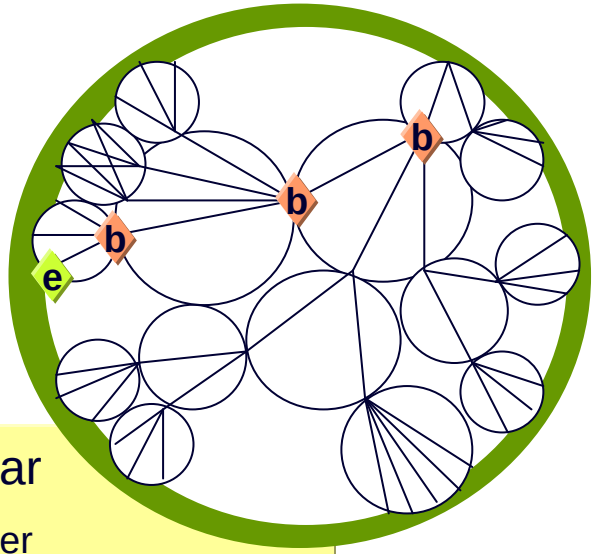


- >50% of comms revenues depend on paths over interconnect, just in UK
- O-E-O at borders will limit growth
  - 10-15yr horizon
- all-optical global internetwork?
  - with n ~ 10^4-10^6 electronic interfaces
- can we avoid store+forward in optics?
  - × label switching (store+forward) doesn't help
  - × use solely edge-edge λ circuits?
    - n^2 λs with most capacity wasted
  - × in a word, no
- best we can do is a mix
  - intra-domain λ circuits
  - but need (optical) packet routers at borders

the challenge

entrusting border packet functions to the edge

- border functions? or entrusted to edge?
  - **b** packet forwarding over n prefixes
  - transport functions:
    - **b** packet buffering
    - **b** active queue management (AQM)
    - **e** packet class scheduling (min 2 at **b**, rest at **e**)
    - **e** token bucket policing of classes
    - **e** flow admission control & policing
    - **e** session border control
    - **e** DDoS & fairness policing
  - **?** policy routing filters
  - **?** stateless / stateful firewalls
- whether optical or electronic
  - doing less at borders scales better
- entrusting critical border protection
  - "it's as much in your interest as mine to do this reliably for me"

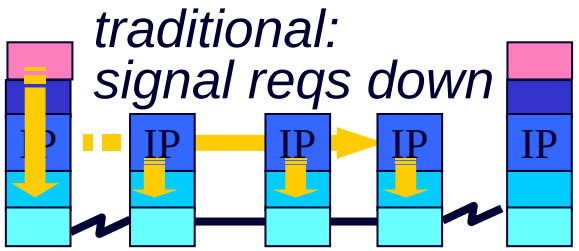


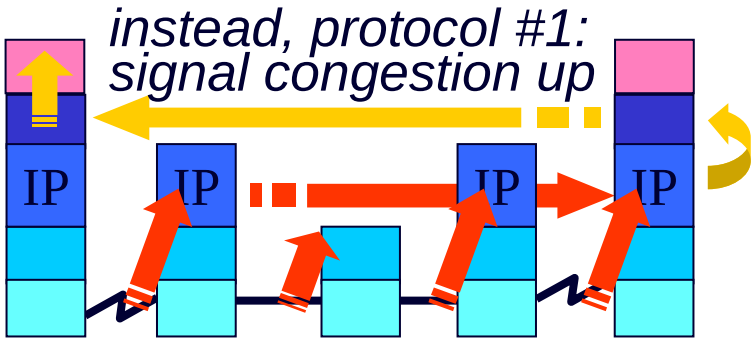
conclusions so far

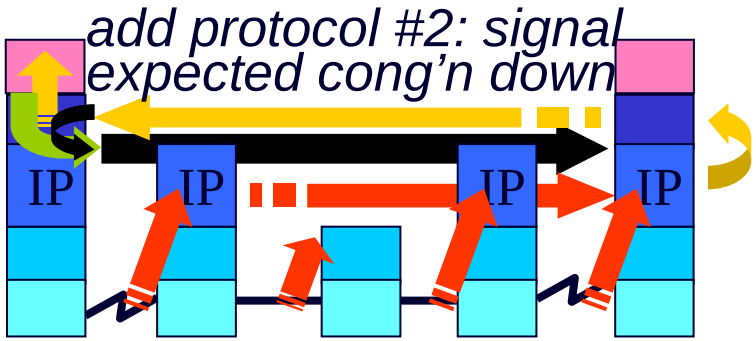
- **b** = must be at border
- **e** = can entrust to edge
- **?** = future research (Trilogy / WISDOM)

two building blocks

for entrusting transport control to edge



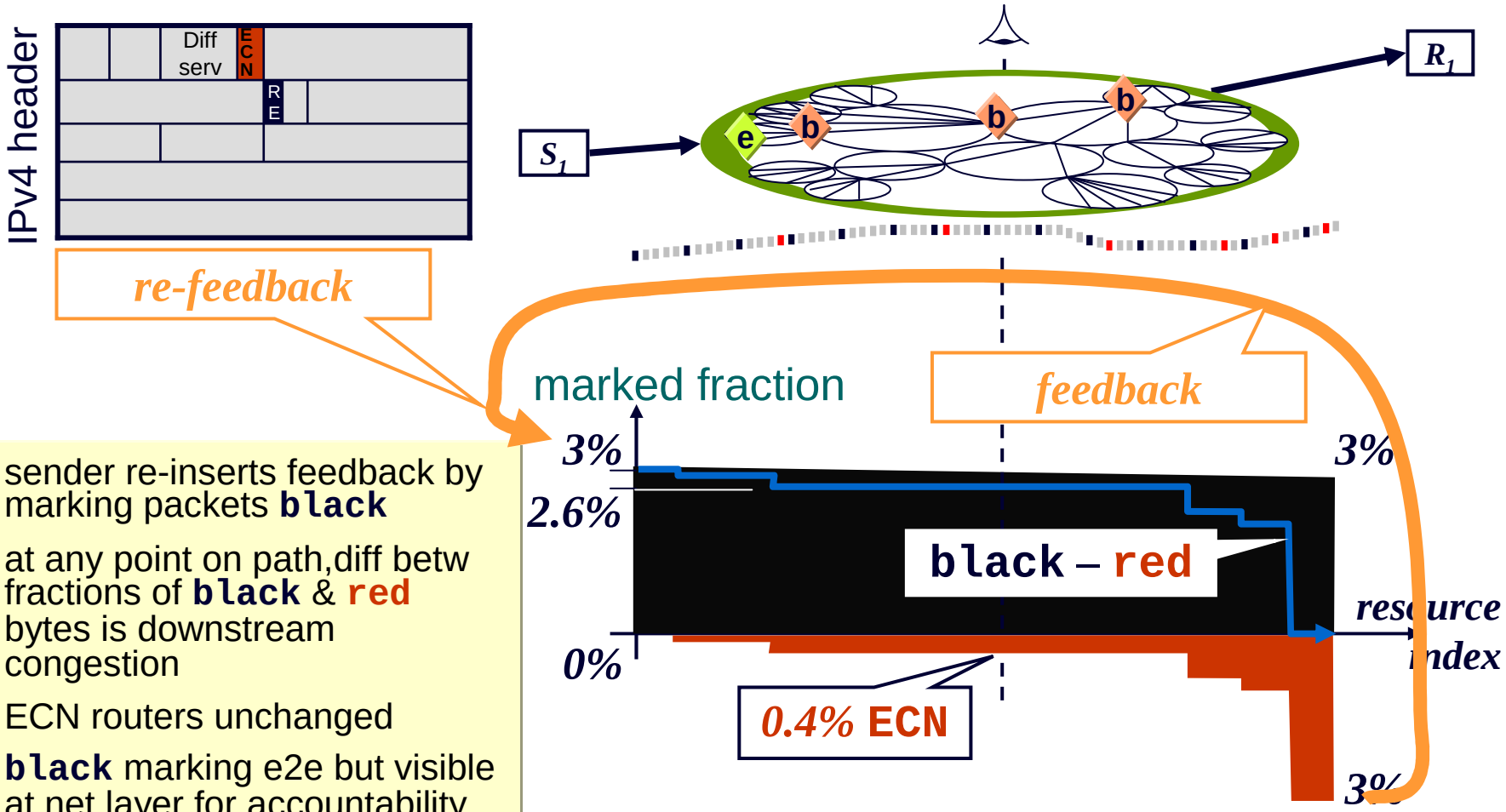




1. (already std) reveal approaching congestion experienced by packets
   - important for other nodes to see congestion, but difficult to detect missing packets
   - ECN = explicit congestion notification flag in IP header
     - or equivalent in lower layer header – propagated up the layers
     - each queue more likely to mark ECN field the longer the queue
   - markings have direct economic interpretation as marginal cost of usage
2. (proposed) reveal congestion that packets *expect* to experience
   - make sent packets declare congestion expected on path, in a second IP header flag
   - network elements don't change this field, but they can read it
   - if expected congestion persistently below actual (cheating), need not forward pkts
   - at start of a flow, sender needs to declare expectation conservatively
   - result: ingress edge can hold sender accountable for congestion that pkts *could* cause

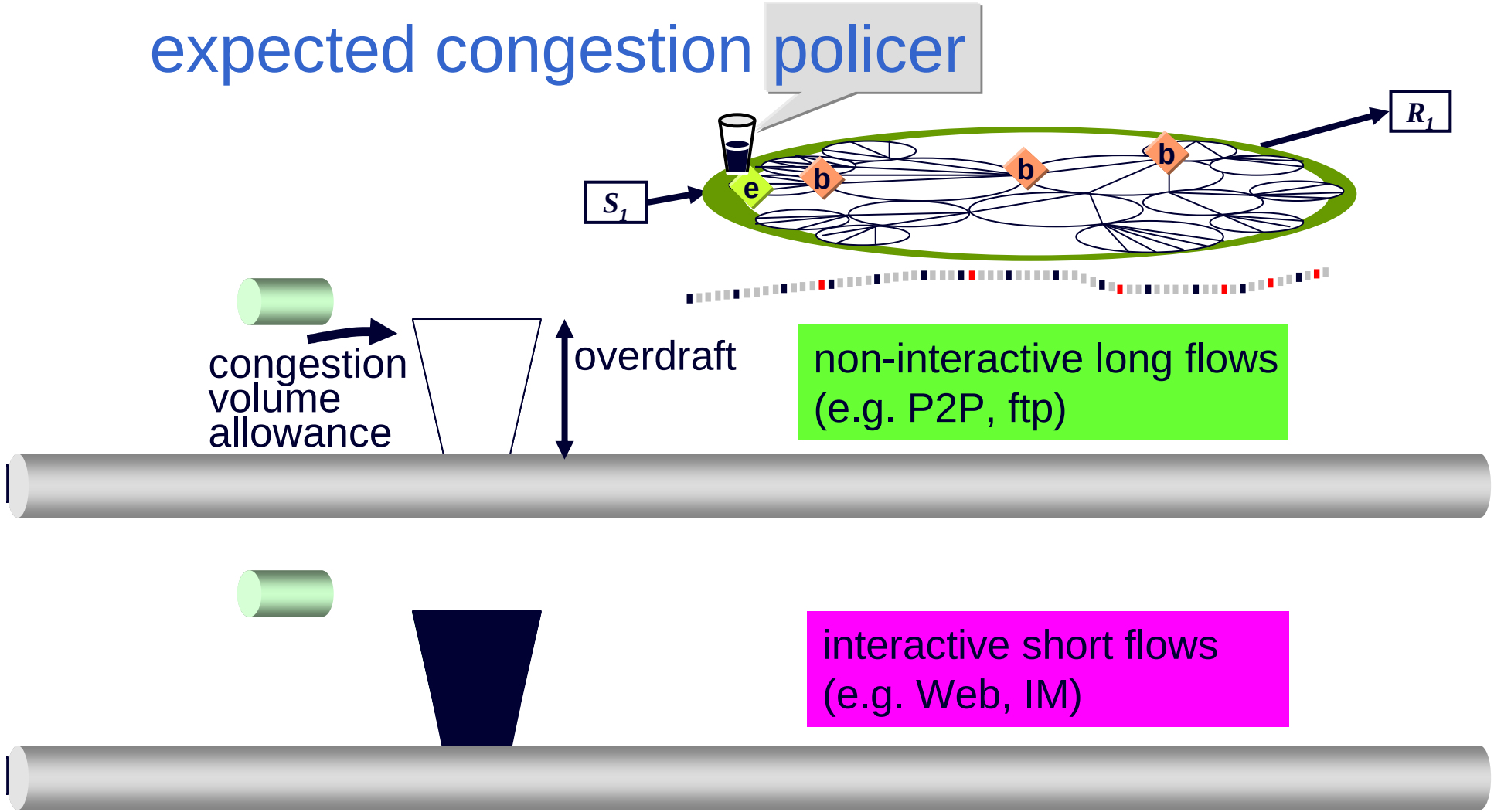
measurable downstream congestion

re-ECN – reinserted ECN feedback



- 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

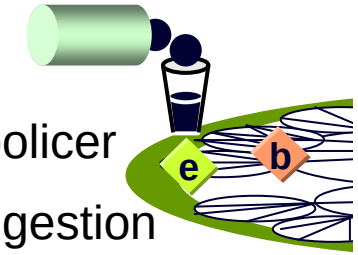
expected congestion policer



two different customers, same deal

edge-controlled differentiated service

- traditional differentiated service
  - scheduler at a congested queue gives premium packets priority
- edge-controlled differentiated service
  - just buy a faster congestion allowance feeding the edge policer
  - premium flow can just send faster, responding less to congestion
  - ECN early warning usually keeps everyone out of drop regime

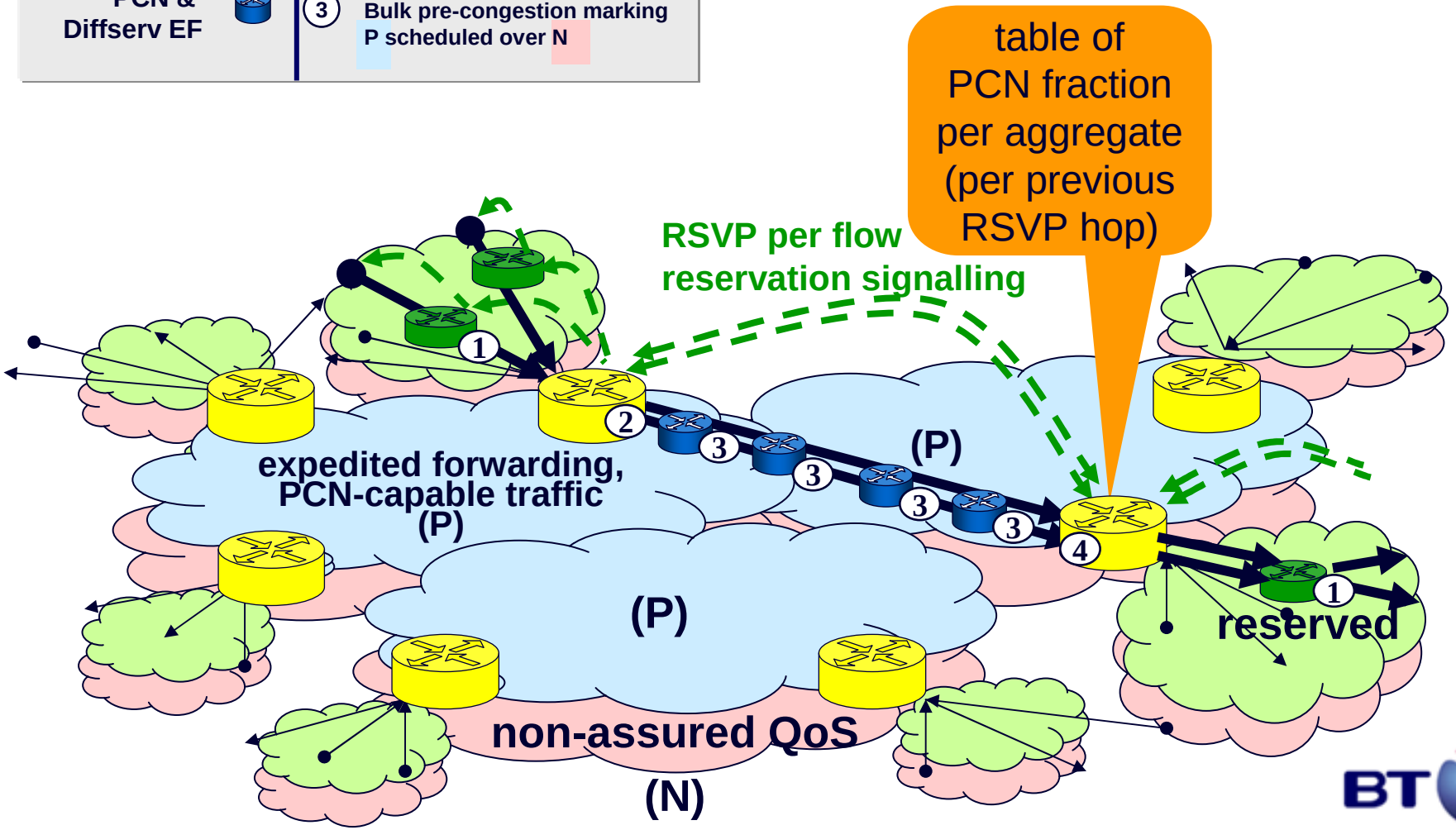


edge admission control

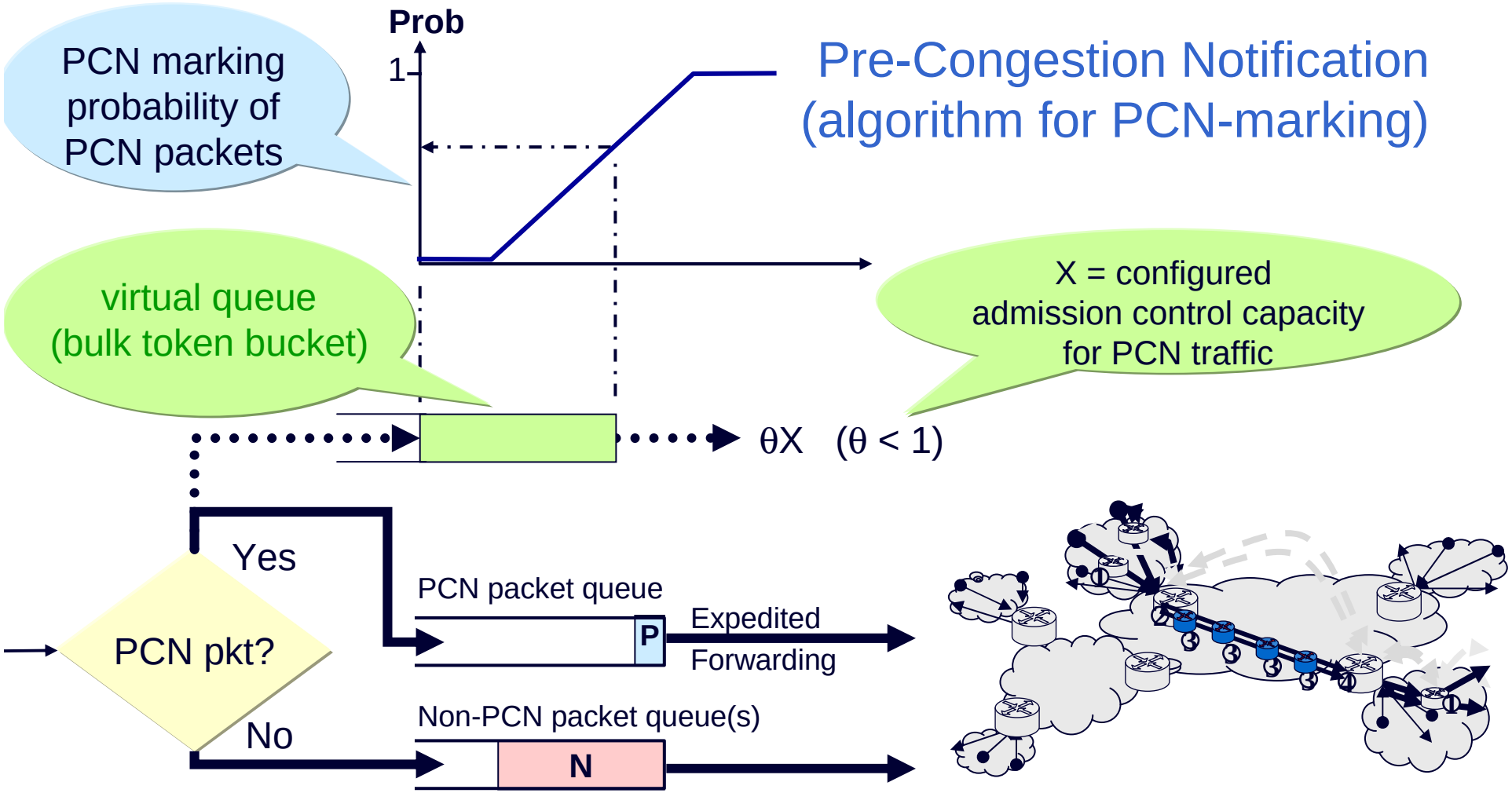
pre-congestion notification (PCN)

highlighting 2 flows

IP routers	Data path processing
Reservation enabled	1 Reserved flow processing
RSVP/PCN gateway	2 Policing flow entry to P 4 Meter pre-congestion per peer
PCN & Diffserv EF	3 Bulk pre-congestion marking P scheduled over N



Pre-Congestion Notification (algorithm for PCN-marking)

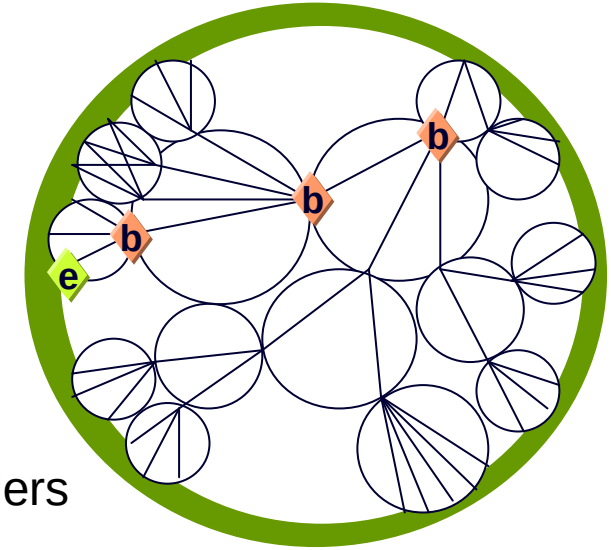


- virtual queue (a conceptual queue – actually a simple counter):
  - drained somewhat slower than the rate configured for adm ctrl of PCN traffic
  - therefore build up of virtual queue is ‘early warning’ that the amount of PCN traffic is getting close to the configured capacity
  - NB mean number of packets in real PCN queue is still very small

further work

- congestion control for hi-rate hi-acceleration flows
  - for stability, trend towards network rate control [XCP, RCP]
    - unlike TCP/IP's endpoint control
  - our research: congestion notification with higher precision per pkt
  - one packet immediately gives congestion state of path
- getting PCN & re-ECN standardised

summary



- optically-assisted packet routers
  - seem essential, esp. at inter-domain borders
- not just route look-ups and buffering
  - packet routers do many transport functions, esp at borders
- most transport functions could be entrusted to edge
  - pre-requisite #1: explicit congestion notification
    - need photonic ECN/PCN mechanism with a virtual queue
  - pre-requisite #2: proposed re-ECN field in IP header

more info

- **These slides <www.cs.ucl.ac.uk/staff/B.Briscoe/present.html#0709ecoc-fid>**
- **Explicit Congestion Notification (ECN) IETF RFC3168**
  - "*Layered Encapsulation of Congestion Notification*" IETF Internet-Draft <draft-briscoe-tsvwg-ecn-tunnel-00.txt> (Jun 2007)
  - "*Explicit Congestion Marking in MPLS*" IETF Internet-Draft <draft-ietf-tsvwg-ecn-mpls-01.txt> (Jun 2007)
- **IETF PCN working group documents <tools.ietf.org/wg/pcn/>** in particular:
  - *Pre-Congestion Notification Architecture*, Internet Draft <draft-ietf-pcn-architecture-00.txt> (Aug'07)
  - *Emulating Border Flow Policing using Re-ECN on Bulk Data,* Internet Draft <www.cs.ucl.ac.uk/staff/B.Briscoe/pubs.html#repcn> (Jun'07)
- **re-feedback project page <www.cs.ucl.ac.uk/staff/B.Briscoe/projects/refb/>**
  - Fixing mindset on fairness
    - Flow Rate Fairness: Dismantling a Religion ACM Computer Comms Rvw 37(2) 63-74 (Apr 07)
  - Overall re-feedback idea, intention, policing, QoS, load balancing etc
    - Policing Congestion Response in an Inter-Network Using Re-Feedback (SIGCOMM'05 – mechanism outdated)
  - re-ECN Protocol Spec and rationale
    - Re-ECN: Adding Accountability for Causing Congestion to TCP/IP IETF Internet Draft (Jul 2007)
  - Using re-ECN with pre-congestion notification (PCN)
    - Emulating Border Flow Policing using Re-ECN on Bulk Data IETF Internet draft (Jun 2006)
  - Mitigating DDoS with re-ECN
    - Using Self-interest to Prevent Malice; Fixing the Denial of Service Flaw of the Internet Workshop on the Economics of Securing the Information Infrastructure (Oct 2006)

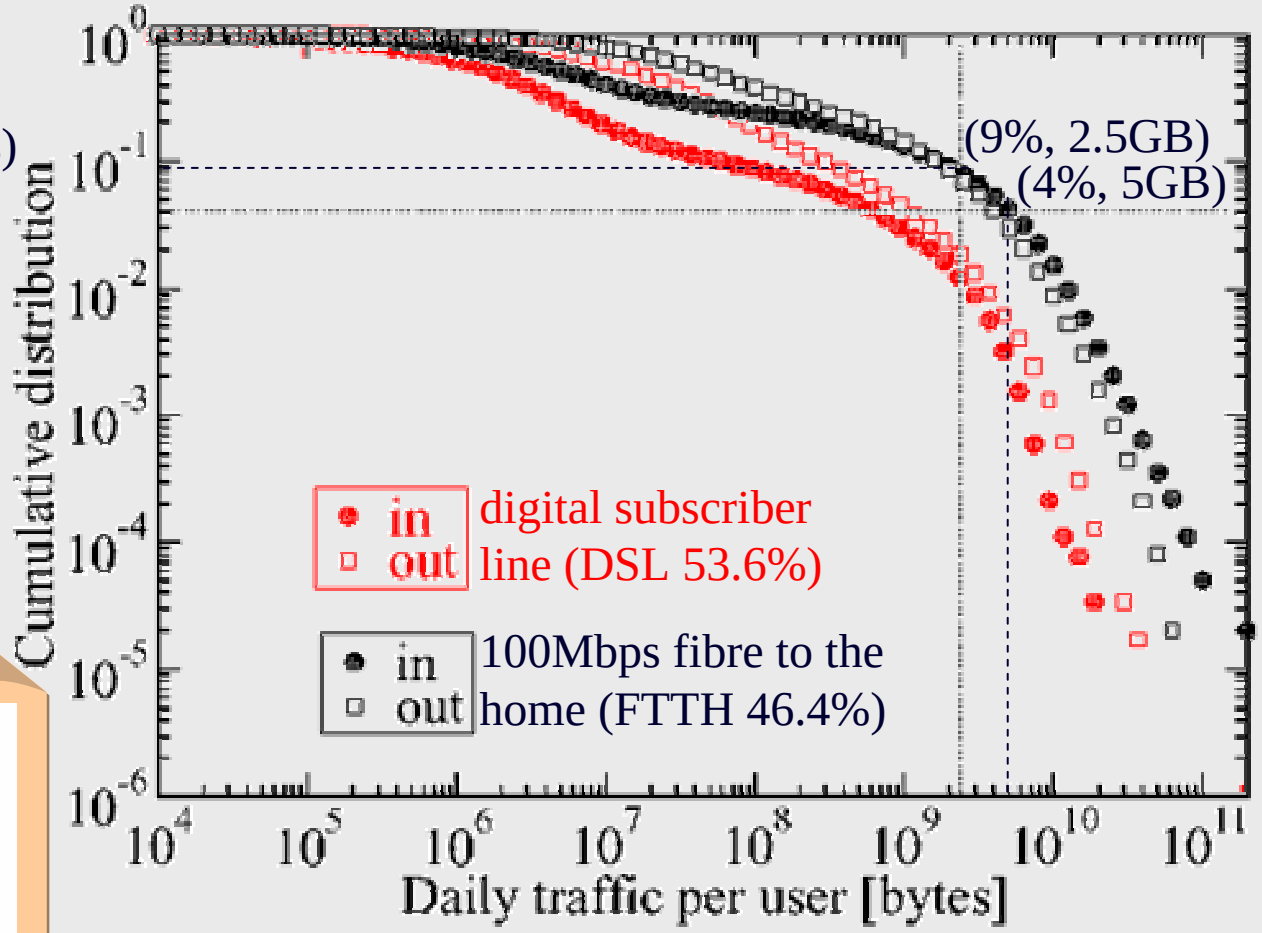
Pushing Packet Processing to the Edge

Q&A

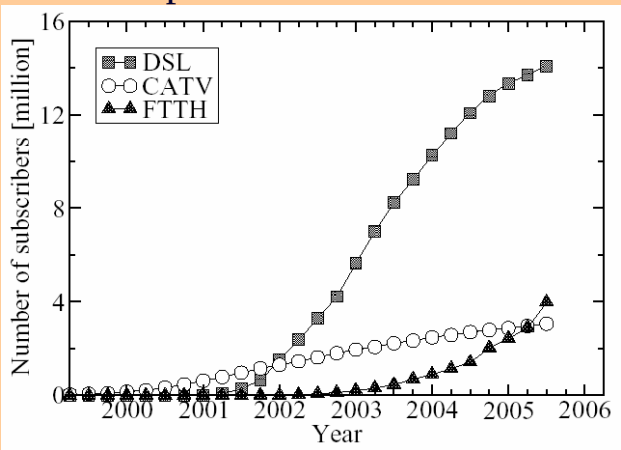
capacity growth will prevent congestion?

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

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



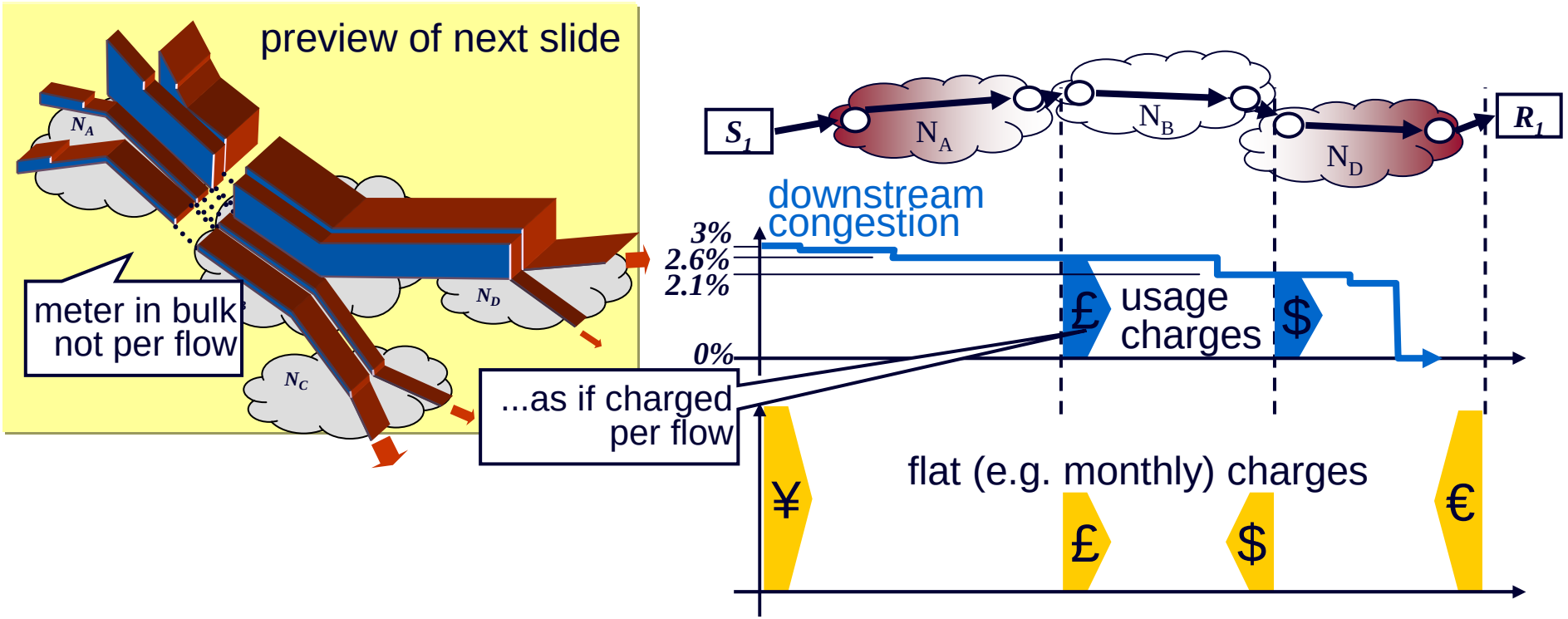
Changing technology shares of Japanese access market



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

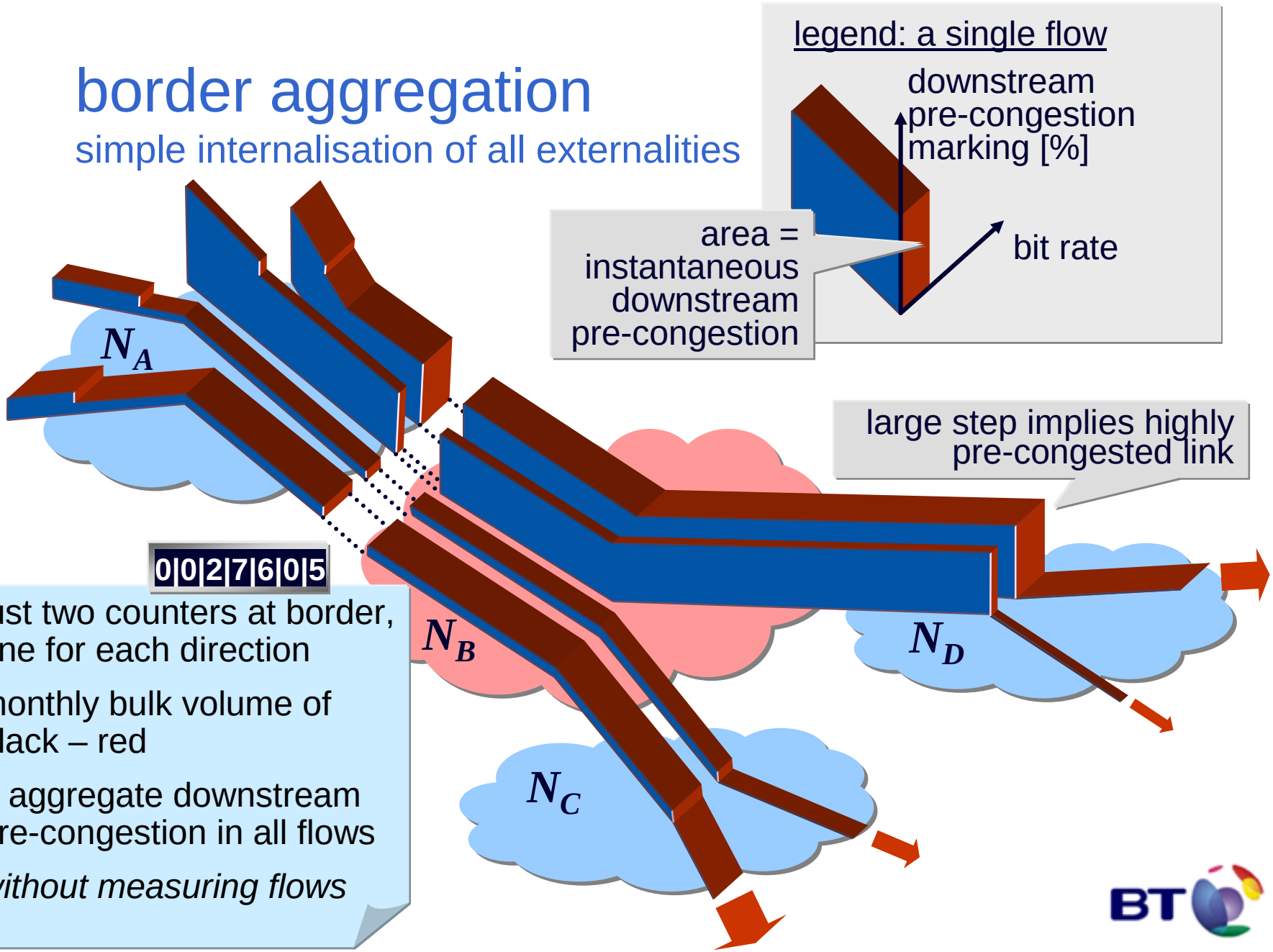
inter-domain accountability for congestion

- 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



just two counters at border, one for each direction

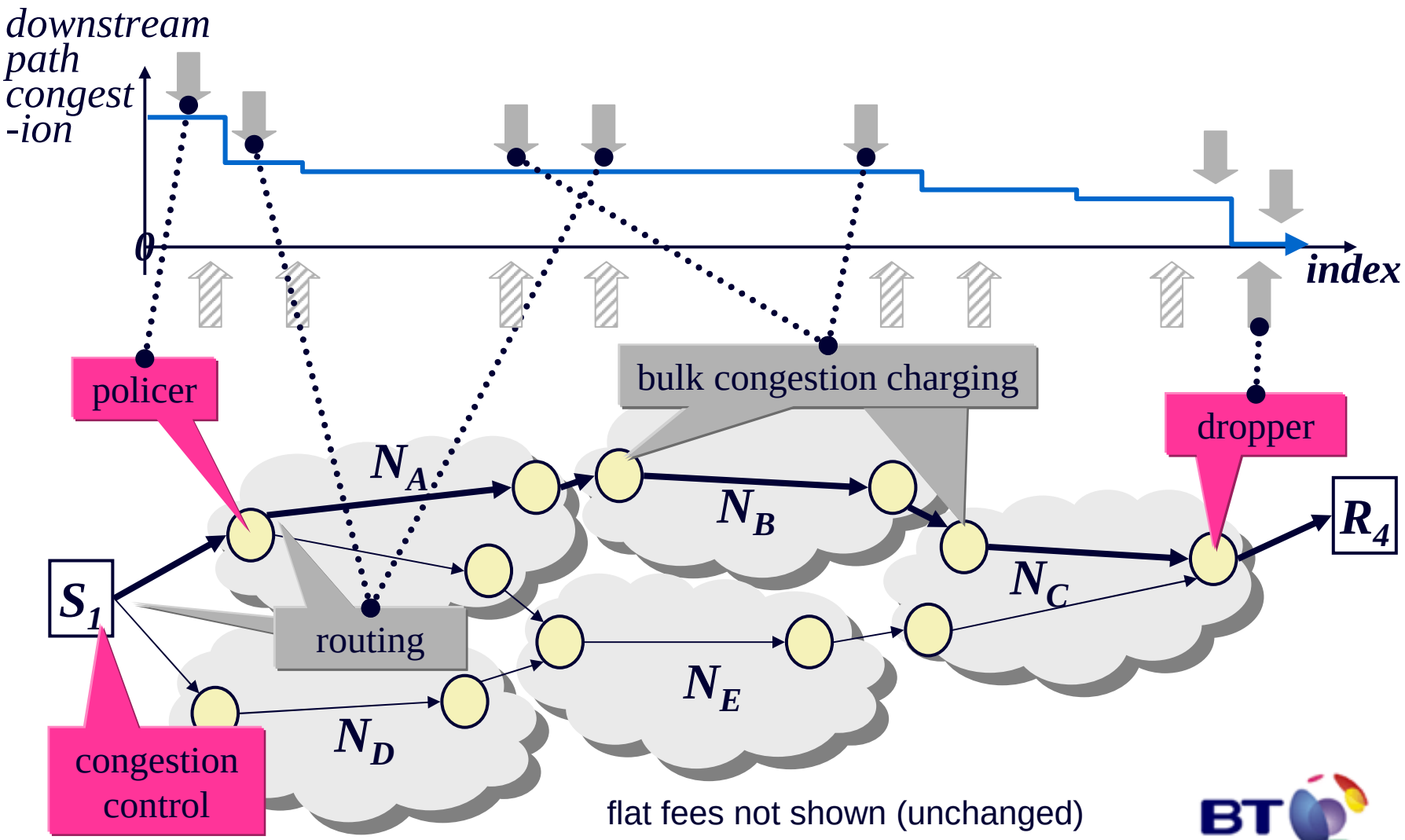
monthly bulk volume of black – red

= aggregate downstream pre-congestion in all flows

without measuring flows

re-feedback incentive framework

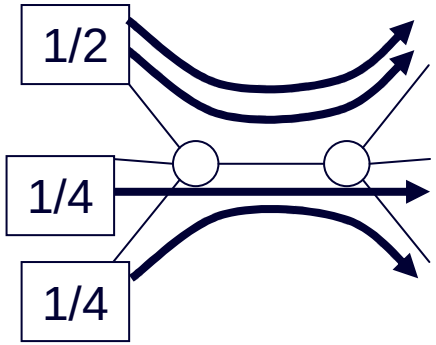
inline resource control functions only at edges of internetwork



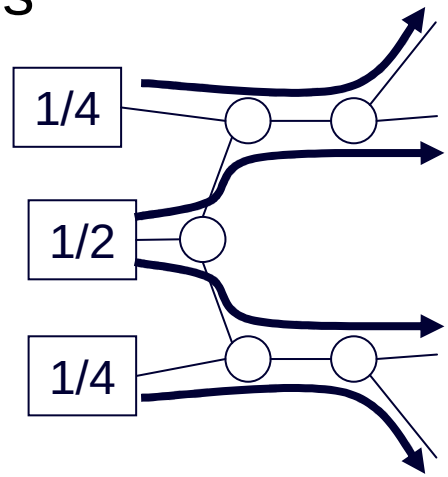
flat fees not shown (unchanged)

flow rate equality (TCP-fairness)

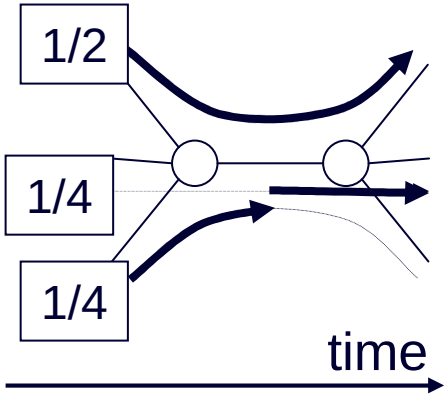
dismantling a religion



- doesn't even address relevant questions
  1) how many flows is it fair for an app to create?
  2) how fast should flows go through separate bottlenecks?
  3) how fast should a brief flow go compared to a longer lasting one?



- myopic
  - across flows, across network and across time



resource sharing

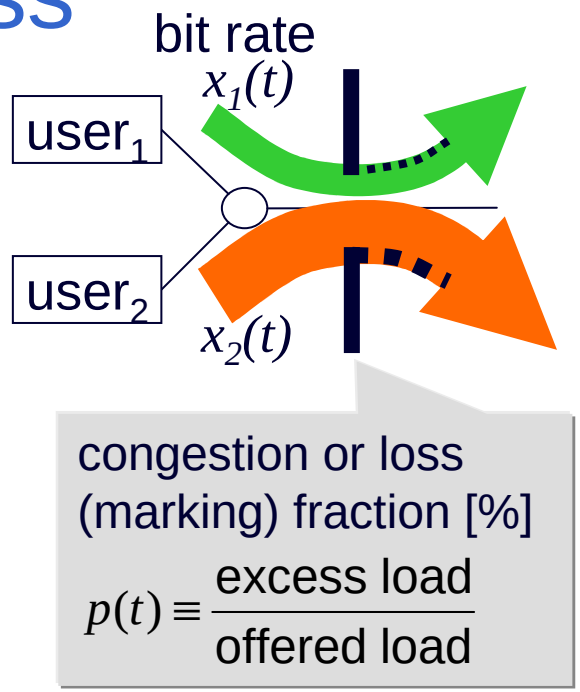
why network elements can’t arbitrate

- useful (ie competitive) resource sharing
  - requires very unequal flow rates
  - requires shares of capacity to depend on user history
- a queue may encounter nearly any user’s traffic
  - can’t be expected to hold history of everyone in the world
  - can’t be expected to synch with every other queue in the world

only alternative

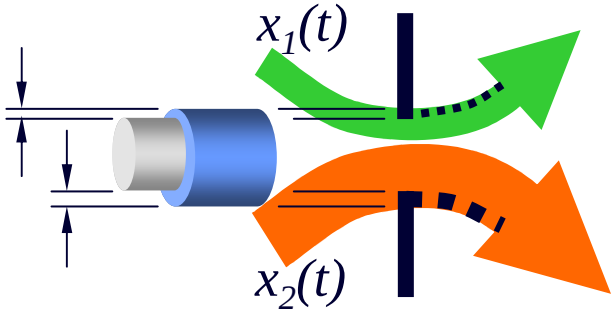
- edge-based control of shares of all queues on path
  - simple inline policing at first interface (electronic)
  - off-line metering at trust boundaries
  - only needs network elements to notify their congestion into traffic
  - fits with E-O-O-O-O-O-E vision

cost accountability / fairness



- cost of your behaviour on others
  - ☒ not your bit rate $x_i(t)$
- but bit rate weighted by the congestion when you sent it
  - ☑ loss (marking) fraction times your bit rate $p(t)x_i(t)$
- bytes you contributed to excess load
  - = your bytes that didn't get through (or didn't get through unmarked)
  - • termed congestion volume [bytes]
- accumulates simply and correctly
  - • across flows, across network paths and across time

calibrating ‘cost to other users’



note: diagram is conceptual

congestion volume would be accumulated over time

capital cost of equipment would be depreciated over time

- 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
- 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