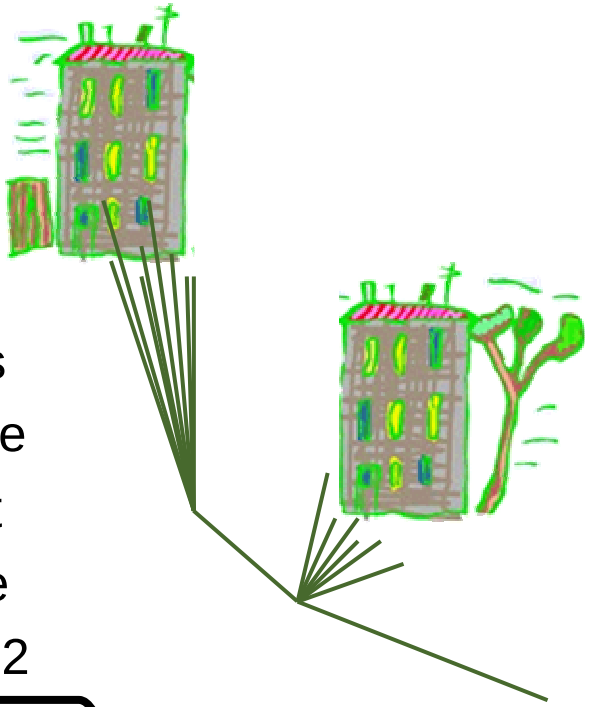
mending the Internet value chain...
…in one bit
Internet capacity sharing & QoS

Bob Briscoe
Chief Researcher, BT

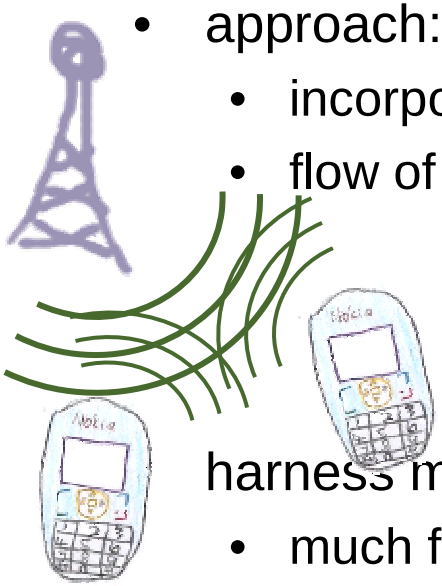
Oct 2009

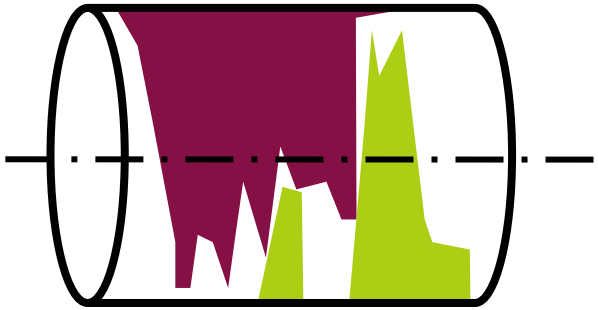
This work is partly funded by Trilogy, a research project supported by the European Community
www.trilogy-project.org

shared capacity



- shared access technology
  - PON, cable, cellular, WiFi, ...
  - huge gains from sorting out multiple access
  - currently in denial about the passage of time
- approach: sort out sharing the whole Internet
  - incorporate sharing access as part of whole
  - flow of info: L1 → L2 → L3 → L4 → L3 → L2





  - harness mutual flexibility
  - much faster when you really need it
- greater value, better quality of experience, simpler
- inability to prevent free-riding kills capacity investment [CFP06]

how to share the capacity of the Internet?

- the job of end-to-end L4 protocols (e.g. TCP)?
  - TCP's dynamic response to congestion is fine
  - but the way it shares capacity is very wrong
- ISP's homespun alternatives have silently overridden TCP
  - result: blocks, throttles & deep packet inspection
  - if it's new, it won't get through (if it's big, it won't either)

- IETF transport area consensus reversed since 2006
  - 'TCP-friendly' was useful, but not a way forward
  - rewrite of IETF capacity sharing architecture in process
  - commercial/policy review in process driven by 'captains of industry'

- approach: still pass info up to L4 to do capacity sharing
  - but using weighted variants of existing congestion controls (weighted TCP)
    - similar dynamics, different shares
  - give incentive for apps to set weights taking everyone into account
    - backed by enforcement – simple ingress policing

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

moving mountains IETF

glossary
IETF Internet Engineering Task Force
IESG Internet Engineering Steering Group
IAB Internet Architecture Board
IRTF Internet Research Task Force

- since 2006 IETF support for TCP capacity sharing has collapsed to zero
  - thought leaders agree TCP dynamics correct, but sharing goal wrong
    - many support our new direction – not universally – yet!
  - rewrite of IETF capacity sharing architecture in process
    - IETF delegated process to IRTF design team
- Oct'09
  - proposed IETF working group: "congestion exposure" (experimental)
  - IESG / IAB allowed agenda time, Hiroshima Nov'09
    - non-binding vote on working group formation
  - >40 offers of significant help in last few weeks; individuals from
    - Microsoft, Nokia, Cisco, Huawei, Alcatel-Lucent, NEC, Ericsson, NSN, Sandvine, Comcast, Verizon, …
- not a decision to change to IP – defer until support is much wider

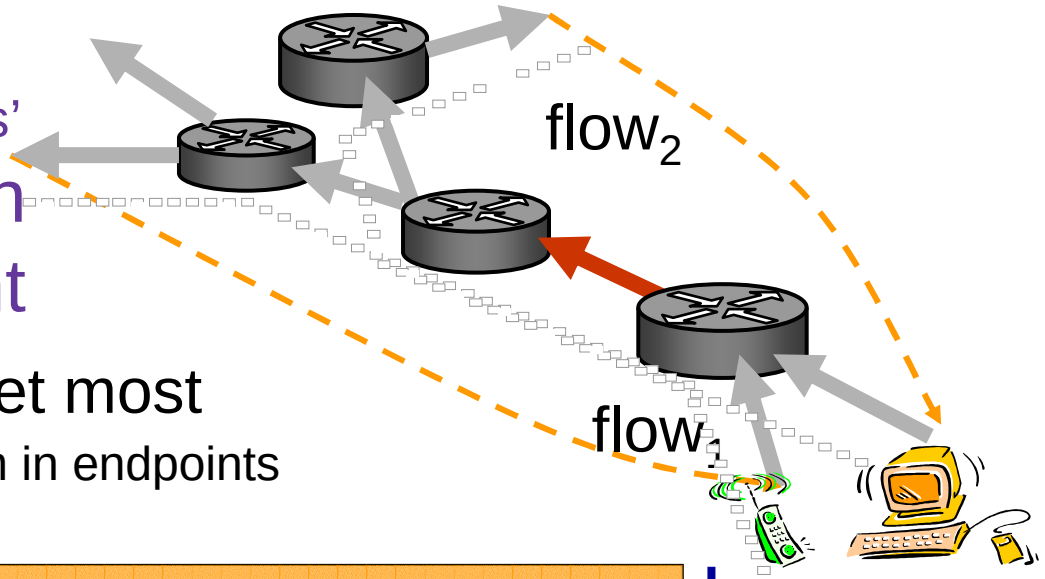
moving mountains ptII

the global ICT industry

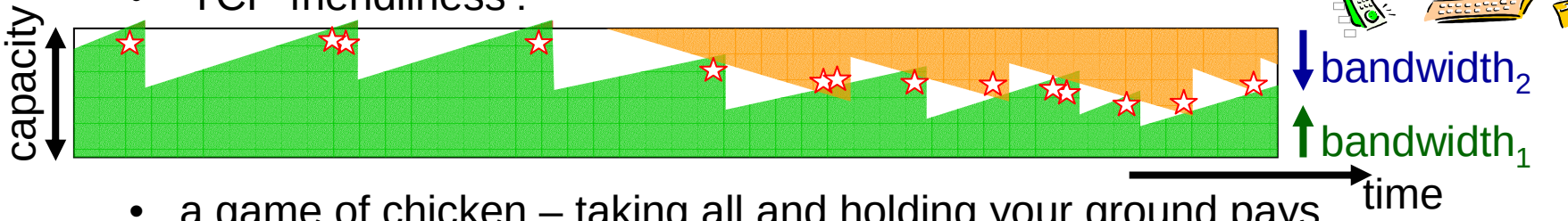
- GIIC: ~50 CxOs of the major global ICT corporations
  - Apr 09: then BT CTO (now Huawei Global CTO) proposed GIIC endorses BT solution
  - commissioners voted for endorsement decision within 30 days of expert review: public policy, commercial & technical
  - 30 Sep 09: favourable expert review in front of and by CxOs
    - all supported, but pointed out known obstacle (ie. ambitious)
  - if endorsed, becomes corporate lobbying position, standards position etc
- technical media coverage (Guardian, ZDnet, PCWorld, c't, …)
  - prompts near-universally reasonable reader postings
    - on broadband speed, quality, pricing, net neutrality!

how Internet sharing 'works'

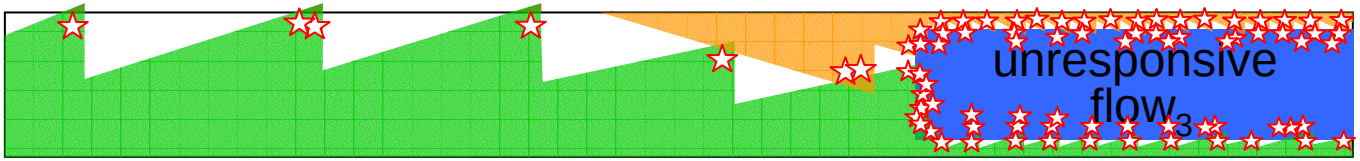
endemic congestion & voluntary restraint



- those who take most, get most
  - voluntarily polite algorithm in endpoints
  - 'TCP-friendliness':

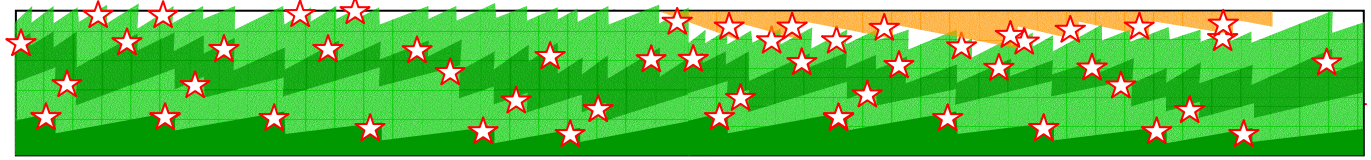


  - a game of chicken – taking all and holding your ground pays



(VoIP, VoD Joost 700kbps)

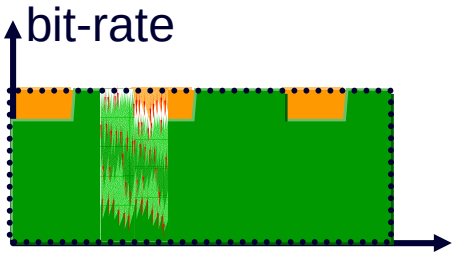
  - or start more 'TCP-friendly' flows than anyone else **(Web: x2, p2p: x5-100)**



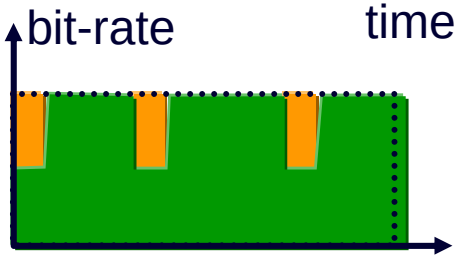
  - or for much longer than anyone else (file transfer x200)
  - net effect of both (p2p: x1,000-20,000 higher traffic intensity)

no traditional sharing approaches harness end-system flexibility… over time

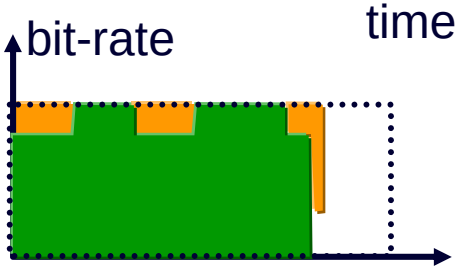
1. TCP



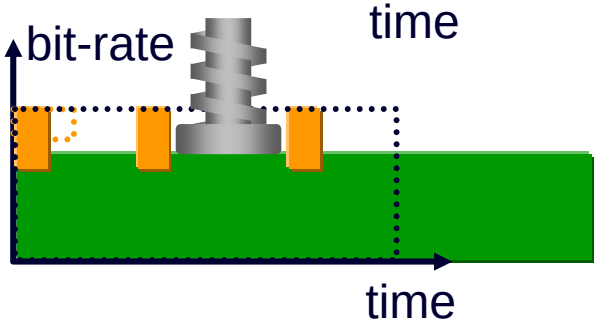
2. WFQ



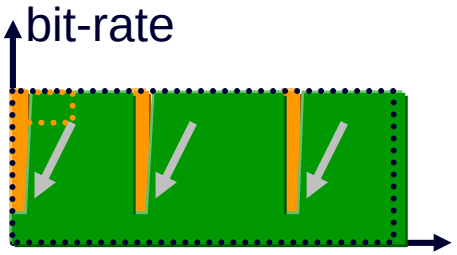
3. volume cap

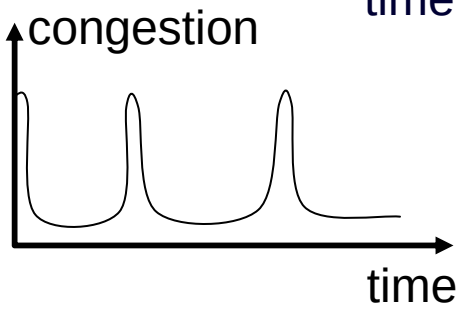


4. DPI



weighted sharing





- light usage can go much faster
- hardly affects completion time of heavy usage

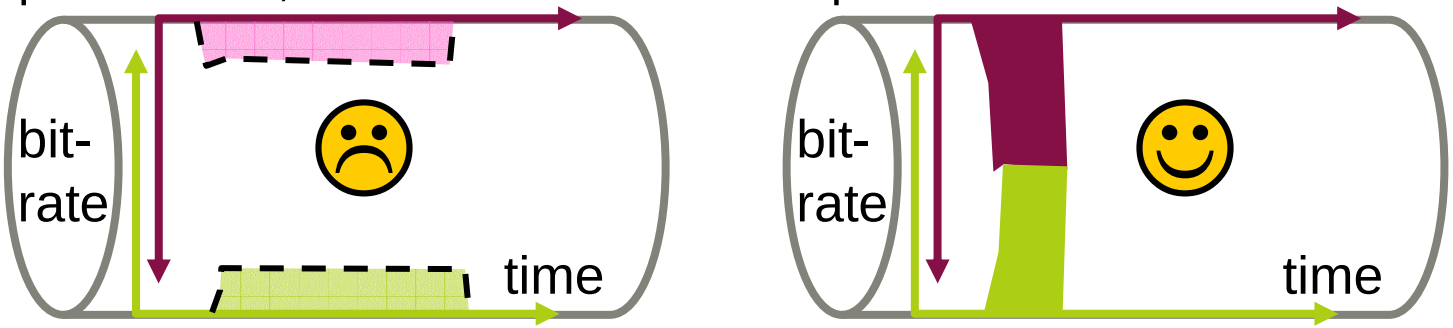
NOTE: weighted sharing doesn't imply differentiated network service

- just weighted aggressiveness of end-system's rate response to congestion cf. LEDBAT

congestion is not evil

congestion signals are healthy

- no congestion across whole path ⇒ feeble transport protocol
  - to complete ASAP, transfers should sense path bottleneck & fill it



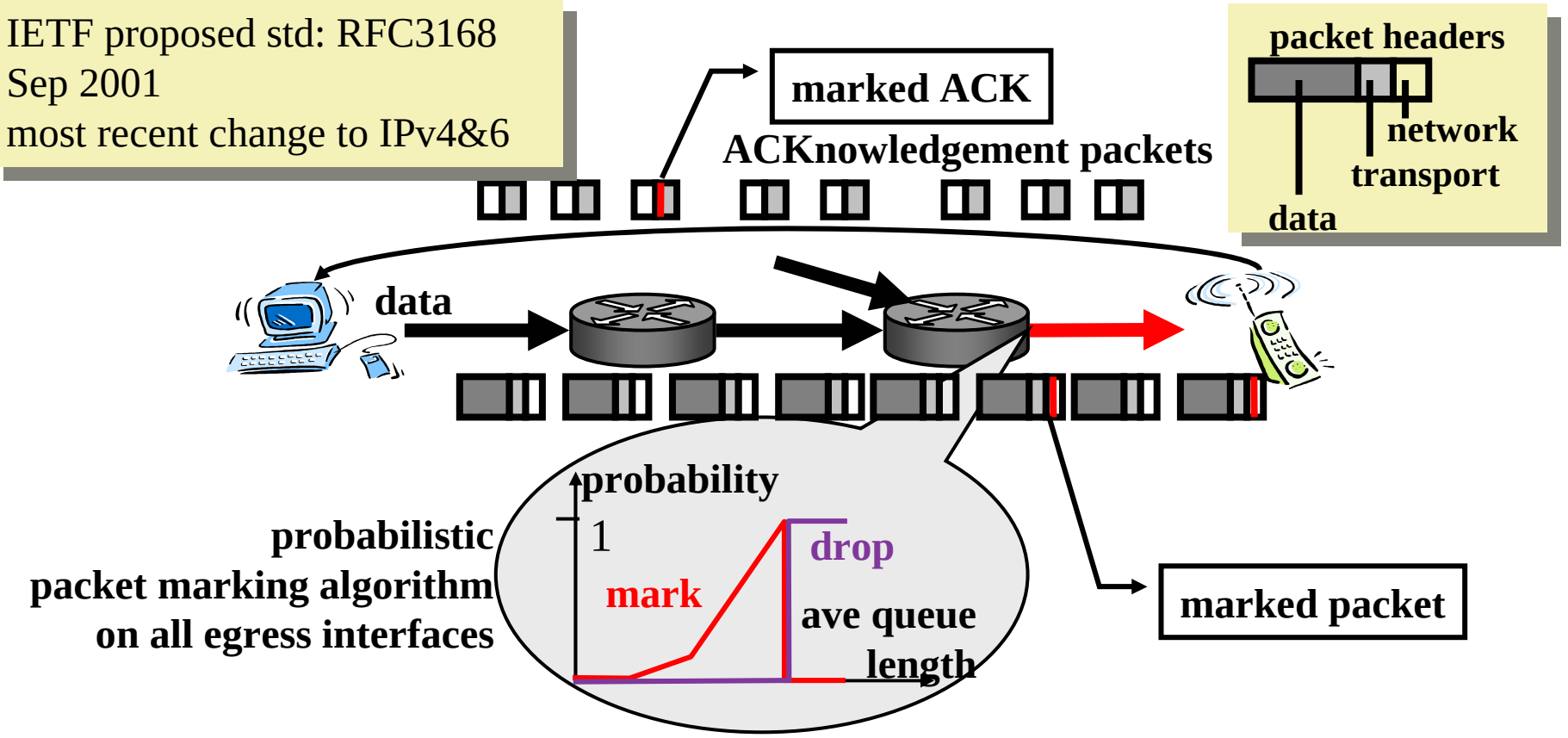
the trick

congestion signal *without* impairment

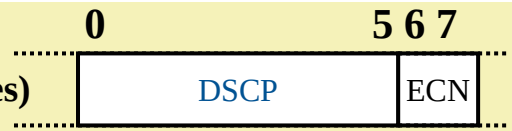
- explicit congestion notification (ECN)
  - update to IP in 2001: mark more packets as queue builds
- then tiny queuing delay and tiny tiny loss for all traffic
- no need to avoid congestion (whether core, access or borders) to prevent impairment

explicit congestion notification (ECN)

IETF proposed std: RFC3168
Sep 2001
most recent change to IPv4&6



00:	**Not ECN Capable Transport (ECT)**
01 or 10:	**ECN Capable Transport - no Congestion Experienced (sender initialises)**
11:	**ECN Capable Transport - and Congestion Experienced (CE)**



bits 6 & 7 of **IP** DS byte

powerful resource accountability metric
congestion-volume

- volume
  weighted by congestion when it was sent

- takes into account all three factors

congestion-volume	TCP	WFQ	Vol	DPI
• bit-rate	✓	✓	✓	✓	✓
• weighted by congestion	✓	~	~	✗	~
• activity over time	✓	✗	✗	✓	✓

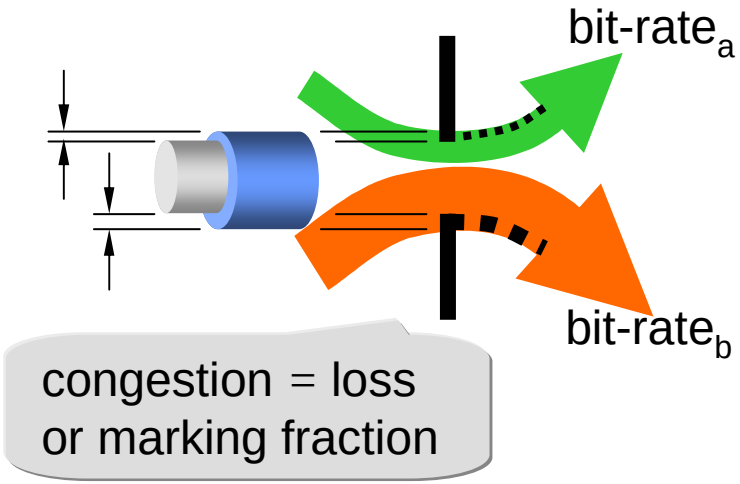
- how to measure
  - volume that is marked with explicit congestion notification (ECN)
  - can't be gamed by strategising machines

- a dual metric
  - of customers to ISPs (too much traffic)
  - and ISPs to customers (too little capacity)

a) cost to other users of your traffic

b) marginal cost of equipment upgrade
  - so it wouldn't have been congested
  - so traffic wouldn't have affected others

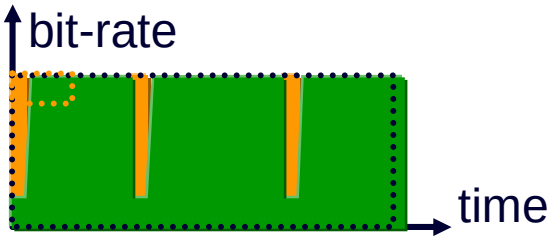
- competitive market matches a) & b)

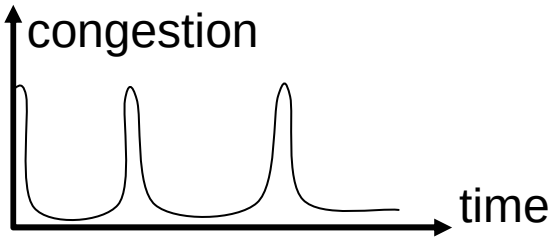


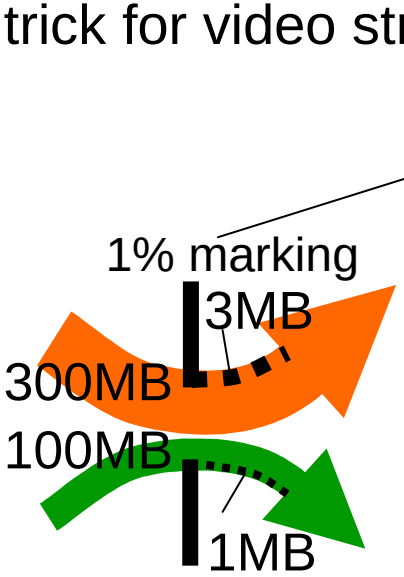
note: diagram is conceptual
congestion volume & capital cost of equipment would be accumulated over time

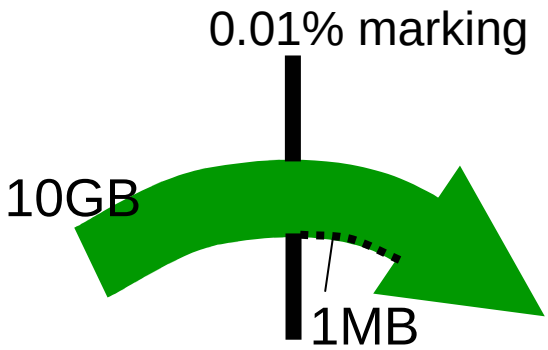
measuring marginal cost

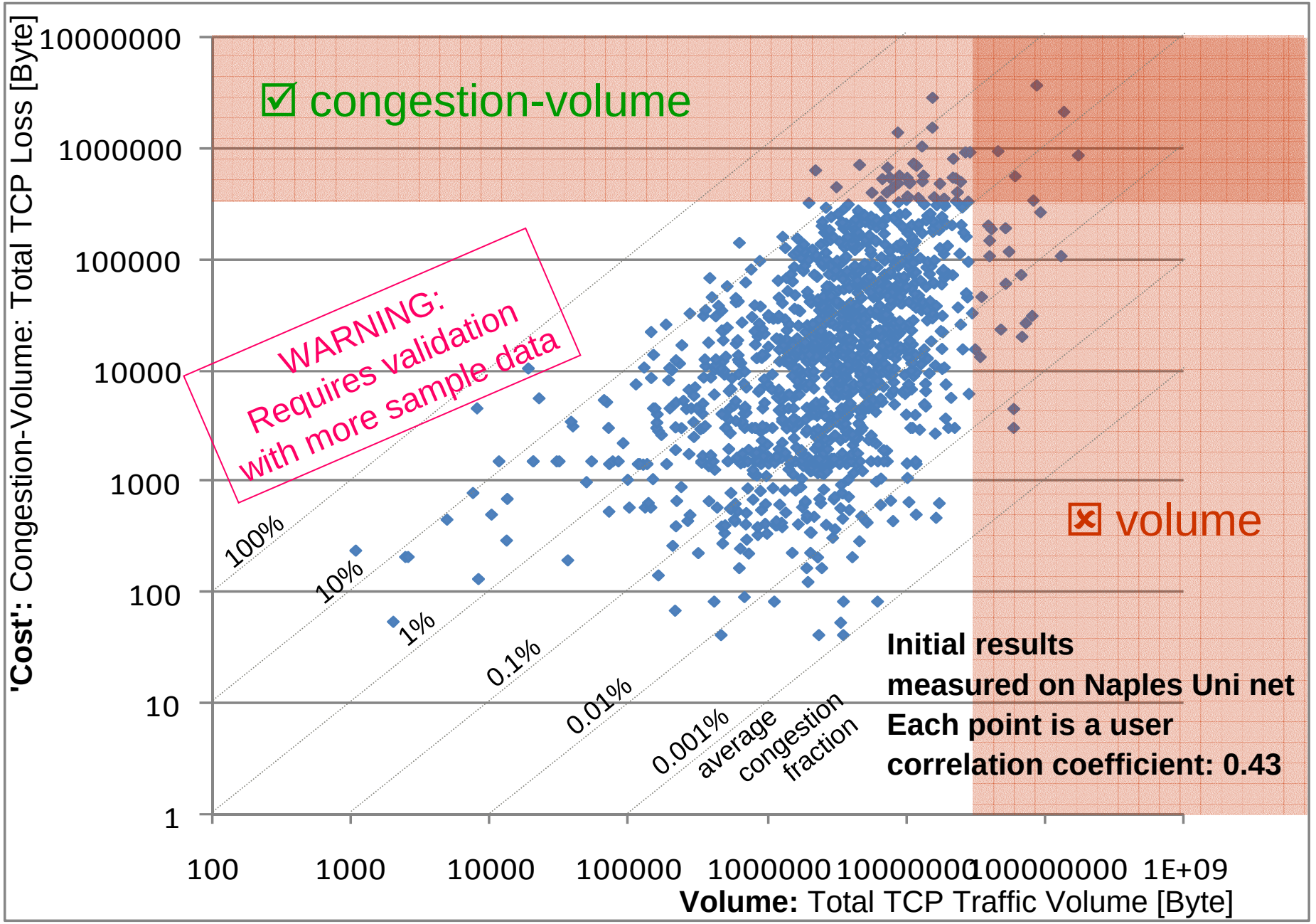
- user's contribution to congestion
  = bytes marked
- can transfer v high volume
  - but keep congestion-volume v low
  - similar trick for video streaming










'Cost': Congestion-Volume: Total TCP Loss [Byte]
10000000
1000000
100000
10000
1000
100
10
1
☑ congestion-volume
WARNING:
Requires validation
with more sample data
☒ volume
100%
10%
1%
0.1%
0.01%
0.001%
average
congestion
fraction
Initial results
measured on Naples Uni net
Each point is a user
correlation coefficient: 0.43
100
1000
10000
100000
1000000
10000000
100000000
1E+09
Volume: Total TCP Traffic Volume [Byte]

if only... ingress net could see congestion...

flat fee congestion policing

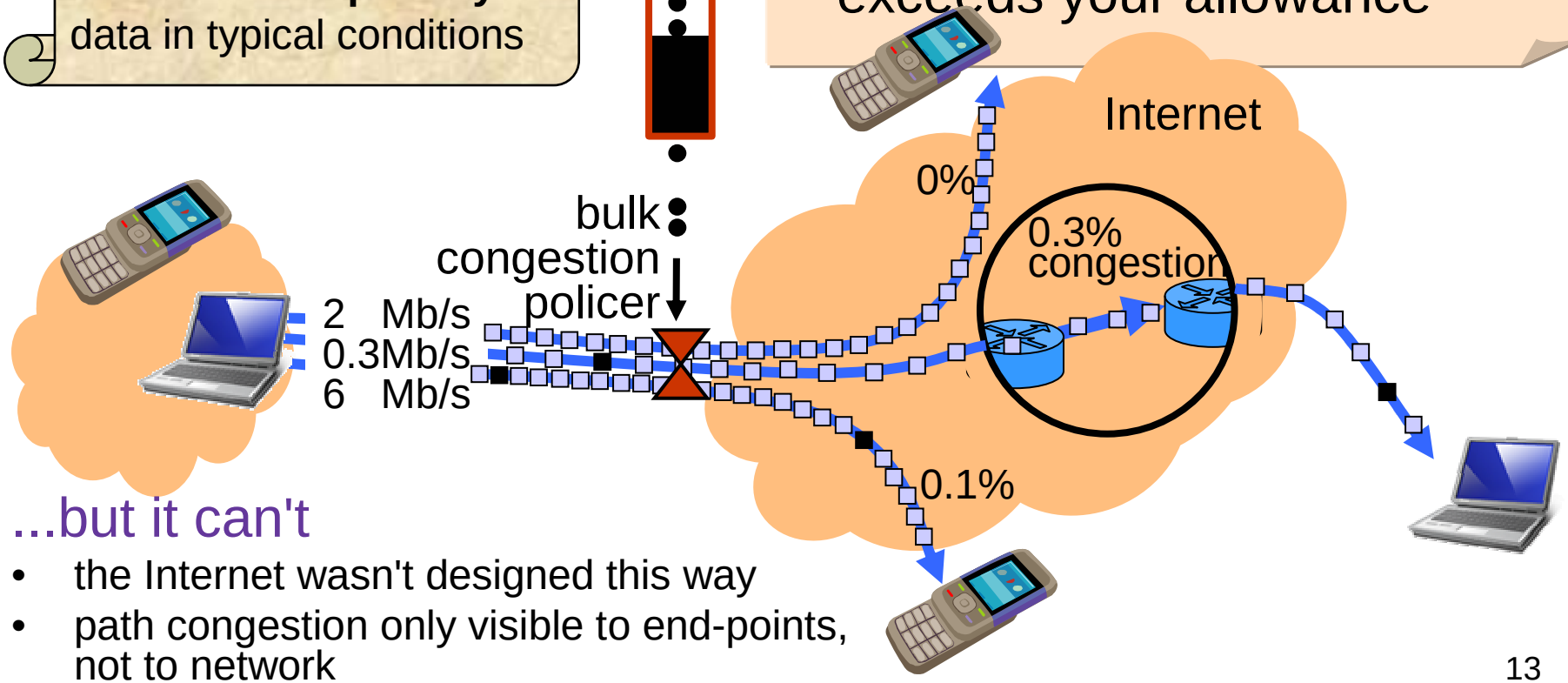
Acceptable Use Policy

'congestion-volume' allowance: 1GB/month

@ €15/month

Allows ~**70GB per day** of data in typical conditions

- incentive to avoid congestion
- only throttles traffic when your contribution to congestion in the cloud exceeds your allowance

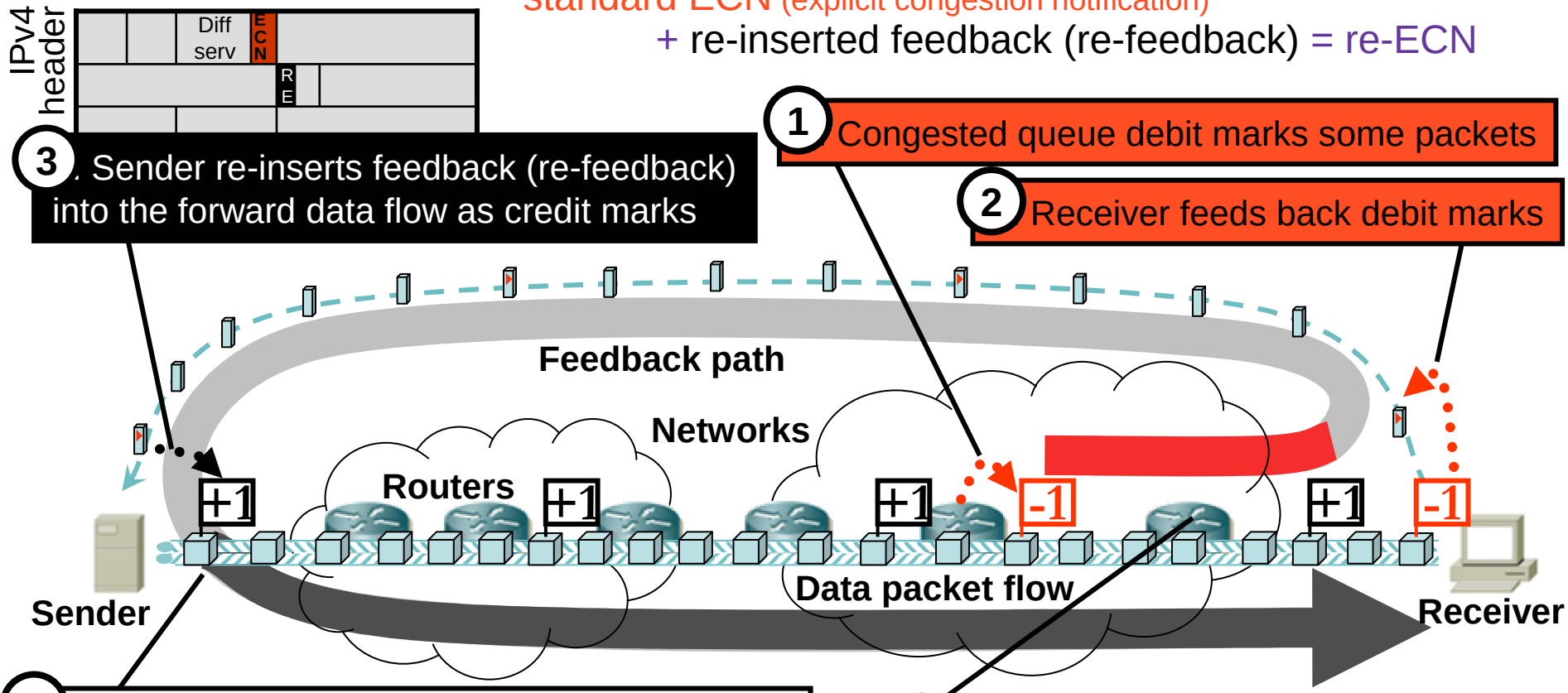


...but it can't

- the Internet wasn't designed this way
- path congestion only visible to end-points, not to network

congestion transparency in one bit

standard ECN (explicit congestion notification)
+ re-inserted feedback (re-feedback) = re-ECN



no changes required to IP data forwarding

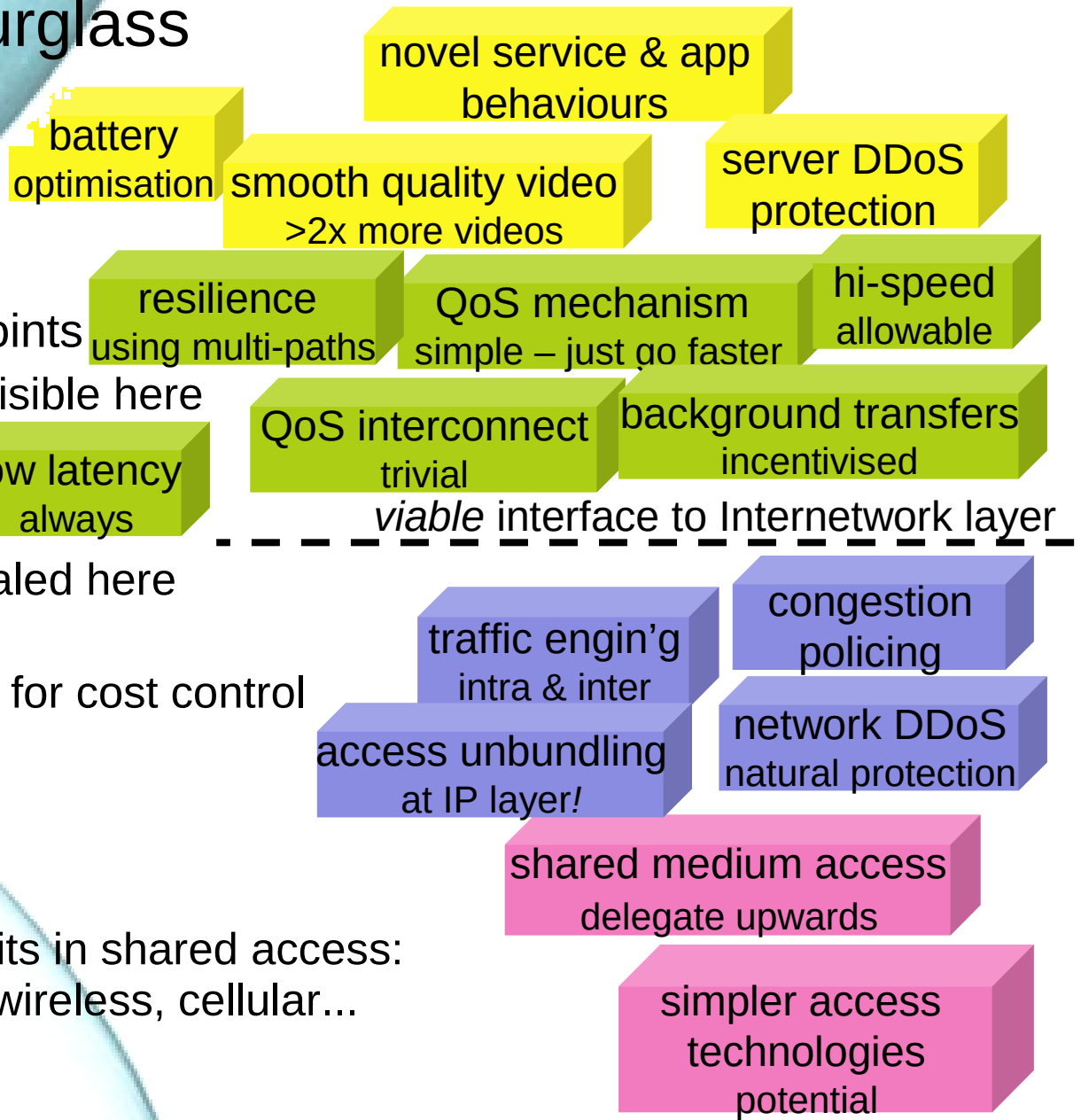
main steps to deploy re-feedback / re-ECN

summary
rather than control sharing in the access links,
pass congestion info & control upwards

- network
  - turn on explicit congestion notification in data forwarding
    - already standardised in IP & MPLS
    - standards required for meshed network technologies at layer 2 (ECN in IP sufficient for point to point links)
  - deploy simple active policing functions at customer interfaces around participating networks
  - passive metering functions at inter-domain borders
- terminal devices
  - (minor) addition to TCP/IP stack of sending device
  - or sender proxy in network
- then new phase of Internet evolution can start
  - customer contracts & interconnect contracts
  - endpoint applications and transports
- requires update to the IP standard (v4 & v6)
  - started process in Autumn 2005
  - using last available bit in IPv4 header or IPv6 extension header

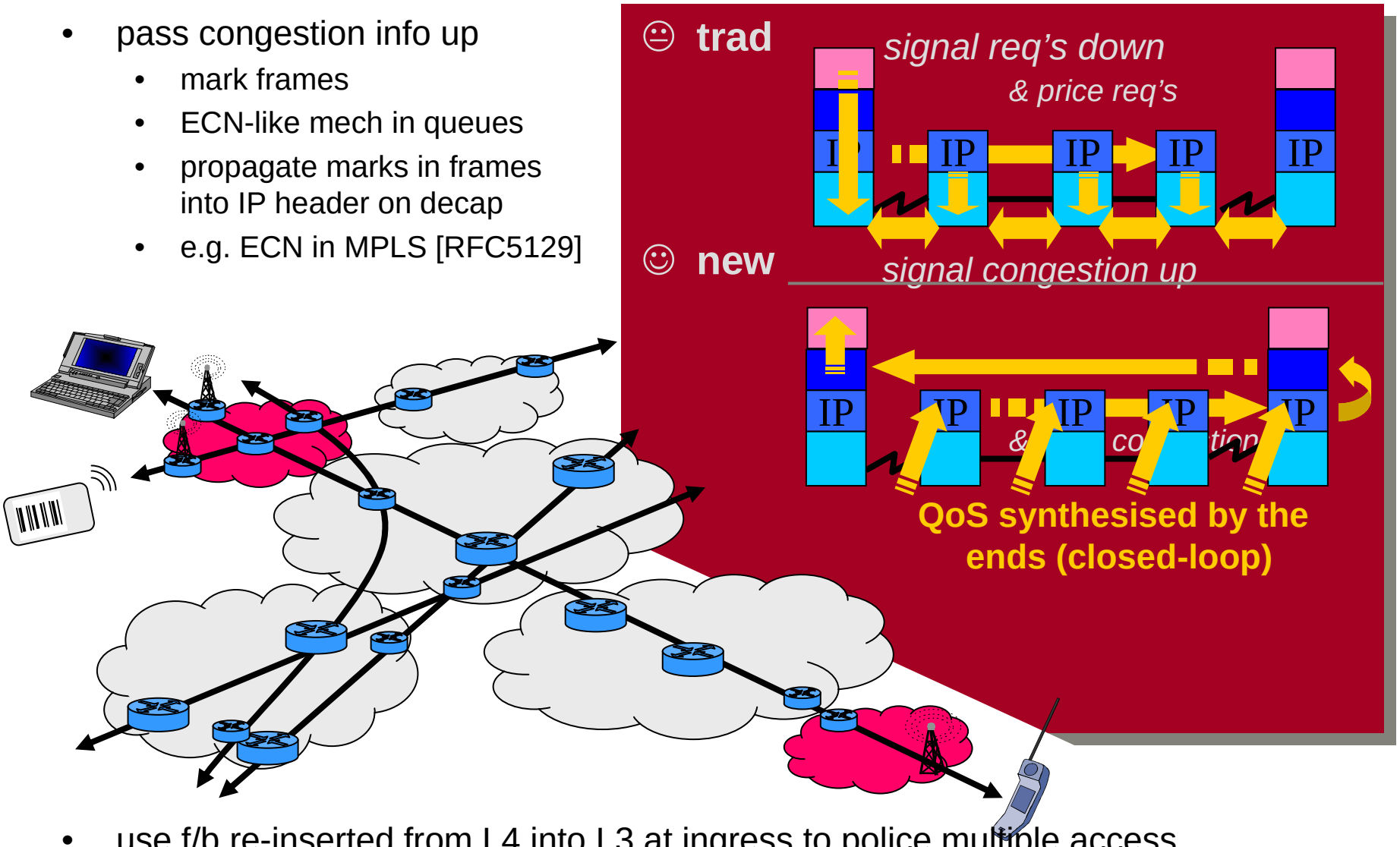
the neck of the hourglass ...but for control

- applications & services
- transport layer on end-points
  - usage costs currently visible here
- internetwork layer
  - once usage costs revealed here
  - ISPs won't need deep packet inspection for cost control
- link layer
  - can remove bit-rate limits in shared access: passive optical, cable, wireless, cellular...



message for layer 2

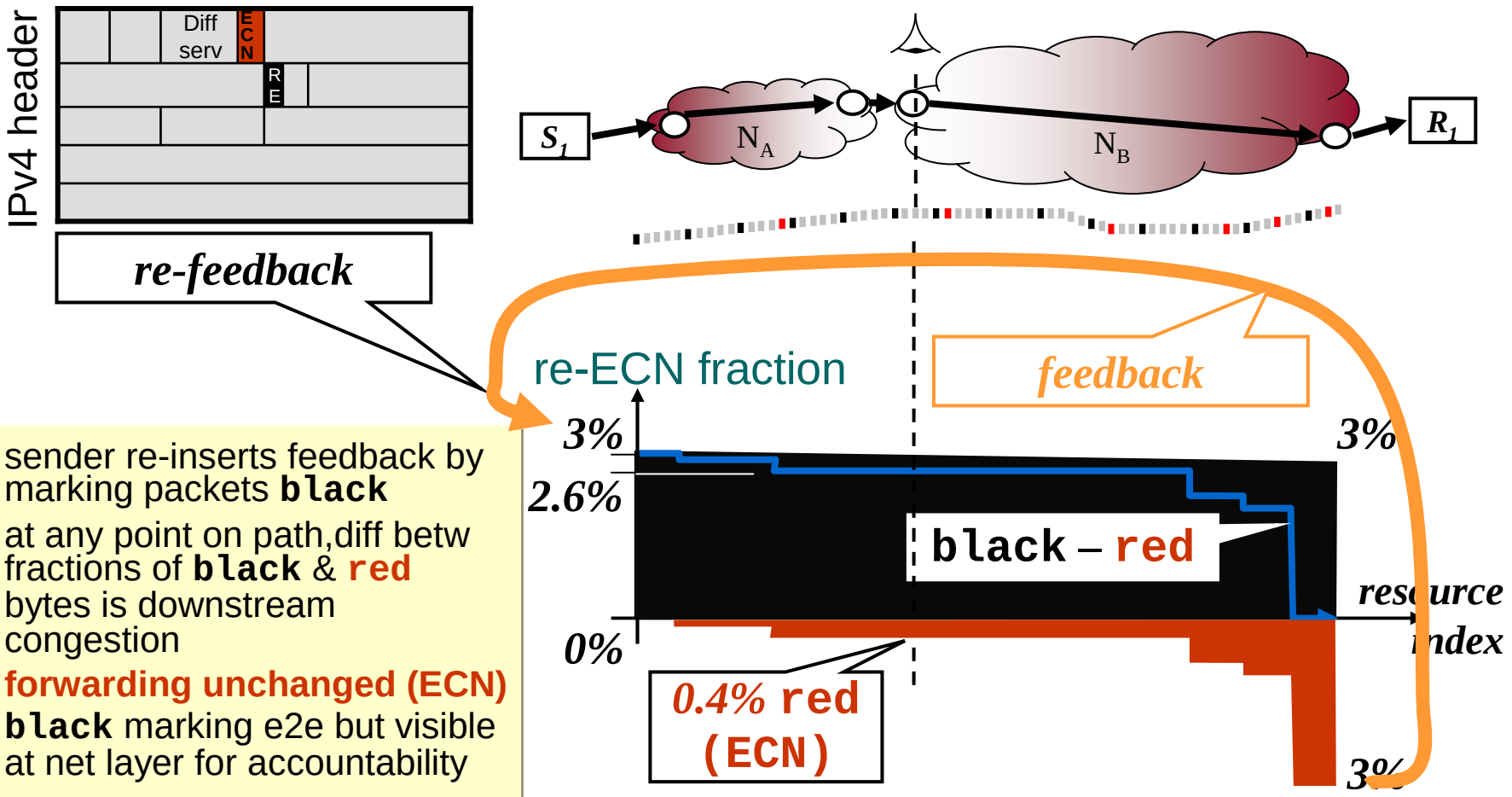
- pass congestion info up
  - mark frames
  - ECN-like mech in queues
  - propagate marks in frames into IP header on decap
  - e.g. ECN in MPLS [RFC5129]



- use f/b re-inserted from L4 into L3 at ingress to police multiple access

congestion exposure with ECN & re-ECN

measurable upstream, downstream and path congestion

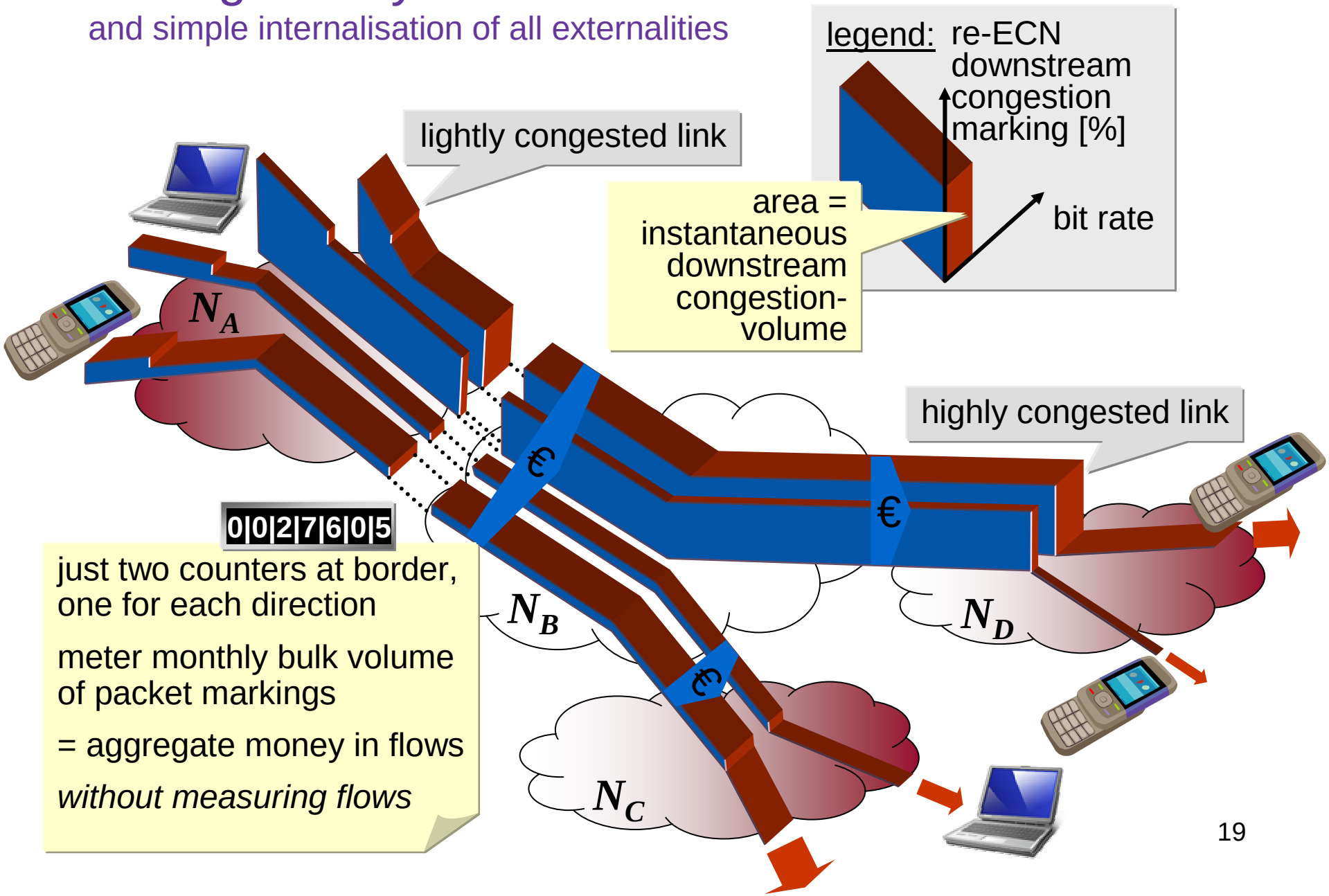


- sender re-inserts feedback by marking packets **black**
- at any point on path,diff betw fractions of **black** & **red** bytes is downstream congestion
- **forwarding unchanged (ECN)**
- **black** marking e2e but visible at net layer for accountability

BT

routing money

and simple internalisation of all externalities



just two counters at border, one for each direction

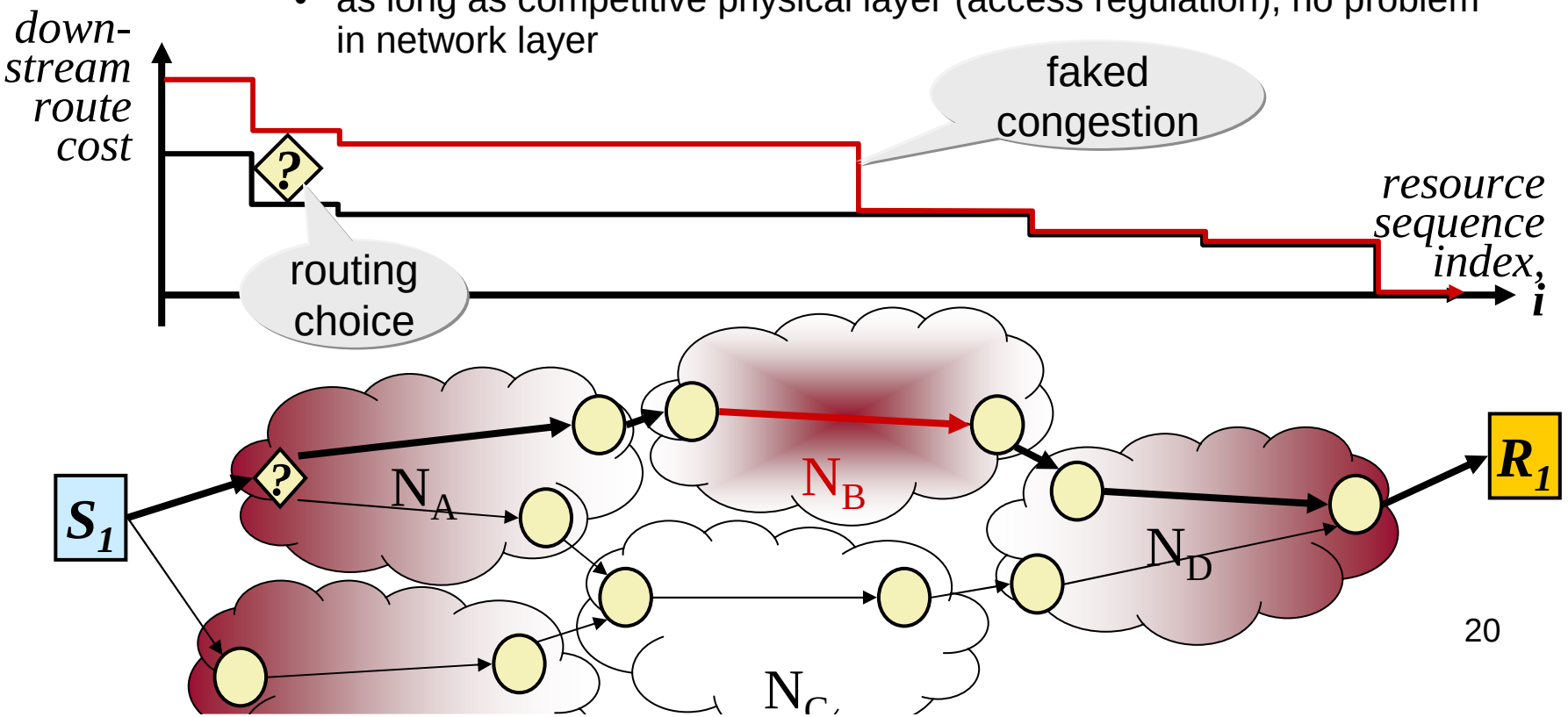
meter monthly bulk volume of packet markings

= aggregate money in flows

without measuring flows

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



best without effort

- did you notice the interconnected QoS mechanism?
  - *endpoints* ensure tiny queuing delay & loss for all traffic
  - if your app wants more bit-rate, it just goes faster
  - effects seen in bulk metric at every border (for SLAs, AUPs)
- simple – and all the right support for operations

summary
mending the Internet value chain

- the invisible hand of the market
  - favours ISPs that get their customers to manage their traffic in everyone else‘s best interests
- incentives to cooperate across Internet value chain
  - content industry, CDNs, app & OS authors, network wholesalers & retailers, Internet companies, end-customers, business, residential

more info...

- The whole story in 7 pages
  - Bob Briscoe, “Internet Fairer is Faster", BT White Paper (Jun 2009) ...this formed the basis of:
  - Bob Briscoe, "A Fairer, Faster Internet Protocol", IEEE Spectrum (Dec 2008)
- Slaying myths about fair sharing of capacity
  - [Briscoe07] Bob Briscoe, "Flow Rate Fairness: Dismantling a Religion" ACM Computer Communications Review 37(2) 63-74 (Apr 2007)
- How wrong Internet capacity sharing is and why it's causing an arms race
  - Bob Briscoe et al, "Problem Statement: Transport Protocols Don't Have To Do Fairness", IETF Internet Draft (Jul 2008)
- re-ECN protocol spec
  - Bob Briscoe et al, “Adding Accountability for Causing Congestion to TCP/IP” IETF Internet Draft (Mar 2009)
- Re-architecting the Internet:
  - The Trilogy project <www.trilogy-project.org>

IRTF Internet Capacity Sharing Architecture design team
<http://trac.tools.ietf.org/group/irtf/trac/wiki/CapacitySharingArch>

re-ECN & re-feedback project page:
<http://bobbriscoe.net/projects/refb/>

Congestion Exposure (ConEx) IETF ‘BoF’: <http://trac.tools.ietf.org/area/tsv/trac/wiki/re-ECN>
subscribe: <https://www.ietf.org/mailman/listinfo/re-ecn>, post: re-ecn@ietf.org

implementation (linux or ns2) bob.briscoe@bt.com

Internet capacity sharing for packets not flows

discuss...

bringing information to the control point

- flat fee policer is just one example...
- huge space for business & technical innovation at the control point
  - cost based, value-cost based
  - bulk, per flow, per session
  - call admission control
  - policing, charging
  - tiers, continuous
  - wholesale, retail
- truly converged architecture
  - can apply different industry cultures
  - through policies at the control point
  - not embedded in each technology