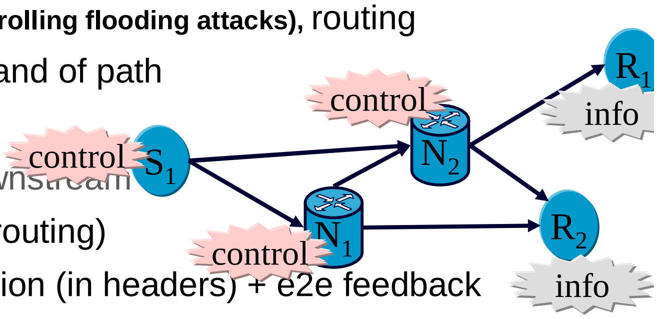
shared control of networks using re-feedback

Bob Briscoe
Arnaud Jacquet, Andrea Soppera,
Carla Di Cairano-Gilfedder &
Martin Koyabe
BT Research, Oct 2004

intro | incentives | apps | deployment | discussion

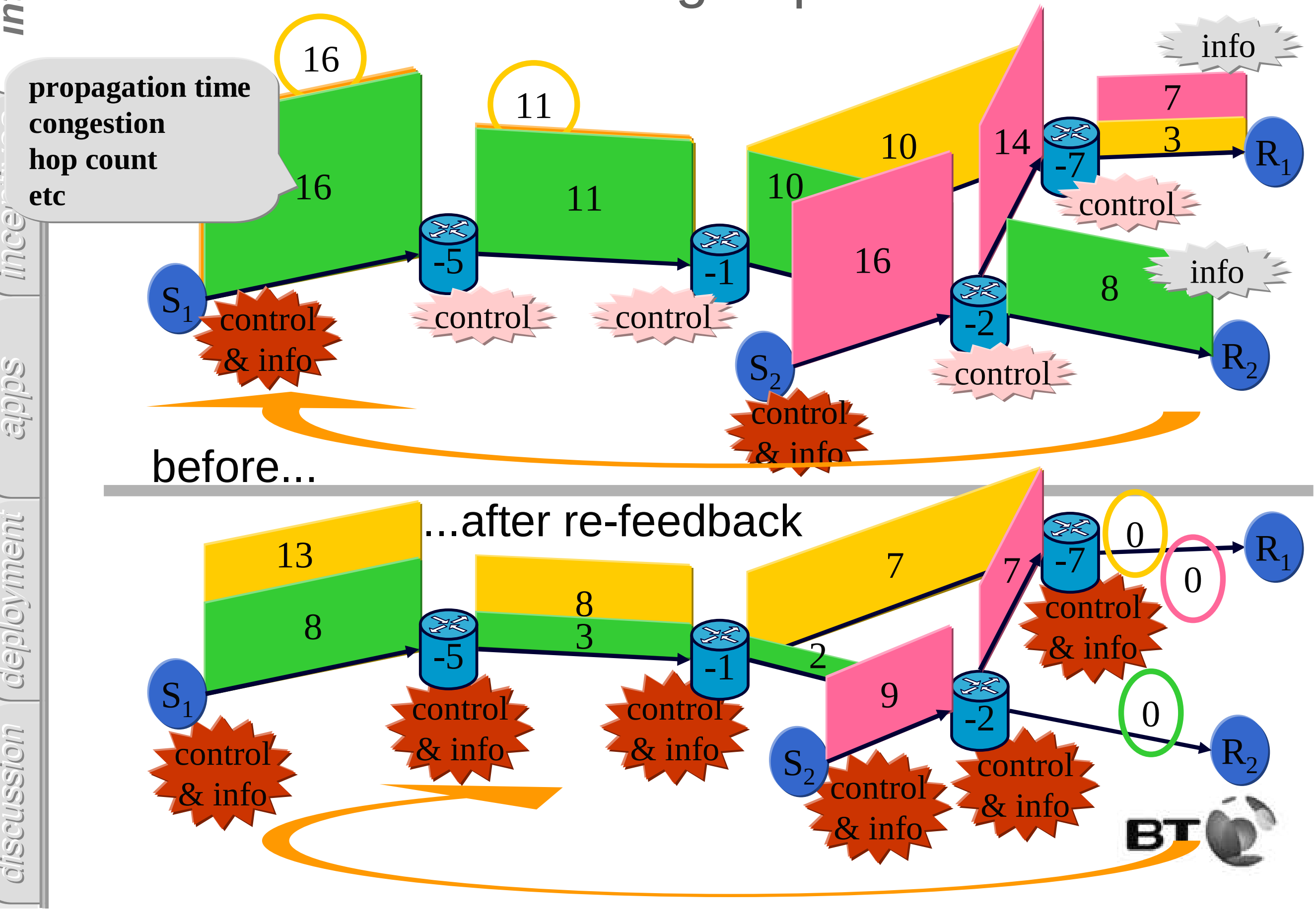
the problem

- context: packet networks
  - focus on Internet **(alternatively sensor nets, p2p, optical packet)**
- path characterisation underlies basics of networking:
  - resource allocation **(incl. controlling flooding attacks)**, routing
  - control: upstream of each link and of path
    - loading, routing
  - information: collected from downstream
    - explicit reverse messages (routing)
    - explicit or implicit accumulation (in headers) + e2e feedback



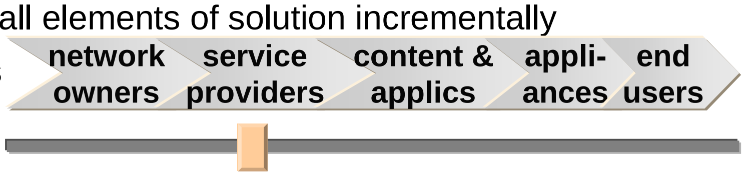
- current architecture embeds who controls what
  - routers route, sources control congestion
  - absolute control corrupts – need to temper or even reverse

downstream knowledge upstream: the idea



contributions

- arrange honesty & responsibility to be dominant strategies
  - even for first packets of a flow
  - without tampering with retail pricing
- downstream information upstream
  - updated within round trip
  - enhance, never reduce, info usefulness to each party
  - overload existing path characterisation data headers (e.g. TTL, ECN)
  - incentives to deploy all elements of solution incrementally
  - no change to routers



- control architecture
  - re-feedback designed for tussle over who controls what
  - Q. who controls the slider? A. socio-economic (market, regulation)
  - sufficient to police others, or to take full control (proxy)

contributions: applications

- congestion control/QoS
  - rate (e.g. TCP) policing
  - differentiated service synthesised from diff. congestion response
  - guaranteed QoS synthesised from path congestion-based AC
  - inter-domain traffic policing emulated by bulk metering
  - incentivise ‘slow-enough-start’
  - first line of defence against flooding
- routing
  - advert validation
  - traffic engineering
  - capability-based routing
- not exhaustive
  - re-feedback intended as enabler

approach

- part of effort to determine new Internet architecture
- determine target, then work out path from legacy

- distributed resource control
- based on network economics
  - recommend mechanism for non-co-operative end-game
    - asymptotic: in practice, some domains may stick before end-game
    - must have mechanisms for end-game in case we arrive there
  - dynamic pricing often used to align incentives (as in previous work)
    - re-feedback saves having to tamper with retail pricing
- work in progress

justifying the approach
– a game is being played out

- retail/end-user
  - flat charging
  - p2p file-sharing
  - usage charging, capping
  - …
  - differentiated QoS
  - policing fairness

time

- wholesale/interconnect
  - flat charging & path length-based BGP
  - CDNs
  - capacity & usage charging
  - peak demand charging
  - smart multipath routing
  - …
  - congestion charging
  - fast smart multipath routing

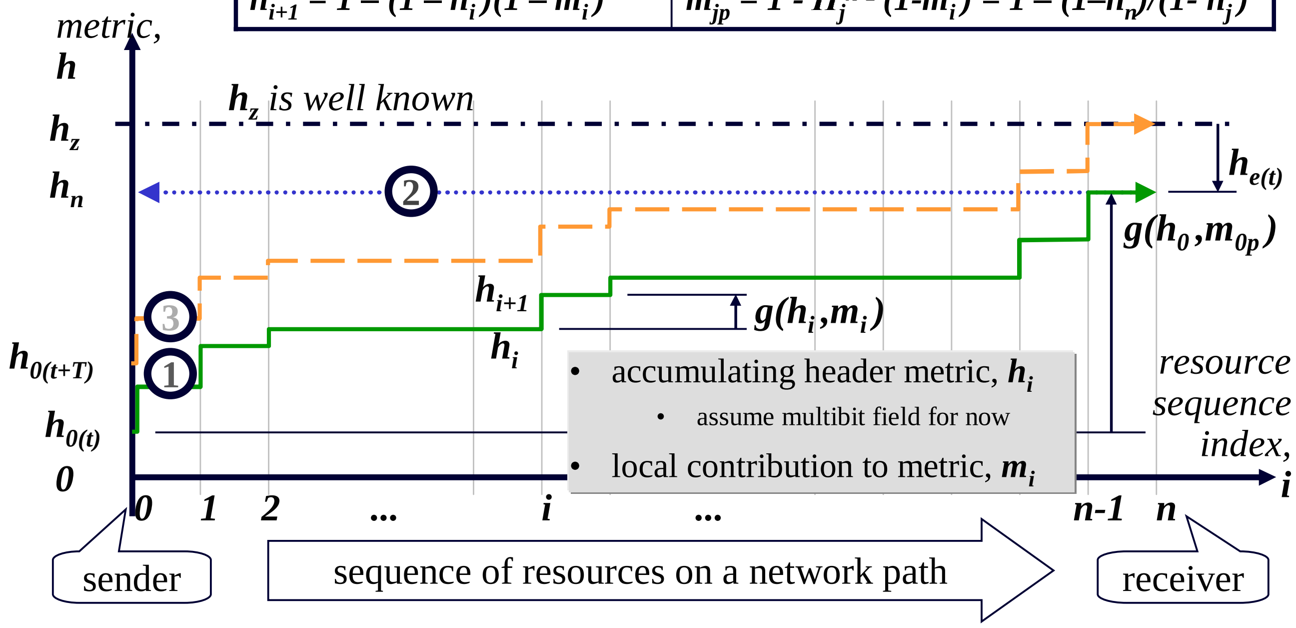
generalised re-feedback

accumulation funcn, $f(h_i,m_i)$	path metric downstream of j (incl. itself)
$h_{i+1} = h_i + m_i$	$m_{jp} = \Sigma_j^{n-1} m_i \quad = h_n - h_j$
$h_{i+1} = h_i - m_i$	$m_{jp} = \Sigma_j^{n-1} m_i \quad = h_j - h_n$
$h_{i+1} = 1 - (1 - h_i)(1 - m_i)$	$m_{jp} = 1 - \Pi_j^{n-1} (1-m_i) = 1 - (1-h_n)/(1-h_j)$

e.g. TTL

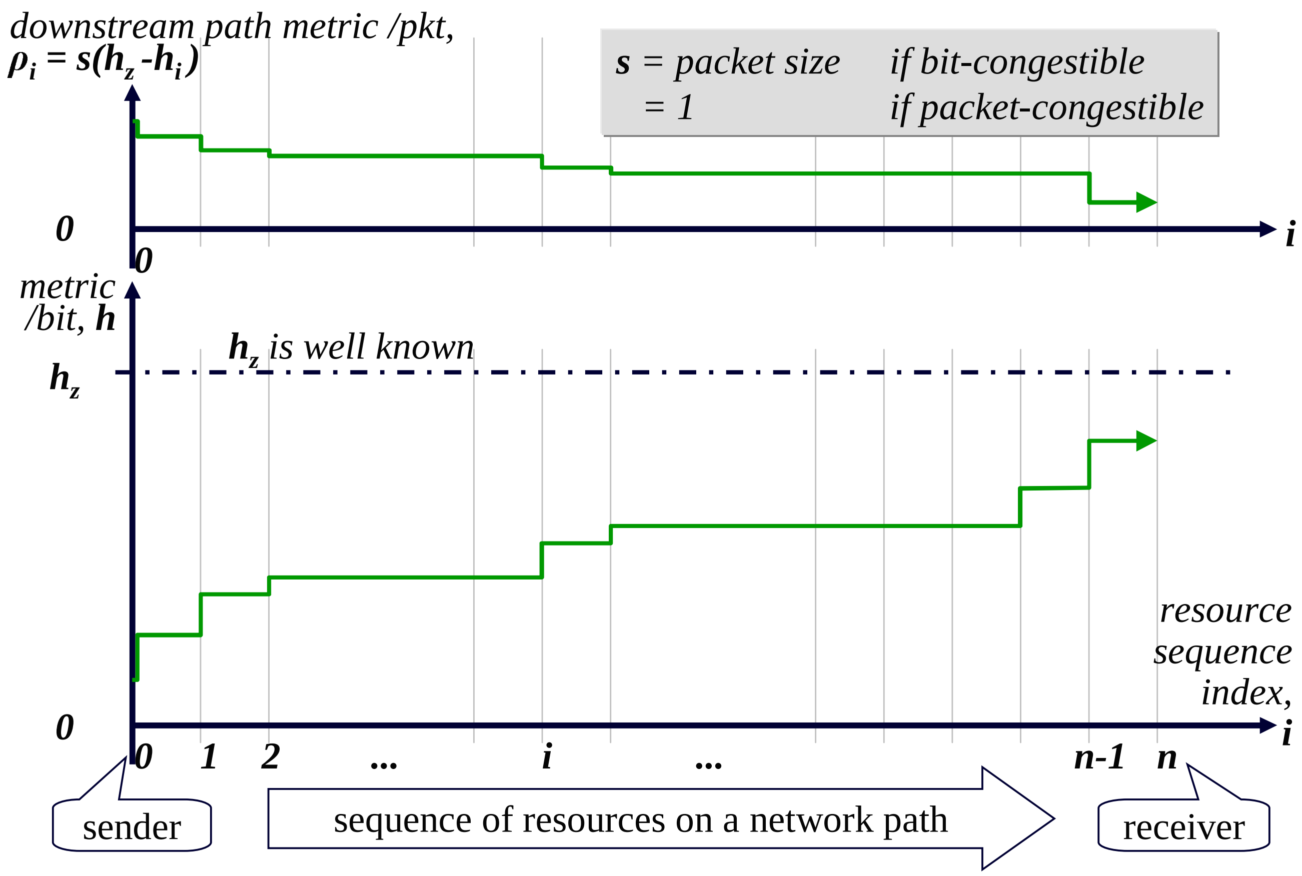
e.g. congestion

because $h_n \rightarrow h_z$ known locally



- accumulating header metric, h_i
  - assume multibit field for now
- local contribution to metric, m_i

normalised re-feedback



congestion protocol terms

- focus on congestion
  - to be concrete
  - for incentives discussion
- $\rho_i = s(h_z - h_i)$ becomes downstream path shadow price (DPSP)
- ECN = Explicit Congestion Notification
- ECL = Explicit Congestion Level
- 're-' = receiver aligned (or re-inserted)

aligned at	binary	multi-bit
sender	ECN	ECL
receiver	re-ECN	re-ECL

- also assume a binary 'certain' flag in packet headers
  - set by sender once received sufficient feedback to set intial metric(s)

definitions

1. The change in congestion, $\Delta E(X_i=1)$, caused by a packet at single resource i is the increase in expectation that the event X_i will occur, if the packet in question is added to the load, given any pre-existing differential treatment of packets.

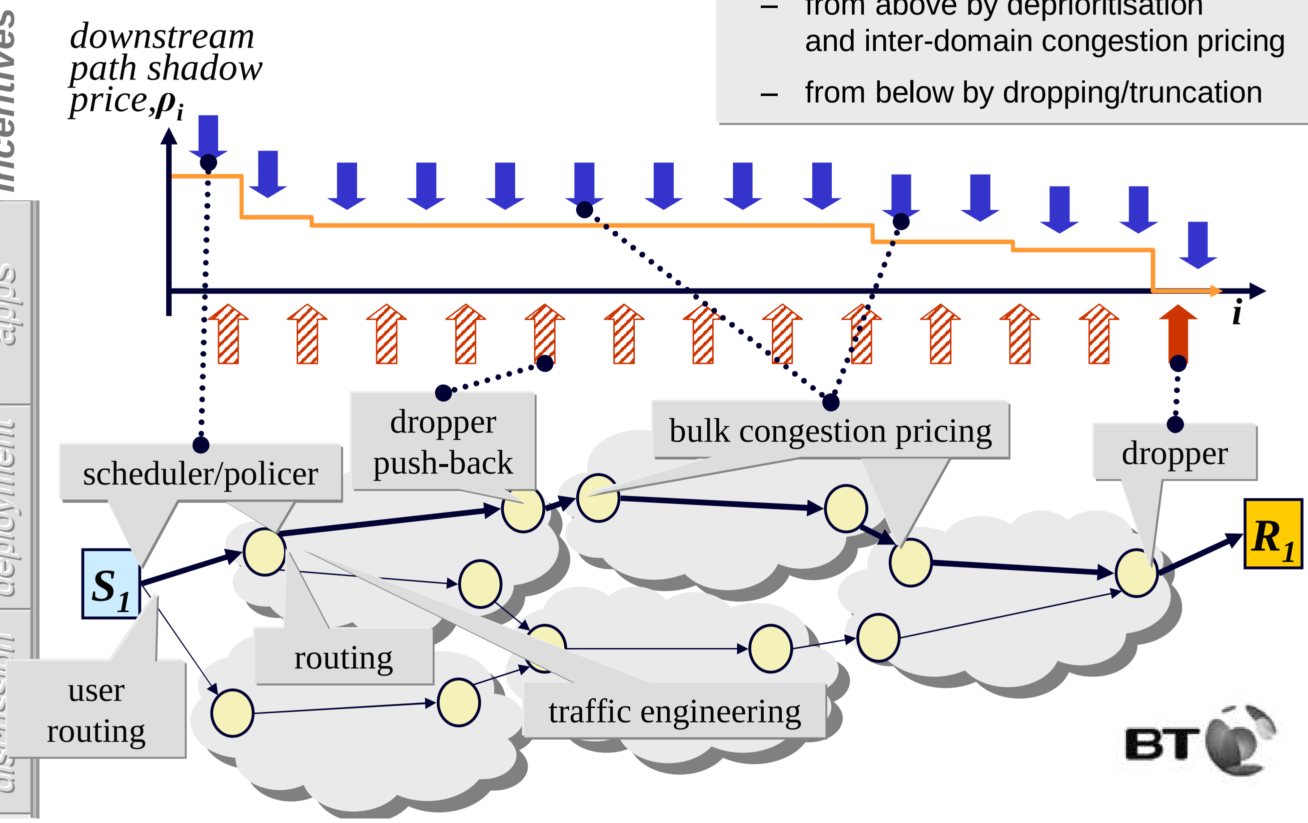
   Where X_i is the event that any packet will not be served to its requirements by resource *i*.
2. The change in path congestion level, $\Delta E(X=1)$, caused by a packet traversing the path is the increase in expectation that the event X will occur if the packet in question is added to the load traversing the entire path, given any pre-existing differential treatment of packets.

   Where X is the event that any packet sharing any resource along the sequence of resources used by the packet in question will not be served to its requirements.

incentive architecture

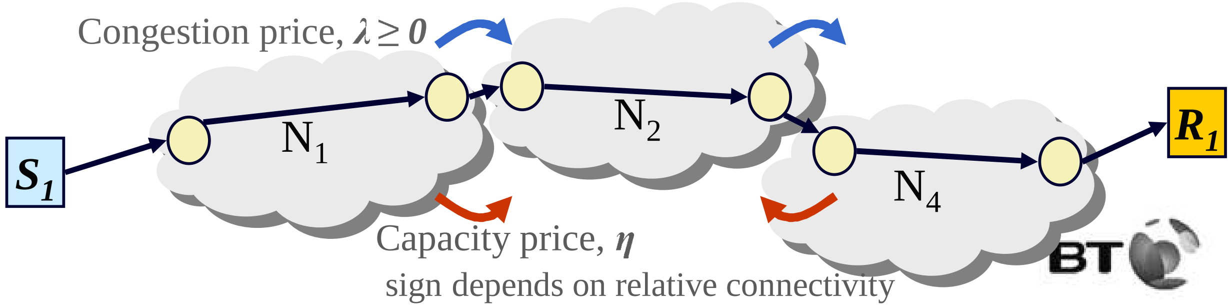
constrain to target at destination

- from above by deprioritisation and inter-domain congestion pricing
- from below by dropping/truncation



inter-domain pricing

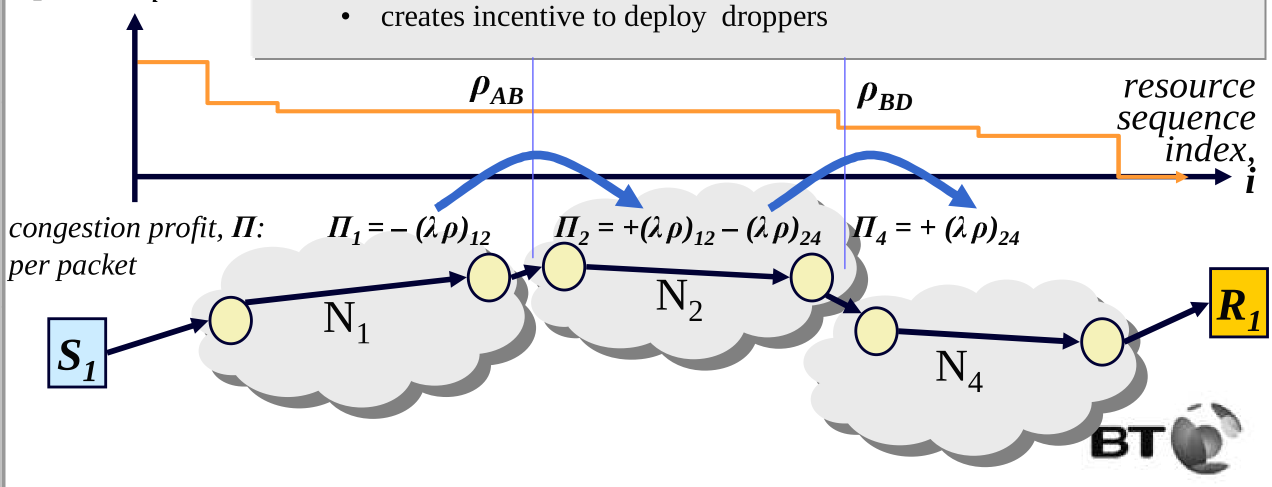
- inter-domain congestion pricing: incentive compatible
  - **emulates border policing but passive & extremely simple**
- sufficient under perfect competition, but …
- …in practice charge by capacity and modulate with congestion
- sending domain pays $\boldsymbol{C = \eta X + \lambda Q}$ to receiving domain (e.g. monthly)
- $\boldsymbol{\eta, \lambda}$ are (relatively) fixed prices of capacity, $\boldsymbol{X}$ and congestion, $\boldsymbol{Q}$ resp.
  - at each interface, separate prices agreed for ingress & egress
  - usage related price $\boldsymbol{\lambda \geq 0}$ (safe against ‘denial of funds’)
  - any receiver contribution to usage settled through end to end clearinghouse



congestion pricing - inter-domain

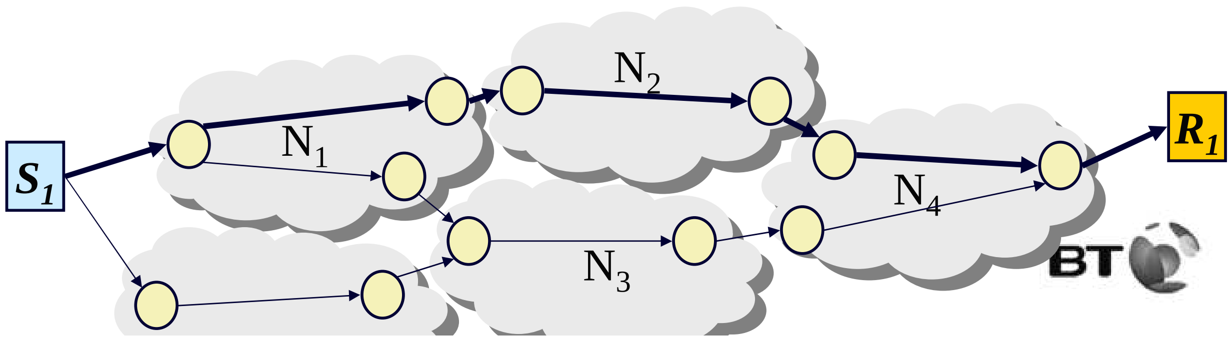
- **“…passive & extremely simple”**
- recall sending domain pays to receiving domain $\boldsymbol{C = \eta X + \lambda Q}$
- congestion charge, $\boldsymbol{Q}$ over accounting period, $\boldsymbol{T_a}$ is $\boldsymbol{Q = \Sigma^{T_a} \rho_i^{+}}$
- $\boldsymbol{\rho_i}$ metered by **single bulk counter** on each interface
- note: negative $\boldsymbol{\rho_i}$ worthless
  - creates incentive to deploy droppers

downstream path shadow price, $\boldsymbol{\rho_i}$

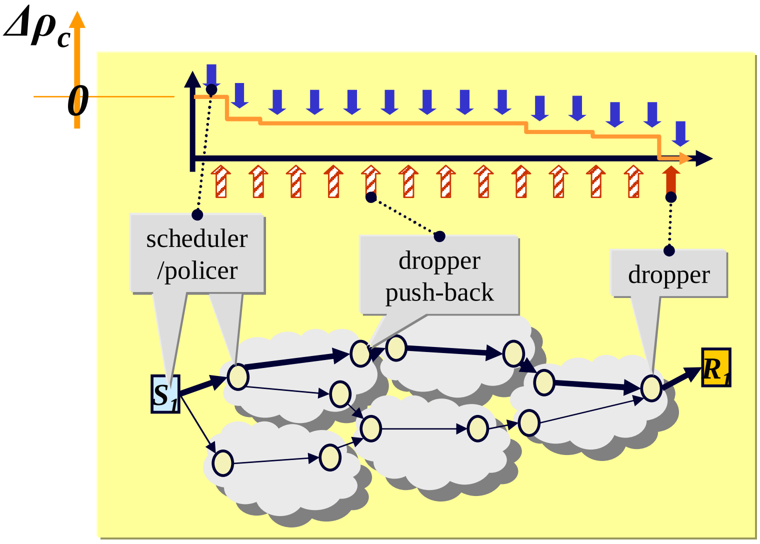


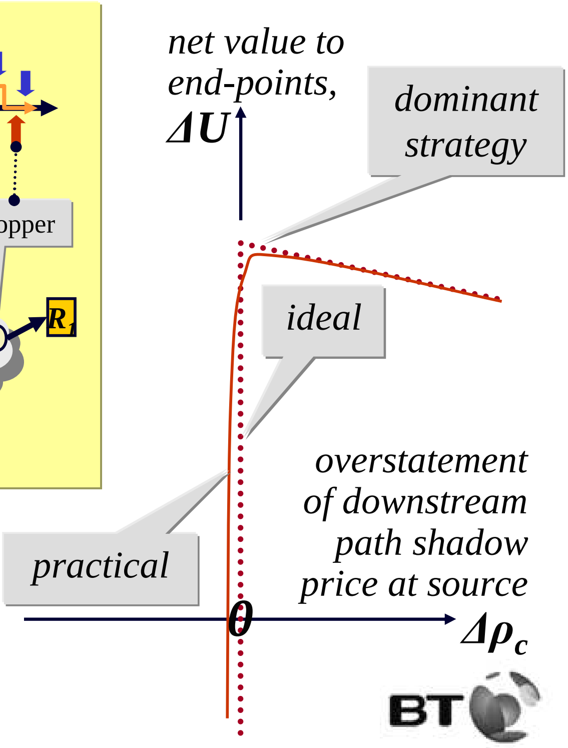
incentive compatibility – inter-domain routing

- why doesn't a network overstate congestion?
  - **msecs**: congestion response gives diminishing returns (for TCP: $\Delta\Pi \propto \sqrt{\Delta\rho}$)
  - **minutes**: upstream networks will route round more highly congested paths
    - by sampling data N_1 can see relative costs of paths to R_1 thru N_2 & N_3
  - **months**: persistent overstatement of congestion:
    - artificially reduces traffic demand (thru congestion response)
    - ultimately reduces capacity element of revenue
- also incentivises provision to compete with monopoly paths



incentive compatibility – hosts

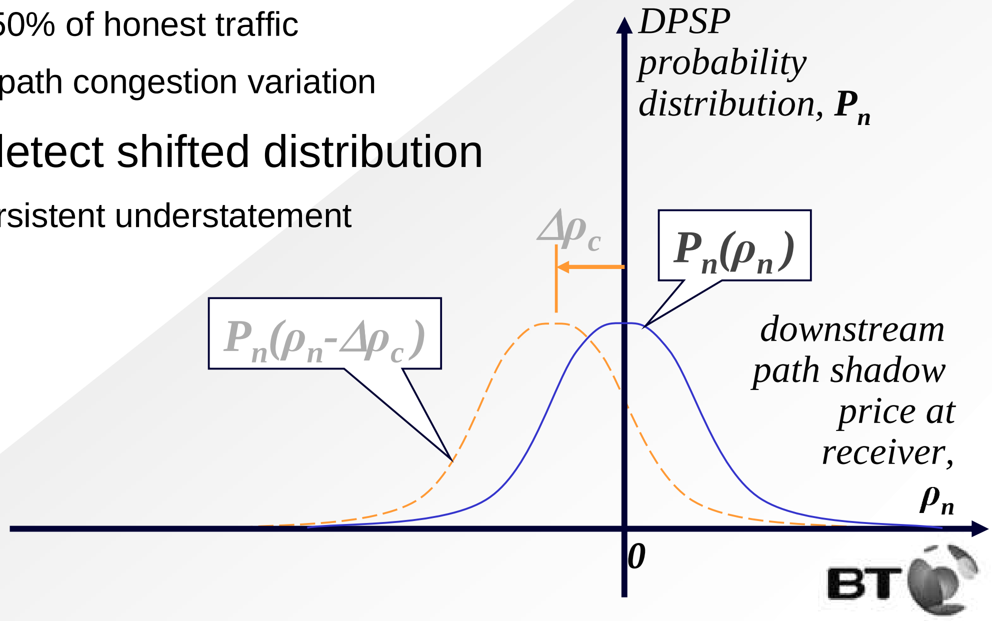




- incentivise:
  - responsible actions
  - honest words

downstream path shadow price at rcvr

- for congestion $m_p \geq 0$
  - congestion being probability *[0,1]*
- naïve: drop ‘negative packets’
  - drops 50% of honest traffic
  - due to path congestion variation
- instead: detect shifted distribution
  - find persistent understatement

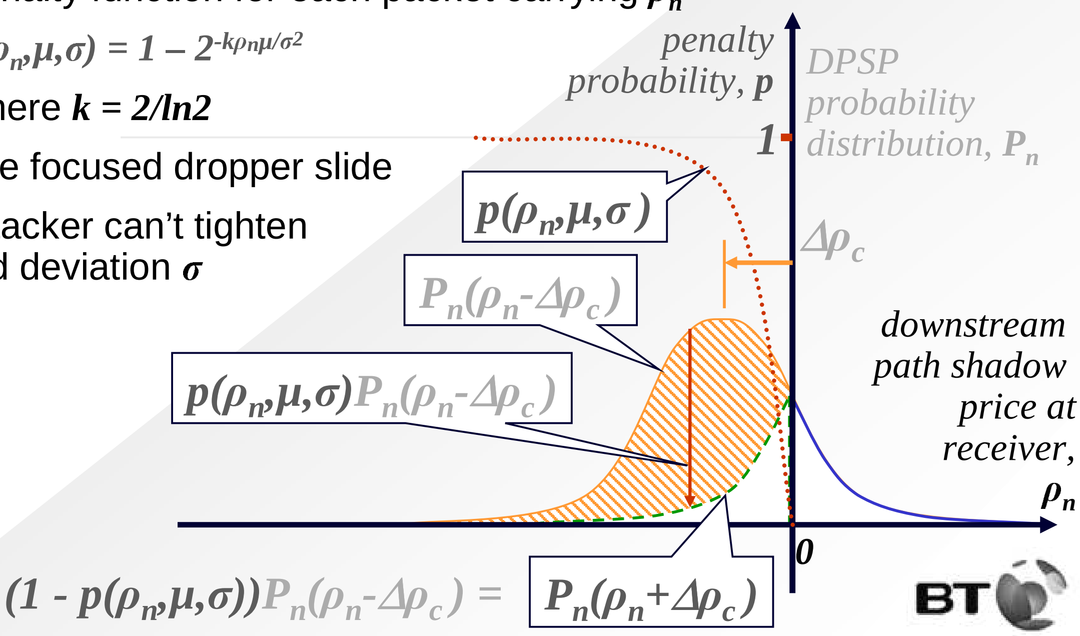


penalising misbehaviour with uncertainty

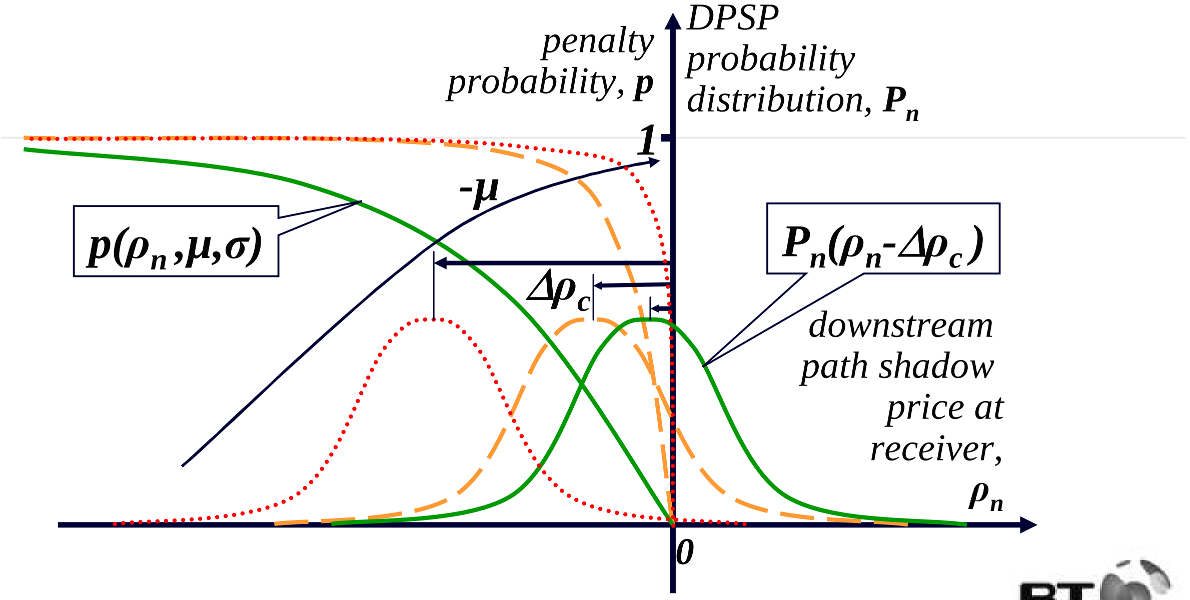
- continuously update μ, the EWMA of ρ_n,
  - not counting any packets flagged ‘uncertain’ with $\rho_n > 0$
- for traffic subset from malicious source, $\mu \rightarrow \Delta\rho_c$
- penalty function for each packet carrying ρ_n

  $p(\rho_n, \mu, \sigma) = 1 - 2^{-k\rho_n\mu/\sigma^2}$

  where $k = 2/\ln 2$
- see focused dropper slide
- attacker can’t tighten std deviation σ



dependence of penalty function on recent history



focused droppers

- use penalty box technique [Floyd99]
  - examine (candidate) discards for any signature
  - spawn child dropper to focus on subset that matches signature
  - kill child dropper if no longer dropping (after random wait)
- push back
  - send hint upstream defining signature(s)
  - if (any) upstream node has idle processing resource
    test hint by spawning dropper focused on signature as above
- cannot DoS with hints, as optional & testable

extending incentives to other metrics

- downstream uncongested delay (emulated by TTL)
  - approximates to ½ feedback response time (near source)
  - each node can easily establish its local contribution
  - identical incentive properties to congestion
    - increasing response time increases social cost
    - physically impossible to be truthfully negative
  - therefore incentive mechanism identical to that of congestion

- assess other metrics case-by-case

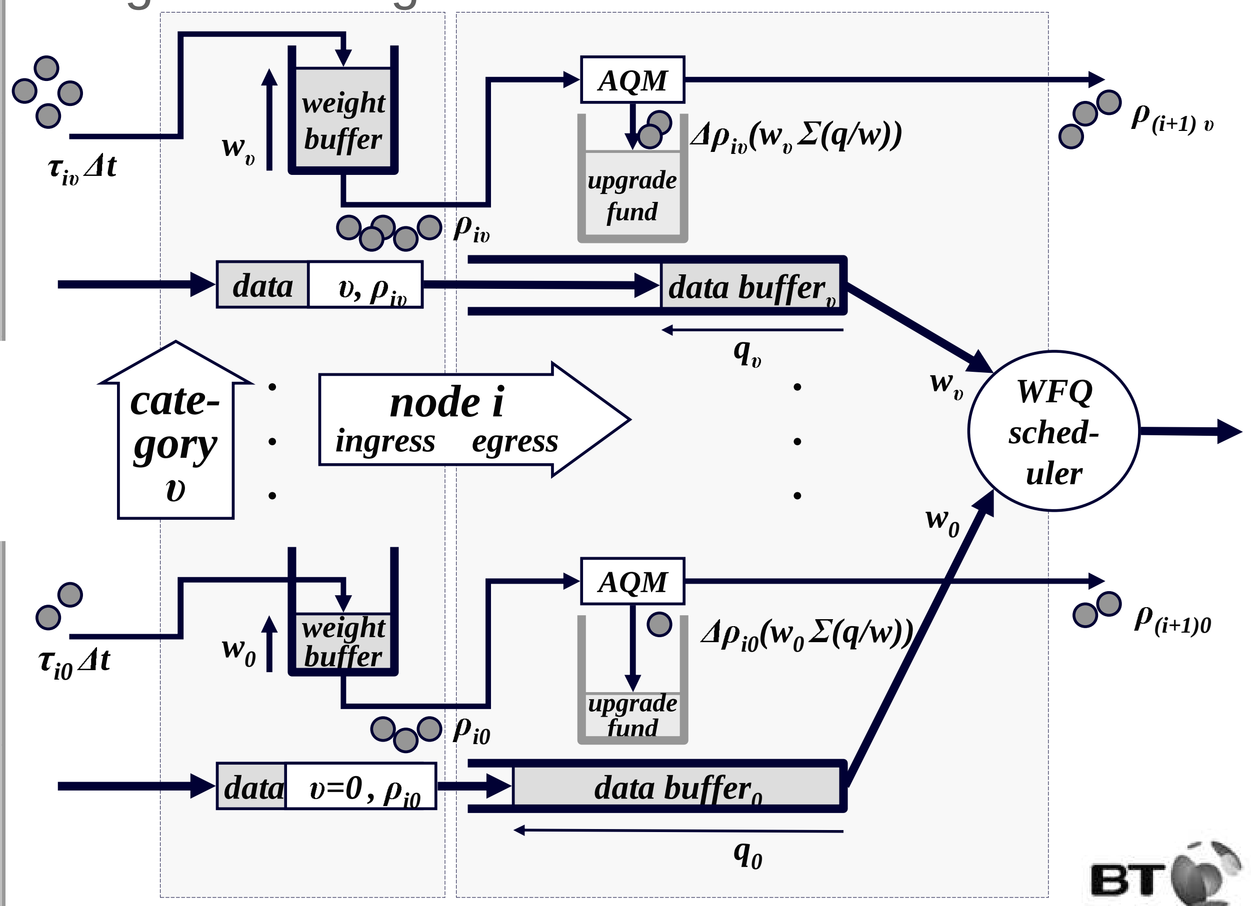
stateless TCP/ECN policer

- rate policing feasible, but TCP policing hard
  - RTT & path loss/marking rate of each flow unknown locally
- TCP congestion avoidance rate converges on

$$\bar{x} \approx \frac{s}{T}\sqrt{\frac{3}{2p}} \quad (p << 1) \qquad \text{ignoring re-transmit timers}$$

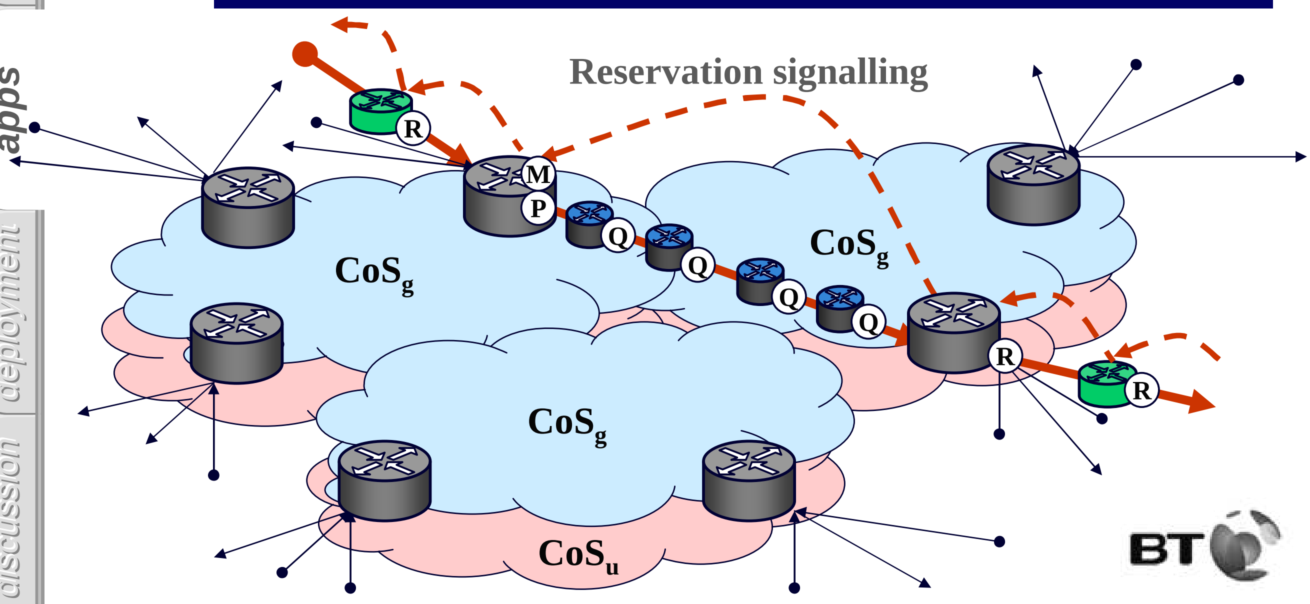
- re-feedback gives truthful values of all these metrics
  - packet headers arrive with prediction of own downstream path
- weight random selection of candidates for drop
  - e.g. Choke-like scheme [Pan00], but based on correct metrics
- inter-domain congestion charging "…emulates border policing"
  - only need TCP policer at first network ingress

congestion weighted differentiated service

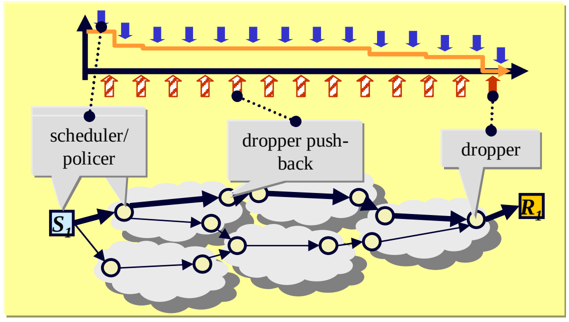


intro incentives apps deployment discussion

IP routers	Data path processing
Reservation-enabled	(R) Reserved flow processing
Guaranteed QoS gateway	(P) Policing flow entry to CoS_g (M) Meter congestion per peer
ECL only	(Q) Bulk ECL marking CoS_g prioritised over CoS_u

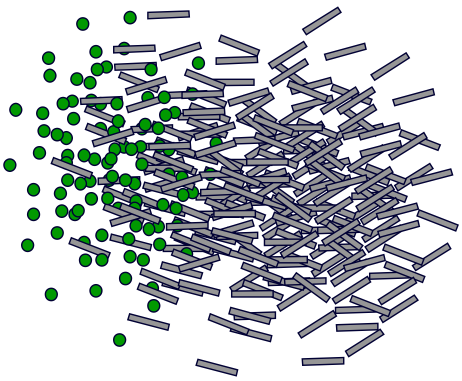


slow-enough-start



- initial value of metric(s) for new flows?
  - undefined – deliberately creates dilemma
  - if too low, may be dropped at egress
  - if too high, may be deprioritised at ingress
- without re-feedback (today)
  - if congested: all other flows share cost equally with new flow
  - if not congested: new flow rewarded with full rate
- with re-feedback
  - risk from lack of path knowledge carried solely by new flow
  - creates slow-start incentive
  - once path characterised, can rise directly to appropriate rate
  - also creates incentive to share path knowledge
  - can insure against the risk (see differentiated service)

single datagram-dominated traffic mix



- current Internet would collapse
  - not designed for all eventualities
  - 10^{12} devices, 10^{9} users, RPCs, sensor nets, event avalanches
- with re-feedback
  - service protected against completely uncorrelated traffic mix
  - demanding users can still insure against risk
- for brief flows, TCP slow start sets rate limit
  - …not technology performance advances
  - with re-feedback, once characterised path, can hit full rate

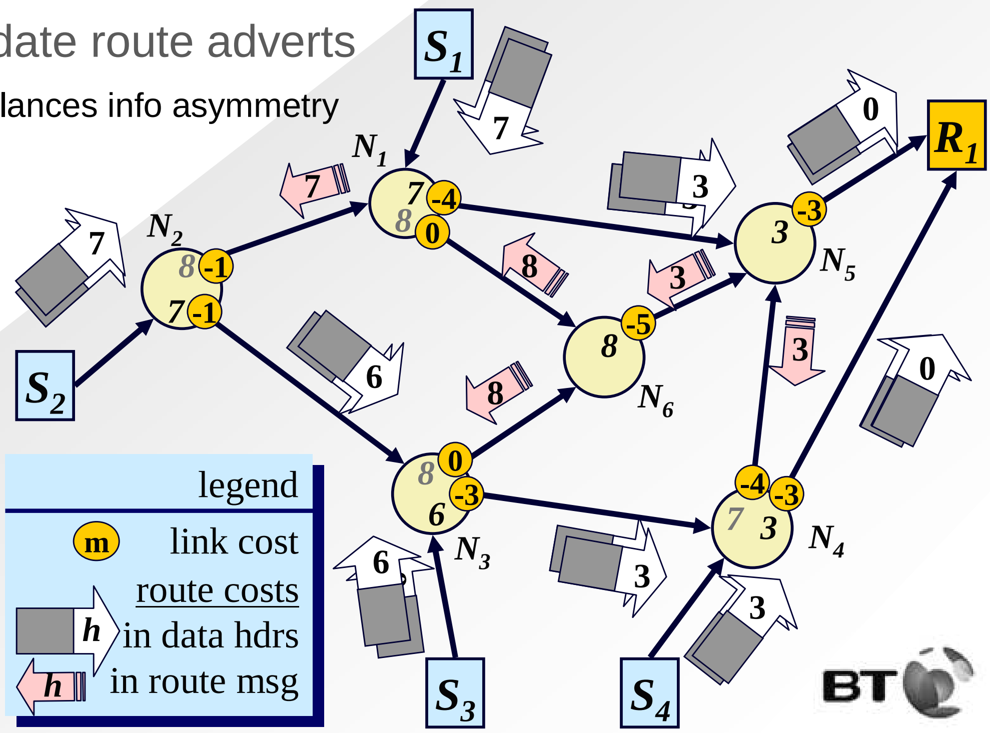
denial of network service protection

- network DDoS causes network congestion (by definition)
- honest sources will increase initial metric
  - which deprioritises their flows relative to uncongested destinations
- if malicious sources don't increase initial metric
  - their traffic will go negative either at the point of attack or before
  - can be distinguished from honest traffic and discarded
  - push back will kick in against persistent attacks
- if malicious sources do increase initial metric
  - scheduler at attacker's ingress will deprioritise attacker
  - only honest sources sharing full path with attackers lose out greatly
- could hijack zombie sources to pay for higher class service
  - incentivises their owners to sort out virus protection
  - marginal cost of network upgrade paid by those that don't!

BT

routing support

- can automate traffic engineering (damped response time)
- can validate route adverts
  - re-balances info asymmetry

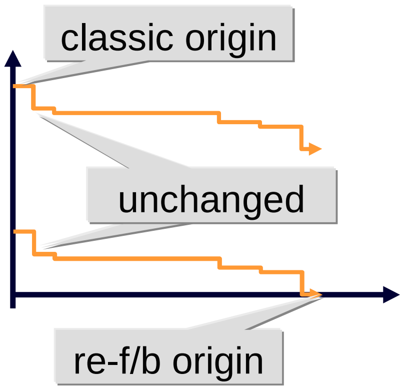


BT

which metrics?

- many applications need niche path metrics
- but which are necessary and sufficient?
  if we were to define a new Internet architecture
  - congestion
  - uncongested delay
- many more possible, but perhaps not necessary
  - explicit loss-rate (esp for wireless)?
  - per bit and per packet congestion?

migration



- (ideal) approach
  - realign metrics around unchanged router path characterisation
  - modify sender and/or receiver stack only
  - network operators add incentive mechanisms to edge routers
  - incentivise incremental introduction of each element
  - still works without each change, but less advantageous
- reasoning:
  - hard to know that no routers on a path haven't been upgraded
- note: migration still very much 'work in progress'

migration: re-ECN

- insufficient codepoints to be sufficiently responsive
  - we know this anyway (e.g. [Ganesh02] or XCP [Katabi02])
- can use the three code-points we have
- multi-bit field: no easy migration
  - effectively impossible (?) with IPv4 (& MPLS!)
  - can use IPv6 hop-by-hop options – added when accuracy needed but needs 32bit header extension for +1bit & 64bit for +(2–32)bit
  - – if any node on path doesn't support multi-bit field, value unreliable
    - detection of this condition possible
    - but little deployment incentive without flag day

Diffserv byte playoff – latest score

Bell heads	6 : 2	Net heads

BT

intro incentives apps **deployment** discussion

migration: re-TTL

- need to avoid interaction with loop detection
  - – set target at destination $h_z = 16$ (say), to allow headroom for path variation without triggering drop due to 'TTL expired'
- need to add feedback in transport layer protocols
  - TCP, RTCP, DCCP, etc.
- need to standardise the unit conversion with time
- issue: TTL is a pretty coarse measure

BT

migration: certain flag

- necessity
  - relays need to average metrics for traffic eng, route validation, dropping etc.
  - uncertain metrics would pollute averages if not flagged
  - more so if traffic matrix becomes dominated by short flows
- can overload certain flag
  - ‘re-feedback capable transport’ flag
  - IPv4 header: bit 49 (reserved but in much demand)
  - IPv6 header: incorporated into header extension for mulit-bit ECN
- incentives as described earlier are arranged
  - to flag certain when you are
  - and not when you’re not

information gains & losses

aligned at	knowledge	sender	relay	receiver
sender	upstream path[1]	-	✓	✓
receiver	upstream path[1]	-	✗[2]	✗[2]
sender	downstream path	✓[3]	✗	-
receiver	downstream path	✓	✓	-

- notes
  1. upstream path knowledge is of little use to anyone for control
  2. both alignments can be included (giving whole path knowledge too)
  3. for TTL, no feedback meant no sender downstream knowledge

deployment incentives

- congestion pricing
  - prevents wasteful investment in resources not targeted at demand
  - initially for access providers to predominantly receiving customers
- policer/scheduler
  - reduces congestion charges to downstream operators
- dropper
  - ensures sufficient congestion charges are paid to receiving access provider by upstream provider to deliver to destination

related work

- MacKie-Mason & Varian “Pricing the Internet” (1993)
  - Smart Market idea of placing bids in packets
  - admitted it was impractical – also poor feedback
- Clark “Combining Sender and Receiver Payments in the Internet” (1996)
  - decrementing payment field in packet – no e2e feedback
  - no separation between technical metric and price to apply to it
- Kelly et al “Rate control for communication networks: shadow prices, proportional fairness and stability” (1998)
  - the game theoretic basis, but with the direction of payment the wrong way round
  - consequently needs retail dynamic pricing
- Savage et al “TCP Congestion Control with a Misbehaving Receiver” (1999)
  - ECN nonce – only effective if sender’s & network’s interests align
- Constantiou & Courcoubetis “Information Asymmetry Models in the Internet Connectivity Market” (2001)
  - describes the inter-domain info asymmetry problem
- Zhu, Gritter & Cheriton “Feedback Based Routing” (2003)
  - dishonest inter-domain routing is better solved by measurement than authentication

further work

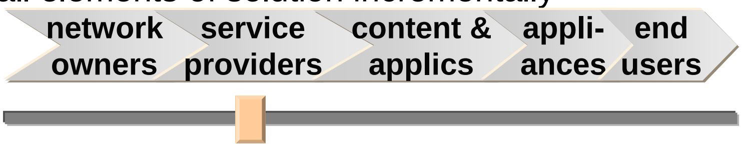
- analysis of accumulation of variation of congestion along a path
  - simulation to validate dropper vulnerability
- formalise game theoretic analysis (largely building on Kelly)
  - adding routing & slow-enough-start
- detail design of applications
  - fairness, slow-start, QoS, routing, DoS (esp dynamic attacks)
- analyse deployment with heterogeneity
  - technical and business
- complete detailed protocol design incl. migration
  - simulation & implementation
- …

discussion

- why aren't networks run like this already?
  - must guess for first packet
  - requires per header storage in sender transport layer
  - without incentive framework, if use info for control, truth incentives distorted
- is the tussle for control in this space strong enough to need re-f/b?
- layering violation?
  - passing info up the layers (ECN) was anathema – is re-feedback 'worse'?
- alternative to route advert authentication?
  - characterises at router layer granularity, not domain layer
  - is this too much info symmetry for operators?
- is characterising only the path your access provider offers sufficient?
  - to empower user choice without loose source routing?
- why isn't even congestion marking being deployed commercially?
- …

contributions

- arrange honesty & responsibility to be dominant strategies
  - even for first packets of a flow
  - without tampering with retail pricing
- downstream information upstream
  - updated within round trip
  - enhance, never reduce, info usefulness to each party
  - overload existing path characterisation data headers (e.g. TTL, ECN)
  - incentives to deploy all elements of solution incrementally
  - no change to routers
- control architecture



  - re-feedback designed for tussle over who controls what
  - Q. who controls the slider? A. socio-economic (market, regulation)
  - sufficient to police others, or to take full control (proxy)

contributions: applications

- congestion control/QoS
  - rate (e.g. TCP) policing
  - differentiated service synthesised from diff. congestion response
  - guaranteed QoS synthesised from path congestion-based AC
  - inter-domain traffic policing emulated by bulk metering
  - incentivise ‘slow-enough-start’
  - first line of defence against flooding
- routing
  - advert validation
  - traffic engineering
  - capability-based routing
- not exhaustive
  - re-feedback intended as enabler