PITTSBURGH SUPERCOMPUTING CENTER

BT

Internet Capacity Sharing Architecture

a design team of the ICCRG

congestion control research agenda

Matt Mathis & Bob Briscoe
PSC & BT

Mar 2010

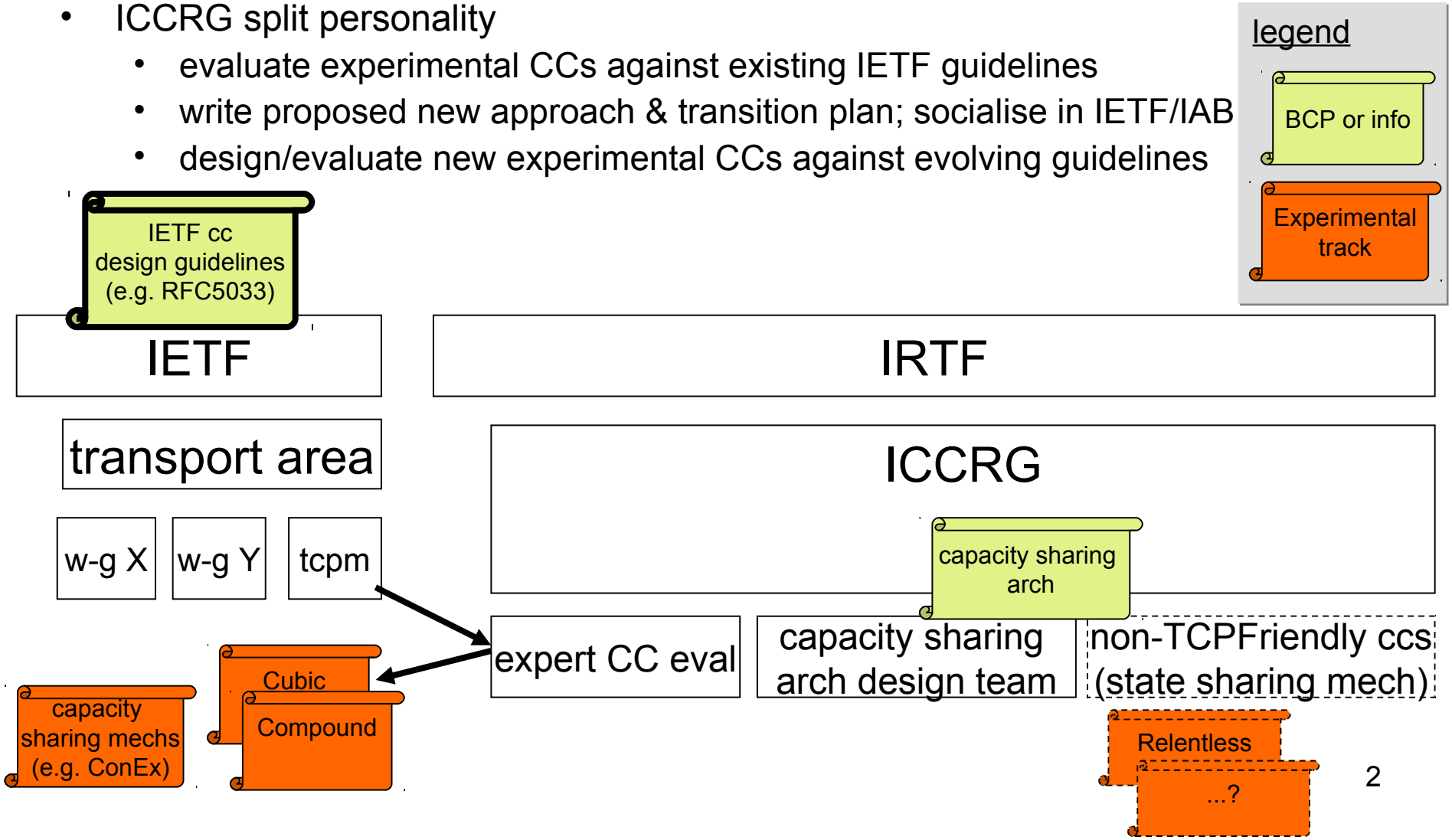
Bob Briscoe is partly funded by Trilogy,
a research project supported by the European Community
www.trilogy-project.org

trilogy

Internet capacity sharing architecture; design team

relation to other ICCRG/IETF activities

- ICCRG split personality
  - evaluate experimental CCs against existing IETF guidelines
  - write proposed new approach & transition plan; socialise in IETF/IAB
  - design/evaluate new experimental CCs against evolving guidelines



work as if Congestion Exposure (ConEx) exists…

- allows us to assume
  - ISPs can count the volume of congestion a user causes
    - = bytes marked with ECN (or dropped)
  - ISPs can incentivise care over contribution to congestion
  - gives license to diversity of individual congestion responses

- challenges us to zoom out to more macro scale
  - flow arrival patterns, flow lengths
    - not just competing flows in congestion avoidance (CA)
  - hi & lo stat mux
- classify research challenges into three areas
  1. scaling transport performance – dynamic range
  2. diversity of congestion responses – weighted etc
  3. predicting congestion patterns & structure

PITTSBURGH SUPERCOMPUTING CENTER

BT

research area #1

scaling transport performance

trilogy

scaling transport performance

briefly recap current received wisdom

w:	window
k:	constant (~3/2)
p:	loss fraction

- TCP CA algo leads to bit-rate of long-running flows:

$$\bar{w} = \sqrt{\frac{k}{p}}$$

- rearranging, bit-rate of identical flows sharing bottleneck increases until loss fraction becomes:

$$p = \frac{k}{\bar{w}^2}$$

- when a set of TCPs each get the bit-rates shown, these loss fractions result, assuming

packet size, s = 1500B
RTT, R = 100ms

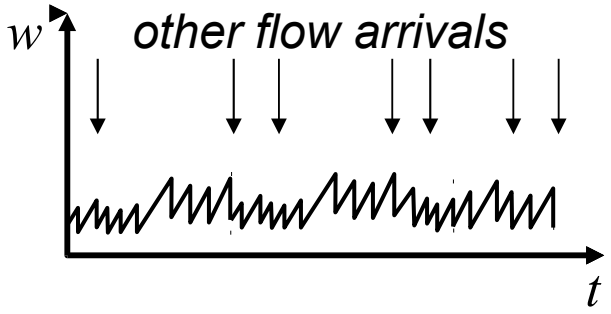
bit-rate	TCP loss fraction	recovery time
1Mb/s	2%	550ms
10Mb/s	0.02%	5.5s
100Mb/s	0.0002%	55s (~1min)
1Gb/s	0.000002%	550s (~9min)

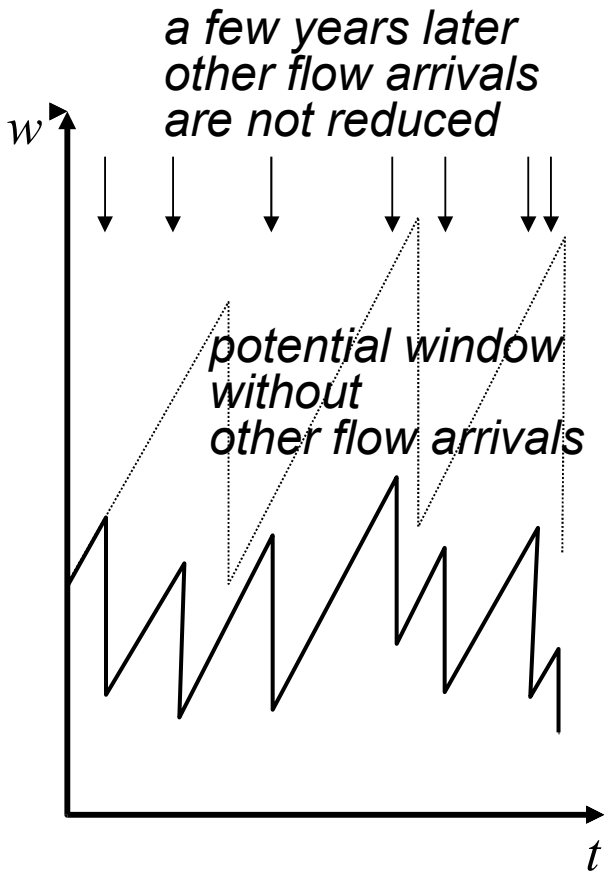
Scripture prophesised this

> "We are concerned that the congestion control noise sensitivity is quadratic in w but it will take at least another generation of network evolution to reach window sizes where this will be significant."

In footnote 6 of:
Jacobson, V. & Karels, M.J., "Congestion Avoidance and Control," Laurence Berkeley Labs Technical Report (November 1988) (a slightly modified version of the original published at SIGCOMM in Aug'88) URL: <http://ee.lbl.gov/papers/congavoid.pdf>

what's the real performance scaling problem?





what's the problem with long recovery times?

- scaling is over 3 dimensions, not just one:
  1. flow rate
  2. # flows
  3. flow size

if #flows through bottleneck does not shrink (2)
and capacity increased so flow rates can grow (1)

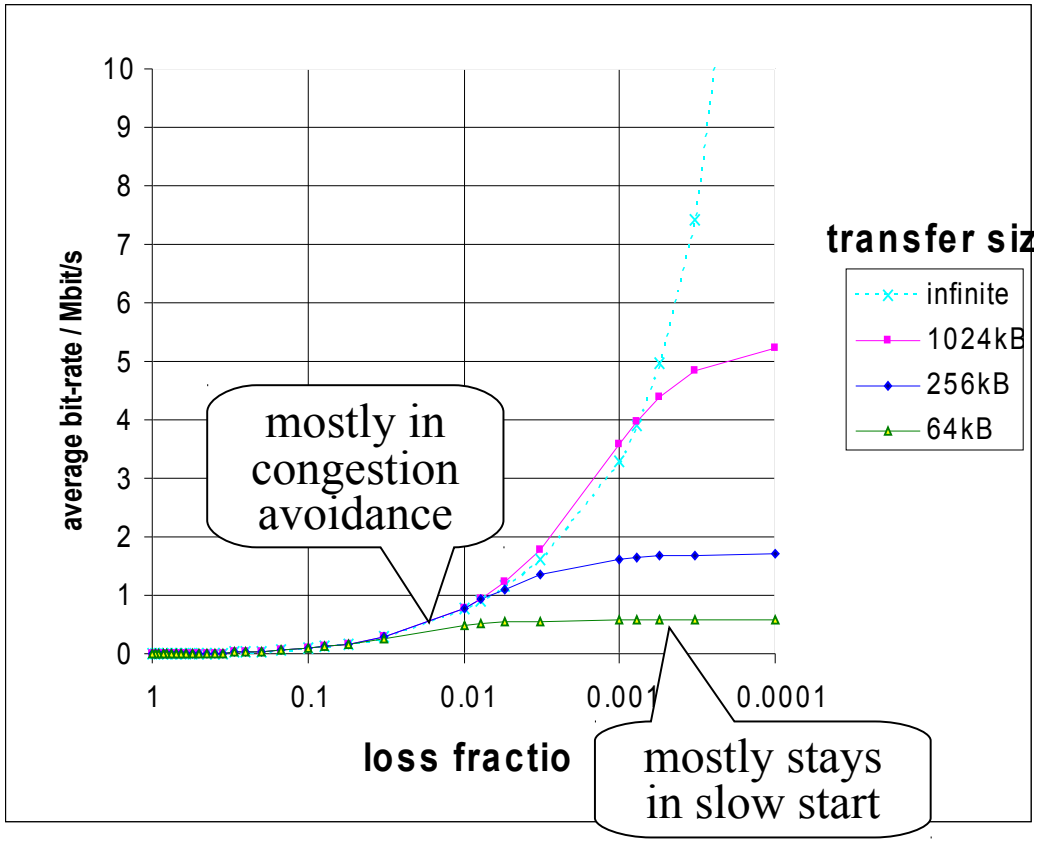
- each flow arrival generates a loss event at the end of slow-start
- window bounded by arrival rate of other flows*
- not by capacity

research focus needs to shift:

- conflicts between slow-start & CA phase
- conflicts between elastic & other transports

* or link bit error rates, esp. wireless but also DSL

what's the real performance scaling problem?



TCP average throughput model for different size flows [Cardwell00]

- scaling over 3 dimensions:
  1. flow rate
  2. # flows
  3. flow size

if flow sizes increase (3)
and capacity increased
so flow rates can grow (1)

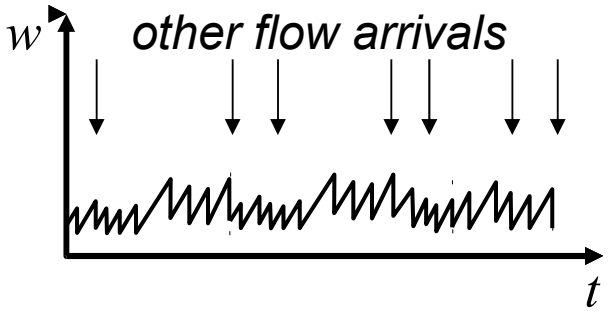
- loss fraction reduces $O(1/w^2)$
- if flow size growth insufficient
- growing proportion of flows limited by slow start
- not by capacity

- motivation for ad hoc tinkering
  - multiple flows, larger IW

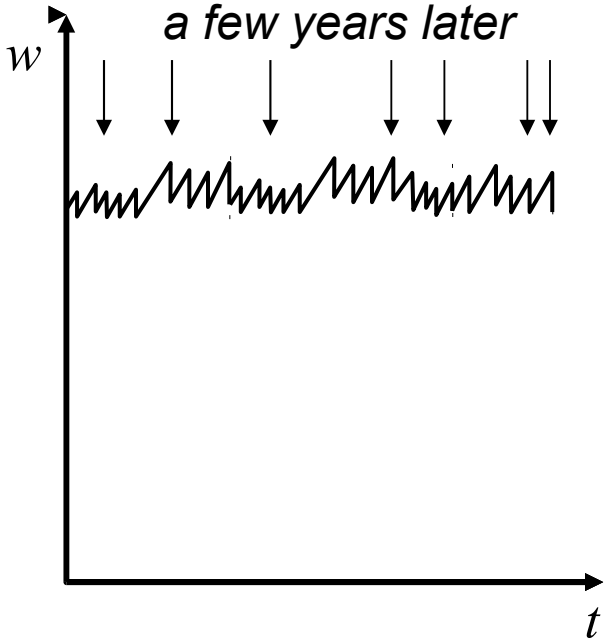
research focus needs to shift:

- mitigating overshoot on start-up

How to scale TCP to any speed



- (A thought experiment about the limiting case)
- Control frequency should not depend on data rate
- For a fixed path, fixed time between losses
- Data between losses is proportional to rate
- Loss probability is inverse of rate
- Model has to resemble data rate $\propto 1/p$



Do we have consensus on this? *

* Outstanding problem: synchronized losses due to drop tail
  - lead to RTT unfairness pathology for $w \propto 1/p^d$ as $d \rightarrow 1$ [Xu04]

network support?

- what new network feature is needed, if any,
  to help e2e transport performance scale?
- challenge #1 in
  “Open Research Issues in Internet Congestion Control”

  <draft-irtf-iccrg-welzl-congestion-control-open-research>

delay sensing – not a panacea

- scaling any of the 3 dimensions upwards drives queuing delay downwards [Kelly00; §2]
  1. flow rate ⇐
  2. # flows ⇐
  3. flow size ⇐
- increasingly hard to distinguish tiny queuing delay from noise

is a scalable congestion control sufficient?

- more aggressive
- and more robust to aggression
- loss *probability* reduces over the years
  - loss *rate* remains the same for the fast transfers

- if a sensitive app (e.g. VoIP) works today
- it should work tomorrow..?

- the challenge
  - high acceleration
  - overshoot when sensing available capacity

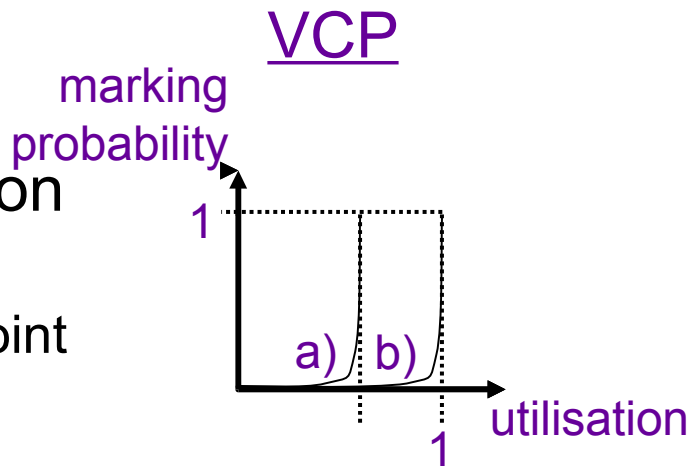
Do we need flow isolation too?

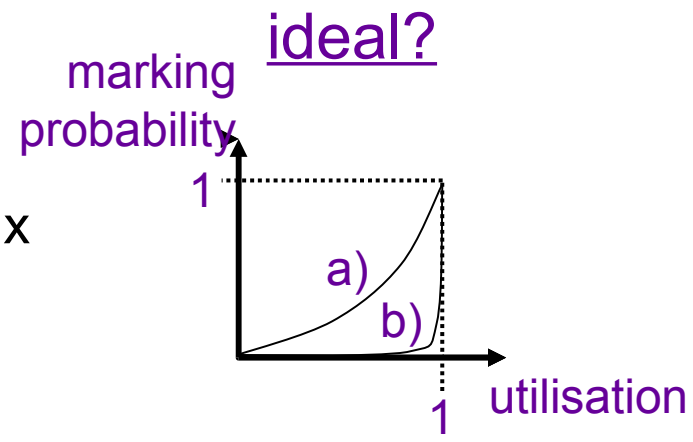
- Isolate traffic such that greedy flows can't harm others
  - Undo “Simple network” assumption
- Requires the network to distinguish between flows
  - Send more signals to aggressive flows
  - Ideally small (short or low rate) flows have predictable rates
- See: draft-livingood-woundy-congestion-mgmt-03
- See: later talk by Matt Mathis
- fundamental conflict with weighted congestion control

or are utilisation hints sufficient network support?

ConEx *and*

- two levels of unary explicit congestion notifications:
  a) bottleneck utilisation: one ECN codepoint
  b) regular ECN
- potential:
  - ConEx creates incentive to avoid b)
  - a) warns that b) is approaching
  - correlation between a) & b) tells transport that bottleneck is low stat mux
  - if a) is partially deployed, not fatal
  - work in progress...





PITTSBURGH SUPERCOMPUTING CENTER

BT

research area #2

diversity of congestion responses

trilogy

research area #2 assuming ConEx deployed

weighted congestion controls

- feasible improvements in completion times?
- limits to the feasible range of weights?
- acceleration independent of weight?
  - convergence
- weight start-up separately or dependent?
  - overshoot?
- not just elastic file-transfer
  - streaming video etc
  - preventing starvation of classic TCP?

- socket API, policy control, etc
- default weight: related to file size?

PITTSBURGH SUPERCOMPUTING CENTER

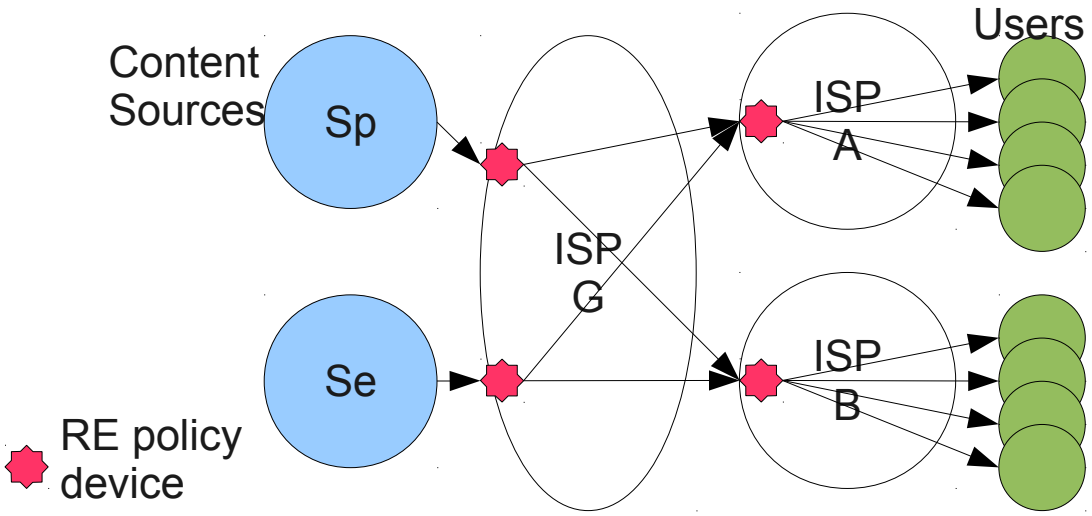
BT

research area #3

predicting congestion patterns & structure

trilogy

Cascaded ISPs



- Policy control at ISP A&B ingress is good
  - It can be used to limit downstream congestion
- Policy control at ISP G's ingress may be problematic
  - No uniform expectation for downstream congestion
  - Unless globally anneal to a uniform congestion level

Problem: Unexpected performance

- Application performance explicitly depends on other users
  - Expected be more erratic than the current net
    - Some people might disagree
  - Especially if users can bid for congestion
    - Most users would prefer stable prices and data rates
- Moves the net away from performance guarantees
  - A big headache for high performance applications
  - Not that we can do performance guarantees today
    - RE-ECN is likely to be quite a bit worse

More predictable performance?

- Re-ECN doesn't change the congestion control
  - explicit dependence on other users unchanged
  - solely enables operator to switch on the significance of minimising congestion
  - likely to encourage shifting of peaks into troughs
- Moves the net towards more assured performance
  - global 'annealing'
- If using network at maximum efficiency
  - can have either stable prices or stable performance
  - if want both, have to pay a constant but higher price
  - or accept lower but consistent service

Which of the two views is probably correct?

Problem: not diagnosable

Point

- Performance depends on things not observable
- User can't tell why any particular marking rate
- Provider sees aggregate marking & data rates
  - No specific information about any particular flow
- Problem may be an unrelated flow that user can't identify
- Out bidding may not be feasible

Counterpoint

- re-ECN gives operator info it doesn’t currently have
  - can locate problems to neighbouring networks
- measuring aggregates is sufficient
  - but nothing to stop looking per flow (e.g. for fault diganosis)

summary: primary research questions

performance scaling

- diminishing performance gain from capacity investment
  - e2e transport is becoming the limit, not transmission capacity
- understand conflicts: slow-start v. CA phase v. other transports
- mitigating overshoot on start-up
  - need to prove whether e2e can be sufficient
  - otherwise flow isolation v. overshoot hints v. …?

diversity of congestion responses - weighted cc

- open research space: whole range of questions

global congestion patterns

- smoother? or more unpredictable?
- reflecting disparities in the global market? or disjoint from them?

references

- [Cardwell00] N. Cardwell, S. Savage and T. Anderson, "Modeling TCP latency,“In Proceedings of IEEE INFOCOM, Tel Aviv, Israel, March 2000. *http://citeseer.ist.psu.edu/cardwell00modeling.html*
- [Wischik07] Wischik, D., "Short Messages," In: *Proc. Workshop on Networks: Modelling and Control* Royal Society (September 2007) http://www.cs.ucl.ac.uk/staff/ucacdjw/Research/shortmsg.pdf
- [Xu04] Lisong Xu, Khaled Harfoush, and Injong Rhee, " Binary Increase Congestion Control for Fast Long-Distance Networks ", In: *Proc IEEE INFOCOM 2004*, pp. 2514-2524 (March 2004)
- [Kelly00] Kelly, F.P., "Models for a Self-Managed Internet," *Philosophical Transactions of the Royal Society* **358**(1773):2335--2348 (August 2000) **http://www.statslab.cam.ac.uk/~frank/smi.html**

PITTSBURGH SUPERCOMPUTING CENTER

BT

Internet Capacity Sharing Architecture

congestion control research agenda

Q&A

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

trilogy