Internet Flow Control - Improving on TCP

Glynn Rogers

The Team - Jonathan Chan, Fariza Sabrina, and Darwin Agahari

CSIRO ICT Centre

Why Bother? - Isn't TCP About as Good as It Gets?

- Well, TCP **is** a very successful protocol
  - Stability of the current Internet is owed to TCP
- But TCP has increasingly evident problems
  - The well known problems with wireless links
  - In congestion avoidance mode, on large delay-bandwidth product links, it can take thousands of RTTs to recover from a packet loss ie an hour or more
  - Indeed on large delay-bandwidth links TCP is unstable ie it can oscillate severely.

Efforts to Improve TCP

- Upgrade TCP by modifying the congestion window increase method
  - Highspeed TCP, Scalable TCP, and BIC TCP
  - These still use packet loss as the signaling mechanism
    - Susceptible to rapid window 'deflation' on large delay links
      - Even with RED - must wait a RTT before effect is seen
- Alternative forms of TCP
  - Use queuing delay as congestion indication
    - Avoids falling over the packet loss 'cliff'
  - TCP Vegas exists but does not appear to be used substantially
  - FAST TCP proposed by Caltech - IETF Draft
    - Uses RTT variation to sense queues
    - 'equation based' - no packet level oscillations
    - Stable flow dynamics

But TCP is Only Part of the Picture

- From a packet by packet perspective TCP is the centre piece
- However from the perspective of long term flows (video etc) -
  - Need to consider the whole flow control system
    - Link bandwidths and utilization
    - Queue management - tail drop, RED in various configurations
    - packet scheduling algorithms
  - Many variants of each in the 'standard model' heterogeneous network
  - TCP window control must cope with this complexity
- Faced with this, perhaps TCP is about as good as it gets!

Does It Matter?

- In this era of easy over provisioning, perhaps not
- But - Hank Kafka, Chief Architect, Bell South, 'guesstimates' that (unicast) Video over the Internet will increase average consumers demand per month by two orders of magnitude
  - This will 'overwhelm current Internet core technology'
  - He sees 'network management/traffic control' as part of the solution

Hank Kafka, "Drivers for Next Generation Networks", Optical Fibre Communications/National Fibre Optics Engineers Conference, Anaheim 2005

So Will Flow Control always be Crippled by Irreducible Complexity?

- Perhaps not!
  - A recent article in Business Communications Review highlighted the popularity of MPLS based VPNs in large enterprise networks
    - Also pointed out the need for complete QoS control on VPNs
    - This has been implemented by some enterprise networks
- Put this together with recent ideas on
  - overlay networks, user controlled lightpaths, virtualisation etc
- Maybe we begin to see the beginnings of a new Architecture
  - Oriented to long lived flows requiring QoS
  - Would sit alongside the current routed network
- Could tap into emerging fundamental ideas on complex systems architecture

John Bartlett and Rebecca Wetzel, "QoS over MPLS - the Complete Story", Business Communications Review, February 2006.

What sort of Components Might this Architecture Need?

- Well, to start with, it would need well defined Classes of Service
  - A Premium Class with
    - Guaranteed rates and latency
    - Full admission control
    - Policing
  - Needs heavy handed management - expensive
- So also need some form of automated rate control
  - as in TCP, ATM's Available Bit Rate concept etc
- But no (or minimal) packet loss for real time flows
  - Suggests queuing delay as congestion indication as in TCP FAST etc

Now We have a Problem with Long Lived Flows

- Cannot rely on short timescale statistical fluctuations in load to maintain queue stability
  - Flows may come and go on longer timescales
  - Queues may grow alarmingly particularly during 'worst case' events
  - On large delay-bandwidth links queues may grow very large in a RTT
  - Result is unacceptable jitter at best

Our Solution - Define an Elastic Class

- Concept first aired in IETF's Integrated Services discussions
- Dimension the network to provide
  1. A Premium Class with reserved bandwidth
  2. An adequately provisioned traditional Best Effort Class
  3. An Elastic Class provisioned to give 'good' service under expected conditions
     - Then two responses to sudden load increase
       1. Short term - sacrifice some Best Effort performance
       2. Longer term - apply control signals to sources
     - Provide incentive by offering a guaranteed minimum rate
- Yes, you have heard all this before somewhere - ATM?

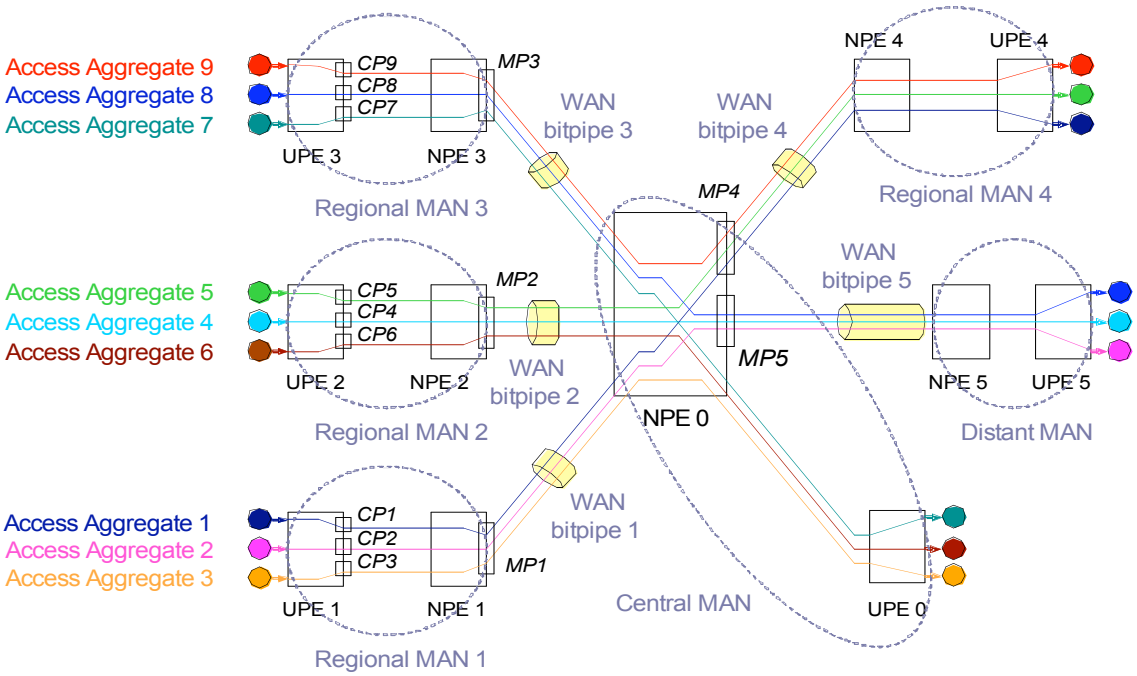
CSIRO

But Have We Avoided the Complexity Issues? - No? - OK, Next Move

- Distinguish between
  1. The issues associated with individual source control and
  2. The control of flow aggregates in the network core
     - This is consistent with the IETF's Diffserv concept
- We have left 1 to the experts (more or less) and focused on 2
  - Partition the overall network into external and internal components
  - Define a Flow Control Architecture based on
    - high speed local networks connected by fixed capacity pipes (MPLS LSPs etc)
    - Measurement Points at the pipe ingress points
    - Control Points at the access points to the internal network
    - Feedback signaled between the Measurement and Control Points
  - The Control Points
    - Calculate the required ingress aggregate rates at the access point
    - Divide this up fairly amongst its connected sources
    - Communicate with the sources

CSIRO

An Example Network

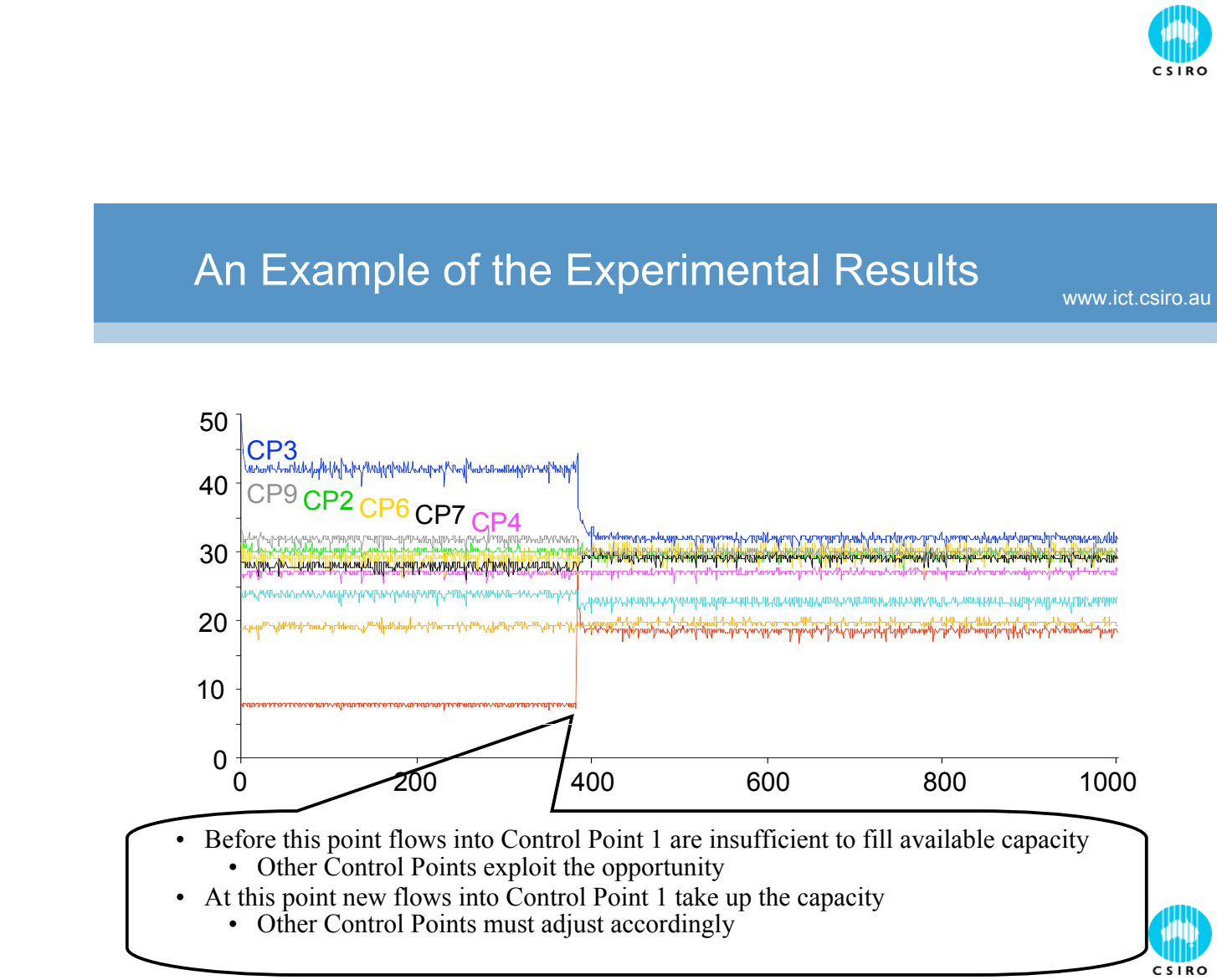


Flow Control

- Distributed, potentially nonlinear, feedback control system
- Major issues of convergence and stability
- Only way to do this properly is to use a mathematical model
  - Generic mathematical model developed at Cambridge by Frank Kelly and colleagues
    - Based on a constrained optimization of the utilization of the network resources
  - We have modified this a little
  - And used it for simulations and an analytical study of stability as well as the design of the flow control system

Does This Actually Work?

- Yes! At least in the small scale network we have been able to investigate
- We have built an NS2 simulation and an experimental network
  - This runs over the CeNTIE research network
  - It has one pipe with a large delay using the CeNTIE Perth link
    - Remember delay is the Achilles heel of feedback control
  - It is based on Linux boxes for advanced functions
    - Eg reading packet counters at intervals of 10's of mSecs
- Agreement between simulation, experimental, and mathematical model results is surprising
  - Because the inevitable 'messiness' of implementation is not captured in the mathematical model
  - Suggests an encouraging degree of robustness



An Example of the Experimental Results

Note: - This is the Internal Network Only What about the External Component?

- The experimental results are based on shaping of aggregate UDP streams at the access points
  - This I hardly a practical method
- We need now to consider the interaction between the Control Points and the individual sources.
- Could control TCP or variants by, for example,
  - Placing a buffer of size > N times the window size at the Control Point and shaping the output
    - Very crude - enormous delay - but it works
  - Varying the receiver window size in the Ack packets
    - Has been proposed in the literature
    - Requires individual flow state at the Control Points
    - Possible but not ideal

CSIRO

A Better Solution - XCP Explicit Control Protocol

- Developed by Dina Katabi (MIT) and proposed by MIT in an IETF Draft
  - XCP provides explicit from routers via a field in a congestion header
  - Rate control is separated from fairness
    - An Efficiency Controller in each XCP router determines the aggregate feedback from bandwidth deficit or excess
    - A Fairness Controller divides this amongst the individual flows
      - Places individual feedback value in Acks
      - But does **not** require per flow state
  - Flow control is 'equation based' and the stability characteristics have been explicitly defined

Dina Katabi, Mark Handley, and Charlie Rohrs, "Congestion Control for High Bandwidth - Delay Networks", SIGCOMM'02, August 2002, Pittsburgh

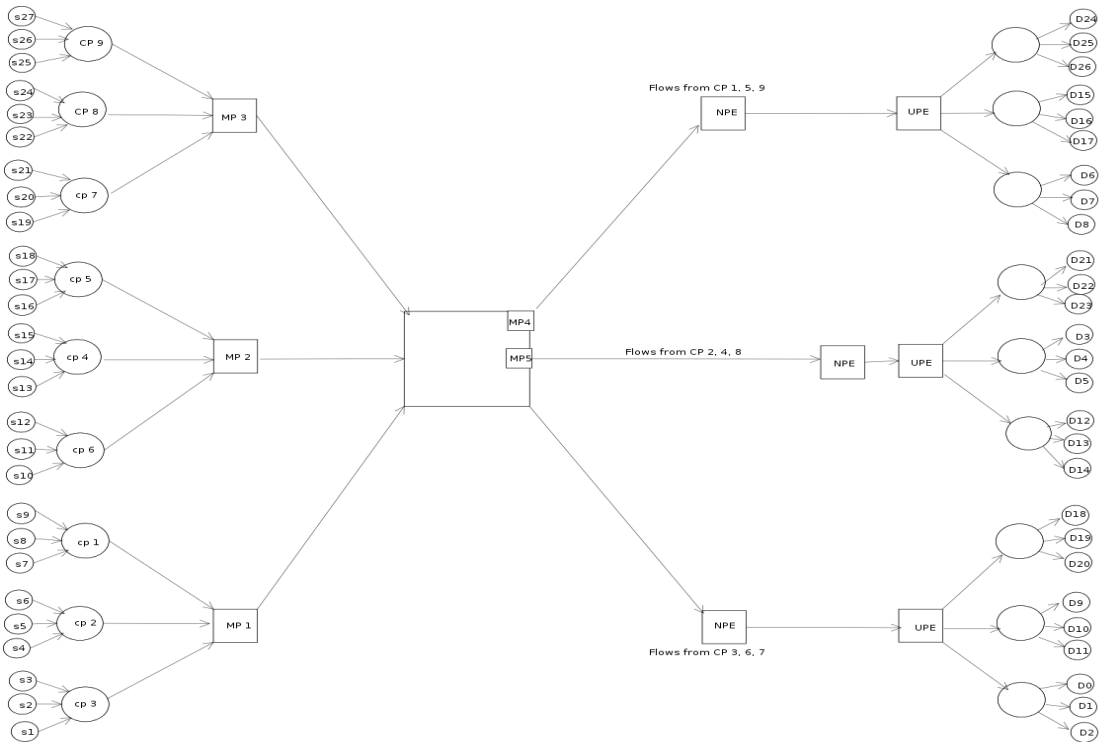
CSIRO

XCP is Just What We Need

- We could use it almost as is
  - The Efficiency Controller is determined by our flow control algorithm
  - The Fairness Controller needed some modification to account for our guaranteed minimum rate
  - Ditto the congestion header
- The Control Point and the internal partition then look like a single XCP router
- We have implemented this as an NS2 simulation of an end to end system
  - We hope to set an experimental implementation as a student project

CSIRO

The Simulation Scenario



CSIRO

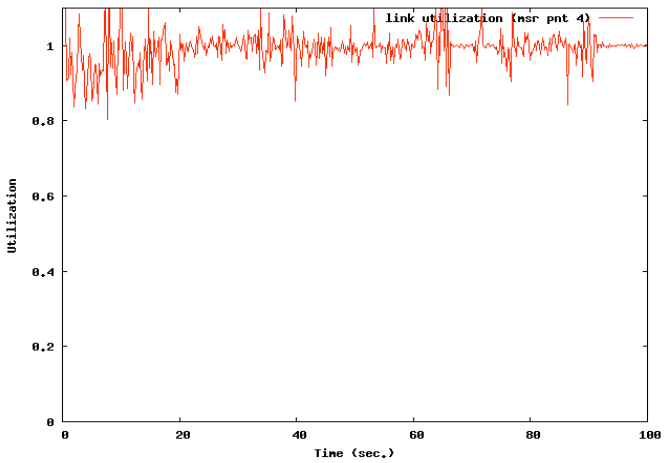
Aggregate Rates at the Control Points

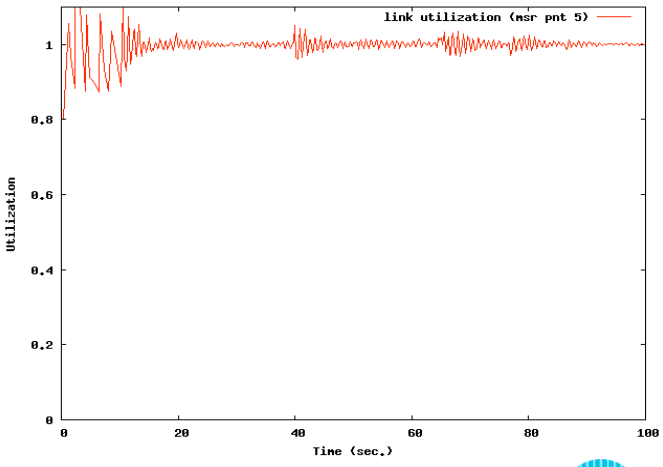
Control Point	Maximum Aggregate Rate (Mbps)	Minimum Aggregate Rate (Mbps)
1	20	12
2	29	26
3	31	25
4	25	23
5	25	14
6	30	20
7	28	16
8	23	15
9	29	20

CSIRO

Ratios of Actual Rates to Target Rates at Measurement Points 4 and 5

- Elastic Class is allocated a target rate at a pipe ingress
- Control system seeks to maintain that target rate





CSIRO