Network Performance Isolation in Data Centres using Congestion Policing

draft-briscoe-conex-data-centre-01.txt

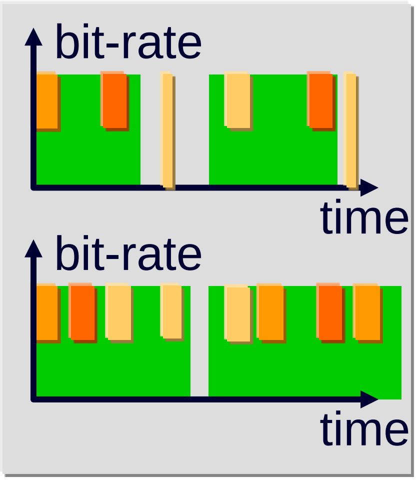
Bob Briscoe, BT

Murari Sridharan, Microsoft

IETF-87 ConEx Jul 2013

Network Performance Isolation in Data Centres using Congestion Policing

- An important problem
  - isolating between tenants, or departments
  - virtualisation isolates CPU / memory / storage
  - but network is highly multiplexed & distributed
- Current solutions
  - assume local interface is the only bottleneck
  - use some form of weighted round robin (or FQ)
  - biases towards heavy hitters (no concept of time)



- Draft is no longer exclusively ConEx
  - title: s/ Congestion Exposure/ Congestion Policing/
  - roadmap: start without ConEx; evolve to exploit gains of ConEx
  - partially solve the problem, then solve it properly with ConEx
- Audience: data centre (private or cloud) people

status of draft

- draft-briscoe-conex-data-centre-01.txt
- Prepared draft-01 in Feb ‘13, but no opportunity to present until now

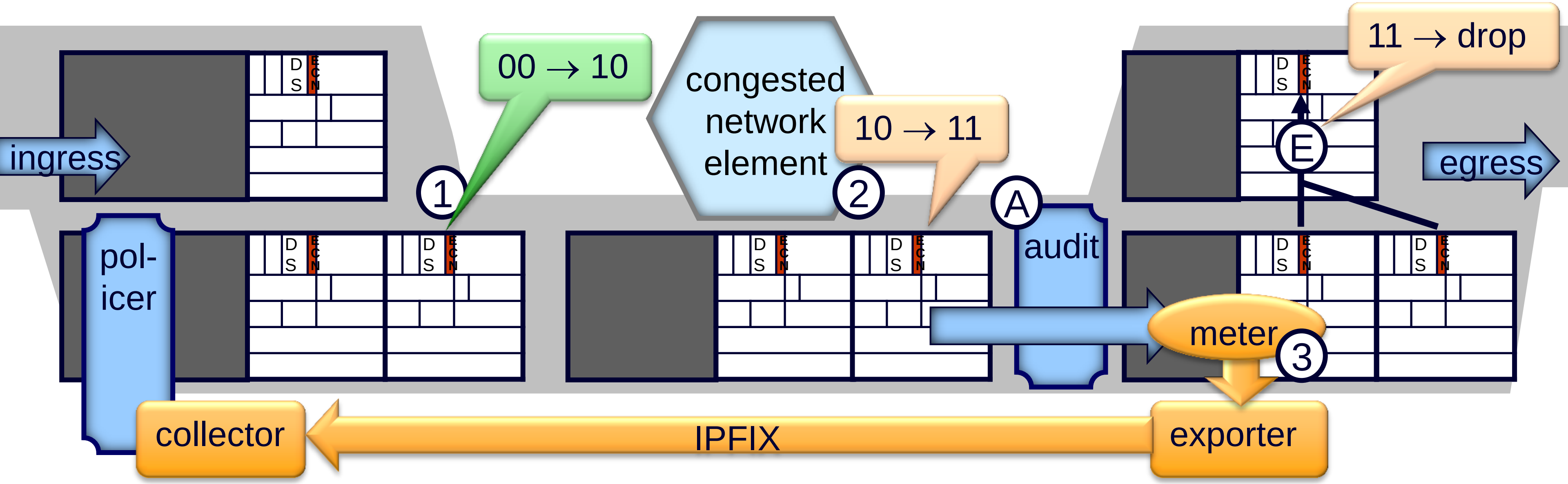
- Cut out huge section (17pp) explaining why congestion policing works
  - Separated out as draft-briscoe-conex-congestion-policing
  - That draft: why / traffic – not specific to data centres
  - This draft: how / engineering – specific to data centres
  - This ‘how draft’ includes a bulleted summary of the ‘why’ draft

- This ‘how’ draft is now a completed write-up of the technology (24pp)
  - Detail design of tunnelling alternative
    - for guest OSs that may not support ConEx or ECN
  - and partial deployment of ConEx solution alongside

- Purpose of this talk
  - seek expert review & WG endorsement
  - before selling in data centre fora

unilateral deployment technique for data centre operator

- exploits:
  - widespread edge-edge tunnels in multi-tenant DCs to isolate forwarding
  - a side-effect of standard tunnelling (IP-in-IP or any ECN link encap)



- for e2e transports that don't support ECN, the operator can:
  - ① at encap: alter 00 to 10 in outer
  - ② at interior buffers: turn on ECN
- defers any drops until egress Ⓔ
- audit Ⓐ just before egress can see packets to be dropped

- for e2e transports that don't support ConEx, the operator can create its own trusted feedback:
  - ③ at decap: *only* for Not-ConEx packets, feedback aggregate congestion marking counters:
    - CE outer, Not-ECT inner = loss
    - CE outer, ECT inner = ECN

designed for evolution to ConEx

- deployable now, unilaterally by data centre operator
  - without ConEx or ECN support in guest operating systems
- but uses ECN or ConEx from any OS that supports either

- advantage of ConEx over tunnelled feedback
  - isolation: ConEx polices short flow congestion & slow-start overshoot
    - tunnel feedback arrives too late to police all this (lacks credit facility)
  - efficiency: tunnel feedback duplicates e2e transport feedback
  - security: ConEx & ECN are inherently bound into the transport flow
    - tunnel feedback would need added message authentication

plans

- intent: present in other working groups at next IETF (e.g. NVO3)
- working group item?

working group input

- review please

Network Performance Isolation in Data Centres using ConEx

draft-briscoe-conex-data-centre-01.txt

Q&A

& spare slides

Features of Solution

- Network performance isolation between tenants
- No loss of LAN-like multiplexing benefits
  - work-conserving
- Zero (tenant-related) switch configuration
- No change to existing switch implementations
  - if ECN-capable
- Weighted performance differentiation
- Simplest possible contract
  - per-tenant network-wide allowance
  - tenant can freely move VMs around without changing allowance
  - sender constraint, but with transferable allowance
- Transport-Agnostic
- Extensible to wide-area and inter-data-centre interconnection

document structure

- Frontpieces (Abstract, Intro)

2. Features of Solution
3. Outline Design
4. Performance Isolation: Intuition
5. Design
6. Incremental Deployment
7. Related Approaches

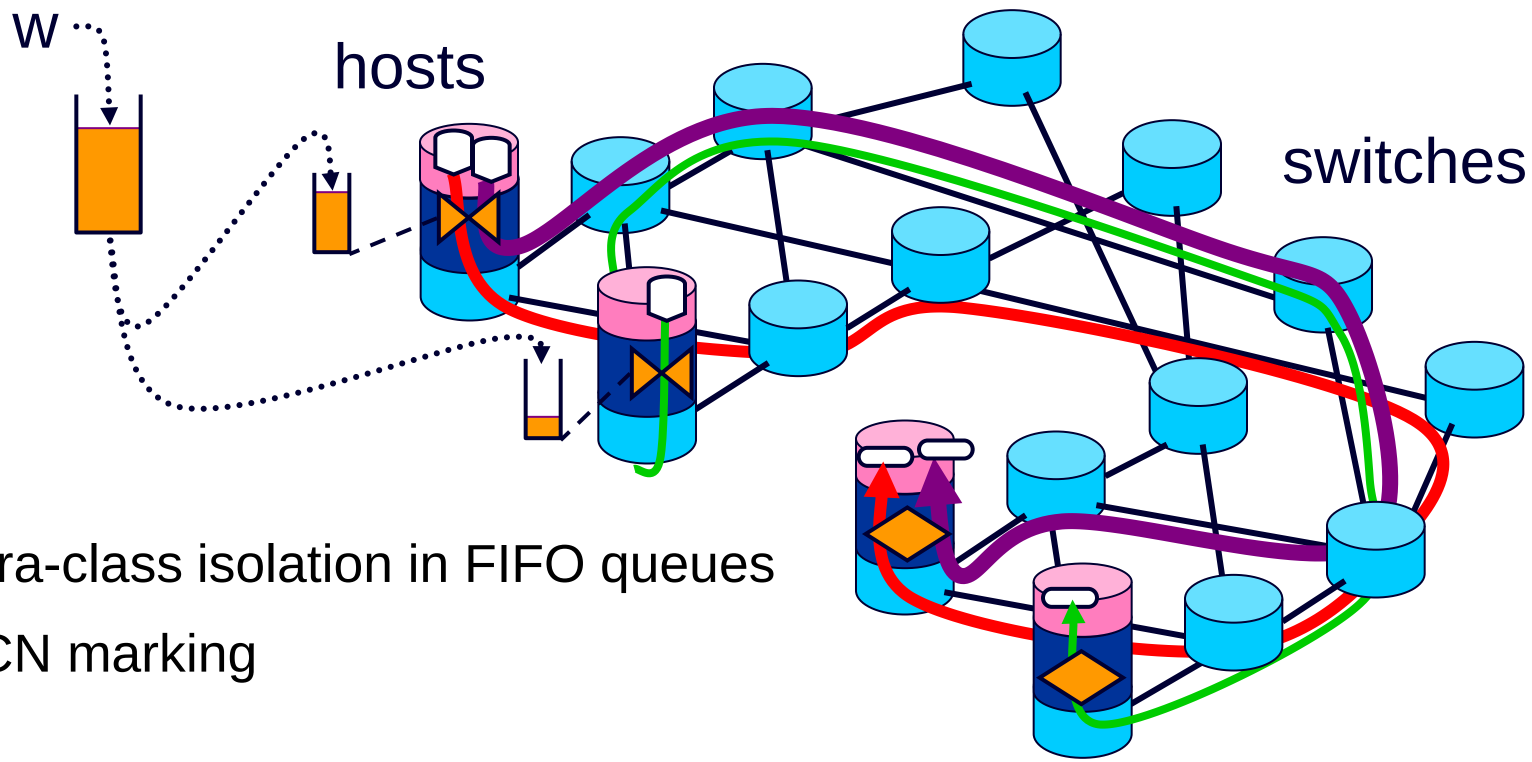
- Tailpieces (Security, Conclusions, Acks)

Outline Design

- Edge policing like Diffserv
  - but congestion policing
- Hose model
- Flow policing unnecessary, but optional

⬡ VM sender
⬭ VM receiver
⋈ congestion policer
◆ audit

guest OS
hypervisor
switching



- intra-class isolation in FIFO queues
- ECN marking