

Network Performance Isolation in Data Centres using ConEx

[draft-briscoe-conex-data-centre-00.txt](#)

Bob Briscoe, BT

Murari Sridharan, Microsoft

IETF-84 ConEx Jul 2012

draft status

- Network Performance Isolation in Data Centres using ConEx
[draft-briscoe-conex-data-centre-00.txt](#)
- new individual draft Jul 2012, requested by w-g Mar 2012
- one of a (growing) set of ConEx deployment arrangement drafts
- largely complete (31pp)
 - another rev to fill a few ToDo's
 - technical ideas complete, but 2 works in progress:
 - detail design of tunnelling alternative for guest OSs that may not support ConEx or ECN
 - parameter setting section
- purpose of this talk
 - generate more interest in readers & reviewers

audience

- data centre (private or cloud) people, not ConEx
 - not written as a way to deploy ConEx
 - rather two ways to solve the isolation problem
 1. ConEx: better (warrants “using ConEx” in title)
 2. tunnelling: works for non-ConEx guest OSs, but inferior
- audience assumed sceptical
 - how it works is simple
 - why it works is outside people’s comfort zones
 - isolate tenants with no per-tenant config on the switches?

document structure

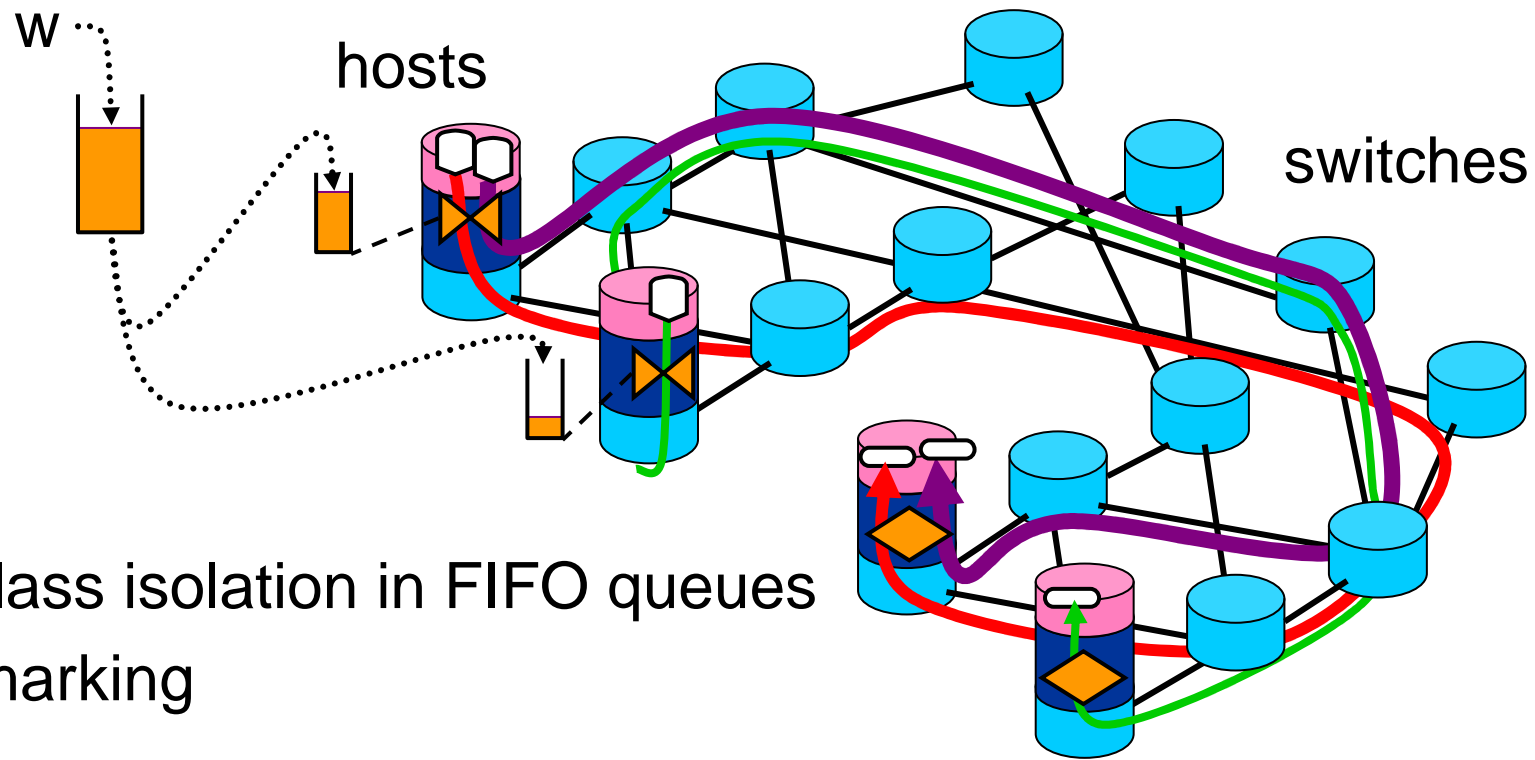
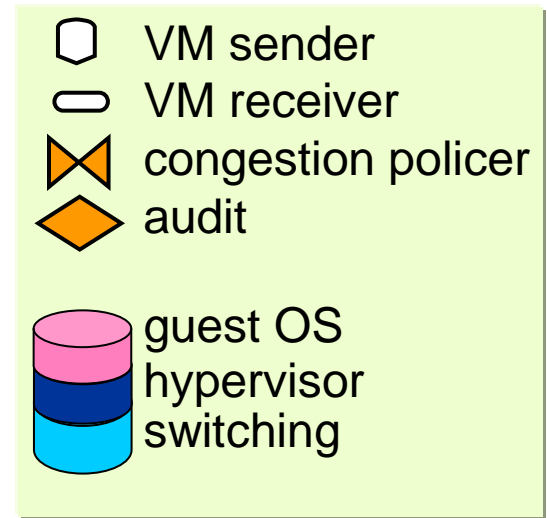
- Frontpieces (Abstract, Intro)
- 2. Features of Solution
- 3. Outline Design
- 4. Performance Isolation: Intuition
- 5. Design
- 6. Parameter Setting
- 7. Incremental Deployment
- 8. Related Approaches
- Tailpieces (Security, Conclusions, Acks)

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

Outline Design

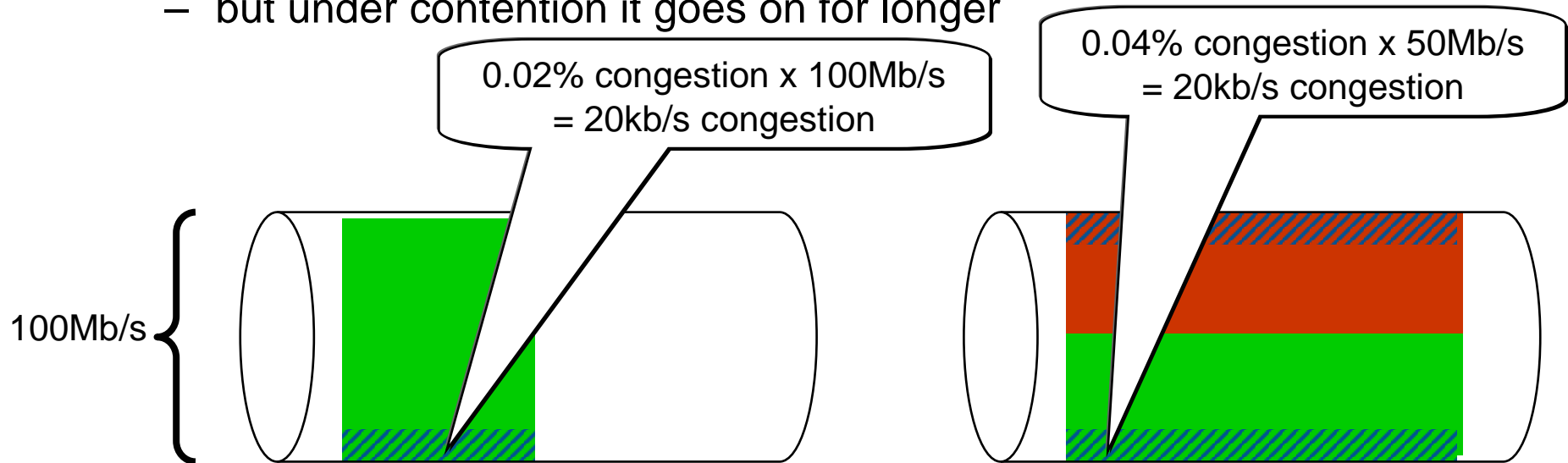
- Edge policing like Diffserv
 - but congestion policing
- Hose model
- Flow policing unnecessary, but optional



- intra-class isolation in FIFO queues
- ECN marking

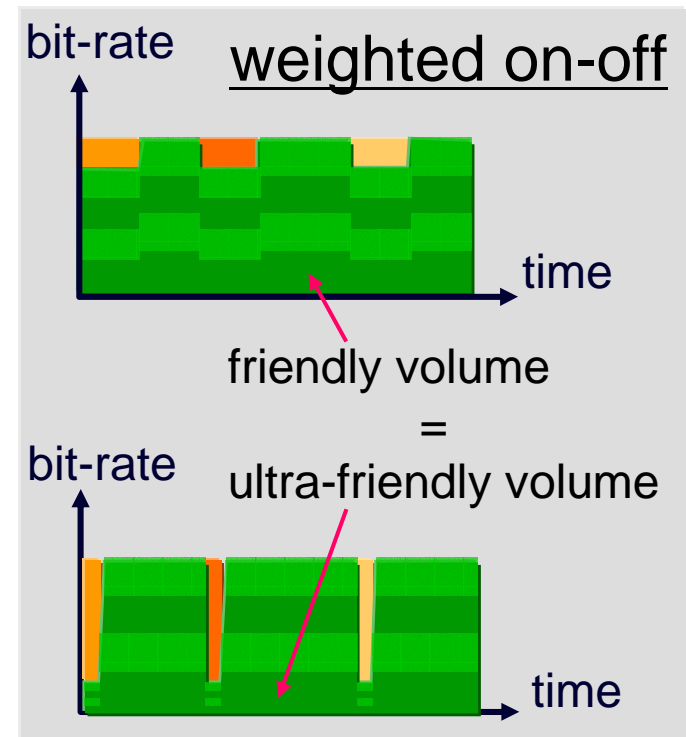
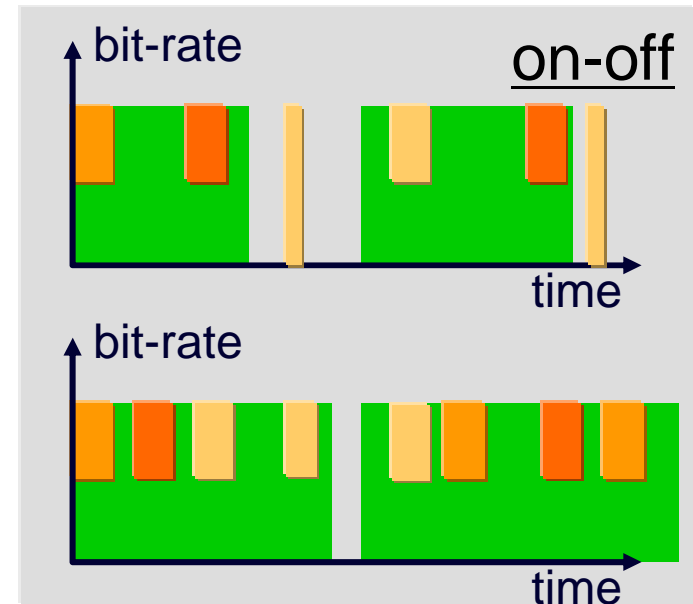
performance isolation intuition

- congestion policer enforces bit-rate, $x = w/p$, where
 - w is a constant for that tenant (the policy)
 - p is % congestion
- similar to a so-called ‘scalable congestion control’
 - but for aggregate (hose) from source, made up of flows
 - TCP is evolving towards this (Compound, Cubic, DCTCP etc)
- property easy-to-say but hard to grasp:
 - same rate of congestion per tenant, however many other tenants capacity is shared with
 - congestion-bit-rate in one flow is the same
 - but under contention it goes on for longer



intuition built-up as follows

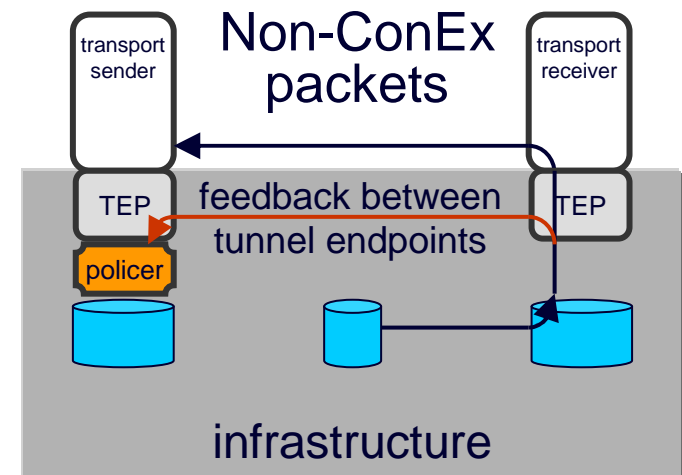
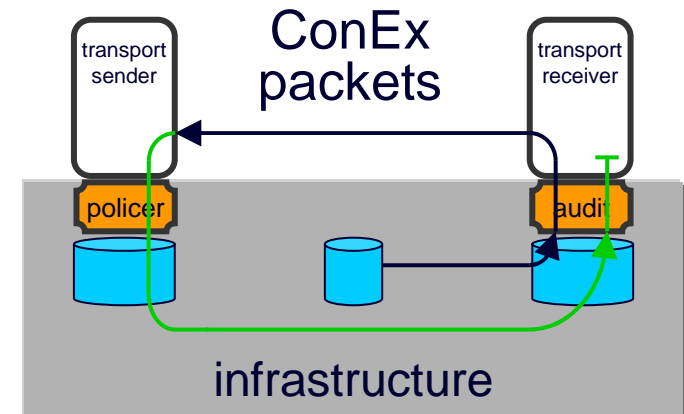
- ‘scalable congestion control’ as boundary case (previous slide)
- single link
 - long running flows, single link
 - similar to (weighted) round-robin
 - **on-off** flows
 - congestion-volume accounts for how often a tenant is *not* ‘on’
 - **weighted on-off** flows
 - longer flows shift away
- network of links
 - congestion-volume allows for how many links tenant is ‘on’ in
- transients



work in progress & innovations

incremental deployment

- Can deploy all infrastructure under control of one administration
- ConEx & ECN depend on sender and receiver in guest OS
 - trusted feedback tunnel back to policer
 - without ConEx or ECN in guest OS
 - under control of DC operator
 - concrete example builds on NV-GRE
- Hybrid
 - non-ConEx packets: feedback tunnel
 - non-ECN packets: feedback tunnel
 - ConEx packets: no feedback tunnel
- tunnel egress, if not-ECT on inner header
 - feedback CE to ingress
 - drops any ECN-marked packets
- tunnelling inferior:
 - less isolation (congestion knowledge delayed by RTT)
 - more complicated (tunnel feedback set-up)
 - less efficient (duplicates TCP feedback)
- reward ConEx at policer for being more efficient?



work in progress & innovations
interconnection

- DC operator buys WAN pipe
 - between data centres
 - to enterprise customers' (uncontended) LANs
- within pipe tenants share as within DC
 - based on loss or ECN at ingress to pipe
 - just another internal link

plans

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

working group input

- could the “intuition” section (16pp) be a stand-alone draft?
 - already well-summarised in the introduction
- review please

Network Performance Isolation in Data Centres using ConEx

[draft-briscoe-conex-data-centre-00.txt](#)

Q&A