

Network Performance Isolation in Data Centres using ~~ConEx~~ Congestion Policing

[draft-briscoe-conex-policing-01](#)

[draft-briscoe-conex-data-centre-02](#)

Bob Briscoe
Chief Researcher, BT
IRTF DCLC Jul 2014



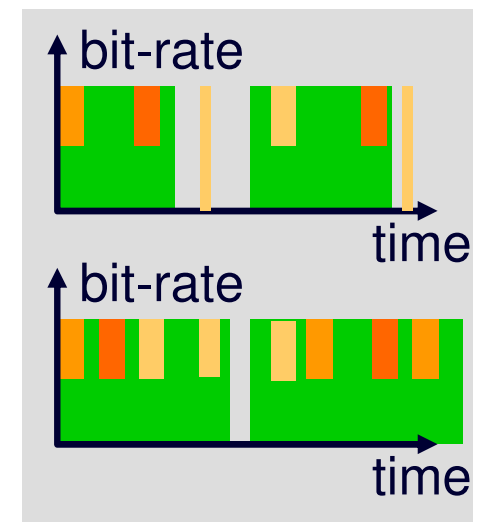
Bob Briscoe's work is part-funded by the European Community under its Seventh Framework Programme through the TrilogY 2 project (ICT-317756)

purpose of talk

- work proposal for the data centre latency control r-g
 - data centre queuing delay control
 - designed for global scope (inter-data-centre,... Inter-net)
 - this talk: adds first step: intra-data centre
 - without any new protocols
- started in the IETF congestion exposure (ConEx) w-g
- generalised for initial deployment without ConEx
 - and even without ECN end-to-end
 - now even without ECN on switches (in slides, not draft)

Network Performance Isolation in Data Centres

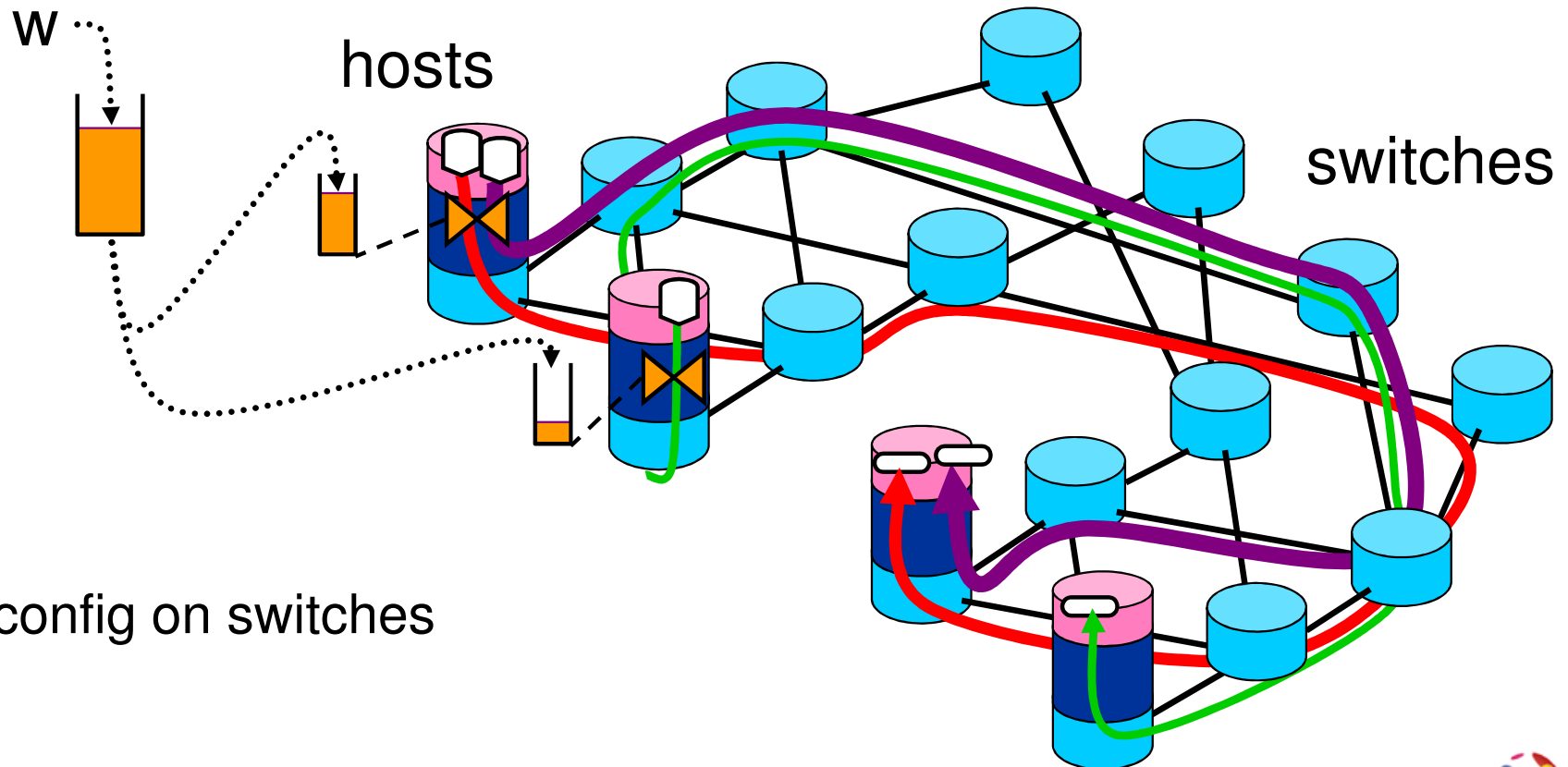
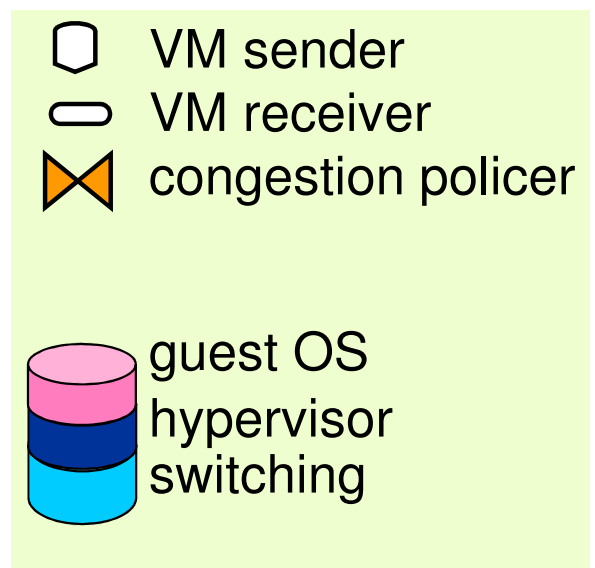
- An important problem
 - isolating between tenants, or departments
 - virtualisation isolates CPU / memory / storage
 - but network and I/O system is highly multiplexed & distributed
- SDN-based (edge) capacity partitioning*
 - configuration churn: nightmare at scale
 - poor use of capacity
- edge-based weighted round robin (or WFQ)
 - More common
 - but biases towards heavy hitters (no concept of time)



Outline Design – First Step

edge bottlenecks by capacity design

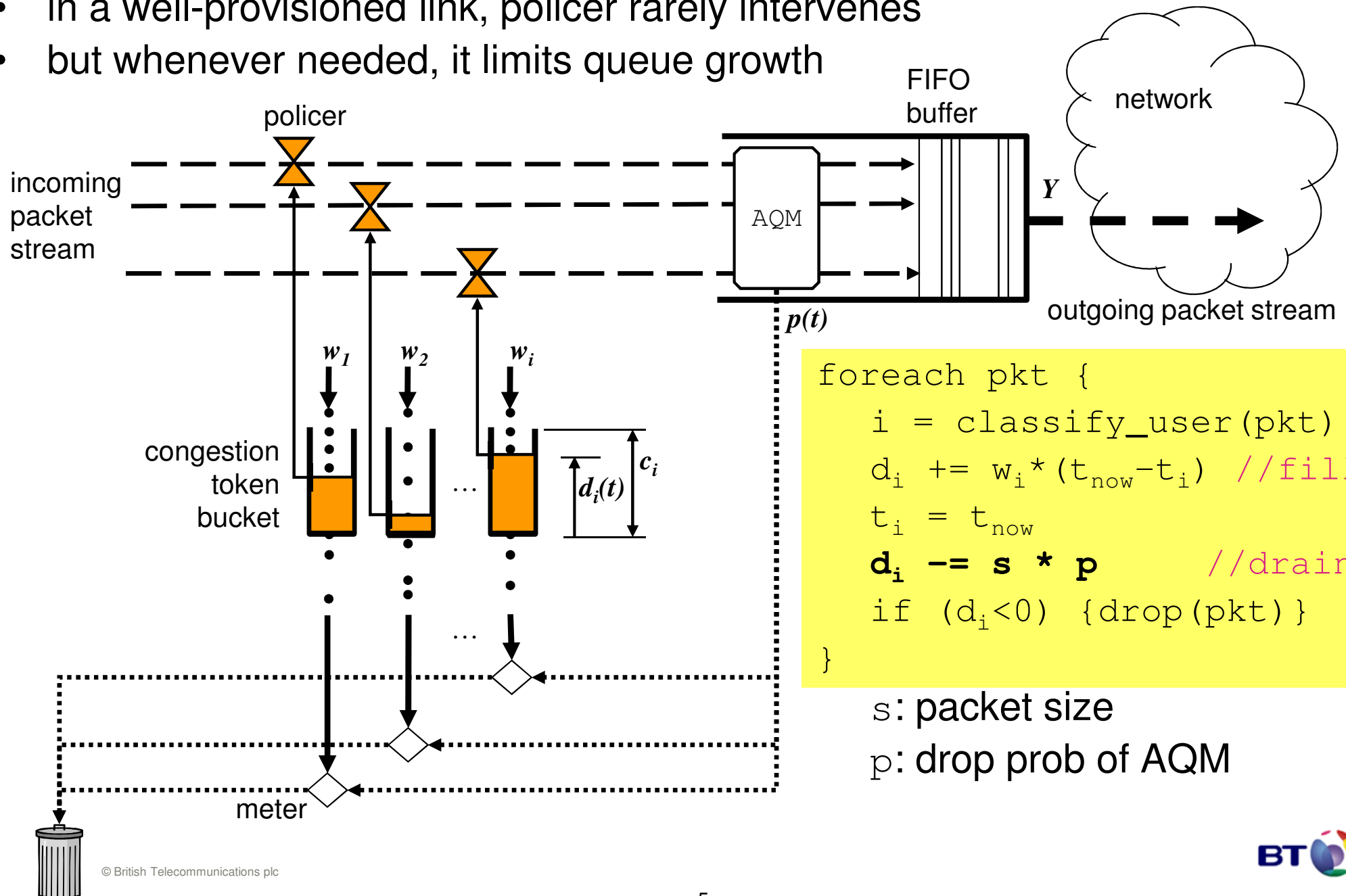
- Edge policing like Diffserv
 - but congestion policing (per guest)
- isolation within FIFO queue



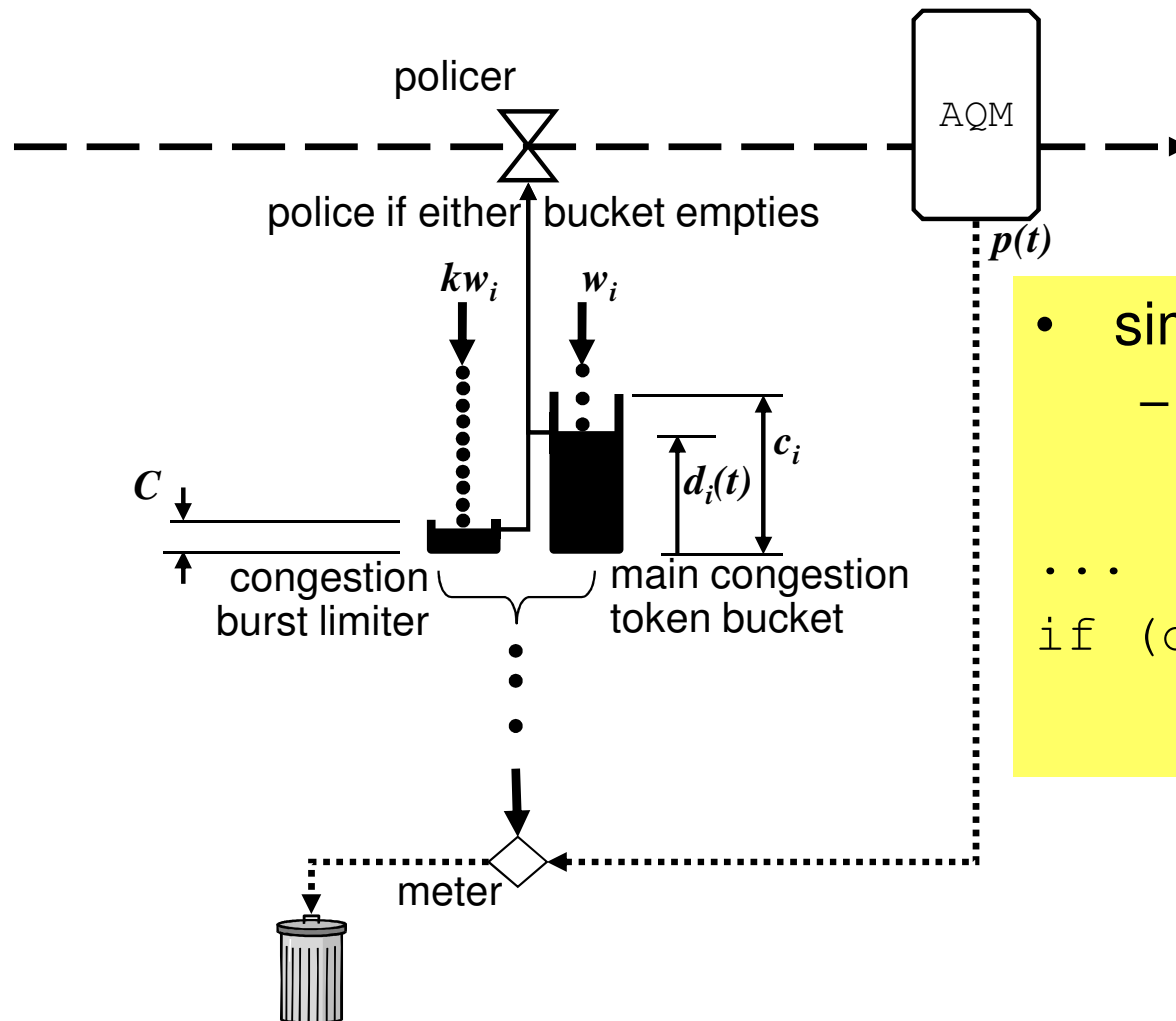
- no config on switches

bottleneck congestion policer

- in a well-provisioned link, policer rarely intervenes
- but whenever needed, it limits queue growth



actually each bucket needs to be two buckets to limit bursts of congestion



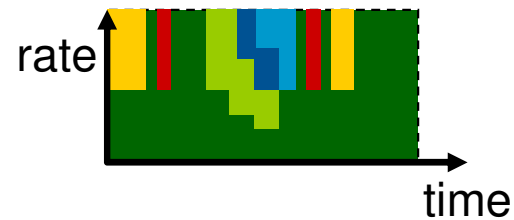
- similar code
 - except 2 token buckets

```

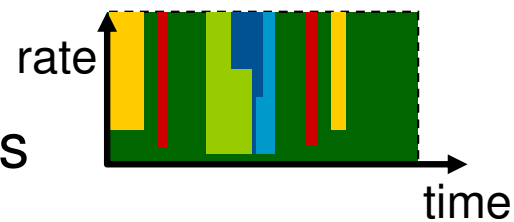
...
if (di1<0 || di2<0)
    {drop(pkt)}
    
```

performance isolation outcome

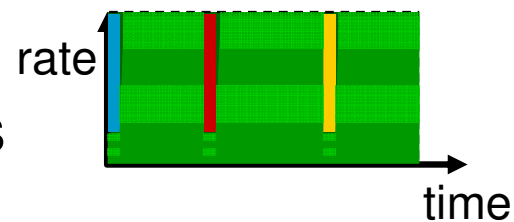
- WRR or WFQ



- congestion policer
– with unequal traffic loads



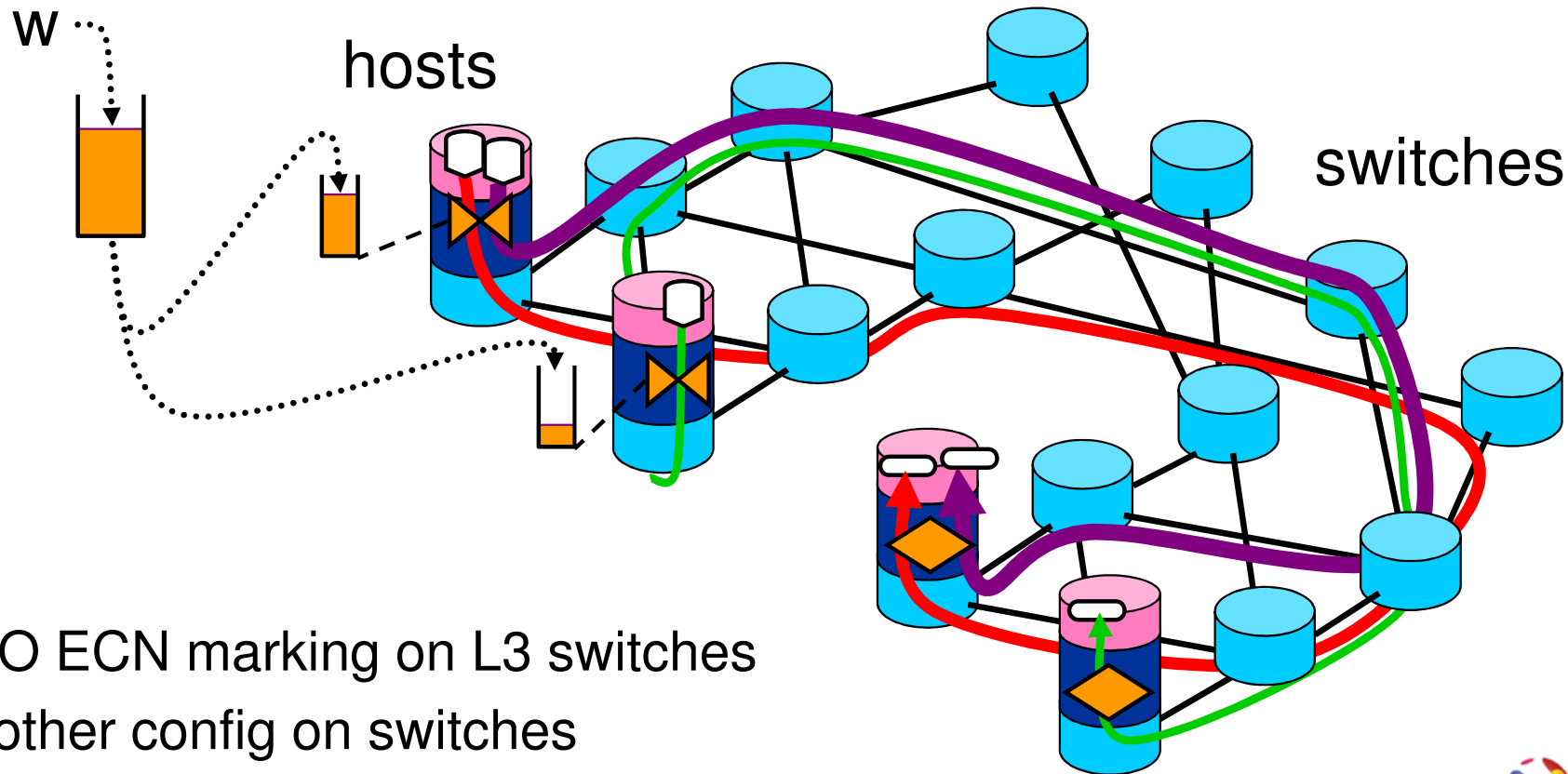
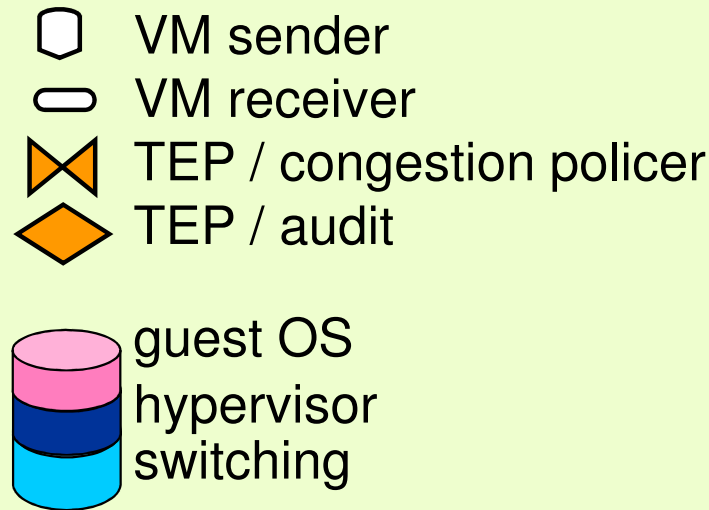
- congestion policer
– treats equal traffic loads equivalently to WRR



Outline Design

edge and core queue control

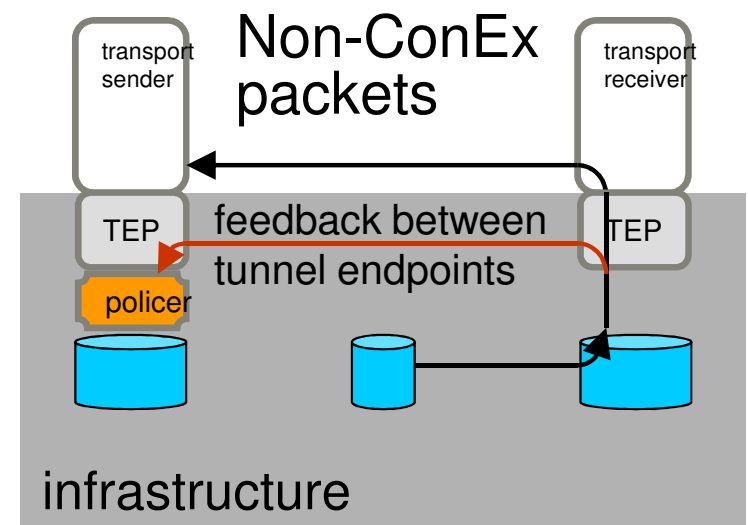
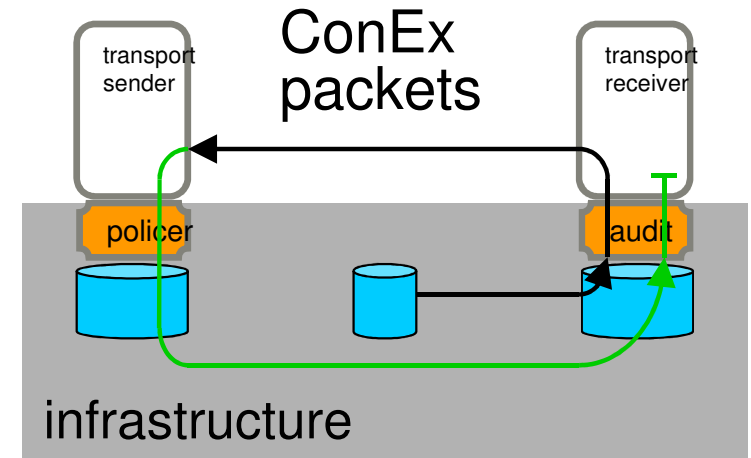
- Edge policing like Diffserv
 - but congestion policing (per-guest)
- Hose model
- intra-class isolation in all FIFO queues



- FIFO ECN marking on L3 switches
- no other config on switches

trusted path congestion feedback

- Initial deployment
 - all under control of infrastructure admin
- ECN on guest hosts: optional
 - ECN enabled across tunnel
- ConEx on guest hosts: optional
 - any ConEx-enabled packet doesn't require tunnel feedback
- details – see spare slide or draft



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
- 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 interconnect

call for interest

- implementation in hypervisors
- evaluation

Network Performance Isolation in Data Centres using congestion policing

[draft-briscoe-conex-policing-01](#)

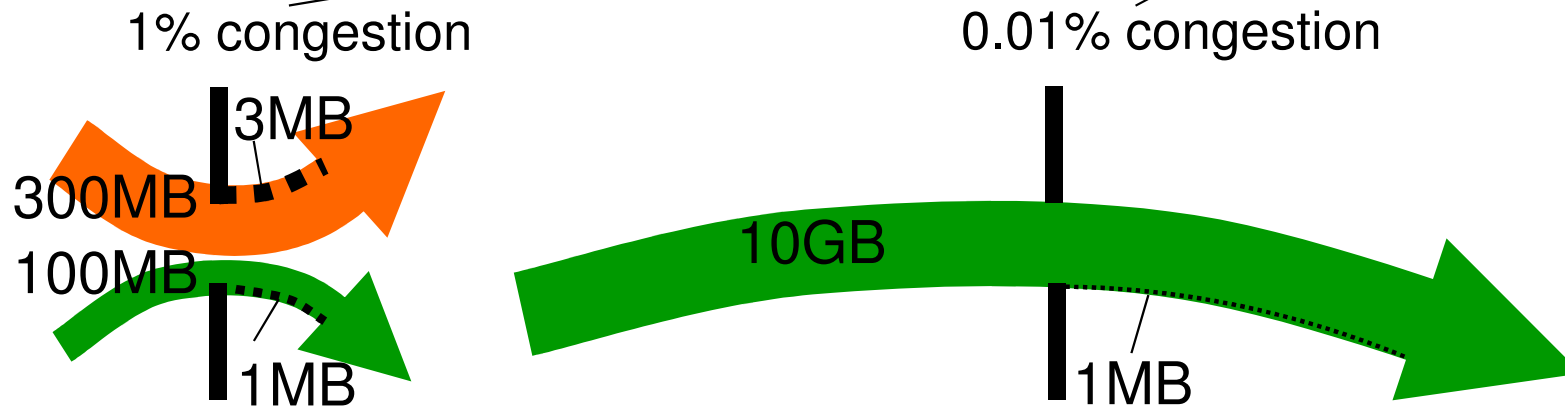
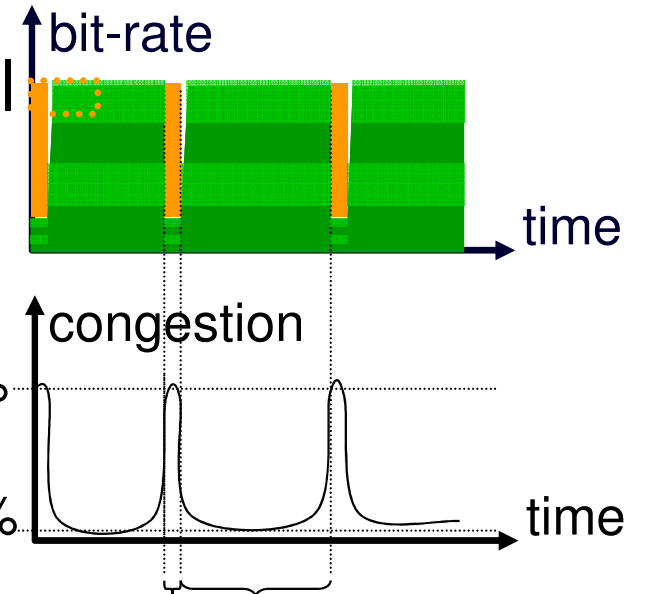
[draft-briscoe-conex-data-centre-02](#)

Q&A

& spare slides

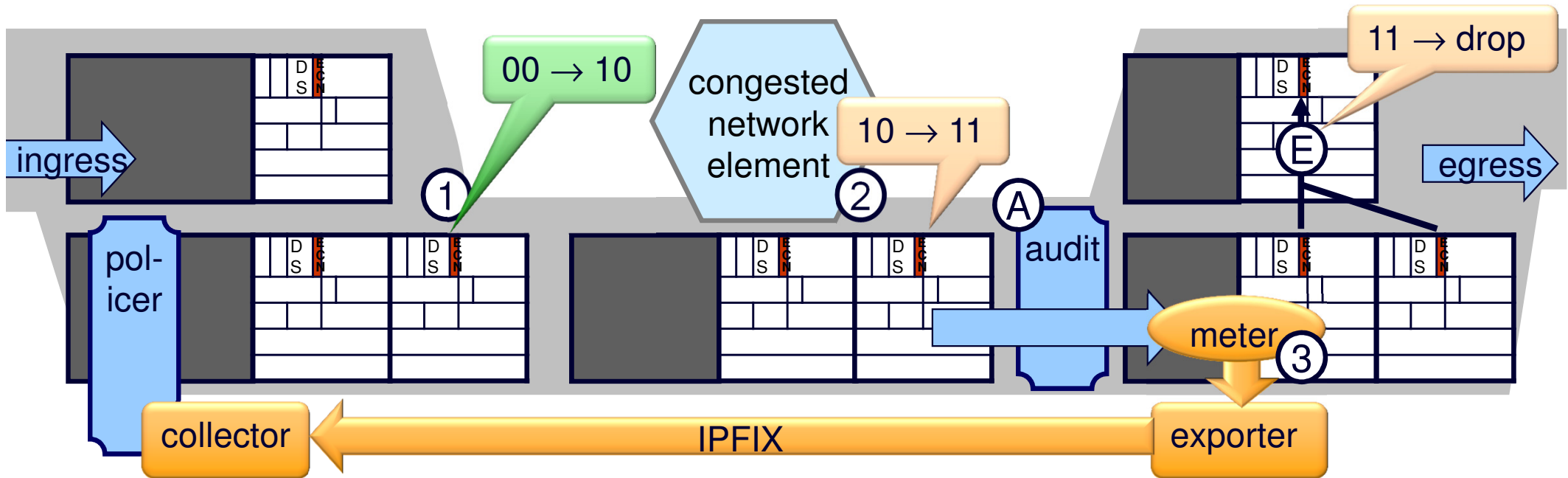
measuring contribution to congestion

- = bytes weighted by congestion level
- = bytes dropped (or ECN-marked)
- = 'congestion-volume'
- as simple to measure as volume



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 (E)
- audit (A) 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