Further considerations on data center congestion control

IETF89@London
denglingli@chinamobile.com

Outline

- Review on TCP CC in Internet DCs
- Discussion on CC in Operator DCs

Recap on E2E Congestion Control

- Internet
  - achieve convergence in multiple round trips
  - fairness among (legacy) flows
- Data center
  - performance highly sensitive to latency variation
  - differentiation rather than fairsharing
    - explict delay/priority indication from apps

TCP Perf. Issues in DCs (1/3)

- Incast collapse
  - Problem: Perf. drops as # of concurrent substasks increases
    - perf. indicator: # of concurrent working servers allowed w/o perf. loss
  - Triggers: timeout/slow start from repeated losses
    - counter measures:
      - » Req1: reduce # of loss/timeout
      - » Req2: mitigate perf. impact from loss/timeout
  - Roots: shallow-buffer@ToR switches
    large # of synchronized short flows
    - counter measures:
      - » enlarge buffer? no good for other issues.
      - » break traffic synchronization? depending on app, may decrease overall delivery performance by adding extra delay.

TCP Perf. Issues in DCs (2/3)

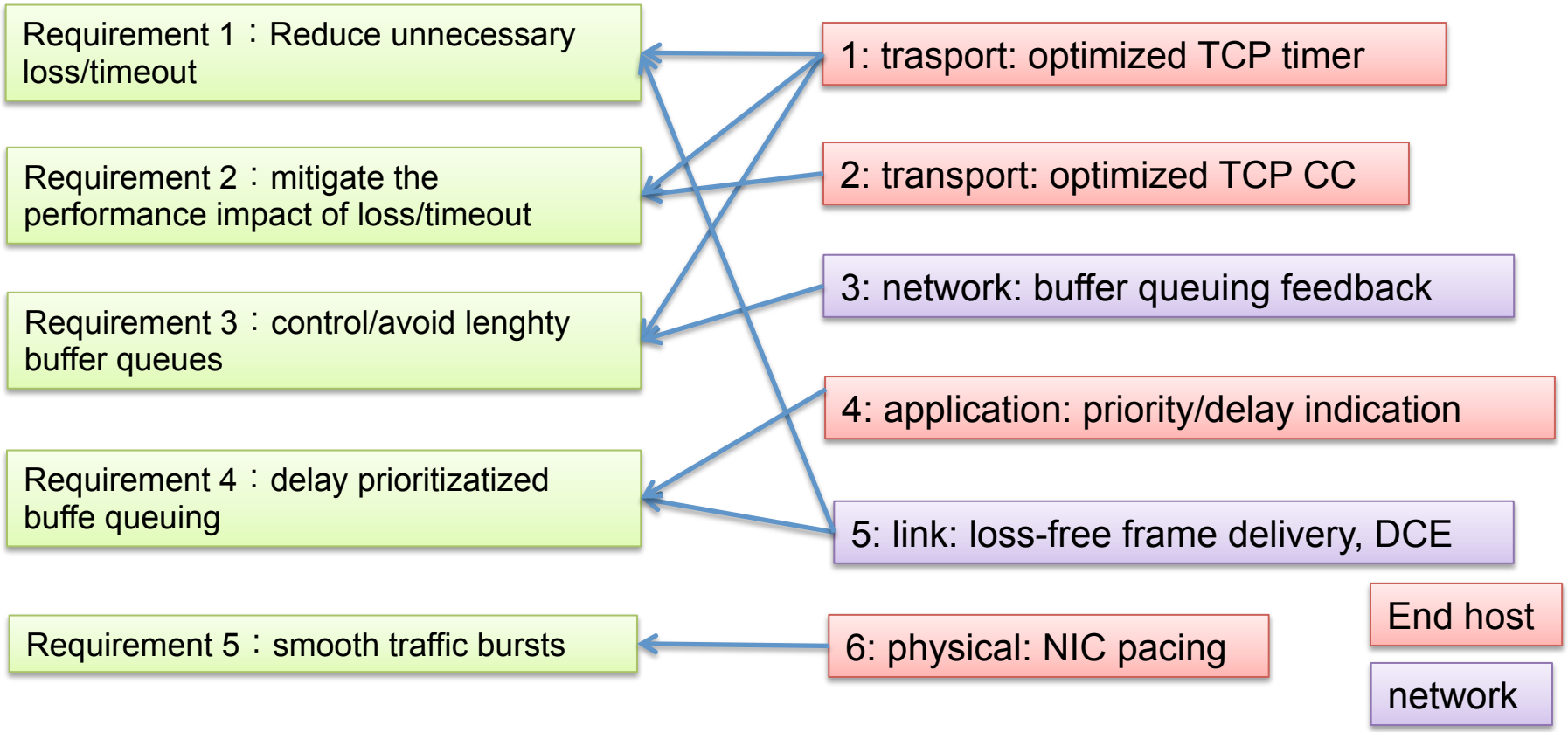
- Long Tail of RTT
  - Problem: delay-sensitive short flows suffer from long RTTs
    - perf. indicator: variation of flow RTTs
  - Trigger: buffer queuing
    - counter measures:
      - » Req3: control the length of or even eliminate buffer queues
  - Root: existence of greedy long flows
    - counter measures:
      - » Req4: delay prioritizatized buffer queuing

TCP Perf. Issues in DCs (3/3)

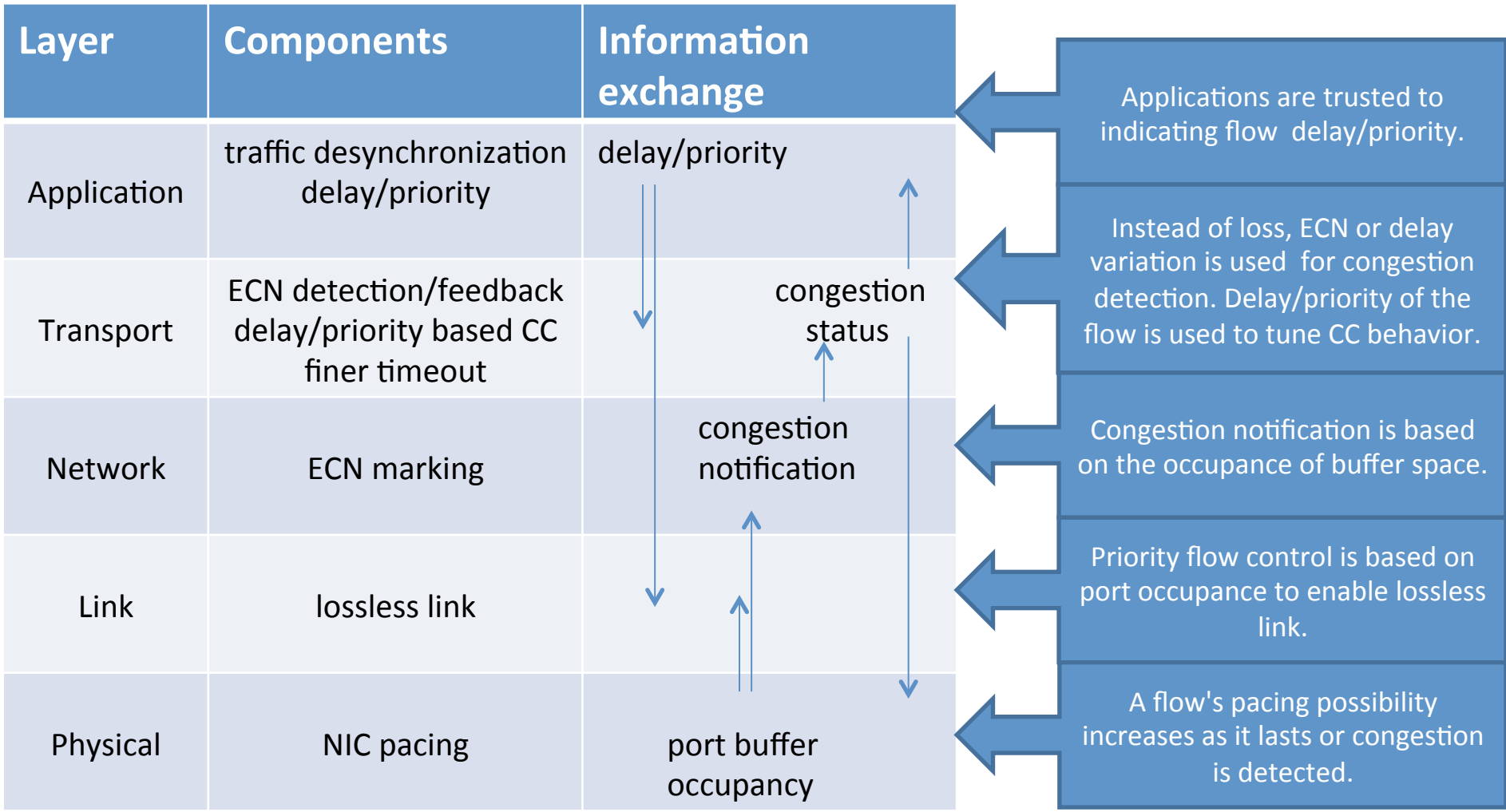
- Buffer Pressure
  - Problem: bursty delay-senstive flows suffer from shortage of available buffer space
    - Perf. indicator: length of buffer queues
  - Trigger: loss from buffer bloat
    - counter measures:
      - » Req3: control/avoid lenghty buffer queues
      - » Req5: smooth traffic bursts
  - Roots: existence of greedy long flows
    burstieness of delay-sensitive flows
    - counter measures:
      - » Req4: delay prioritizatized buffer queuing

TCP Perf. Enhancement in DCs

Since both hardware device and software stacks is usually highly customized by a single DC onwer, there have been various private solutions for these issues, including cross-layer, cross boundary (network+end host) hybrid ones.



Generalized Cross-layer e2e CC in DCs

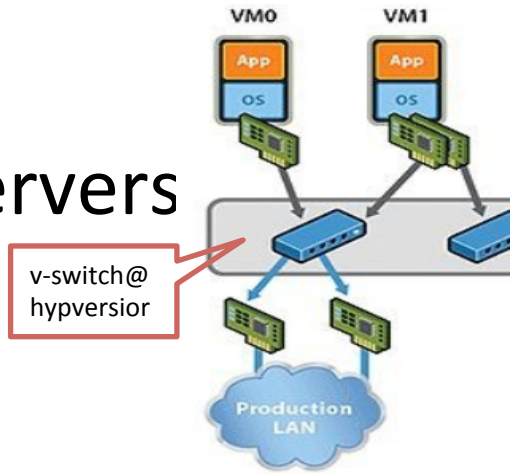


There has been work on more accurate ECN feedback for DCs. draft-ietf-tcpm-accecn-reqs-05
Proposal to add latency/priority specification into transport API. draft-deng-taps-datacenter-01

Network Operator's Data Centers

- Internet Service Resource Introduction
  - reduce opex for interworking traffic
  - increase UoE by feeding requests locally
  - multi-tenant resource sharing
- Network Funcation Virtualization (NFV)
  - reduce capex for dedicated hardware/software
  - ensure reliability goals through pooling/migration
  - increase managability by centralized routing control

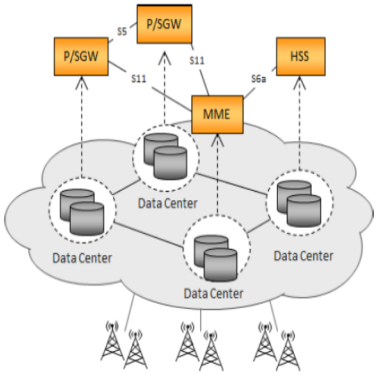
Example-1: V-IDC

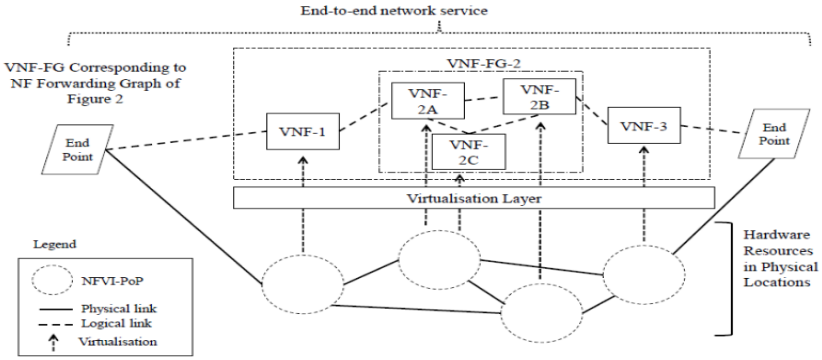


- providing VM instead of physical servers
  - It seems easy if we map the solution with virtual switch architecure provided by local hypervisor and apply e2e cc via unmodified VMs, but does it?
- latency-bounded traffic types in a V-IDC
  - frond-end production traffic among VMs
    - e.g. small web site hosted as a virtual-dc tenant in operator's IDC
    - issue1: latency drifting with VM timer
  - VM management traffic among physical servers
    - e.g. life-cycle management, migration, etc.
    - issue2: latency specification and cc by hypervisors rather than VM

Example-2: Network Function Virtualization

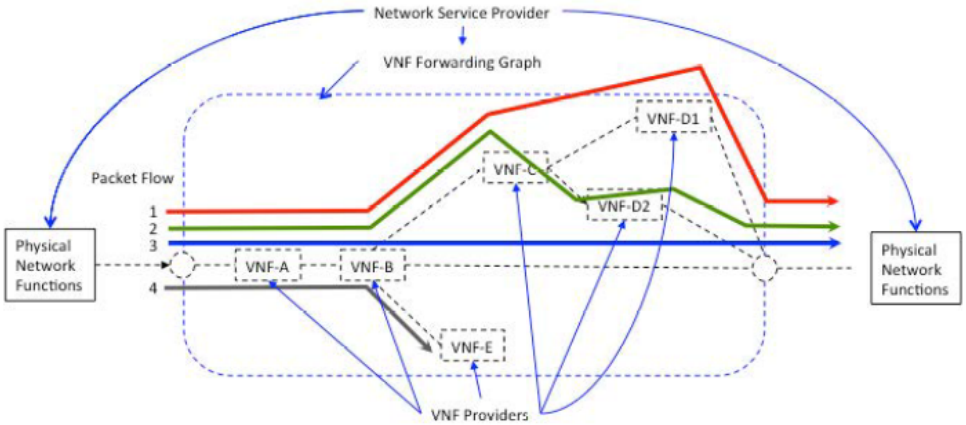
- V-EPC: tunneled traffic across DCs





•Core network infrastructure are virtulized and running on commodity platforms.

- SFC: dynamic tunneled traffic within DCs



•Flow specific network service functions are decoupled from dedicated network device, as chained VMs on the path of tunneled traffic.

•DPI/charging/PEP/header enrichement/ NAT/FW/etc.

- Issue3: A NFV DC is only a segment of the e2e data traffic.

Congestion Management in Operator's DCs

- Similarities to SP IDC
  - low-cost general platform servers/switches
  - performance sensitive to internal congestion
- Differences from SP IDC/Internet
  - VM vs physical
    - hypervisor be a potential congestion/mangement point
  - Tunnel vs E2E
    - e2e CC is suspected to be not responsive enough
  - Policying vs fairsharing
    - more intelligence network provision through flexible control

Discussion: What may help?

- Case 1: hypervisor-involved distributed CC
  - specify VM-hypervisor(v-switch) & hypervisor-hypervisor interaction for CC in terms of VM-VM/host-host traffic
- Case 2: segment CC for tunneled e2e traffic
  - enable intermediary congestion feedback/control for tunneling traffic by hypervisor/VM
  - explore its interaction with e2e CC
- Case 3: status exposure for centralized management
  - instead of VM, more accurate/objective congestion feedback can be provided by hypervisors

Open discussion invitation

- Topic
  - the state-of-art solutions for DC CC
  - relevant work/considerations on virtualized DC CC
  - anything you would like to share
- Tentative arrangement
  - Friday 13:00-14:00, room: TBD
- Contact: denglingli@chamobile.com
- Food/Drink provided:-)