draft-ietf-mpls-forwarding-02

MPLS Forwarding Compliance and Performance Requirements

Curtis Villamizar (OCCNC)
Kireeti Kompella (Contrail)
Shane Amante (Level 3)
Andrew Malis (Verizon)
Carlos Pignataro (Cisco)

Note: Authors believe this version is ready for WGLC.

Two Parts to Presentation Slides

- Problem addressed by this work

- Backup Slides - not presented
  - (solution oriented)

Motivation

- Initial Motivation
  - Common mistakes among chip makers with limited MPLS experience
- Later Motivation
  - Missed requirements among chip makers and system makers
  - High cost of not getting it right for -
    - chip makers - system makers - deployed base

High cost of not getting it right

- cost to chip vendor
  - may be transitioning from Layer-2 only to +IP to +MPLS
  - mistakes may result in respin (costly) or redesign (worse)
  - system designers don't want the older (buggy) chip
- cost to system vendor
  - may need a chip upgrade or even worse change chip sets
  - customer (SP or other) may not want the older cards
  - may result in large scale free or low cost card swap
- cost to deployed base
  - too often problems are found after deployment
  - bugs can hinder deployment of new capabilities or services
  - may be stuck with bugs if caught after evaluation period
  - some faulty access equipment may be around for a long time

Scope

- In scope
  - MPLS forwarding
  - base PW forwarding + CW and sequence
  - MPLS OAM + MPLS-TP OAM
  - multipath and load balancing entropy
  - recommendations on fast path vs slow path OAM
  - DoS protection
- Out of scope
  - specific PW AC and NSP
  - PW applications such as various forms of VPN
  - load balancing of tunneling protocols within IP
  - MPLS over other (ie. GRE, L2TP, UDP)
  - implementation details

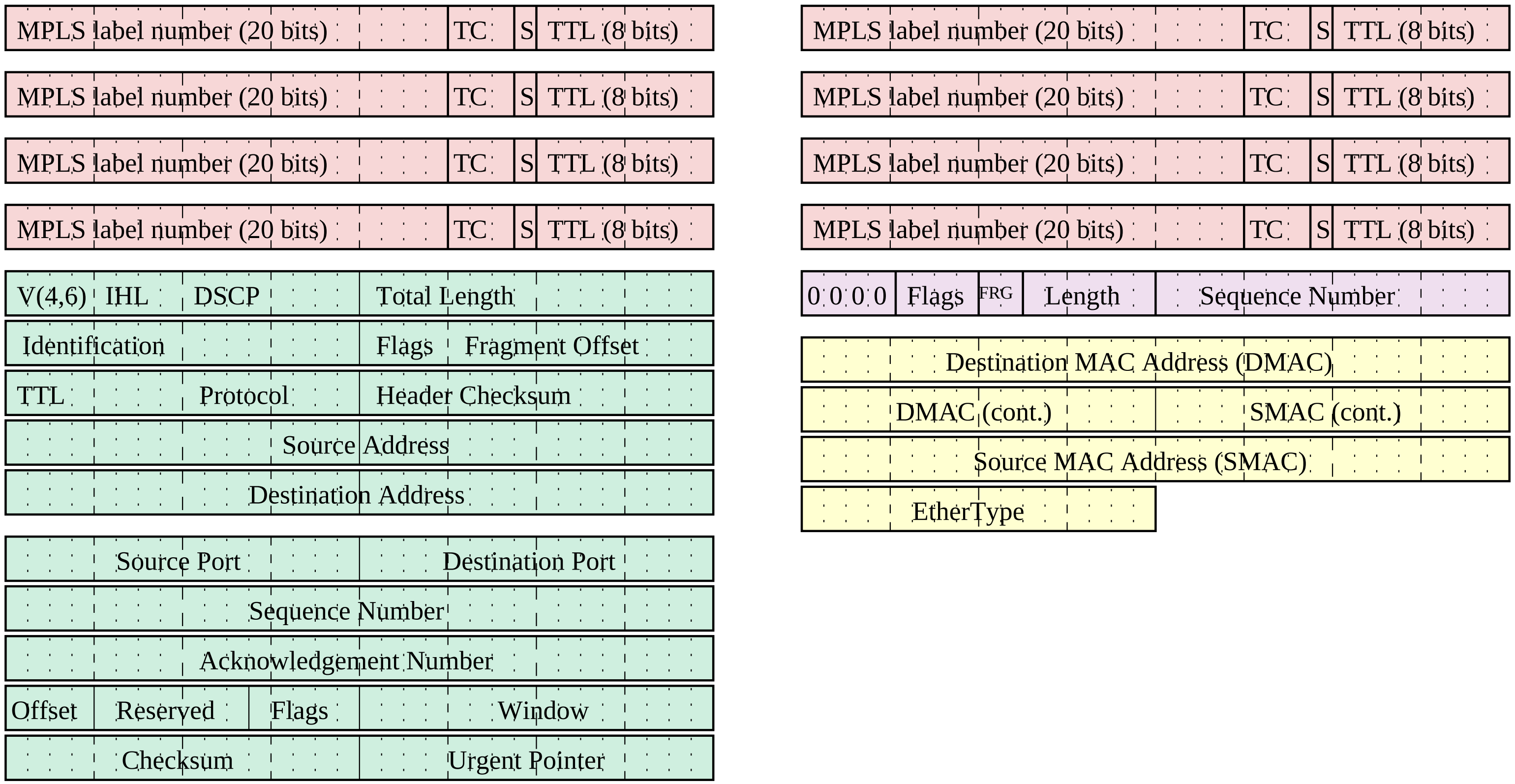
Spotlight on Specific Problems

- Deep Stack Problems
- Lack of PW CW support in edge equipment
- Small Packet Burst Tolerance
- Packet Size Performance Sawtooth
- DoS and OAM Hardware Assist

Deep Stack Problems

- Most severe problems occur with poor multipath implementations
- PHP insures that at most one POP or SWAP is needed.
- (OTOH MPLS-TP mandates use of UHP)
- To get adequate load split, entropy from multiple label entries is needed (preferably all label entries), plus IP headers if present.

Deep Stack - What's wrong with this picture?

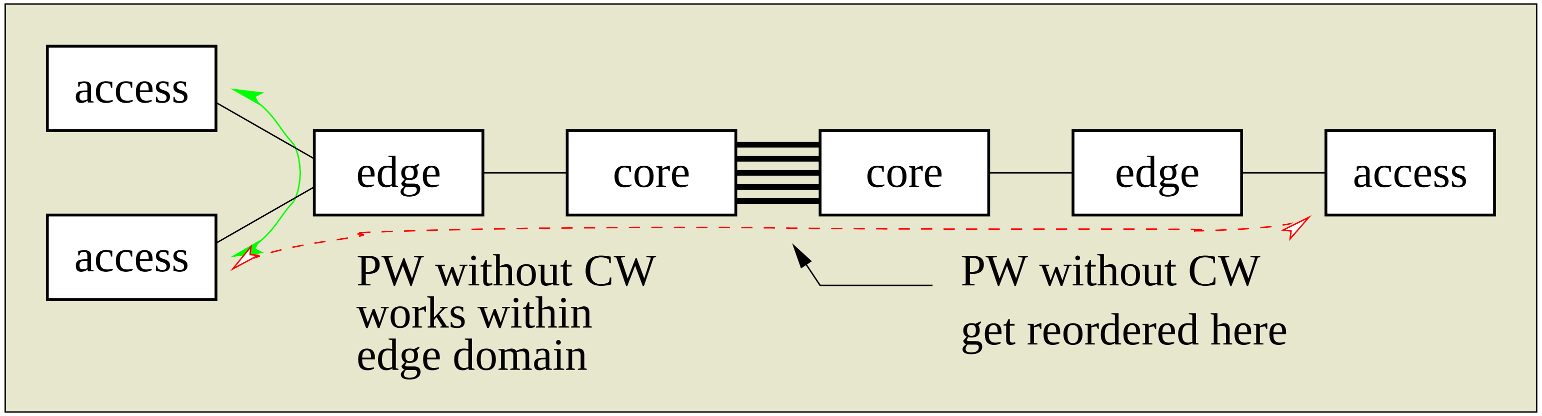


hint: nothing is wrong, except for a few chip makers

Deep Stack Examples

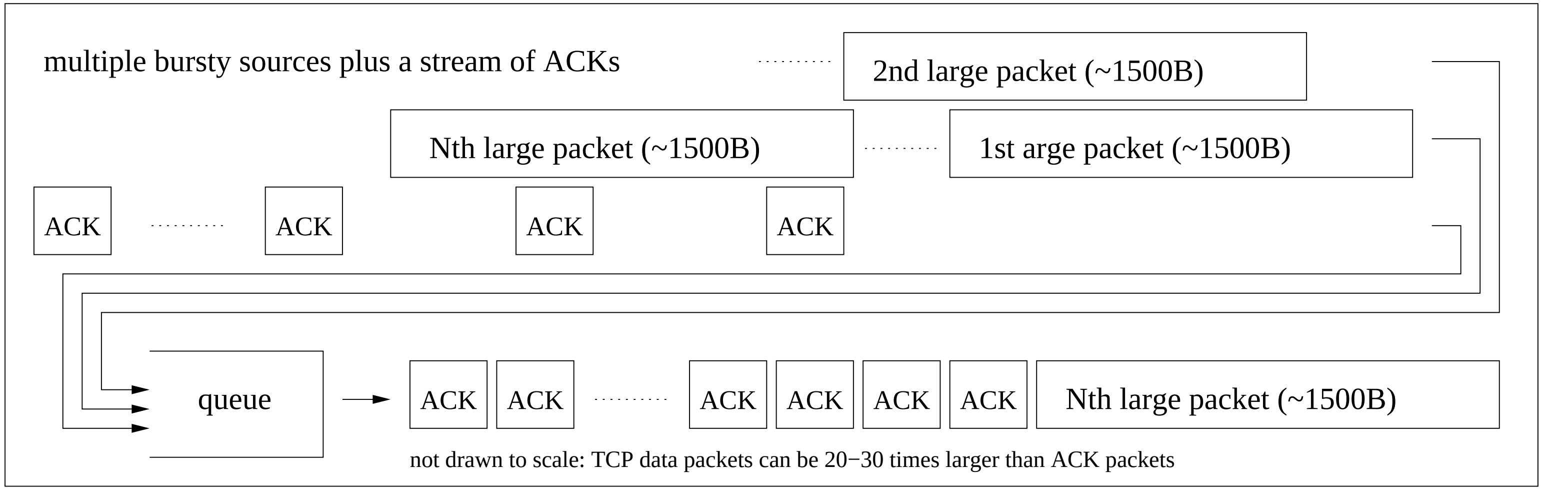
- Stacks with three or four labels:
  - (3) RSVP-TE, ELI, EL, (IP payload)
  - (3) LDP, PW, fat-PW, (CW + PWE3 payload)
  - (4) RSVP-TE, ELI, EL, L3VPN, (IP payload)
  - (4) FRR, RSVP-TE, LDP, L3VPN, (IP payload)
- Stacks with more that four labels:
  - (5) RSVP-TE, LDP, ELI, EL, L3VPN, (IP payload)
  - (5) FRR, RSVP-TE, LDP, ELI, EL, (IP payload)
  - (6) PSC-1, ELI, EL, RSVP-TE, ELI, EL, (IP payload)
  - (8) PSC-1, ELI, EL, RSVP-TE, ELI, EL, LDP, L3VPN (IP payload)
  - (10) FRR, PSC-1, ELI, EL, RSVP-TE, ELI, EL, LDP, PW, fat-PW, (CW + PWE3 payload)
- label stacks can get larger than 2-3 labels
- where encountered, these will not be "rare occurances"

Lack of PW CW support in edge equipment



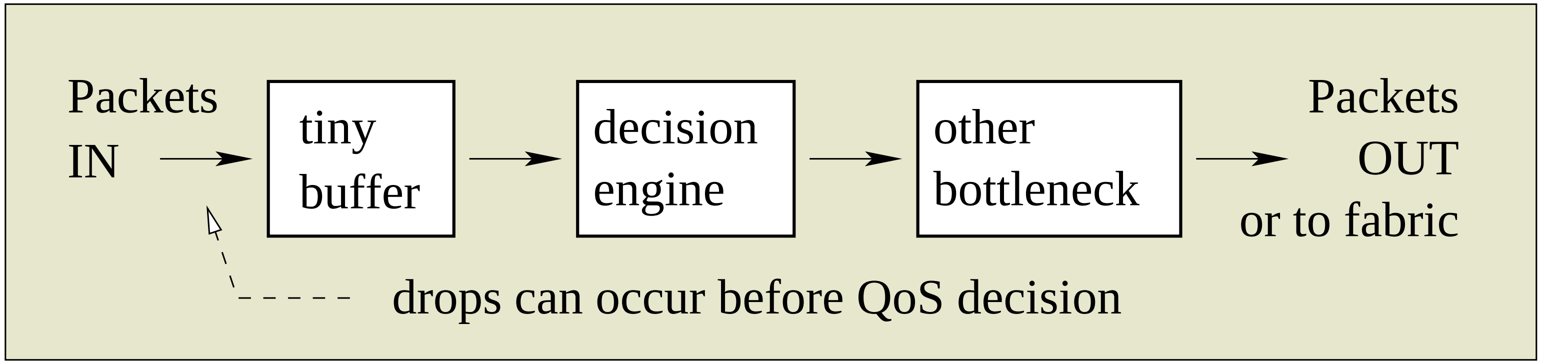
- network cores need to use multipath due to high core to core capacities
- PW from access going through same edge may work fine
- PW passing through core will experience packet reorder if CW is not used

Cause of Small Packet Bursts



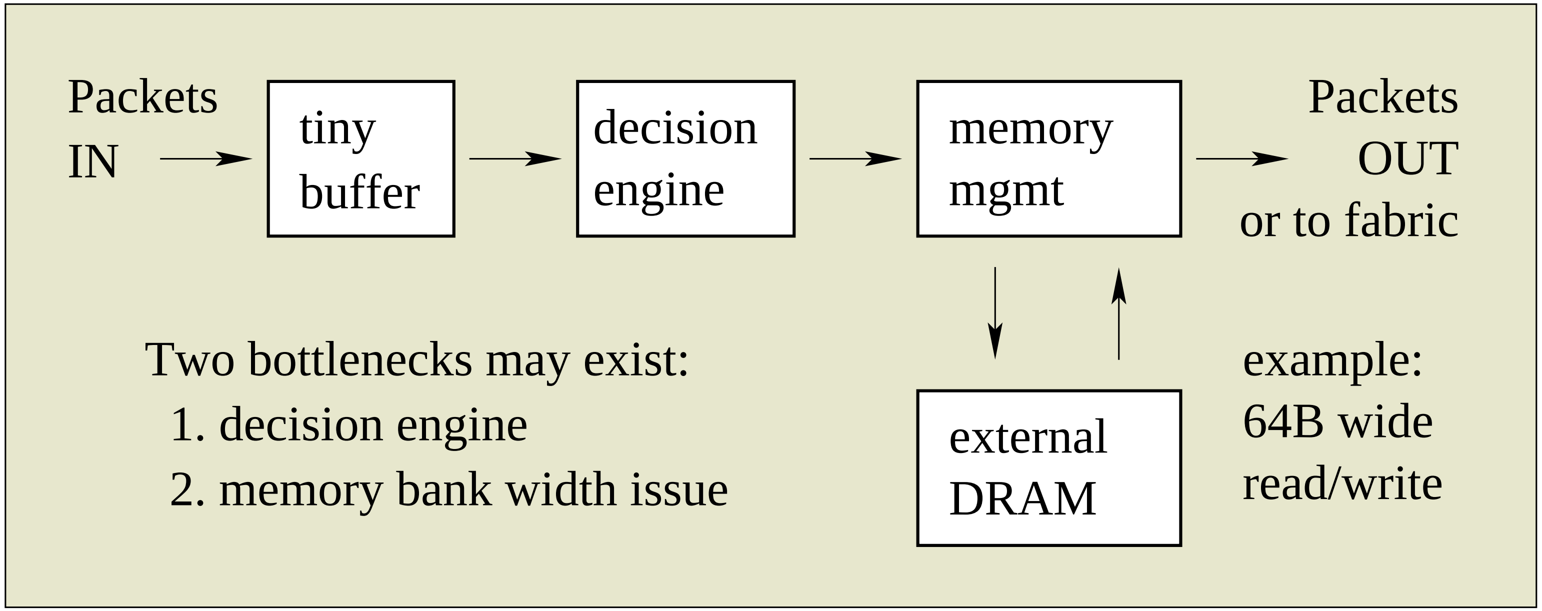
- Above is a simplistic example capable of creating a burst.
- The phenomenon is known as "TCP ACK Compression".
- Multiple streams of evenly spaced ACKs and multiple streams of bursty TCP data (for example during slow start) can cause large bursts.
- Bursts up to 200 TCP ACKs (40 byte) have been observed in service provider networks.

Small Packet Burst Tolerance



- QoS agnostic drops can occur before QoS decision is made.
- A bottleneck downstream can have the same effect if it backpressures the decision process.

Packet Size Performance Sawtooth

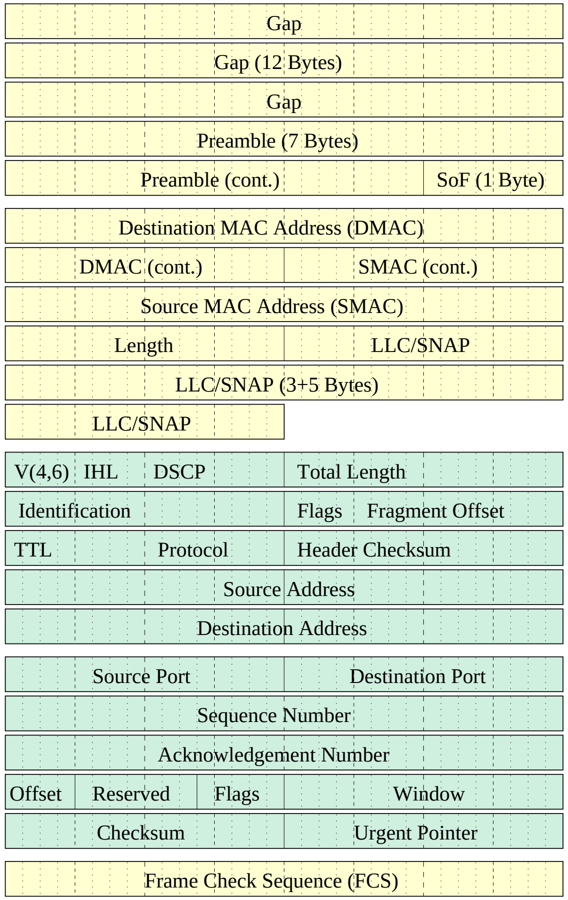


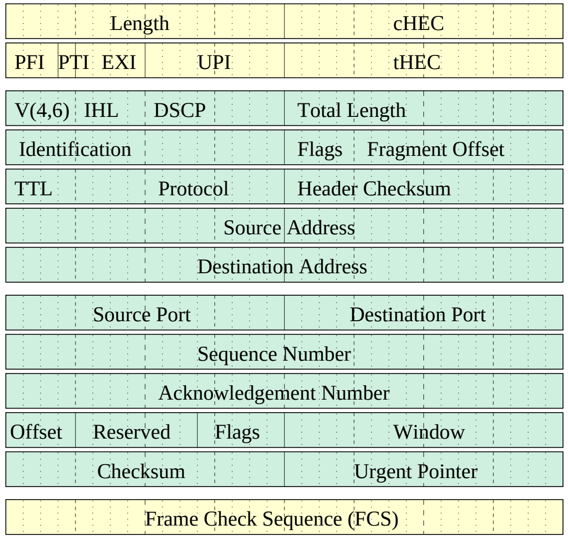
- Result is a sawtooth in max Mpps vs packet size graph
- Does it matter? Maybe not if memory management can cache and buffer bursts rather than backpressure

Packet Size Performance Sawtooth - example

- Example (made up but somewhat realistic):
  - decision engine speed 6.9 nsec (145 Mpps)
  - one packet enters decision pipeline per 6.9 msec
  - memory limit - one 64B wide read/write per 4.6 nsec
- 100G Ethernet with 802.3 (high overhead 46B)
  - 12 B gap, 7 B preamble, 1 B start of frame
  - 6 B DMAC, 6 B SMAC, 2 B length, 8 B LLC/SNAP, 4 B FCS
  - 46 B overhead + 40 B payload = 86 B
  - 7.14 nsec / 40 B pkt = 140 Mpps (@ 103.125 Gb/s)
- GFP/ODU4 (low overhead 12B)
  - no gap, no preamble, no start of frame
  - 8 B headers, 4 B FCS
  - 12 B overhead + 40 B payload = 52 B
  - 3.97 nsec / 40 B pkt = 252 Mpps (@104.782 Gb/s)

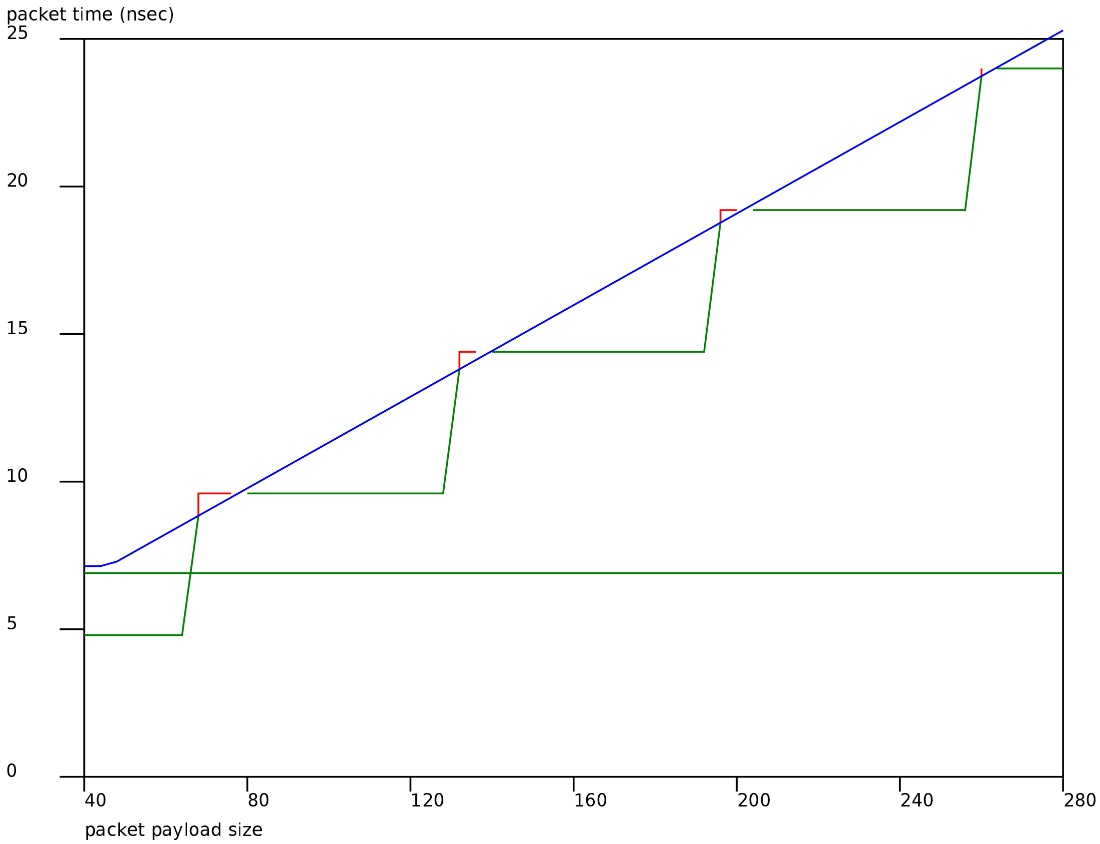
Performance Sawtooth - Encapsulation Efficiencies



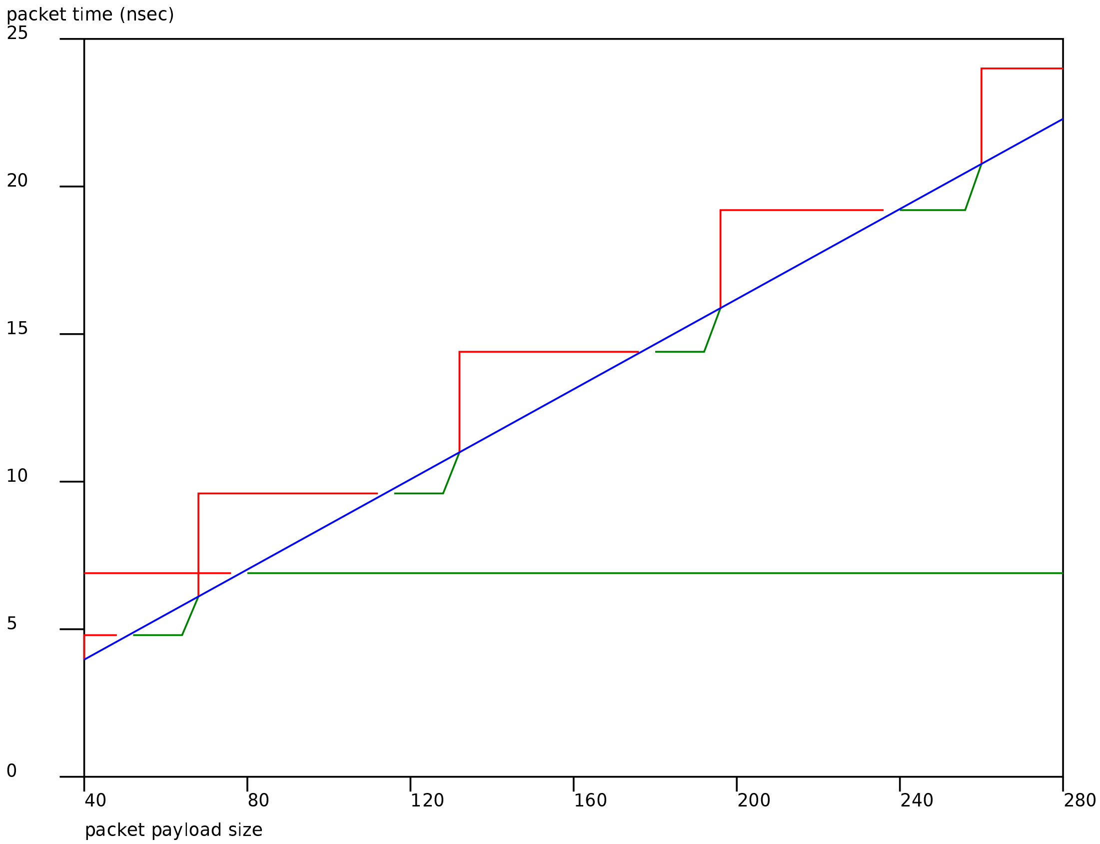


Useful UPI values:
- UPI 0x0d = GFP−F MPLS
- UPI 0x0f = GFP−F ISIS/CLNP
- UPI 0x10 = GFP−F IPv4
- UPI 0x11 = GFP−F IPv6

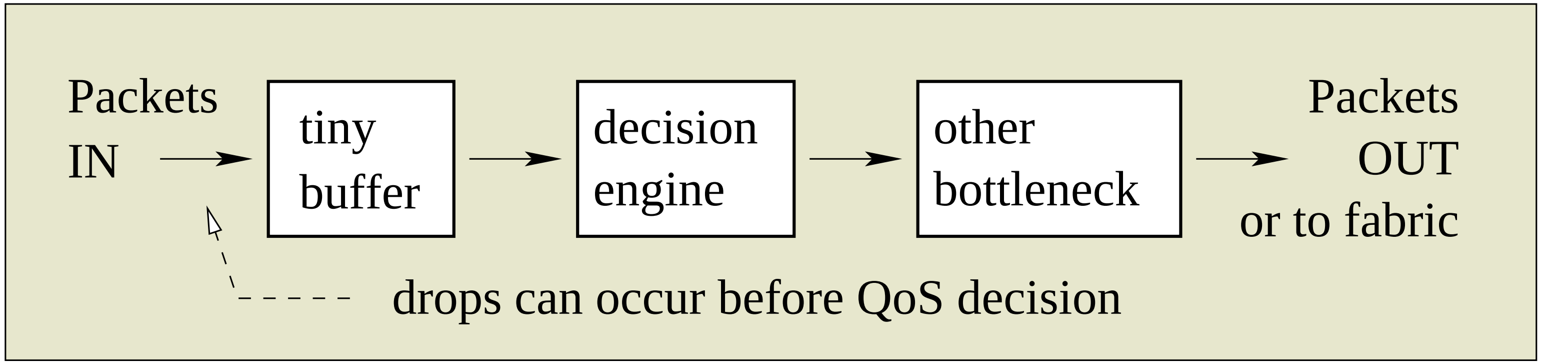
Performance Sawtooth - prior example - 100GbE



Performance Sawtooth - prior example - GFP/ODU4

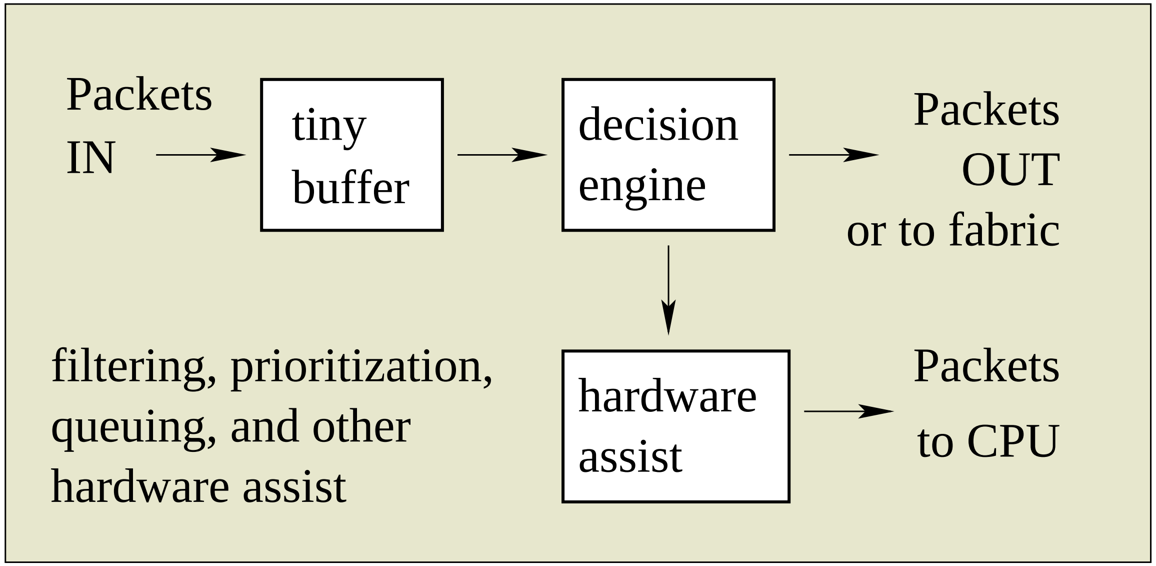


Small Packet Burst Tolerance & QoS



- QoS agnostic drops can occur before QoS decision is made.
- The packets that get dropped may include high priority traffic which is highly drop sensitive.
- A small buffer to deal with bursts of small packets avoids this problem. (Correst value of "small" is an exercise for the audience).

DoS and OAM Hardware Assist



- Packet rate to CPU has to be limited for some types of traffic.
- Filtering is needed to get rid of obviously bogus traffic during DoS.
- General purpose CPU is easily swamped in high volume attacks or major OAM misconfiguration.

Discussion

- anyone read this or prior versions?
- comments and/or flames?
- questions?

BACKUP SLIDES

- No intention to present the remaining slides
- May refer to specific slides if relevant to questions/discussion

Basics - Base

- Base - RFC3031 + RFC3032 + RFC3209
- TTL processing - RFC3443
- MPLS Explicit NULL - RFC4182
- Diffserv - RFC3270 + RFC4124 + RFC5462
- MPLS ECN - RFC5129
- G-ACh and GAL - RFC5586
- link layer codepoints - RFC5332
- PW ACH - RFC5085; MPLS G-ACh - RFC5586
- Entropy Label - RFC6790

Basics - MPLS Special Purpose Labels

- label values 0-15 - RFC3032
  - IANA: Multiprotocol Label Switching Architecture (MPLS) Label Values
- draft-ietf-mpls-special-purpose-labels
  - IANA: Extended Special Purpose MPLS Label Values

Basics - MPLS Differentiated Services

- base - RFC2474 + RFC2475 + RFC5462
- E-LSP and L-LSP - RFC3270
- class-type (CT) mapping to TC-¿PHB - RFC4124

Basics - Time Synchronization

- NTP and PTP are important
- PTP over MPLS - draft-ietf-tictoc-1588overmpls
- this work may be changing and needs to be watched

Basics - Uses of Multiple Label Stack Entries

- lists many uses of multiple labels in label stack
- practical cases now exist for four or more
- theoretical scenarios can reach eight or more

Basics - MPLS Link Bundling

- early and limited MPLS multipath - RFC4201
- all-ones component spreads traffic like ECMP (using hash)
- other mode places each LSP on a specific component

Basics - MPLS Hierarchy

- of interest is Packet Switch Capable (PSC) - RFC4206
- four levels of hierarchy PSC1-PSC4 (plus implied PSC-0)

Basics - MPLS Fast Reroute (FRR)

- two modes "detour" and "bypass" - RFC4090
- detour explicitly signals path from PLR to merge
- bypass uses bypass LSP and is far more common
- bypass requires use of platform label space

Basics - Pseudowire Encapsulation

- arch - RFC3985
- control word (CW) - RFC4385 (motivation in RFC4928)
- VCCV - RFC5085 (associated channel in RFC4385)
- pseudowire sequence number is useful for some payload types

Basics - Layer-2 and Layer-3 VPN

- impact on midpoint LSP within scope
- L2VPN and L3VPN add a label
- encap/decap and VRF at LER is out of scope

MPLS Multicast

- layer-2 encaps clarification in RFC5332
- signaled using RSVP-TE [RFC4875] or LDP [RFC6388]
- RSVP-TE uses root initiated join
- LDP uses leaf initiated join (more like IP multicast)
- where to replicate is an local matter but needs careful thought
- LSR may be leaf, replicating, or bud wrt a P2MP LSP
- MP2MP similar but with multiple senders possible

Packet Rates

- dropping packets is bad! (duh)
- number of packets per second depends on packet size
- long bursts of small packets (about 40-48 byte) common
- ethernet rounds to 64, but not everything is ethernet
- need small buffer before decision engine
- to avoid dropping high priority traffic need -either-
  - handle sustained 40 byte (plus label) packets -or-
  - absorb bursts of small packets before decision engine

Multipath

- very important for large SP - important for others as well
- adequate balance requires adequate entropy
- entropy from stack alone is insufficient - look for IP headers
- common practice is to reinspect for entropy at each hop
- entropy label may simplify task of midpoint LSR

Pseudowire Control Word

- PW CW support is essential for LSR at all tiers
- PW without CW get out-of-order when crossing multipath in core
- not supporting CW will not earn friends

Large Microflows

- Large microflows (ie: Gb/s to tens of Gb/s) are trouble for multipath
- active management of the hash space is local issue and out of scope

Pseudowire Flow Label

- some PW types are OK with reordering if microflows stay ordered
- examples are Ethernet and FR
- flow label (fat-pw) allows multipath
- fat-pw preserves order of microflows
- avoids large microflow problems

MPLS Entropy Label

- like PW flow label entropy label helps with multipath
- RFC6790 defined entropy label indicator (ELI) and EL
- entropy label allows ingress to extract entropy
- save deep packet inspection at midpoint LSR
- allows truncation of label stack inspection

Fields Used for Multipath Load Balance

- four subsections
  - MPLS Fields in Multipath
  - IP Fields in Multipath
  - Fields Used in Flow Label
  - Fields Used in Entropy Label
- too little time to go into details on this

MPLS-TP and UHP

- Egress UHP POP, counter, then lookup, then another counter
- Using PSC hierarchy can result in multiple lookup, POP, count per packet
- performance impacts if this isn't done right

Local Delivery of Packets

- packets sent to local general purpose CPU can swamp it
- hardware support is needed to protect CPU
- prevents accidental and malicious (DoS, DDoS) outage

DoS Protection

- filtering in hardware before sending to CPU
- GTSM is special filtering - RFC5082
- involved topic - see draft - basics covered

Extent of OAM Support by Hardware

- MPLS OAM, PW OAM and MPLS-TP OAM discussed in draft
- OAM can swamp a general purpose CPU
- hardware support or assist recommended for some OAM flavors

Number and Size of Flows

- some hardware can't handle very large microflows
- some hardware can't handle huge number of microflows
- both problems are bad - latter may be worse

Use of RFC 2119 Keywords in this draft

- RFC2119 all upper case keywords used when:
  - stating a requirement that comes from an existing RFC
  - implied requirement needed to conform to existing RFC
  - clearly marked "advice" with strong reasons given

Are there omissions?

- hopefull not but it would help if WG thought about this

Potential Topics of Discussion

- in scope vs out of scope
- use of RFC2119 language in an informational document
- reasons for recommending small packet burst tolerance
- details of recommendations on multipath
- DoS and OAM hardware assist
- would profiles be overkill?
  - core vs edge vs access vs enterprise vs data center, etc