CSIG Telemetry At Layer 2

Simple and Effective In-band Network Signals for Efficient Traffic Management in Datacenter Networks

Paul Bottorff (HPE)
Brad Karp (Google, LLC)
Jai Kumar (Broadcom)
Ramesh Sivakolundu (Cisco)
Paul Congdon (Congdon Consulting, LLC)

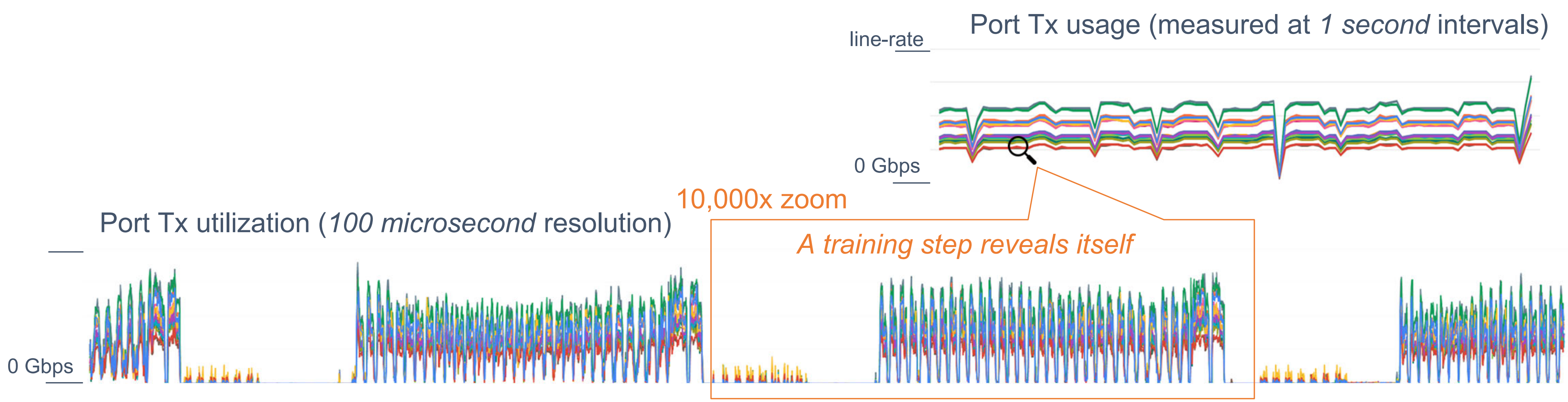
IEEE 802.1 Meeting
January 19, 2026

CSIG as General-Purpose Telemetry

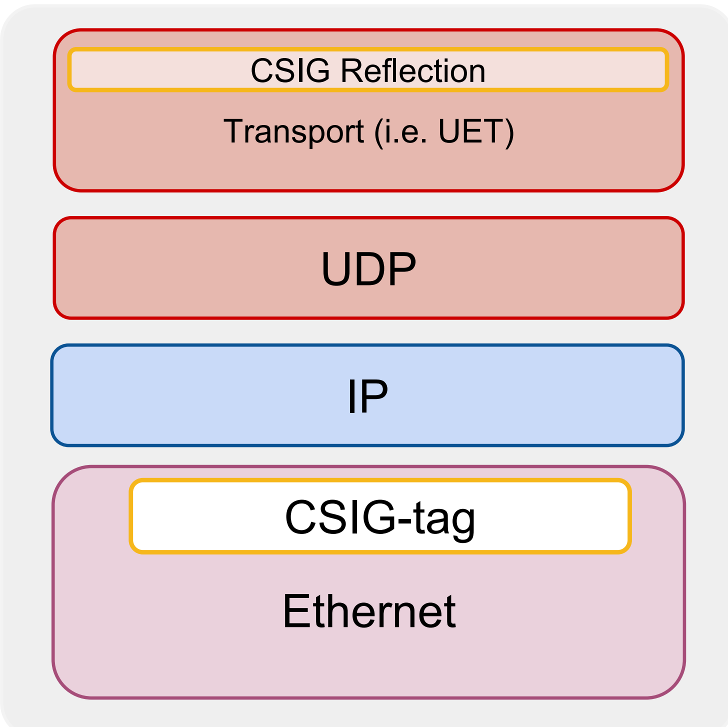
- Many Possible Use Cases, for instance:
  - End-to-End Congestion Control (requires integration with L4)
  - Path Selection and Load Balancing
  - Debugging and Fault Isolation
  - Traffic Engineering
  - SLA Compliance
- Currently, CSIG is used to provide precision telemetry at the transport layer to improve congestion control, path selection, and load balancing.
- However, since the telemetry is collected and transmitted at the L2 layer it could be used for general-purpose telemetry at any of the L2, L3 or L4 layers.
- Providing CSIG support at L2 enables the use of CSIG in environments with any transport, current of future.
  - Even if the transport does not support CSIG, telemetry applications such as debugging and fault isolation are still possible uses of CSIG.

High-resolution network signals are *necessary*

- Accurately detecting congestion *locally* on a switch requires signal measurements at sub-millisecond timescales
- Real-world example from a GPU ToR at Google:
  - Shifting from 1-second to 100-μsec telemetry exposes the fine-grained, repeating congestion patterns and idle gaps inherent to AI workloads

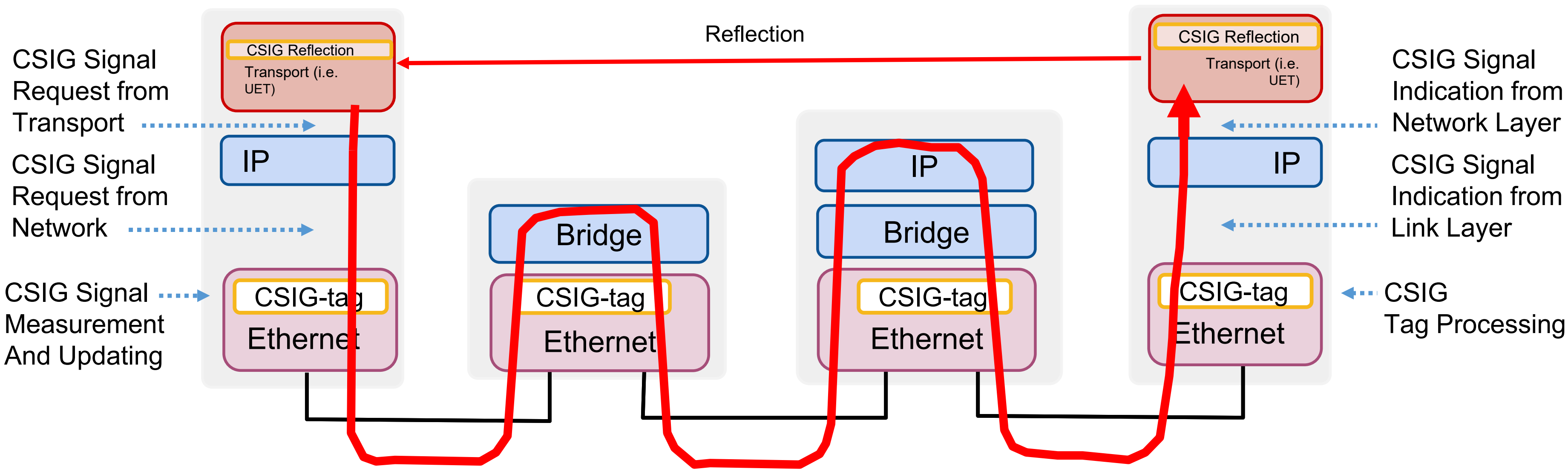


CSIG: Practical & Effective In-band Telemetry Protocol



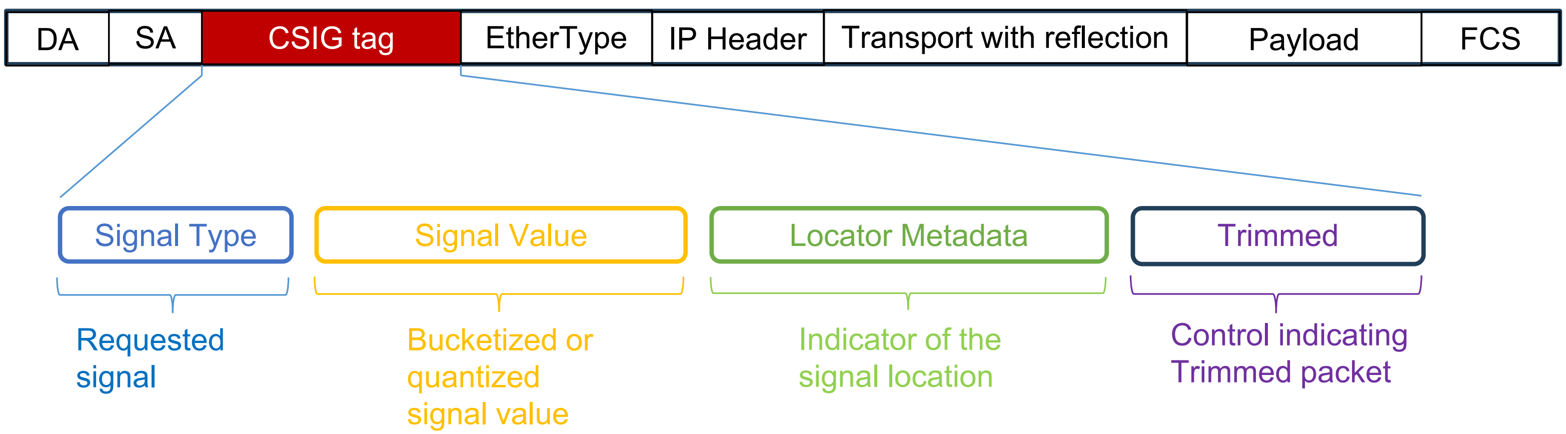
- Provides fixed-length simple summaries from the path bottlenecks
- Designed for Congestion Control, Traffic Management and Network debuggability use-cases
- Designed for brownfield deployment with backward compatibility / interoperability
- Link to UEC Draft 0.50 from UEC liaison is in public domain- https://github.com/opencomputeproject/OCP-NET-UEC-CSIG
- CSIG differs from Connectivity Fault Management (CFM) in that it is an In-band technique carried within data packets.

CSIG: L2 telemetry supporting multiple layers



- Here the CSIG signal requests come down the stack from transport and network layers to the data link layer where CSIG tag processing occurs
- The CSIG telemetry is encoded in an L2 tag on the wire
- The CSIG telemetry tag is an in-band signal which can be placed on every packet
  - Each packet can have a single CSIG tag which specifies a single signal type (single type of measurement)
  - CSIG tags are fixed size 4 or 8 octets long tags

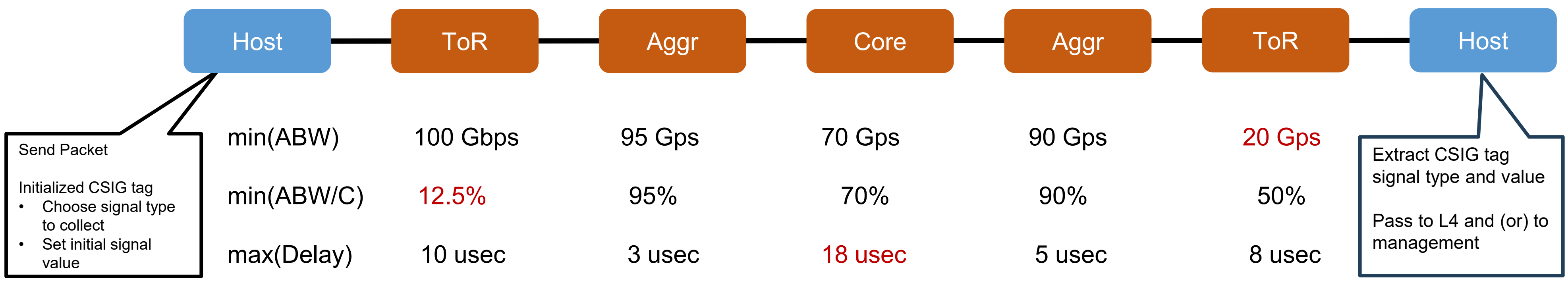
CSIG uses an L2 tag to transmit telemetry



One of:
- Min(ABW): Minimum Available Bandwidth
- Min(ABW/C): Minimum Available Capacity
- Max(Delay): Maximum Per-Hop Delay
- Max(nQD): Maximum Normalized Queue Depth

- Signals are computed at high-resolution timescales on each switch.
- The signal is measured for each hop and updated only if the measure is either greater or less than, depending on signal type, the current CSIG signal value.
- At the receiving end CSIG provides the measure of the bottleneck’s value.

CSIG End-to-End:
Origination, Switch Forwarding Behavior, Reflection



- Each switch compares the local signal value for signal type in tag with signal value in tag; conditionally overwrites value in tag with local value according to aggregation function in signal type

min(ABW): Minimum Available Bandwidth

Absolute minimum available bandwidth in bps across all switch ports traversed along a given packet's path through the fabric

- Signal type = 0
- Associated Math Function 'min'
- Init Value = all ones (max value)
- Algorithm Used: 'raw BW' computation
- Per-egress port information
  - Actual Tx Bytes on the wire

min(ABW/C): Minimum Available Capacity

Normalized minimum available bandwidth as a percentage across all switch ports traversed along a given packet's path through the fabric

- Signal type = 1
- Associated Math Function 'min'
- Init Value = all ones (max value)
- Algorithm Used: 'raw BW' computation
- Per-egress port information
  - Actual Tx Bytes on the wire
  - Dropped and truncated packets not considered in the computation
  - Burstiness of the traffic (queue congestion) not captured by this signal

This signal is robust when there are port speed mismatches along the packet path. In a uniform-speed fabric, this signal does not add utility over min(ABW).

Raw ABW Algorithm

$m = Traffic\ in\ bits\ measured\ over\ a\ time\ interval\ t$
$p = port\ speed (bps)$

$$rate\ r = \frac{m}{t} (bps)$$
$$Available\ BW = (p - r)$$
$Quantized\ Available\ BW :: Q(p - r)$ –Here the ABW computation is quantized to fit in the CSIG tag signal field

$Before\ packet\ egress\ the\ CSIG\ signal\ value\ for\ this\ packet\ (pkt)\ is\ updated\ to:$
$$pkt \rightarrow minBW = \min(Quantized\ Available\ BW, pkt \rightarrow minBW)$$

All network devices in a CSIG Domain must be configured with the same value for t and the same quantization range

m is the tx bits seen on the wire as accounted for in the port statistics
Any overhead from Link Layer is not included in ABW algorithm
ABW should account for data frames that can be controlled by the sender for a given flow rate
[MAC generated frames e.g. PAUSE. LLR, CBFC frame or PFC frame are not included]

Raw ABW/C Algorithm

$m = Traffic\ in\ bits\ measured\ over\ a\ time\ interval\ t$
$p = port\ speed$

$rate\ r = m/t$
$Consumed\ Load = (^r/_p)$ * 100
$Available\ Load = 100(1\ - (^r/_p))$
$Quantized\ Available\ Load :: Q(Available\ Load)$ –Here the ABW/C computation is quantized to fit in the CSIG tag signal field

$Before\ packet\ egress\ the\ CSIG\ signal\ value\ for\ this\ packet\ (pkt)\ is\ updated\ to:$
$pkt \rightarrow minLoad = \min(Quantized\ Available\ Load, pkt \rightarrow minLoad)$

All network devices in a CSIG Domain must be configured with the same value of t and the same quantization range

m is the tx bits seen on the wire as accounted for in the port statistics
Any overhead from Link Layer is not included in ABW algorithm
ABW should account for data frames that can be controlled by the sender for a given flow rate
[MAC generated frames e.g. PAUSE. LLR, CBFC frame or PFC frame are not included]

max(Delay): Maximum Per-Hop Delay

Maximum delay in nanoseconds among observed per-switch delays at switch elements traversed by a given packet's path through the fabric

- Signal type = 2
- Associated Math Function 'max'
- Init Value = 0
- Algorithm Used: Observed value in the switch pipeline
- Per-packet information
  - Forwarding pipeline delay is included
  - Link-layer (MAC/PHY) delay is not included

CSIG Domain need not be time-synchronized

max(nQD): Maximum Normalized Queue Depth

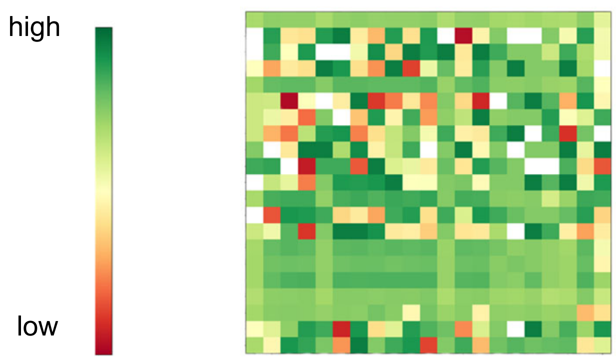
Maximum among queue depths as percentage occupancy at each successive queue traversed along a packet's path, each measured at dequeue time

- Signal type = 3
- Associated Math Function 'max'
- Init Value = 0
- Algorithm Used: Observed value in the switch pipeline
- Functions like a multi-bit ECN on egress: check queue depth at dequeue time, but record normalized depth value rather than a boolean result of threshold comparison
- Per-packet information
- Leading indicator of congestion
- Tolerates non-uniform buffer size across the UE fabric

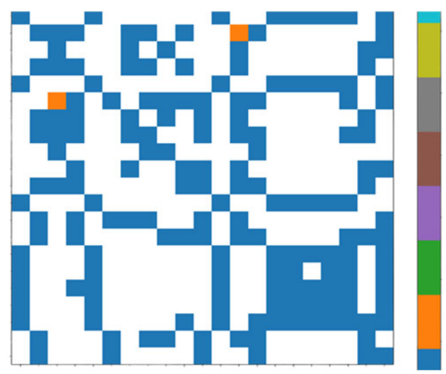
Example: CSIG Telemetry for Tensor Processing Unit (TPU) Training at Google: Fine-Grained Congestion Observability

- CSIG min(ABW) telemetry for an ML training job maps even brief intervals of congestion in the fabric, enabling:
  - ML training job refinement: “Can my training job perform more communication without encountering congestion, and if so, where?” “Where is my training job congesting the fabric?”
  - Fabric capacity planning: “Is the fabric adequately provisioned for the models I am running?”

Telemetry from a production Tensor Processing Unit ML training job at Google:



Heatmap of mean available bandwidth observed by TPU job’s flows



Most frequent (mode) location of bottleneck experienced by TPU job’s flows, only among heavily loaded bottleneck links

Each small square summarizes data for packets between one portion of cluster (x coord) and another (y coord)

CSIG: Simple, Fine-Grained, Efficient Telemetry

- CSIG is a practical and highly effective protocol providing very fine-grained telemetry.
- The small fixed-length 4 and 8 byte CSIG tags incur negligible bandwidth overhead and support line-rate telemetry on every data packet at switches and end hosts while avoiding complicated variable-length header processing.
- CSIG has a broad scope for uses across congestion control, load balancing, scheduling, debugging and fault isolation, traffic engineering, and SLA Compliance.
- Use cases such as Debugging and Fault Isolation, Traffic Engineering, and SLA Compliance could make measurements starting and ending at lower network levels.
  - A management entity could control CSIG initiations at the L2 layer
  - This could be done in a way that didn't interfere with CSIG requests from L4
- An option to initiate and terminate CSIG at L2 would greatly expand the utility of CSIG, since operation at a L2 would support all current and future transport (L4) protocols.

Thank You

Backup

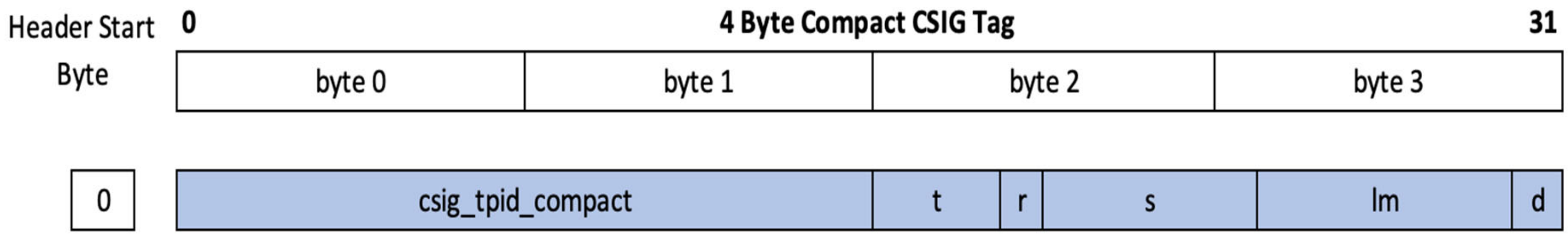
CSIG Entity Functions

- Insert and Delete CSIG tags from frames
  - Operation of these functions depends on the service request, CSIG entity capabilities, management parameter settings, and LLDP negotiation state.
- Measurements
  - Available Bandwidth (over a specified measurement interval at egress)
  - Residence Time (for this frame, measured from ingress to egress MSAP)
  - Current Queue Depth (for this frame at frame egress time)
- Updating CSIG tags
  - Compare the ingress signal value with the calculated signal value over this hop
  - Update the signal value if the calculated signal value is greater than (or less than) the ingress signal value (signal types determine if the measure is greater of less than), otherwise forward the CSIG tag unchanged from the ingress value.
  - If the signal value was updated, then also update the location metadata.
  - If packet was trimmed at this hop then set to freeze updating.

Thoughts on CSIG initiation and L2 implementation

- Current use cases initiate CSIG tagging at the transport (L4) layer and use the telemetry to improve transport function.
  - min(ABW) Optimal path selection: Choose paths with greater available bandwidth (instead of lumping together all paths that are non-ECN-marked)
  - min(ABW/C) Fast adjust of cnwd: Fast ramp up of cwnd within measured available spare capacity, to avoid overshooting (also proposed in HPCC++ Internet Draft)
  - max(Delay) Avoid estimated path's baseRTT: Inaccurate measure if queues are already built up
  - max(nQD) provides a multi-bit ECN allowing indication of both over and under utilization
- Other use cases such as Debugging and Fault Isolation, Traffic Engineering, and SLA Compliance could make measurements starting and ending at lower network levels.
  - A management entity could control CSIG initiations at the L2 layer
  - This could be done in a way that didn't interfere with CSIG requests from L4
  - An option to initiating and terminating CSIG at L2 would greatly expands the utility of CSIG, since operation at a L2 doesn't depend on a transport (L4) protocol supporting CSIG capabilities (i.e. it could be run at L2 or L3 under any current or future transport).

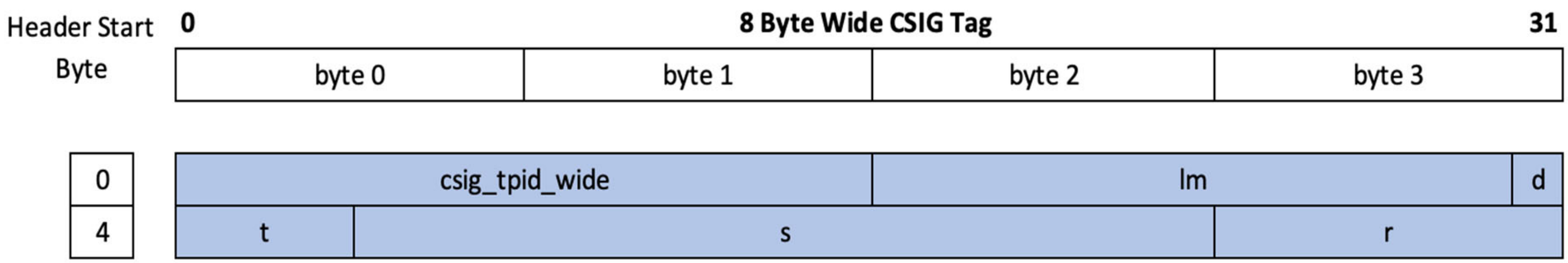
Two L2 Tags: the 4 byte tag is aligned to VLAN tag fields



Bit Offset	Width (in bits)	Field Name	Comments
0-15	16	csig-tpid-cpct	New Ethertype allocated by IEEE
16-18	3	t	Type of Signal
19	1	r	Reserved
20-24	5	s	Quantized Signal Value
25-30	6	lm	Locator Metadata
31	1	d	Do not update (Packet Trimmed) ("(D)ropped")

- The alignment with VLAN tags is critical to enable retrofitting some existing switches (by changes in firmware and microcode) to support CSIG.
- This has enabled the current largescale deployments at Google.

Two L2 Tags: the 8 byte tag provides fine grained measures



Bit Offset	Width (in bits)	Field Name	Comments
0-15	16	csig-tpid-wide	New Ethertype allocated by IEEE
16-30	15	lm	Locator Metadata
31	1	d	Do not update (Packet Trimmed) (“(D)ropped”)
32-35	4	t	Signal Type
36-55	20	s	Quantized Signal Value
56-63	8	r	Reserved

- With silicon developments currently in progress it will be possible to implement 8 byte CSIG tagging to provide fine grained measurements.

One signal is carried in each tag

t	Signal	Profile	Aggregation Function	Comments
0	ABW	base	min	Available bandwidth per port
1	ABW/C	base	min	Relative available bandwidth per port
2	Delay	base	max	Per-hop delay
3	nQD	extended	max	Queue depth normalized by port speed

- The signal is measured for each hop and updated only if the measure is either greater or less than, depending on signal type, the current CSIG signal value.
- At the destination transport the CSIG signal value is the value of the hop with either the minimum or maximum value.
- Transport is responsible for generating a collection of signal measurement types it needs for it to control congestion and manage multiple network paths.