aruba
a Hewlett Packard Enterprise company

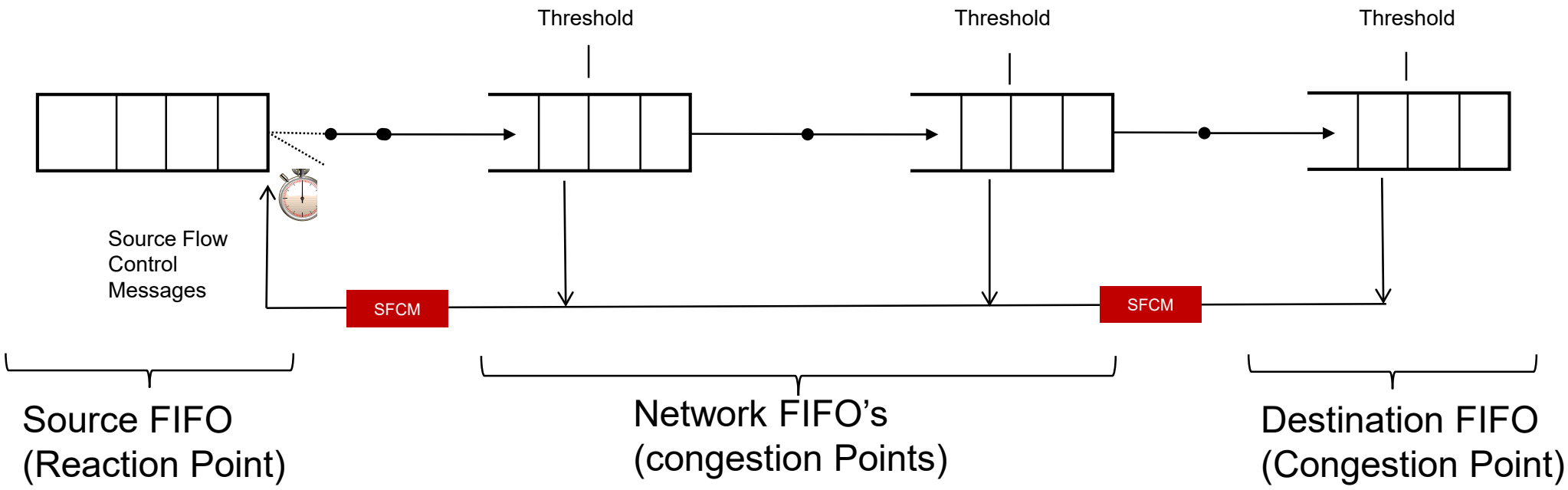
Data Center Source Flow Control

Paul Bottorff (Office of Aruba CTO)

March 2022

Basics Concepts

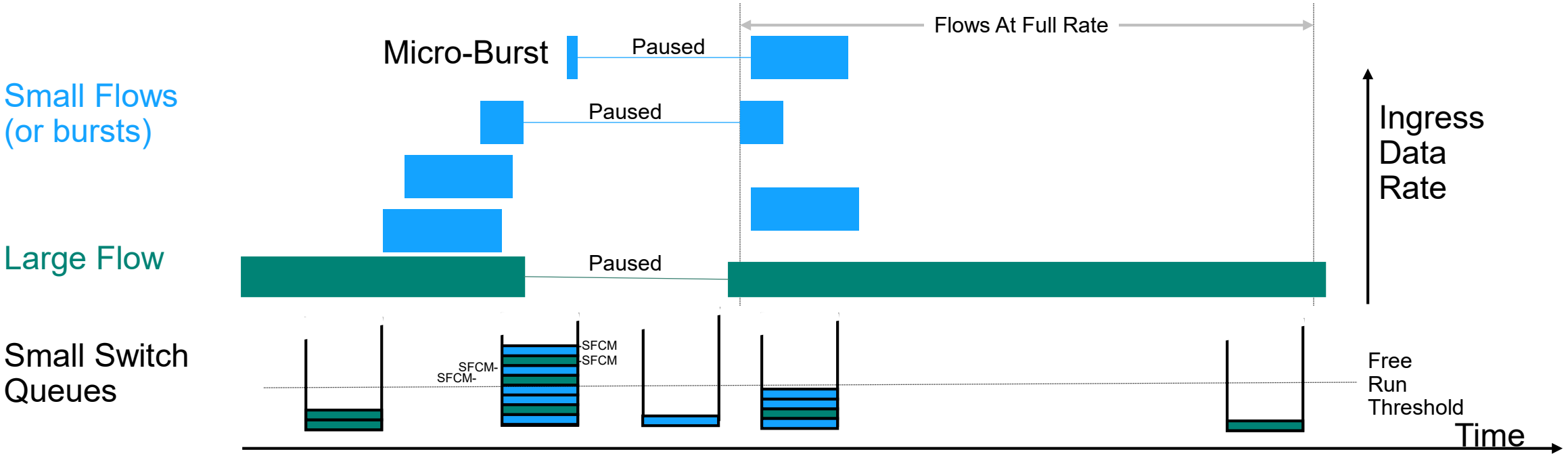
Source Flow Control (SFC) Basics



- Source Flow Control Messages(SFCMs) are generated at each network and destination FIFO for all packets above the threshold.
- Each SFCM message contains a pause interval determined by the time to drain the congested FIFO
- Whenever a SFCM message is received at a source FIFO reaction point it pauses transmitting for the pause interval time specified in the SFCM message
- If the source FIFO is already halted when a SFCM is received, the source FIFO resets the halt time to the maximum of the remaining pause interval or the new pause interval
- A SFC reaction point transmits at the full speed, as determined by other mechanisms such as DCQCN, until it receives a SFCM at which time it stops entirely for the specified pause interval

aruba
a Hewlett Packard Enterprise company

Micro-Burst with Source Flow Control



– Here a Micro-Burst starts to overload a small switch buffer rising above the free run threshold

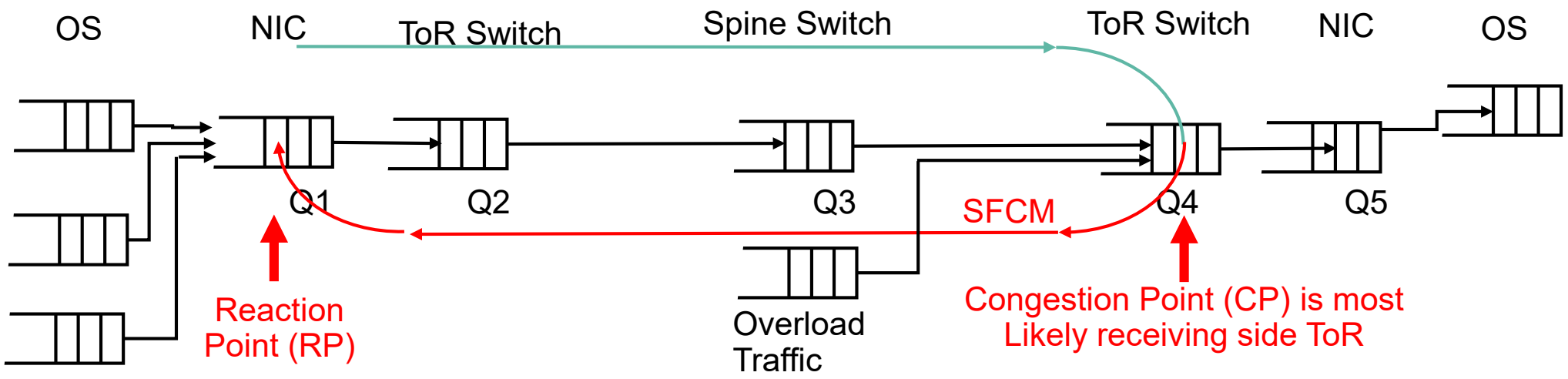
– All flow sources respond to Source Flow Control Messages (SFCM) stopping their transmissions for enough time to prevent packet discard and to allow the switch buffer to drain

– After the switch buffer empties all flow sources return to transmitting at full speed

– All flows buffered in the switch continue to egress during the period when the source is paused

aruba
a Hewlett Packard Enterprise company

SFC Reaction Time ~ < 10 usec



– For a DC POD the wire time for 100 meters is ~= 500 nsec which is small relative to switch pipeline delays

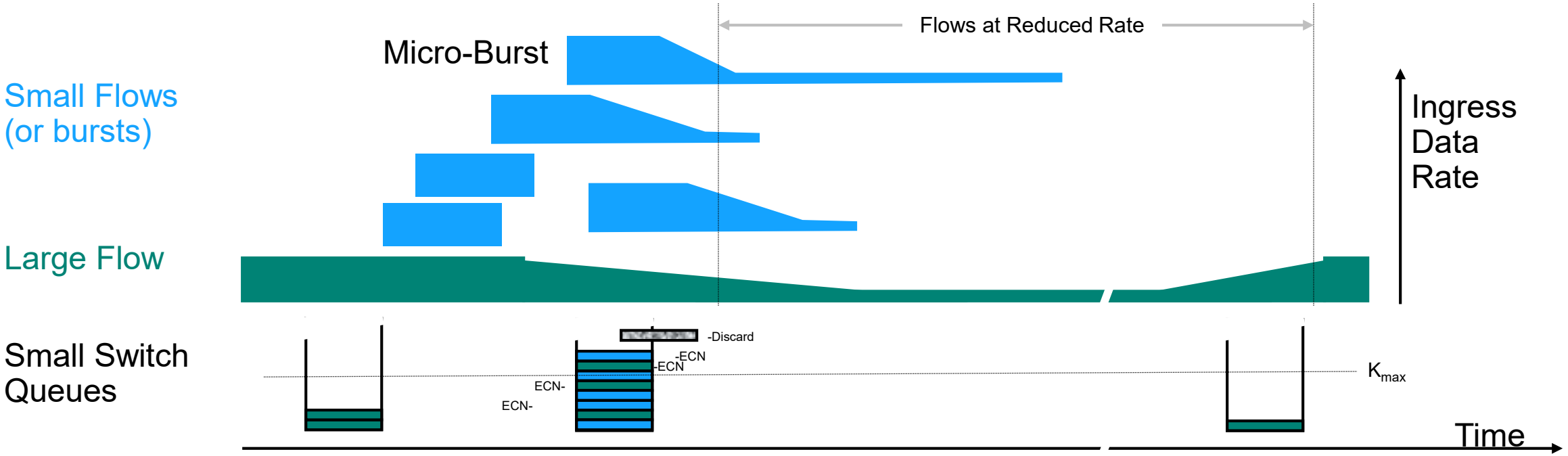
– Assuming each switch chip and NIC adds roughly 1 usec of pipeline latency (ToRs are 1 chip, spines are 3 chips)
  – The SFCM transmission time of ~50 nsec for a 128 byte SFCM packet at 25 Gbit is a relatively insignificant addition
  – Transmission of 1500 byte packets takes about 500 nsec at 25 Gbit and 125 nsec at 100 Gbit

– The SFC reaction time is roughly the wire time + network time + offload adapter latency roughly 7 usec
  – If the congestion point is at the local ToR then the SFC reaction time is roughly 2 usec
  – If the congestion point is at the spine switch then the SFC reaction time is roughly 3.5 usec

– SFC reaction time to fully stop the RP is ~< 10 usec (at 25Gbit and 75% efficiency ~<= 23 Kbytes)

aruba
a Hewlett Packard Enterprise company

Micro-Burst with DCQCN Congestion Management



– Here a Micro-Burst overloads a small switch buffer resulting in packet discarding

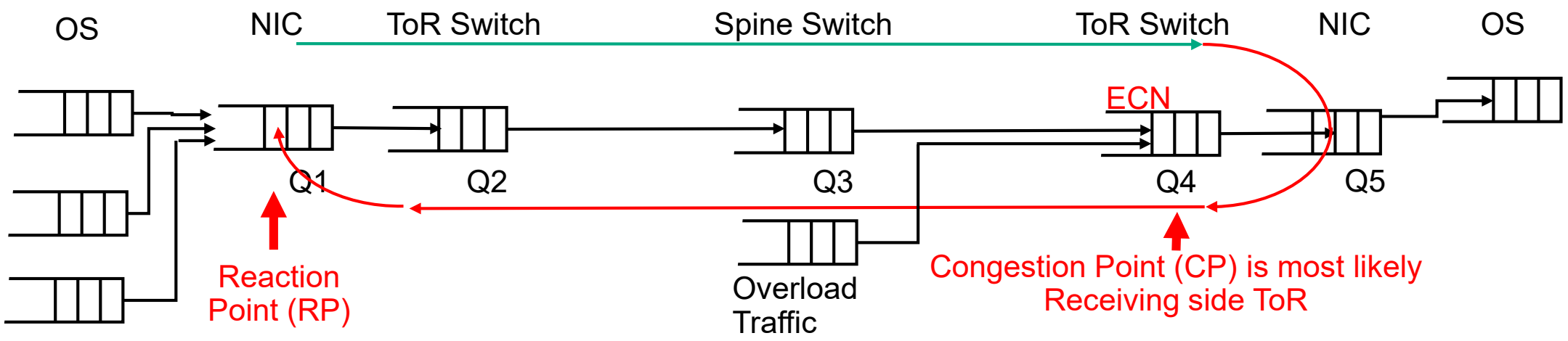
– The flows sources all respond to the ECNs and packet discards by reducing their data rate

– The large (green) flow's rate reduction lasts well past the micro-burst resulting in lowered bandwidth

– The small (blue) flows in the micro-burst don't react fast enough to prevent discards causing latency increases

aruba
a Hewlett Packard Enterprise company

DCQCN Reaction Time ~ >= 100 usec (10-100X SFC)

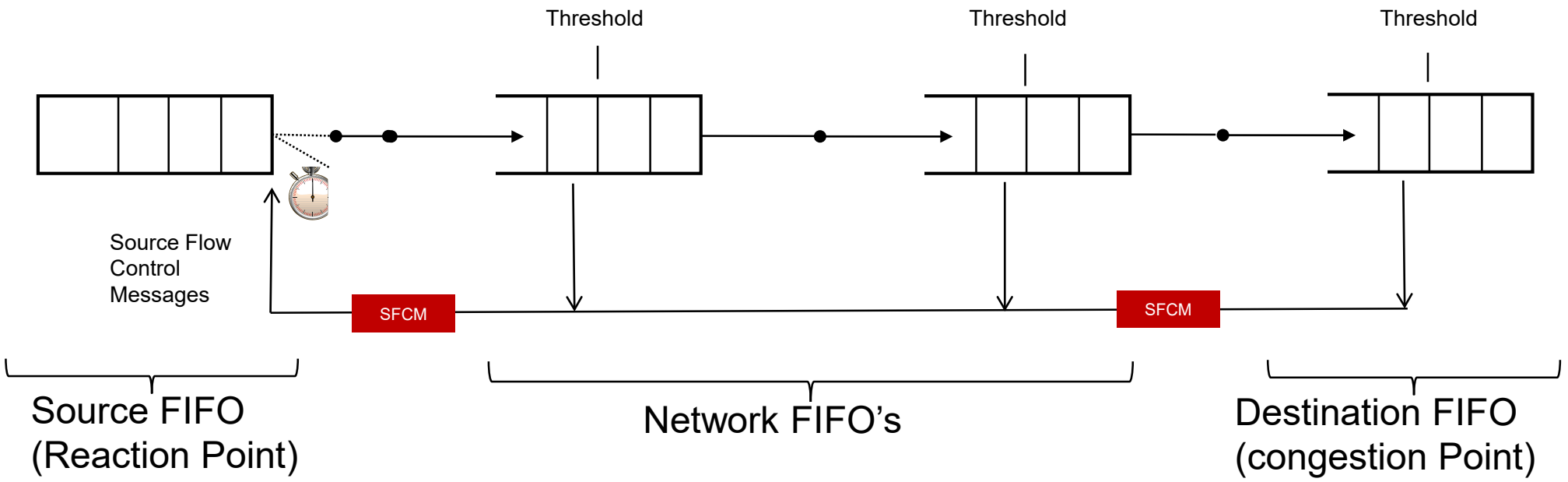


- The time for a single ECN to travel from Congestion Point to Reaction Point is the travel time through the congested queue + network latency + offload adapter latency + wire time
- The ECN congested queue delay for a 125K queue with a drain rate of 25 gigabit at 75% efficiency is roughly 55 usec (assumes only a single congested queue)
- If we assume each switch chip and NIC adds roughly 1 usec of latency (ToRs are 1 chip, spines are 3 chips) and 500 nsec transmission time for a 1500 byte packet at 25 Gbit
- The DCQCN reaction time is roughly the (wire time + network time + offload adapter latency) + buffer delay ~= 64 usec
  - Note: since the buffer delay is the dominant term the reaction time for a CP at the local ToR is close to the same as a CP at a remote ToR
- The DCQCN algorithm generates congestion notification at most every 50 usec and the fastest it reduce the rate is by half, therefore in the very best case it takes two steps to reduce the rate to ¼
- DCQCN reaction time to reduce to ¼ at RP is ~> 100 usec (at 25Gbit and 75% efficiency ~>= 230 Kbytes)

aruba
a Hewlett Packard Enterprise company

SFC Examples

Source Flow Control (SFC) Basics

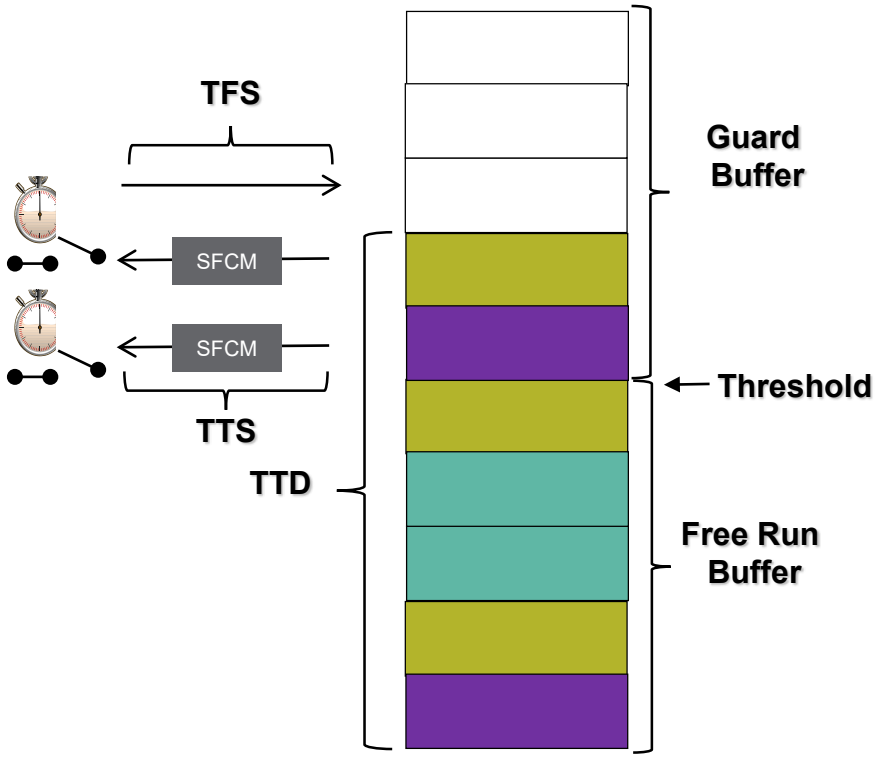


– Source Flow Control Messages (SFCMs) are generated at each network and destination FIFO for all packets above the threshold

– Each SFCM message contains a pause interval determined by the time to drain the congested FIFO

– Whenever a SFCM message is received at a source FIFO reaction point it pauses transmitting for the pause interval time specified in the SFCM message

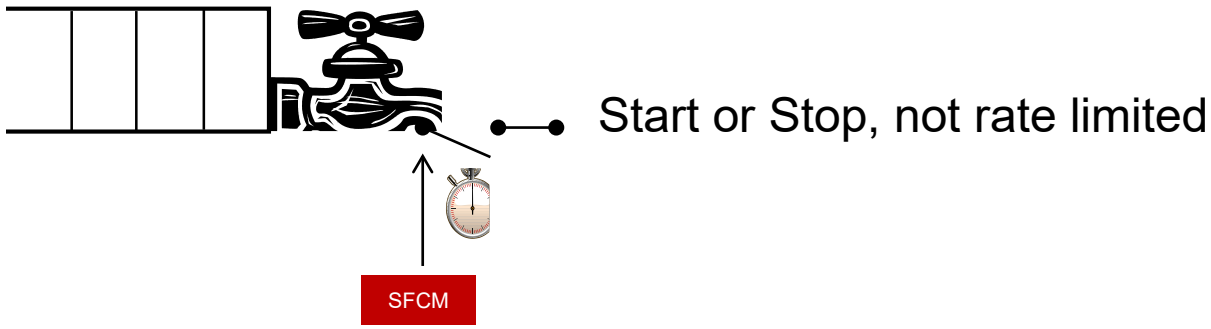
Congestion Point (CP) - Pause Intervals



- Flows Green, Purple and Blue are merging into an Congestion Point

- Below the Threshold packets pass without generating SFCM
- Past the Threshold every new packets generates a SFCM
  - As an enhancement duplicate SFCMs could be filtered at the Congestion Point
- Calculation of the Pause Interval (PI) has three components:
  - Time To Drain (TTD) the congested FIFO
  - Time To Source (TTS) from the congested FIFO
  - Time From Source (TFS) to the congested FIFO
- TTD can be calculated from the number of octets in the FIFO and the current FIFO bandwidth
- TTS is the latency for delivering a SFCM from the congested FIFO to the source FIFO
- TFS is the latency for delivery of traffic from the source FIFO to the congested FIFO

Reaction Point (RP) – Operation Rules

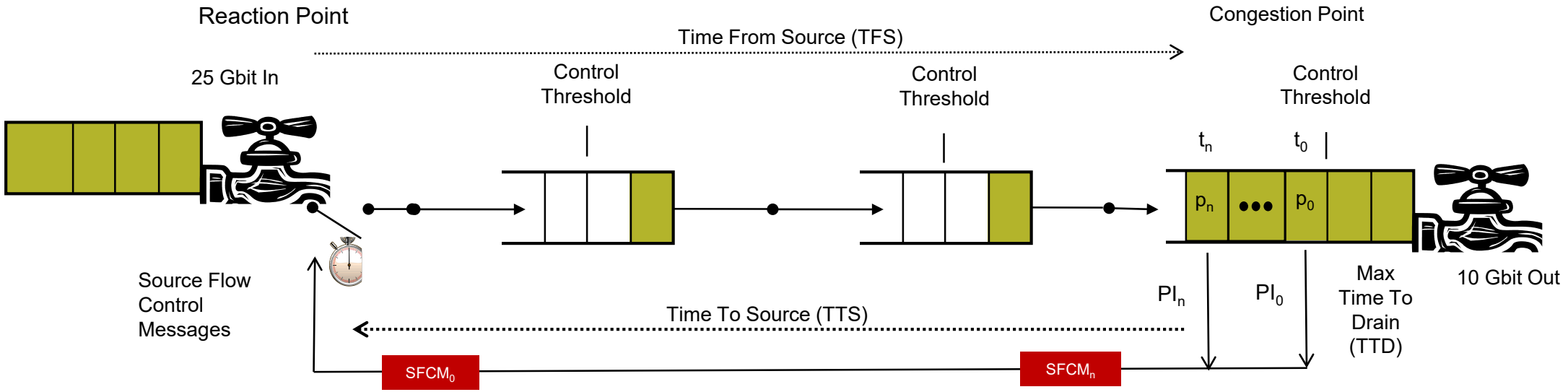


– A SFC Reaction Point transmits at the full speed, which may be determined by other mechanisms such as DCQCN, until it receives a SFCM at which time it pauses transmitting for the pause interval

– After pausing transmission for the pause interval the reaction point resumes transmitting at full speed

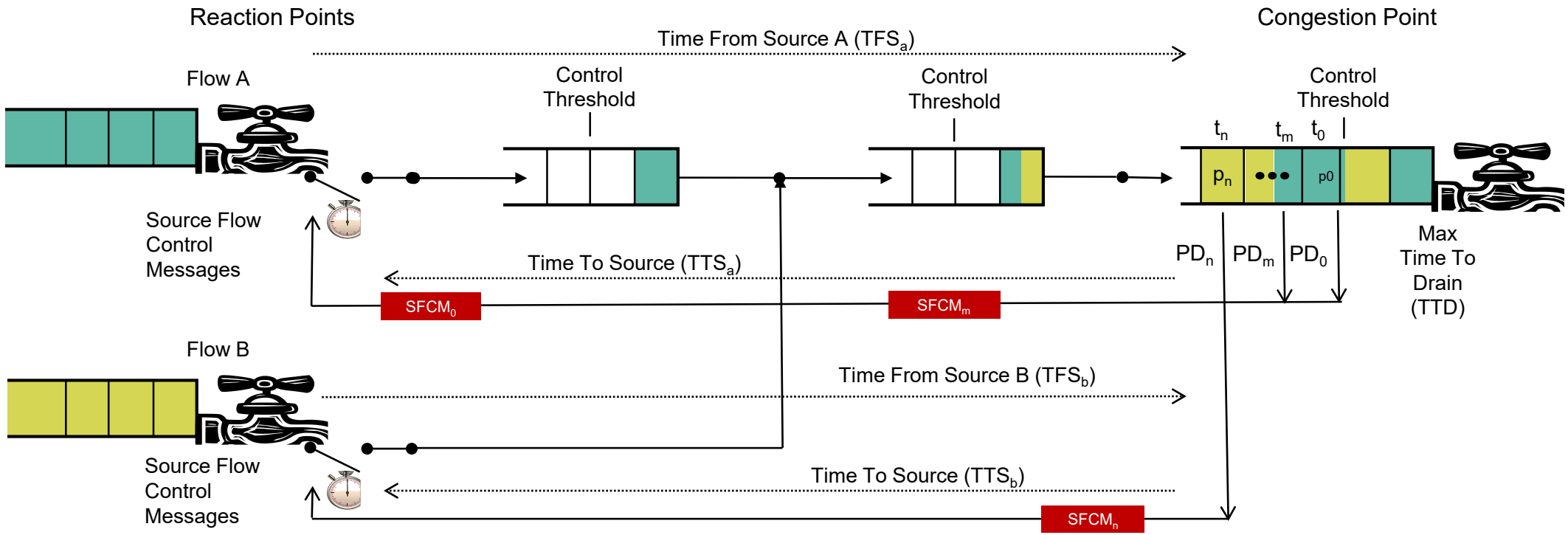
– If the reaction point is already pausing when a SFCM is received, the reaction point resets the pause time to the maximum of the remaining Pause Interval or the new Pause Interval

Rate Mis-match Example



- At time t_0 the Congestion Point receives a packet p_0 which pushes the FIFO depth past the Control Threshold causing $SFCM_0$ to be sent to the Reaction Point with Pause Interval PI_0
- At time t_0 + TTS the Reaction Point receives the $SFCM_0$ and starts pausing for PI_0
- Past time t_0+TTS $SFCM_1$ - $SFCM_n$ arrive at the Reaction Point restarting the Pause Interval with PI_1 - PI_n
- At time t_n ~= t_0 + TTS + TFS traffic from the Reaction Point will stop arriving at the Congestion Point until t_n + (TTS+TFS+PI_n) given:
  - The Source FIFO is delivering constantly at it's maximum capacity (i.e. 10 Gbits)
  - All Congestion Points are operating below their Control Thresholds except the Destination FIFO
- Ideally the Pause Interval seen at the Congestion Point will be sufficient to drain the FIFO. Assuming the drain time is TTD = TTS+TFS+PI_n and solving for PI_n we have PI_n = TTD - (TTS + TFS) which is independent from the sourced bandwidth
- If 0 is used for (TTS + TFS) there is no over-run risk, however the throughput is reduced proportionate to (TTD-(TTS+TFS))/TTD

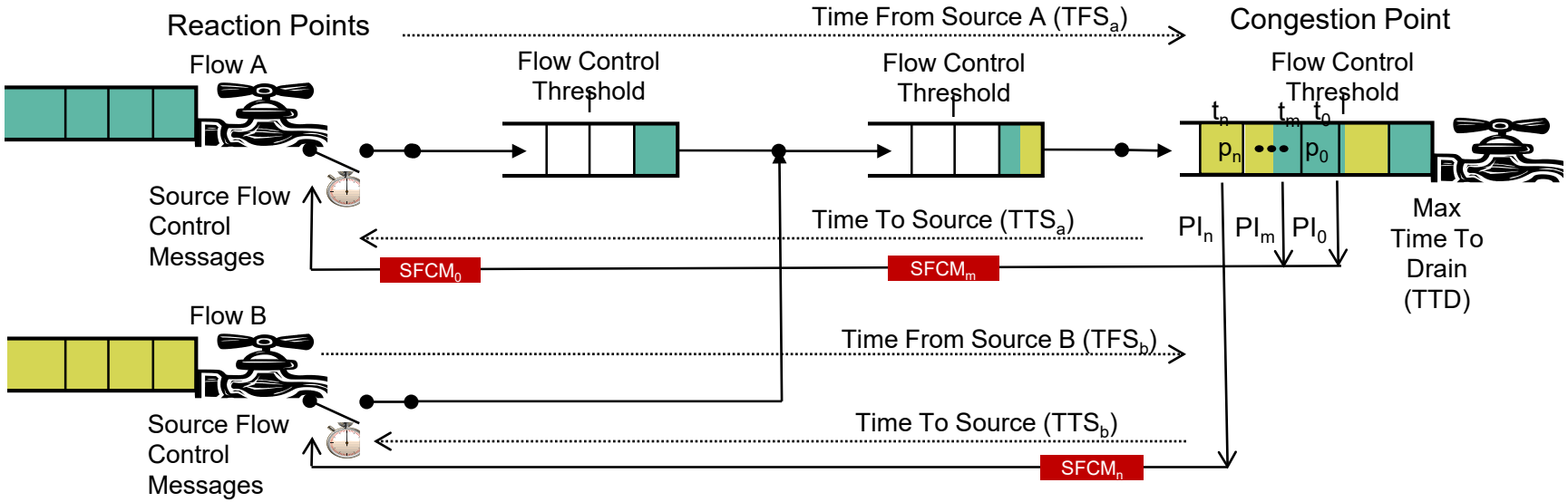
Flow Merging Example



– At time t_m the Congestion Point receives the last packet p_m from flow A which is over the Threshold causing $SFCM_m$ to be sent to the Reaction Point A with Pause Interval PI_m

– At time t_n the Congestion Point receives the last packet p_n from flow B which is over the Threshold causing $SFCM_n$ to be sent to the Reaction Point B with Pause Interval PI_n

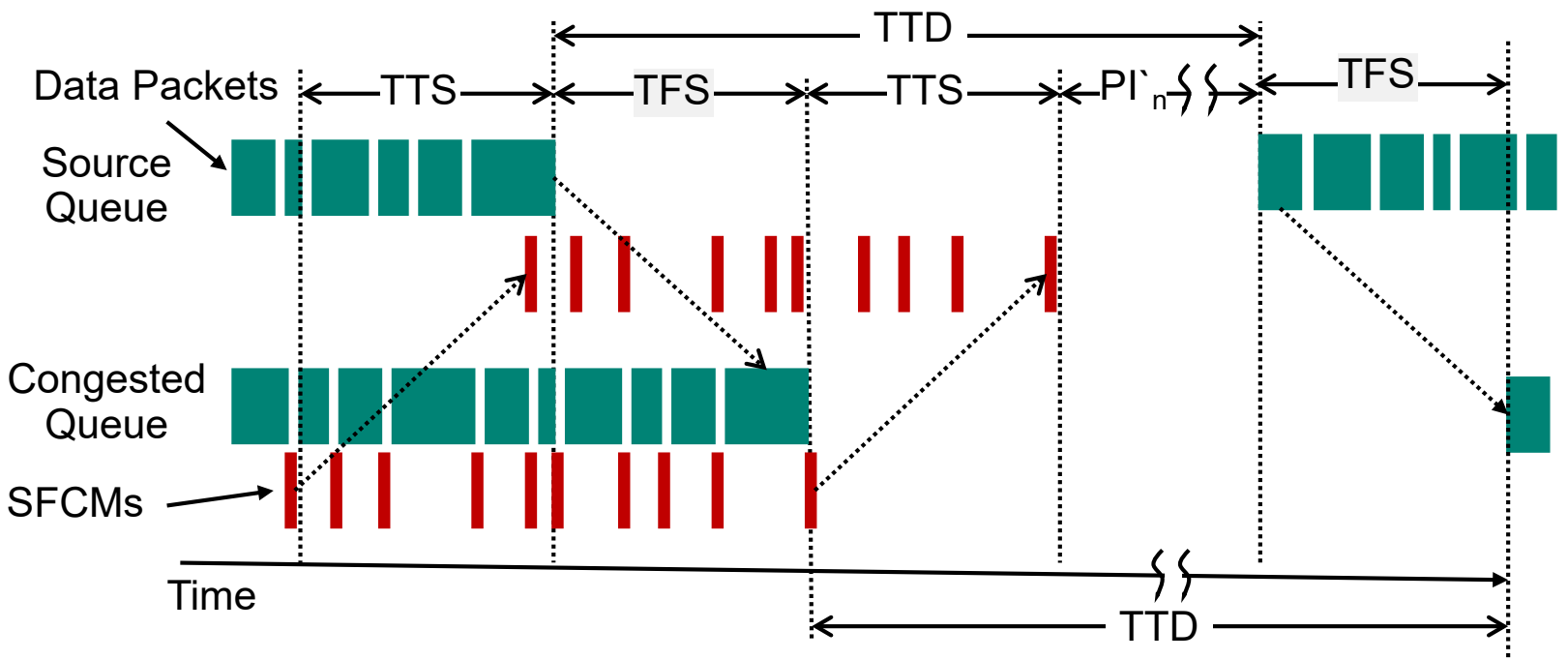
Flow Merging Example Continued



- Reaction Point A receives $SFCM_m$ at time tm + TTS_a and starts pausing for PI_m
- Reaction Point B receives $SFCM_n$ at time tn + TTS_b and starts pausing for PI_n
- At time t_m ~= t_0 + TTS_a + TFS_a traffic from the Reaction Point A will stop arriving at the Congestion Point until t_m + (TTS_a+TFS_a+PI_m)
- At time tn ~= t0 + TTSb + TFSb traffic from the Reaction Point B will stop arriving at the Congestion Point until t_n + (TTS_b+TFS_b+PI_n)
- The Pause Interval seen at the Congestion Point from Reaction Point A is approximated by taking TTD_m = TTS_a+TFS_a+PI_m. Solving for PI_m we have: PI_m = TTD_m - (TTS_a + TFS_a)
- The Pause Interval seen at the Congestion Point from Reaction Point B is approximated by taking TTD_n = TTS_a +TFS_a +PI_n . Solving for PI_n giving PI_n = TTD_n - (TTS_a + TFS_a)

aruba
a Hewlett Packard Enterprise company

SFC Timing Detail

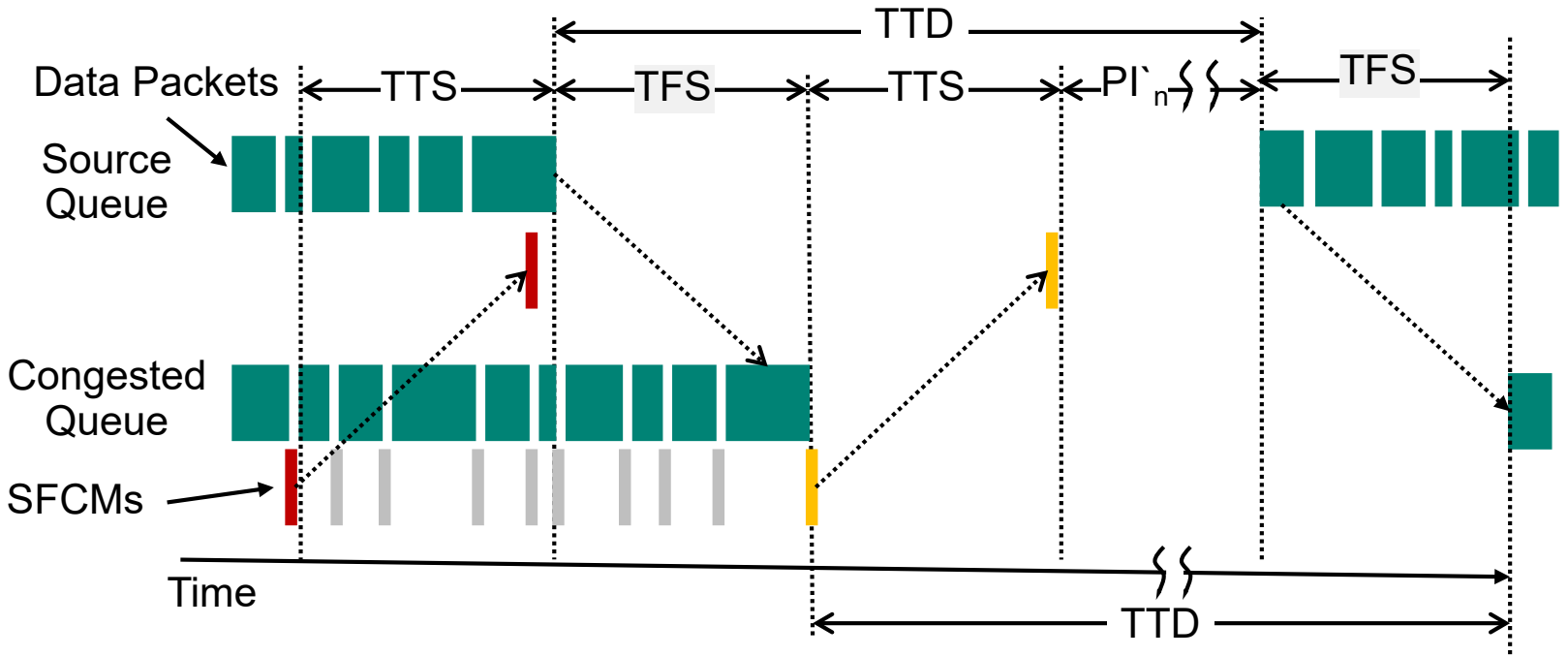


- When the last packet from the source arrives at the congested queue the source has already been paused for TFS
- The last overload packet received at the congested queue will generate a SFCM which is received at the source queue TTS after the last overload packet
- At the time the last SFCM is received at the source the source has already paused for ~TFS+TTS (excluding SFCM wire times)

- Ignoring TTS and TFS results in pausing the source queue longer than necessary resulting in decreased bandwidth
- For a 25 Gbit source at 75% link bandwidth we have 34 Kbytes in transit over the time TTS+TFS. With a 125 Kbyte buffer this represents 25% bandwidth loss per flow.
  - With small buffers per Gigabit (Gbits/(Kbytes at Congest Point)) the percentage will get worse because the threshold will need to be proportionately smaller
- The term TFS+TTS is small enough to be ignored, however providing options for correcting TFS+TTS is desirable
  - An ideal solution measures TTS using a time stamp in the SFCM and then approximates TFS ~= TTS. The reaction point then reduces PI to PI` by subtracting the TTS+TFS approximation. This solution gives a precise correction independent from topology.
  - A more pragmatic solution is to set a reserve buffer in the congested queue of 10-30% when determining PI. The reserve buffer size then depends on the location in the topology and the source data rate.

aruba
a Hewlett Packard Enterprise company

SFCM Filtering



- When the last packet from the source arrives at the congested queue the source has already been paused for TFS
- The last overload packet received at the congested queue will generate a SFCM which is received at the source queue TTS after the last overload packet
- At the time the last SFCM is received at the source the source has already paused for ~TFS+TTS (excluding SFCM wire times)

- Most of the SFCMs for a specific destination can be removed except:
  - The first SFCM for the destination which is necessary to start pausing the reaction point
    - When filtering all but the first SFCM the pause interval does not need to compensate for TTS+TFS
  - The last SFCM for the destination which is useful (but not essential) for precise pause interval timing
    - Filtering all but last is difficult since there is no way to identify the last packet
  - Any SFCM which occurs longer than the pause interval from the last SFCM transmitted
    - Should not happen if the threshold is a lot higher than the bytes transmitted during TTS+TFS

Comparisons

High Performance Buffer/Congestion Control Protocols

	PFC	DCQCN	SFC
Reaction Time (Time To Stop RP Egress)	Slowest >50-200 usec	Middle > 100 usec	Fastest Between 2-10 usec
Time To Stop CP Ingress	Fastest 2 usec	Slowest > 100 usec	Middle-Fast Between 2-15 usec
Reaction Point Location	Per Class	Per Flow	Per Class / Per Interface / Per Flow
Handles Short Lived Flows	Yes	No	Yes
Handles Long Lived Flows	No	Yes	Yes
Micro-Burst Control	Yes	Poor	Yes
Congestion Spread Free	No	Yes	Yes
Deadlock Free	No	Yes	Yes

– Estimates based on operation within a data center pod with end stations connected using 25 Gbit Ethernet

– These protocols are not mutually exclusive and may be combined if desired

aruba
a Hewlett Packard Enterprise company

Thank You

aruba
a Hewlett Packard
Enterprise company

Backup Slides