QoS Catalyst 6500 platformon

Balla Attila
CCIE #7264
balla.attila@synergon.hu

Tartalom

SYNERGON INFORMATION SYSTEMS PLC.

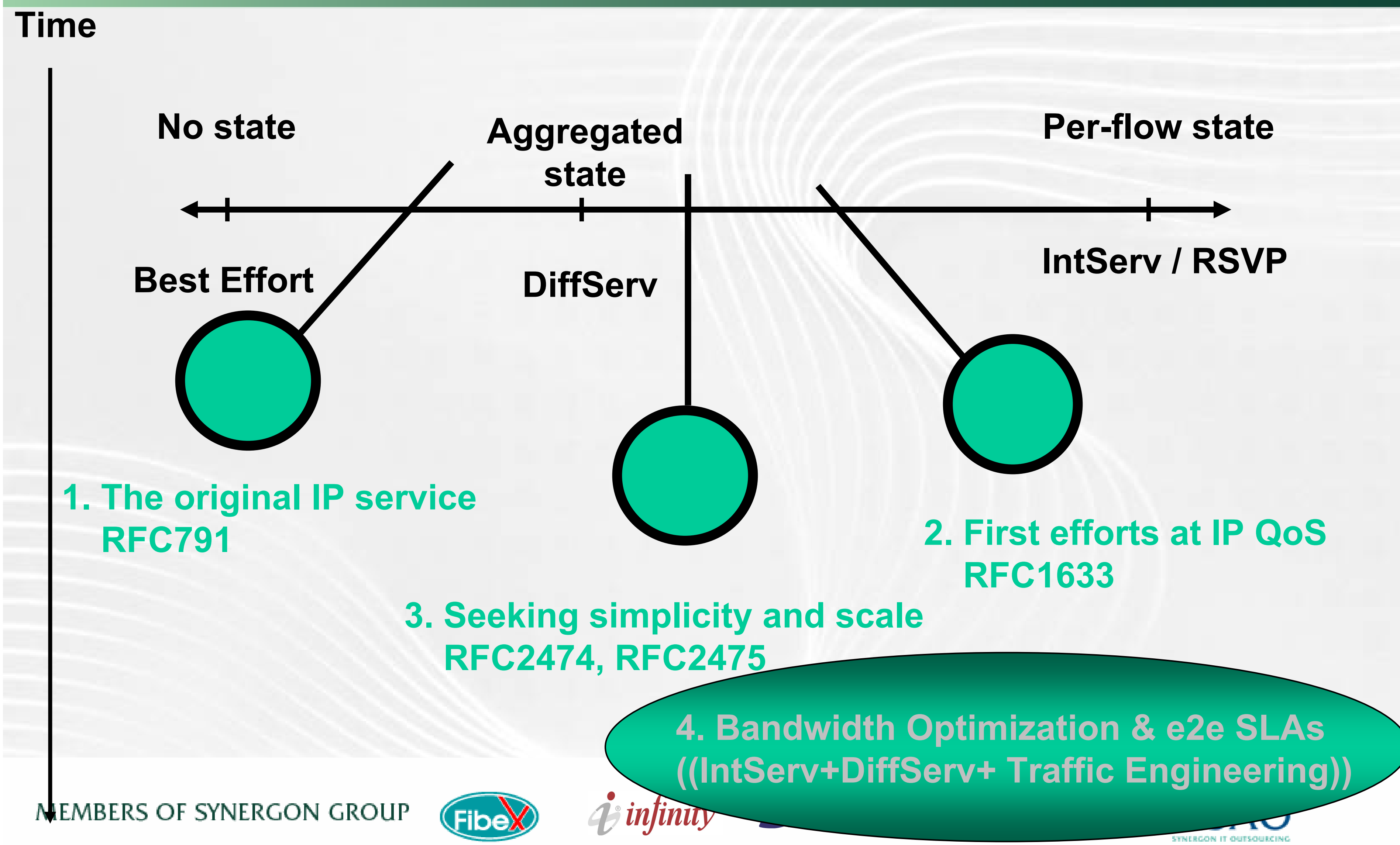
- Bevezető
- QoS modellek
- Catalyst 6500 és QoS
  - Interface Modules

Bevezető

SYNERGON INFORMATION SYSTEMS PLC.

- Miért van szükség QoS-re?
  - Sávszélesség
  - Torlódás
- Miért pont Catalyst 6500-n?
  - Egyik legelterjedtebb L3 switch
  - Többféle Supervisor
    - Sup2: PFC2+MSFC2
    - Sup32: PFC3B+MSFC2a
    - Sup720: PFC3[A|B|BXL]+MSFC3
      - PFC-3C, 3CXL

QoS Models


Time
No state
Aggregated state
Per-flow state
Best Effort
DiffServ
IntServ / RSVP
1. The original IP service RFC791
2. First efforts at IP QoS RFC1633
3. Seeking simplicity and scale RFC2474, RFC2475
4. Bandwidth Optimization & e2e SLAs ((IntServ+DiffServ+ Traffic Engineering))

Simple DiffServ Recipe

- Edge
  - Classification
  - Marking/Coloring
  - Optional policing/shaping
  - Congestion avoidance
    - WRED
  - Congestion management
    - Queuing
- Core
  - Congestion avoidance
    - WRED
  - Congestion management
    - Queuing

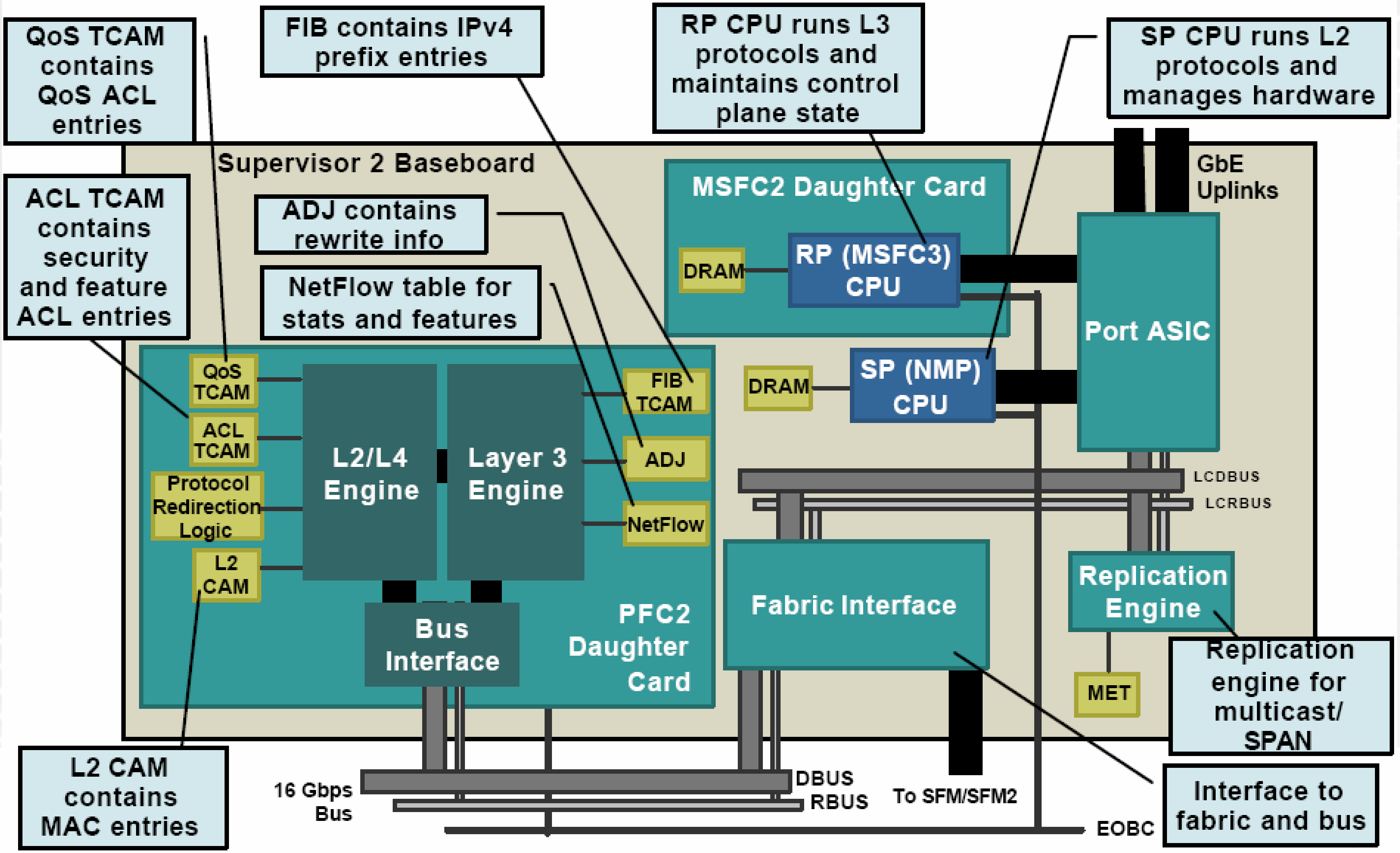
Emlékeztető

- Catalyst 6500 architektúra
  - 2004. Salgóbánya

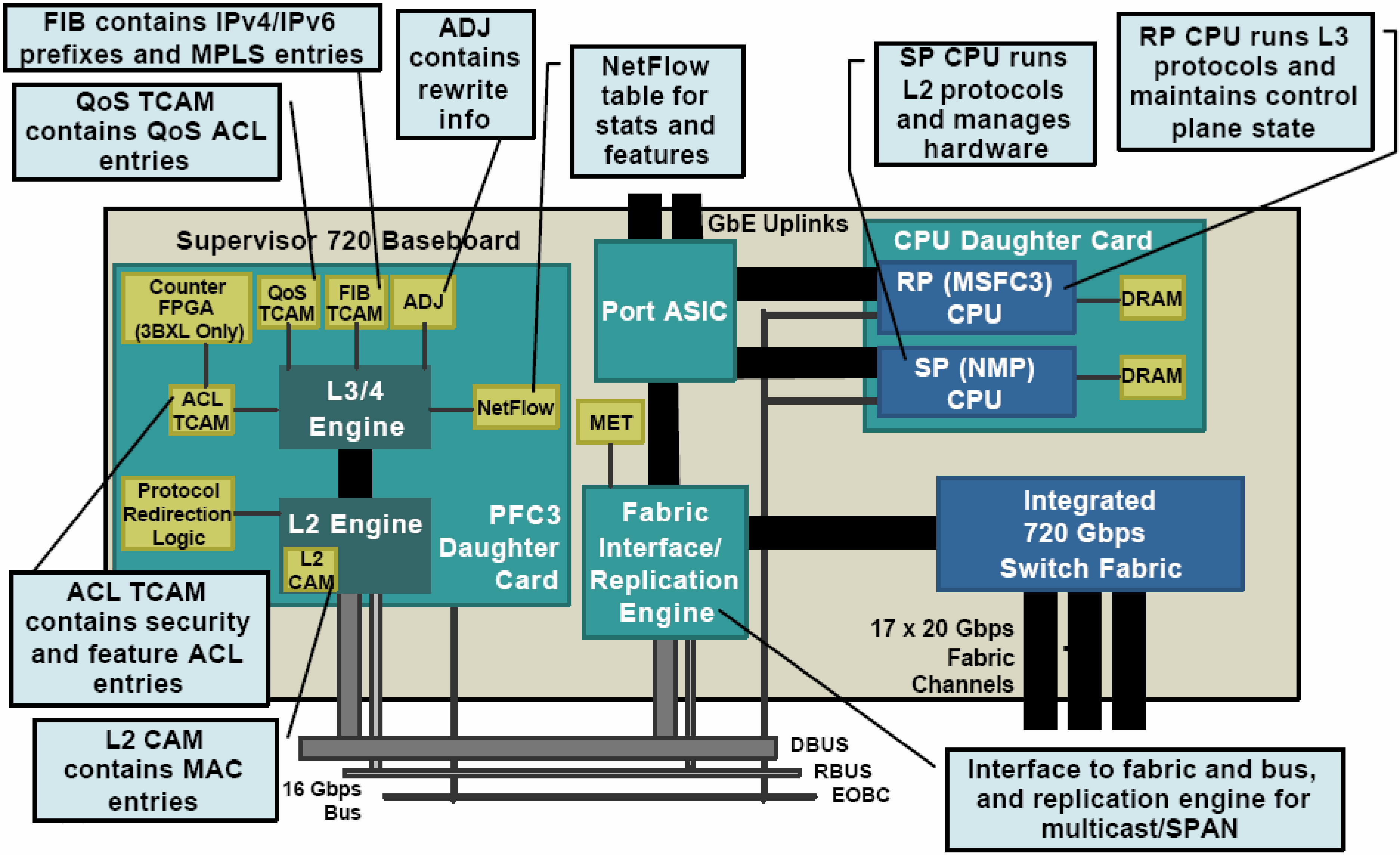
- Sok funkció HW-ből
  - L2 forwarding, L3 forwarding, ACL, Netflow
  - Policy Feature Card
    - CAM, TCAM
    - Nagysebességű memória

- Mi a helyzet a QoS-sel?

Supervisor Engine 2 – PFC2



Supervisor Engine 720 – PFC3



Policy Feature Card

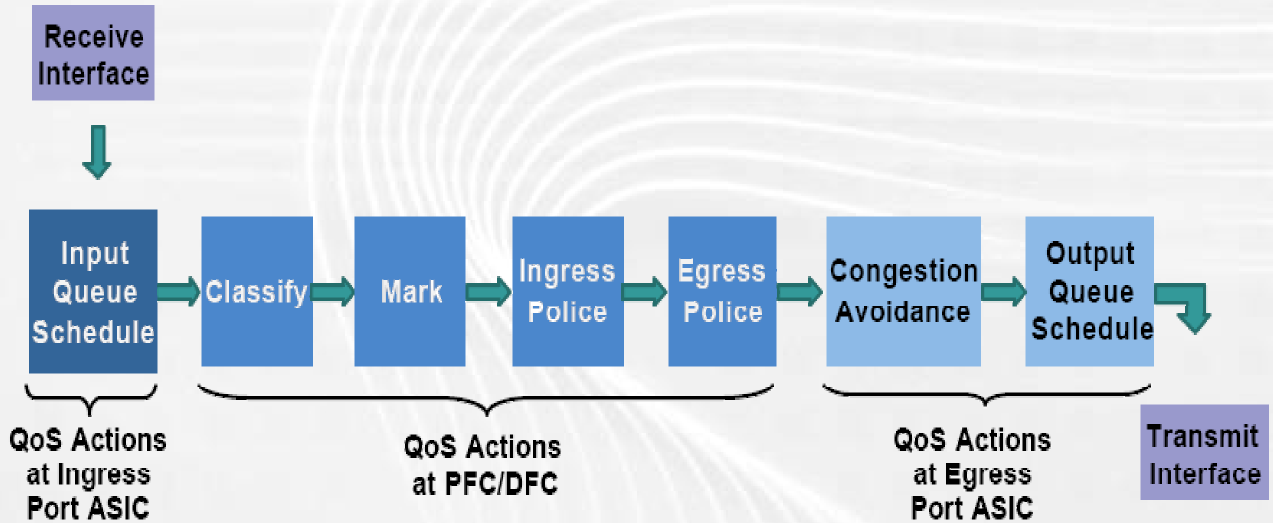
- Daughter card for supervisor engine
- Provides the key components enabling highperformance hardware packet processing
  - 15/30Mpps
- Supervisor 2 supports PFC2
- Supervisor 720 supports:
  - PFC3A
  - PFC3B
  - PFC3BXL

Key Hardware-Enabled Features:

- Layer 2 switching
- IPv4 unicast forwarding
- IPv4 multicast forwarding
- Security ACLs
- QoS/policing
- NetFlow statistics
- PFC3 Also Supports:
  - IPv6, MPLS, Bidir PIM, NAT/PAT, GRE/v6 tunnels

Cat6500 QoS Model

- Actions at ingress
  - Scheduling
  - CoS overwriting

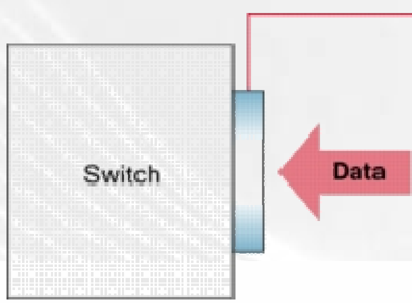


- Actions by PFC
  - Classification L2/L3/L4
  - Policing
  - Mark down

- Actions at Egress
  - Rewrite ToS
  - Scheduling – each queue has configurable size and thresholds
  - WRED, Tail-drop

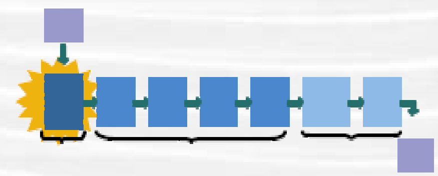
QoS Modes

- Disabled – by default
  - CoS, DSCP, IP Precedence values are preserved
- Enabled
  - `mls qos`
  - Trust port
    - `mls qos trust`
  - PFC3 assigns a priority to each frame
    - Based on QoS Policies
    - Based on CoS, DSCP, IP Precedence
  - Rewrite CoS, DSCP, IP Precedence fields
    - `no mls qos rewrite ip dscp`
- Queueing-only
  - `mls qos queueing-only`
  - CoS, DSCP, IP Precedence values are preserved



Trust Setting	Result
Un-rusted	CoS/ToS Set to Zero
Trust CoS	CoS/ToS Maintained
Trust-IP Precedence	CoS/ToS Maintained
Trust DSCP	CoS/ToS Maintained

Input Queue Scheduling



- Input scheduling only performed if port configured to trust COS
  - Scheduling based on input COS
  - Implements tail-drop thresholds
  - Thresholds at which packets with different COS values are dropped

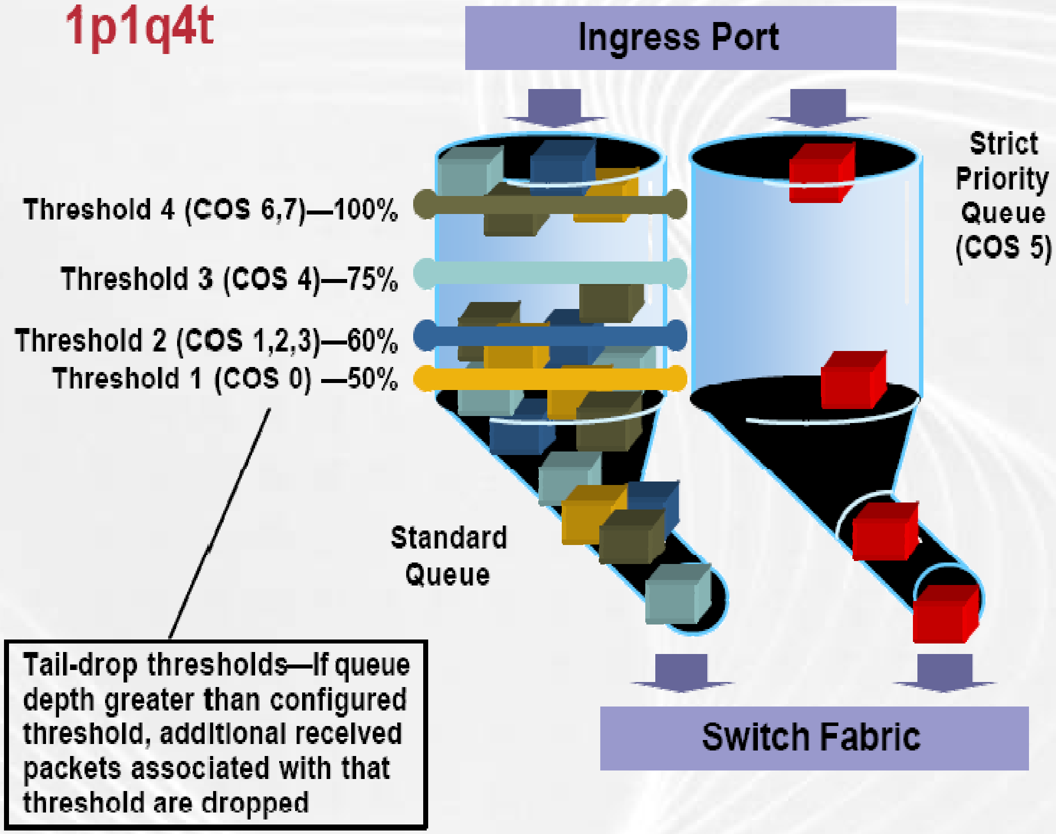
- Queue structure example: 1p1q4t
  - One strict-priority queue, one standard queue with four tail-drop thresholds
  - Some Line Cards support WRED
  - `show queueing interface GigabitEthernet1/1`
  - `rcv-queue threshold`

FAQ: What Are The Buffer Sizes and Queue Structures for the Different Modules?
http://www.cisco.com/warp/public/cc/pd/si/casi/ca6000/prodlit/buffe_wp.pdf

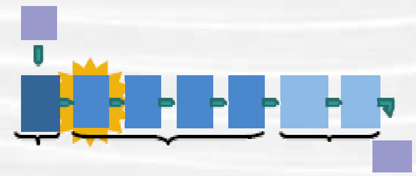
Input Queue Scheduling Details

SYNERGON INFORMATION SYSTEMS PLC.

1p1q4t

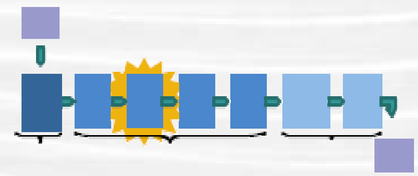


Classification



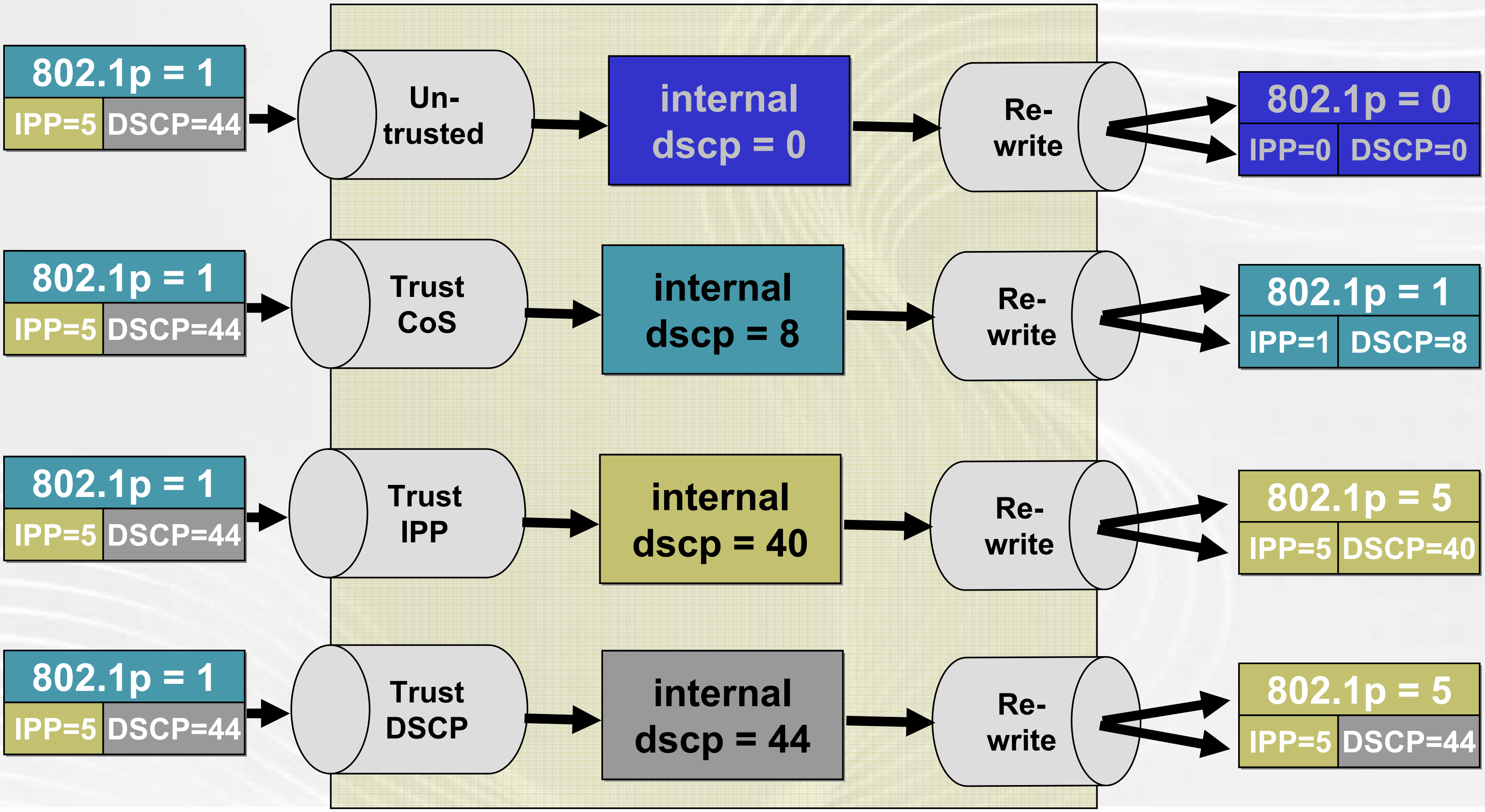
- Selects traffic for further QoS processing
  - Marking
  - Policing
- Based on
  - Port trust
  - QoS ACLs
  - Policing mark-down

Marking



- Untrusted port
  - Set a default QoS value
- Trusted port
  - Use the marking (COS, precedence, DSCP) provided by upstream device
- QoS ACLs
  - Set QoS values based on standard or extended ACL match

Trust


802.1p = 1
IPP=5
DSCP=44
Un-
trusted
internal
dscp = 0
Re-
write
802.1p = 0
IPP=0
DSCP=0
802.1p = 1
IPP=5
DSCP=44
Trust
CoS
internal
dscp = 8
Re-
write
802.1p = 1
IPP=1
DSCP=8
802.1p = 1
IPP=5
DSCP=44
Trust
IPP
internal
dscp = 40
Re-
write
802.1p = 5
IPP=5
DSCP=40
802.1p = 1
IPP=5
DSCP=44
Trust
DSCP
internal
dscp = 44
Re-
write
802.1p = 5
IPP=5
DSCP=44

QoS ACLs

- Used to classify traffic based on Layer 3 and Layer 4 information
- Hardware support for standard and extended IPv4 and MAC QoS ACLs
- Use QoS TCAM and other ACL resources to classify traffic for marking and policing
- Dedicated QoS TCAM
  - 32K entries/4K masks
- Share other resources (LOUs and labels) with security ACLs
  - `show tcam counts`

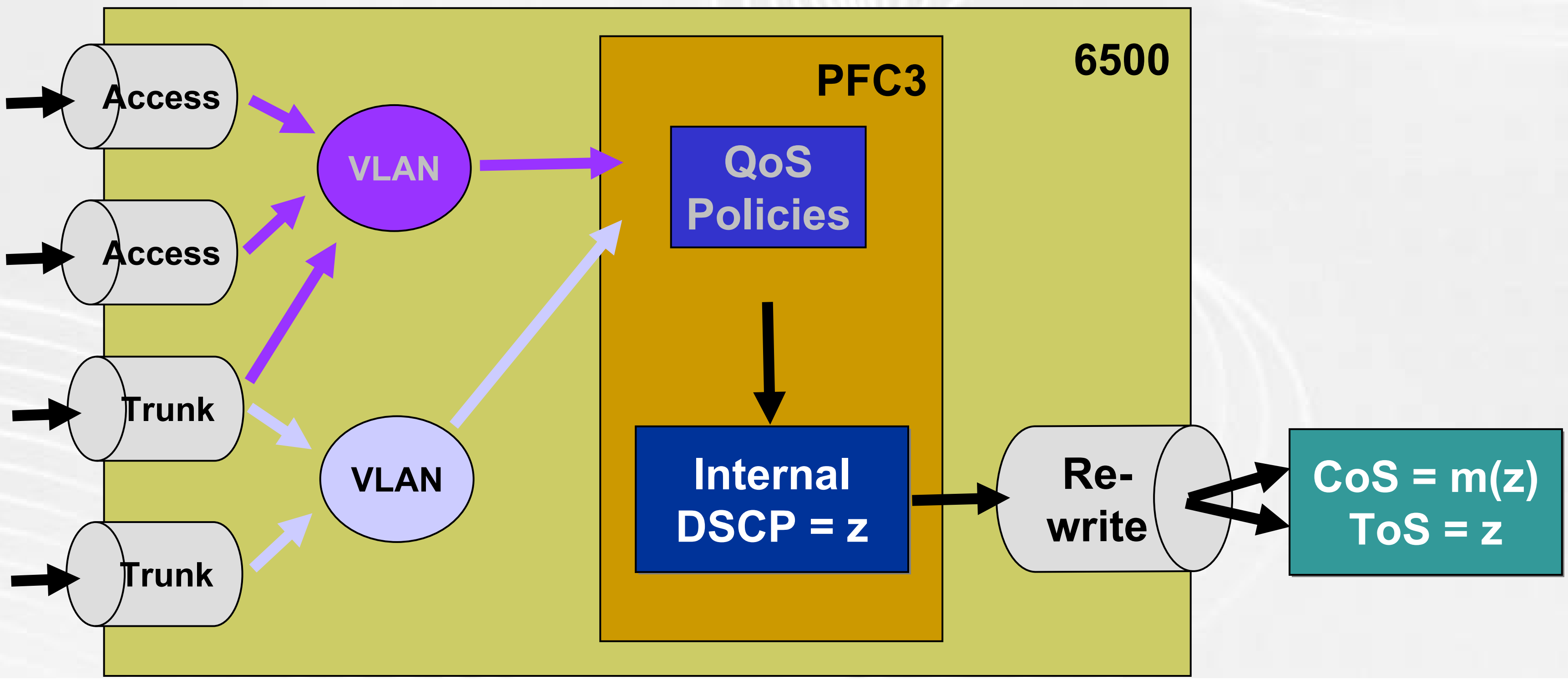
Marking with QoS ACL

- Marking is implemented with the MQC set and police commands
- It does not use the CAR rate-limit command
- Supported by PFC3
  - set commands
    - `set ip precedence`
    - `set ip dscp`
    - `set mpls exp`
  - police commands
    - `set-prec-transmit`
    - `set-dscp-transmit`
    - `set-mpls-experimental-imposition-transmit`
    - `policed-dscp-transmit`

Vlan Based QoS

SYNERGON INFORMATION SYSTEMS PLC.

Each physical Catalyst port can optionally be configured for VLAN-based QoS. In this case, a service-policy is applied to the VLAN interface.



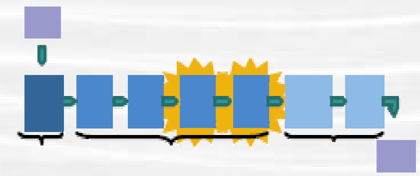
Vlan Based QoS

If a physical port is not configured for VLAN-based QoS, its traffic will not be included in VLAN-based QoS, even if it has traffic in that VLAN.

```
interface GigabitEthernet4/1
  description Customer facing interface
  switchport mode access
  switchport access vlan 100
  mls qos vlan-based

interface GigabitEthernet4/2
  description Customer facing interface
  switchport mode access
  switchport access vlan 100
  mls qos vlan-based

interface GigabitEthernet6/1
  description Core facing interface
  switchport mode trunk
  switchport trunk allowed vlan 100

interface vlan 100
  service-policy input markdown-ip
```

Policing

SYNERGON INFORMATION SYSTEMS PLC.



- Defines a policy for traffic on a port or VLAN, based on the rate at which traffic is received
- Based on a classic token bucket scheme
  - Tokens added to bucket at fixed rate (up to max)
  - Packets with adequate tokens are “in profile”: packet transmitted, tokens removed from bucket
  - Packets without adequate tokens are dropped or marked down
- Leaky Bucket Model
  - Burst ~ depth of the bucket
  - Rate ~ hole in the bucket
- Dual Leaky Bucket
  - PIR, MaxBurst

Note! PFC2 uses Layer 3 packet size; PFC3 uses Layer 2 frame size

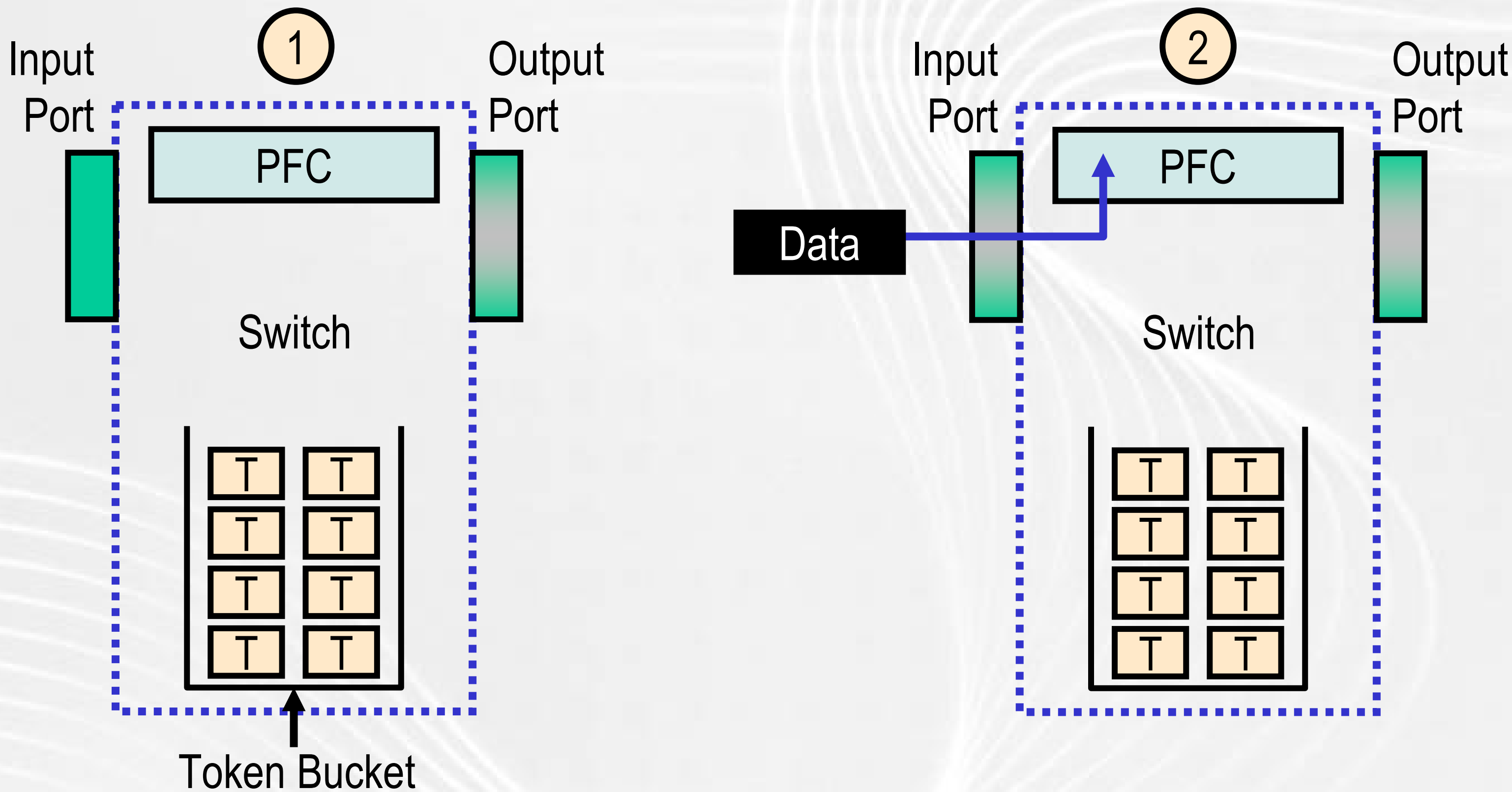
Policing Actions

SYNERGON INFORMATION SYSTEMS PLC.

- In-Profile traffic
  - Forward
- Out-of-Profile traffic
  - Mark-Down
    - Modifying the priority
    - Forwarding
  - Drop
- Hardware interval
  - 0.25msec

Policing I.

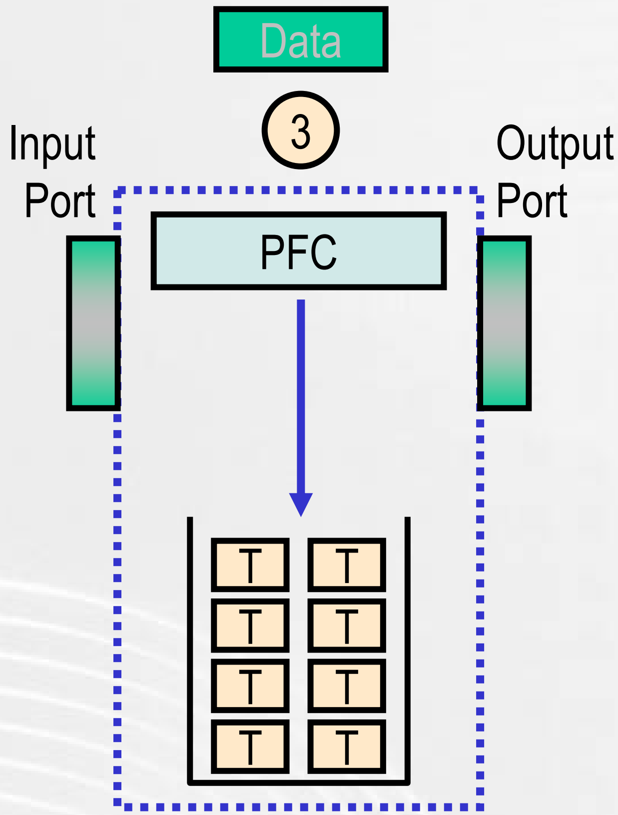
Policing uses a concept of a token bucket for policing data – essentially data is only sent when tokens exist in the bucket The following will try to explain how this works…



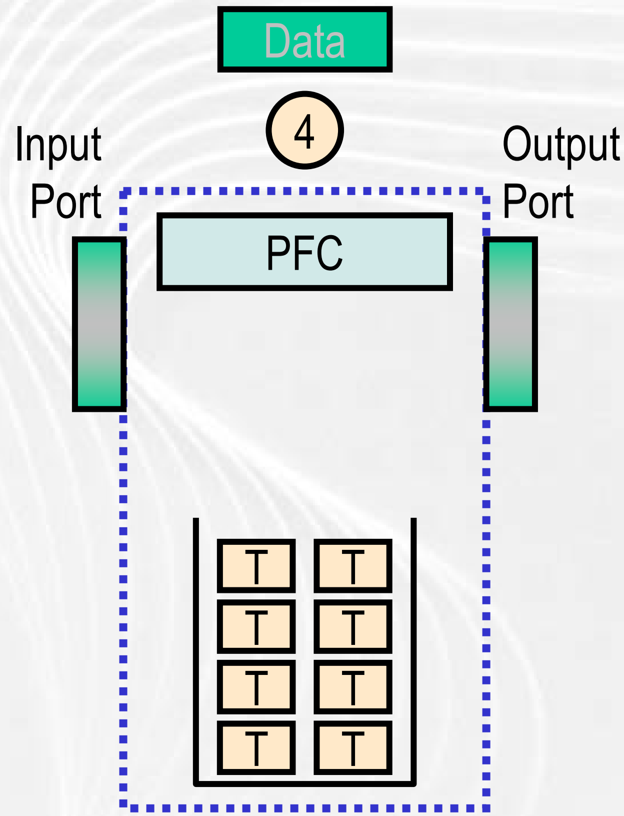
1. At time interval T0, the bucket is loaded with a full complement of tokens

2. When a packet arrives at the PFC, the number of bits that make up the packet are counted

Policing II.

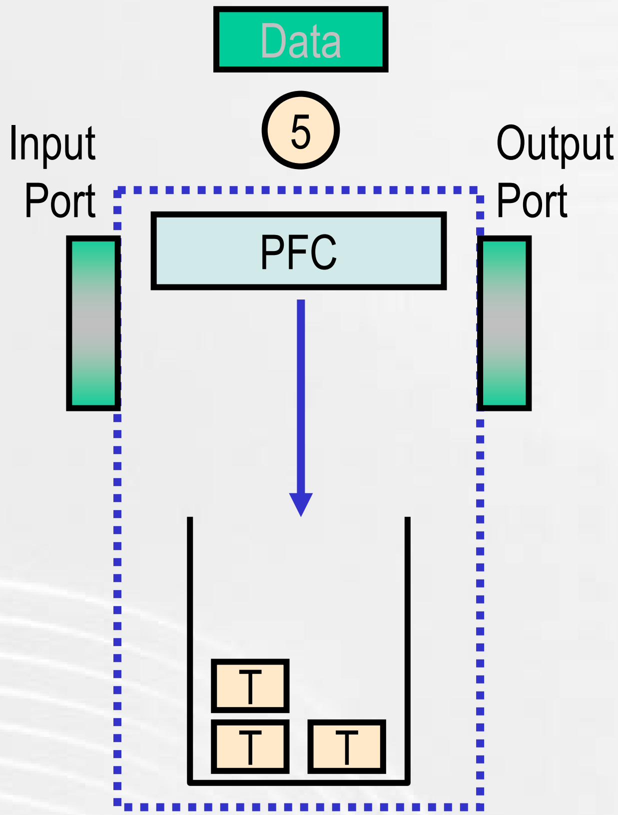


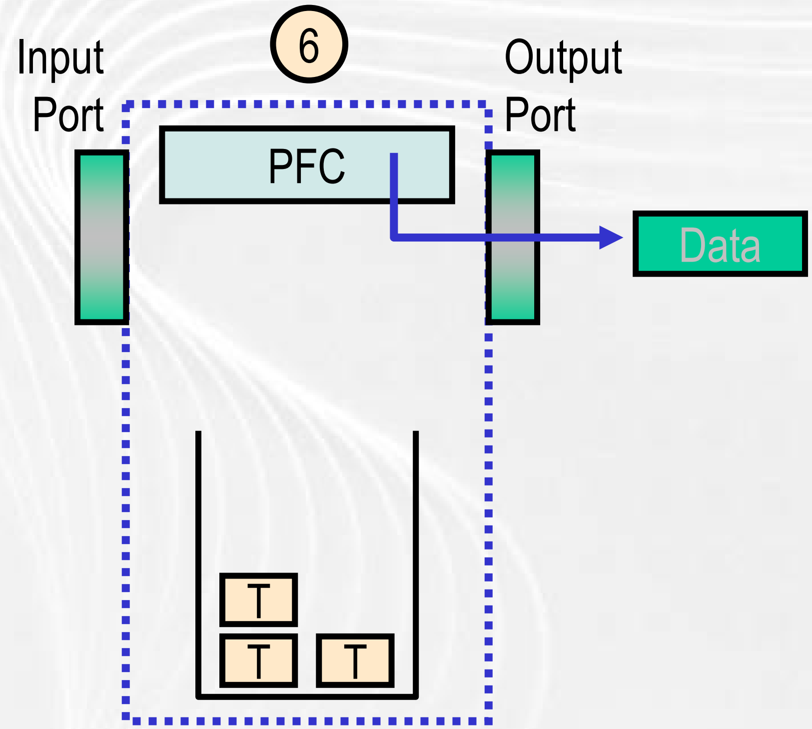
3. The PFC checks the token bucket



4. If the number of tokens in the bucket is >= the number of bits in the packet, the packet can be forwarded – if not, the packet is dropped

Policing III.



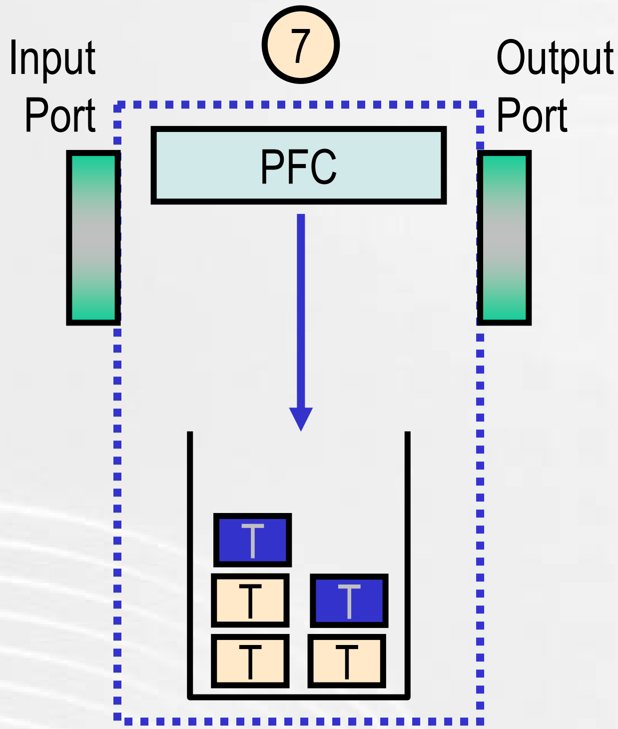


5. The tokens are removed from the bucket

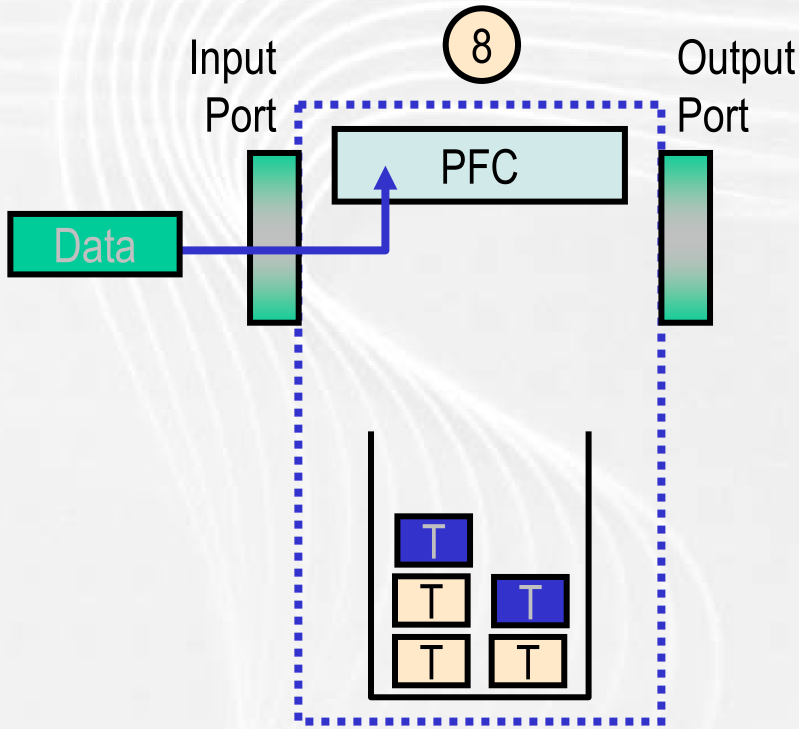
6. The packet is sent by the PFC to its onward destination – other packets will be forwarded in that time interval is enough tokens exist

Policing IIII.

SYNERGON INFORMATION SYSTEMS PLC.



7. At the end of the time interval, the token bucket is replaced with a new complement of tokens



8. The next packet will only be forwarded if there are enough tokens in the bucket – and so goes the cycle

Policing Example

```
police 100000000 26000
 conform-action set-dscp-transmit
 exceed-action drop
```

- policed rate of 100Mb/sec
  - REPLENISHMENT RATE every 1/4000th of a second = RATE / Interval = 100,000,000 / 4000 = 25,000 tokens every 1/4000th of a second
  - Bucket Depth = BURST = 26,000 tokens

Policing Example

SYNERGON INFORMATION SYSTEMS PLC.

Arrival rate is 1Gig/s @ 64byte packets

Time Interval	Bits clocked in interval	Tokens at Start of interval	Number of bits that can be sent	How many packets can be sent?	Tokens at end of interval	Number of bits that are dropped
T1	250,000	26,000	25.600	50	400	224,400
T2	250,000	25,400	25,088	49	912	224,912
T3	250,000	25,912	25,600	50	312	224,400
T4	250,000	25,312	25,088	49	224	224,688
And so on	…	…	…	…	…	…

Policing Details

SYNERGON INFORMATION SYSTEMS PLC.

- Aggregate policers – Bandwidth limit applied cumulatively to all flows that match the ACL
  - Example: All FTP flows limited in aggregate to configured rate
- Microflow policers – Bandwidth limit applied separately to each individual flow that matches the ACL
  - Example: Each individual FTP flow limited to configured rate
  - Leverages NetFlow table
  - Ingress only
  - Single leaky bucket model
- Policing action may reclassify and remark certain traffic
- Supervisor 2 and Supervisor 720 support INGRESS policing, on a per-switchport, per-Layer 3 interface, or per-VLAN basis
- Supervisor 720 also supports EGRESS aggregate policing on a per-VLAN or per-Layer 3 interface basis

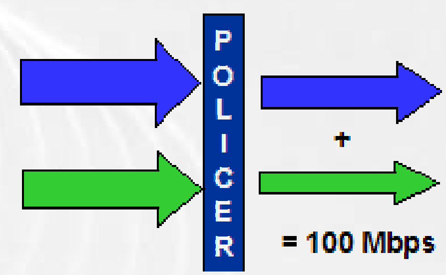
Aggregate Policer

- Per interface
  - Single interface
  - Single class

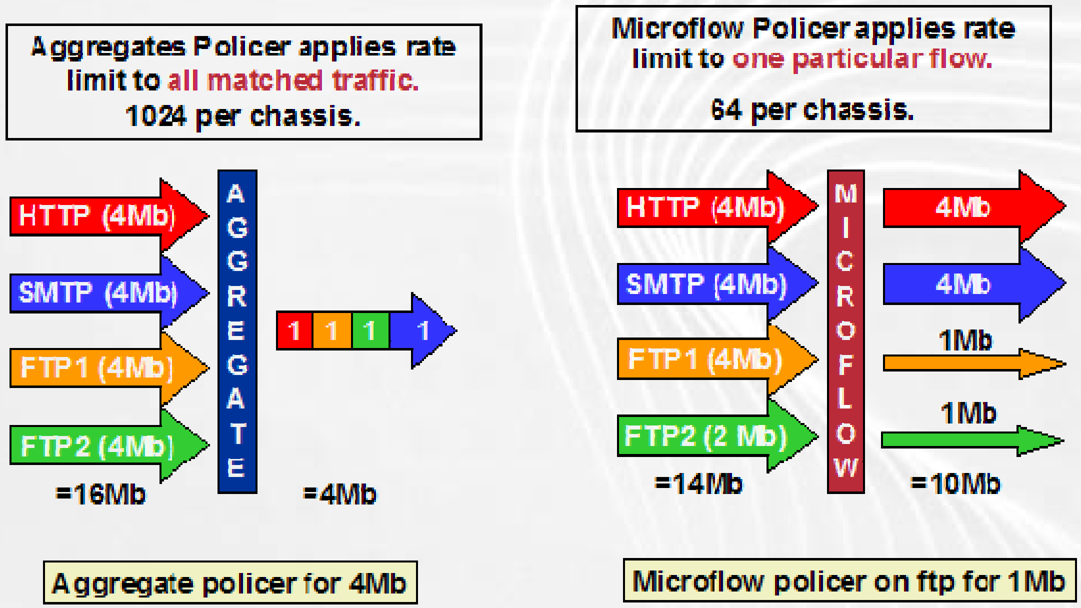
One interface per policer
One class per policer
All flows in the class

- Shared/Named
  - Multiple interface
  - Multiple Class

Several interfaces
One class
All flows



Aggregate vs Microflow Policer



Uninterested traffic will bypass the flow policer

User Based Rate Limiting

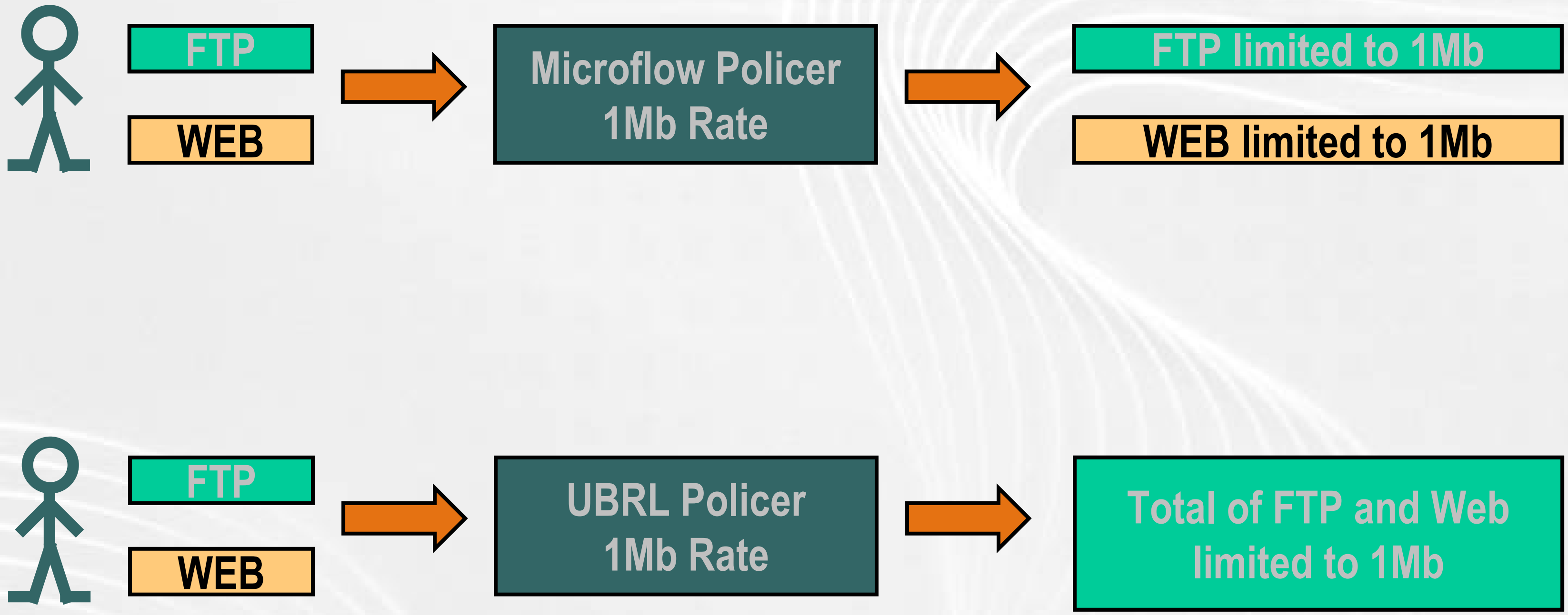
SYNERGON INFORMATION SYSTEMS PLC.

- PFC3 feature
  - PFC1 & PFC2 single flow mask
    - When a microflow policer is enabled, other processes that use the flow mask also have to use the same full flow mask.
  - PFC3 four flow masks (two of them are reserved)

- Based on Netflow
  - Source only
    - Supported by PFC3 only
  - Destination only
  - Destination – Source
  - Full

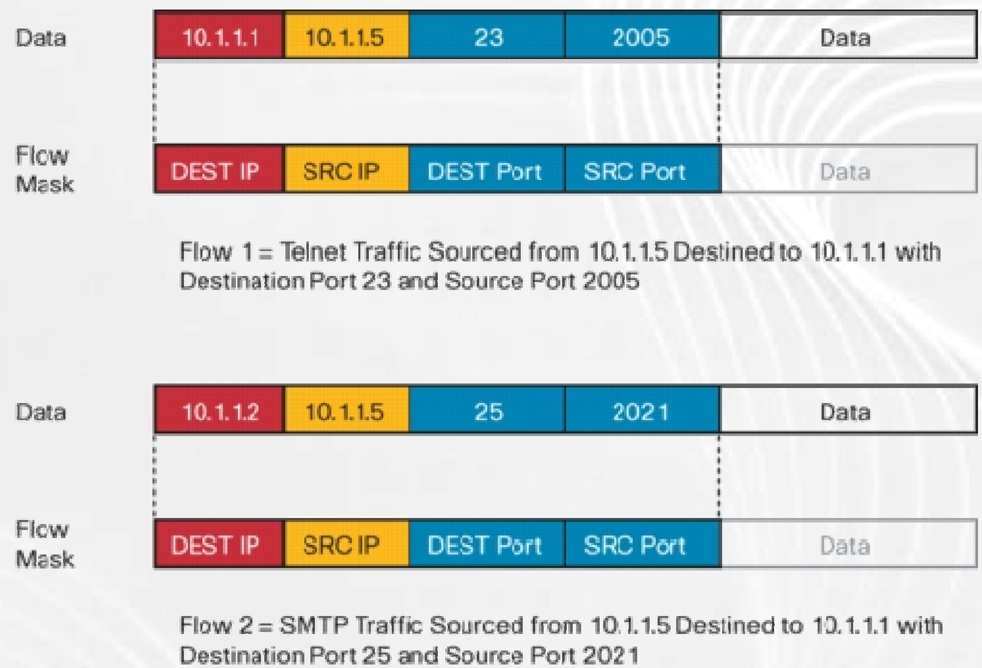
UBRL vs Microflow

SYNERGON INFORMATION SYSTEMS PLC.



Netflow Masks

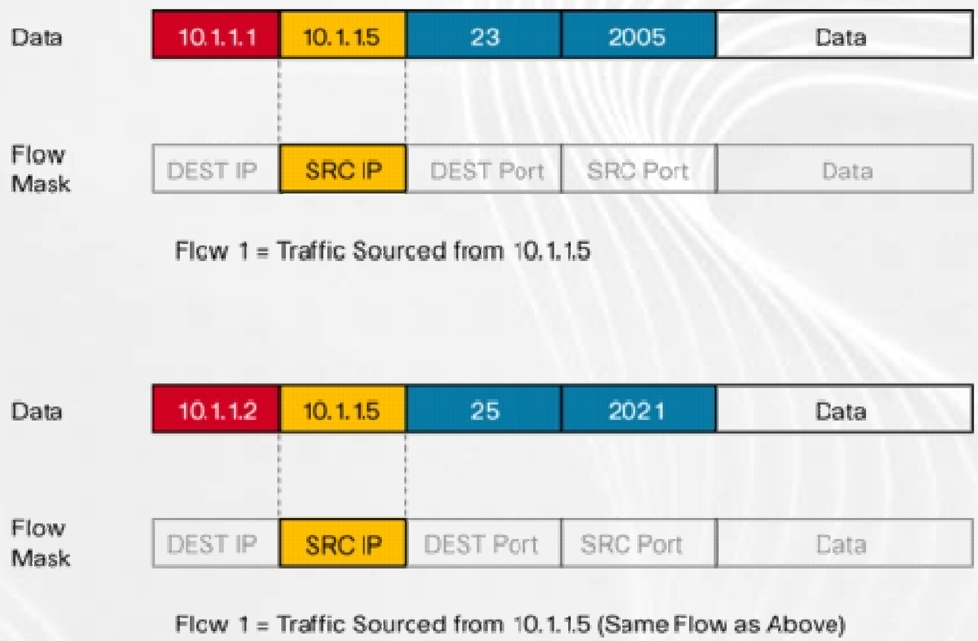
Full Flow Mask



A given user who initiates a Telnet session and accesses an e-mail server would initiate two separate flows

Netflow Masks

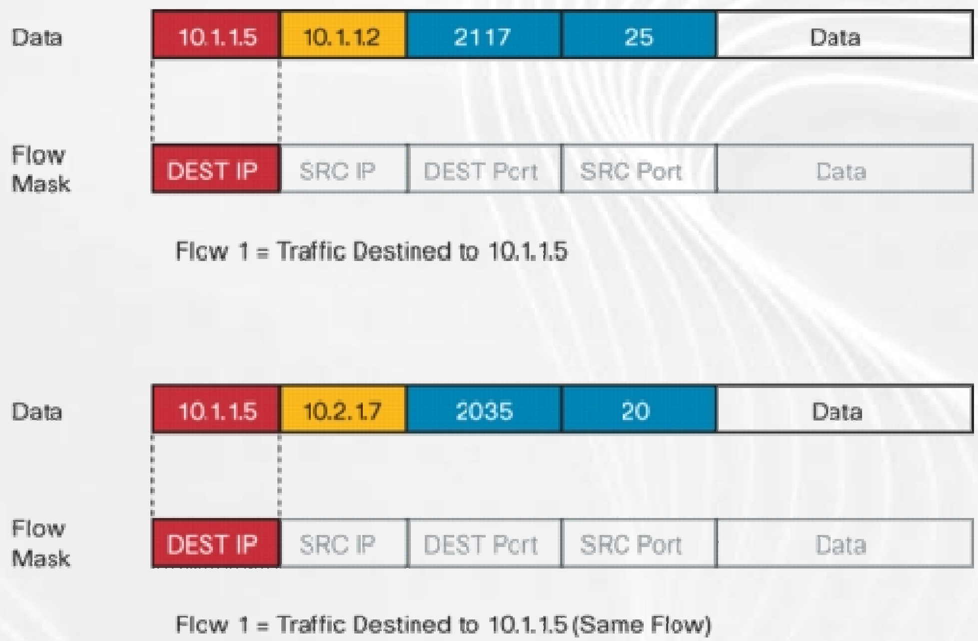
Source IP only



The same user who initiated a Telnet and e-mail session would now be seen as initiating a single flow

Netflow Masks

Destination IP only



In both cases, traffic from each server is considered to be part of the same flow as the mask is, only using the destination address as the unique flow identifier

UBRL Configuration – Example

- Two policers
  - Uplink interface
    - Connected to the Internet
    - Input policer, flow-mask dst-only
  - Downlink interface
    - Connected to the Users
    - Input policer, flow-mask src-only

UBRL Configuration – Example

SYNERGON INFORMATION SYSTEMS PLC.

Applied to user ports
Source only Flow

Applied to uplink ports
Destination only Flow

```
access-list 101 permit ip 10.10.10.0
   0.0.0.255 any

class-map Users-Outbound
 match access-group 101

policy-map Users-Outbound
 class Users-Outbound
  police flow mask src-only 100000 …

int range fast4/1-48
 service-policy input Users-Outbound
```

```
access-list 102 permit ip any
   10.10.10.0 0.0.0.255

class-map Users-Inbound
 match access-group 102

policy-map Users -Inbound
 class Users-Inbound
  police flow mask dest-only 100000 …

int gig 3/1
 service-policy input Users-Inbound
```

Output Policer

- Not supported by PFC2
- Based on received frame
  - Vlan interface
  - L3 port
  - Cannot be applied to switchport
    - PFC3 knows only the egress Vlan and Line Card
- Implemented Parallel with Input Policy
  - By default
  - You can enable sequential processing on PFC-3B & PFC-3BXL
- An output policy is instantiated 1+N times, where 1 represents the PFC3 and N represents the number of DFCs
  - `show mls qos ip`
- Aggregate Policer only

Configuring Policing – MQC

SYNERGON INFORMATION SYSTEMS PLC.

Aggregate Policer

```
Router(config)#mls qos aggregate-policer <name>
```

Class-map

```
Router(config)#class-map class-map-name
Router(config-cmap)# match <ip precedence | ip dscp | access-group>
```

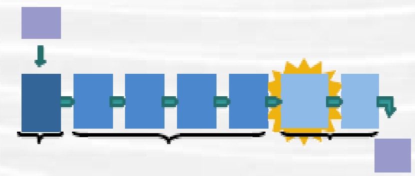
Policy-map Class Action

```
Router(config)#policy-map policy-map-name
Router(config-pmap)#class class-name
Router(config-pmap-c)#police <<flow|aggregate> | set | trust>
```

Service-policy

```
Router(config)#interface interface-name
Router(config-if)#service-policy <input | output> policy-map-name
```

Congestion Avoidance



Weighted Random Early Detection (WRED):

- Congestion AVOIDANCE mechanism
- Weighted because some classes of traffic are more important or sensitive than others
- Random in that the packets to discard are randomly chosen within a class
  - Which classes are more subject to discards is configurable
- Prevents global TCP window synchronization and other disruptions

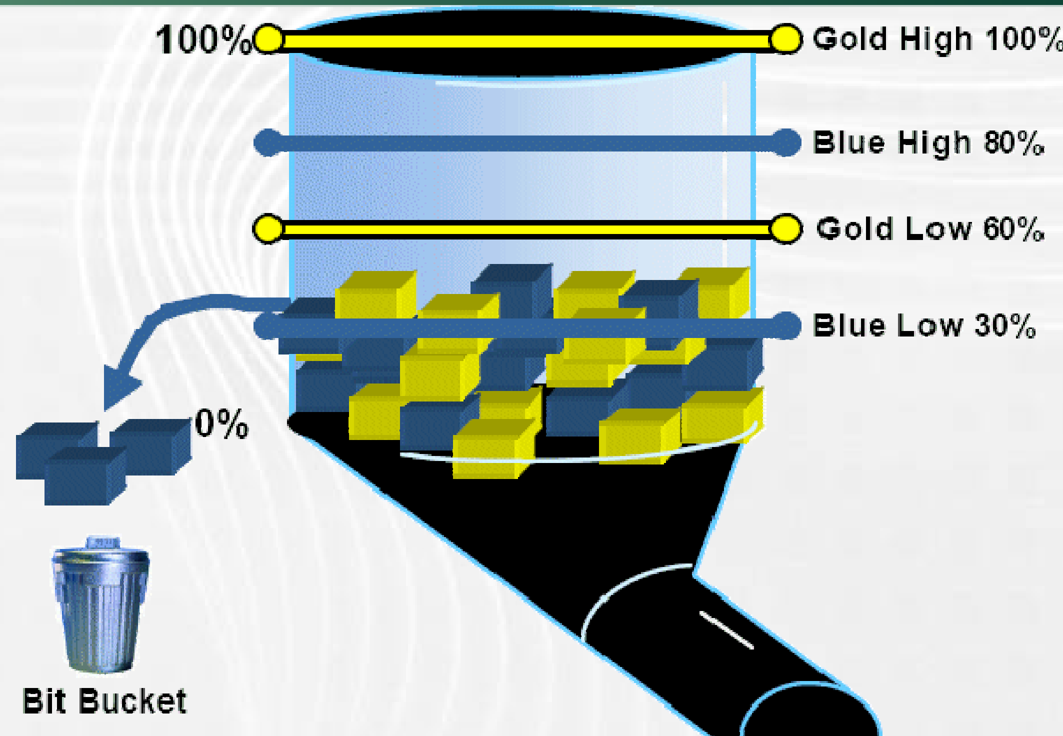
WRED Thresholds

- Each queue has multiple WRED thresholds
- Low threshold is the point at which random discards will begin for a particular class
- High threshold is the point at which tail-drop for the particular class begins
- As buffers fill…
  - Rate of discards increases for traffic associated with lower thresholds
  - Higher thresholds are reached, and new traffic classes are subject to random discards

WRED Operation

SYNERGON INFORMATION SYSTEMS PLC.

- Two classes,
  two thresholds each:
  - ➢ Gold
    - – 100% high
    - – 60% low
  - ➢ Blue
    - – 80% high
    - – 30% low



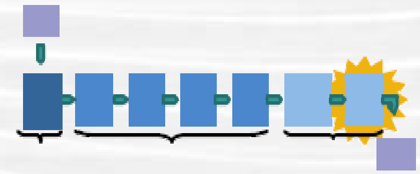
- When queue depth exceeds 30%, some random blue packets are dropped
- When queue depth exceeds 60%, drop rate for blue packets increases and gold packets become subject to random drops
- When queue depth exceeds 80%, tail-drop occurs for blue packets (all exceed packets dropped), and drop rate for gold packets increases

WRED Configuration Commands

SYNERGON INFORMATION SYSTEMS PLC.

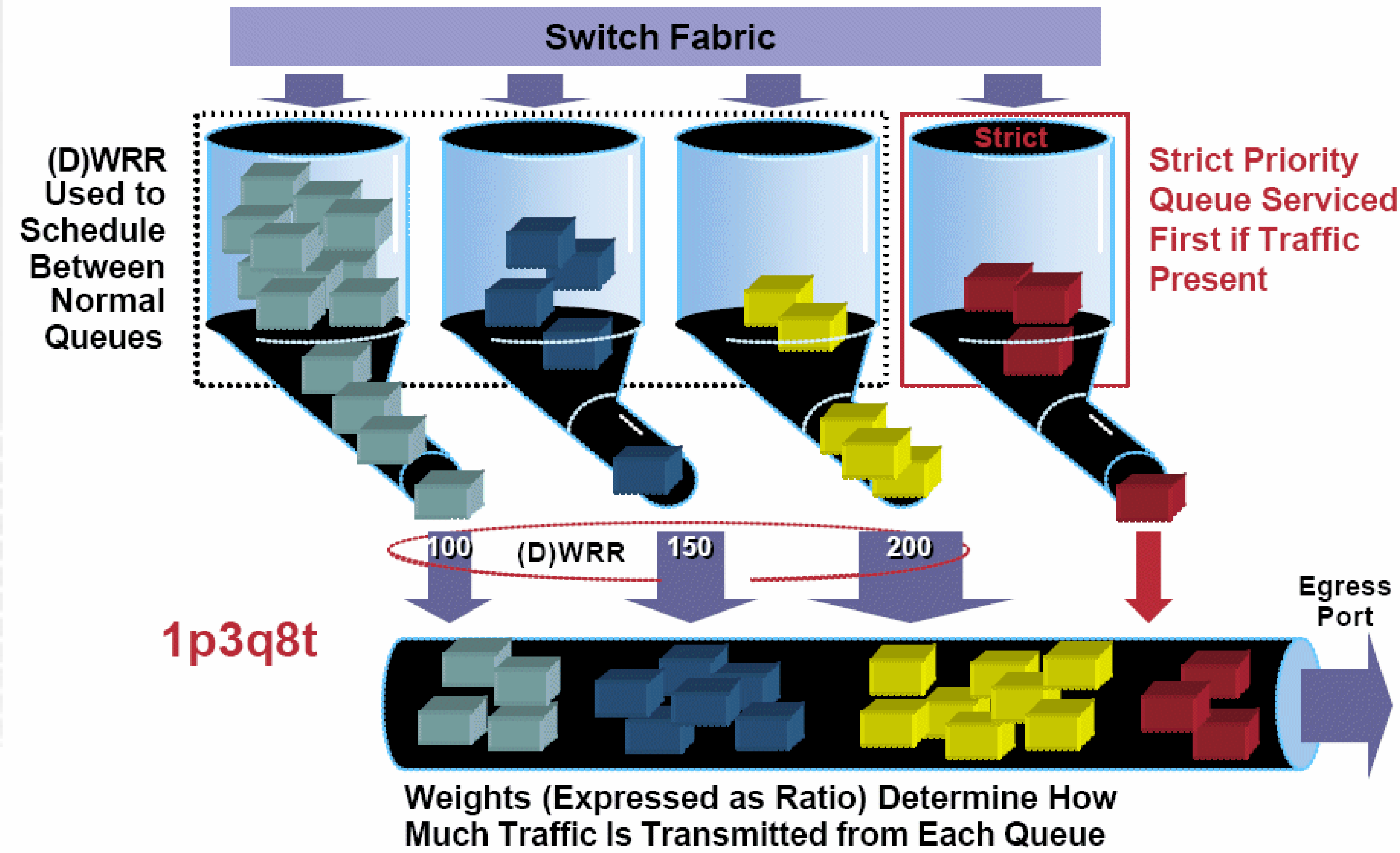
```
wrr-queue random-detect queue-id

wrr-queue random-detect
   {max-threshold | min-threshold}
   queue-id threshold-percent-1 ... threshold-percent-n
```

Output Queue Scheduling



- Scheduling based on COS
- Implements tail-drop or WRED thresholds
- Queue structure example: 1p3q8t
  - One strict-priority queue, three standard queues with eight WRED thresholds each

Output Queue Scheduling Operation



Output Queuing I.

SYNERGON INFORMATION SYSTEMS PLC.

- Weighted Round Robin (WRR)
  - Uses ratio to determine number of packets to transmit from one queue before moving to the next queue
  - Higher weight = more packets transmitted from that queue
  - Unfair with variable-length packets in different queues

- Deficit WRR
  - Also uses ratio, but tracks bytes in each queue using deficit counter
  - Packet(s) transmitted during queue servicing only if size of next packet to transmit is <= deficit counter Deficit counter "refreshed" at beginning of each queue servicing period
  - Results in fair scheduling over time
  - 1p3q1t, 1p2q1t, 1p3q8t, 1p7q4t, and 1p7q8t

Output Queuing II.

- Shaped Round Robin
  - Sup32 only, Integrated port only
  - WS-X6708-3C, 3CXL
    - 200MB buffer/port
  - Shaping instead of policing

- Strict Priority Queuing
  - Only when the strict priority queue is empty will the scheduling process recommence sending packets from WRR queues

Output Queuing Configuration Commands

```
wrr-queue [bandwidth | shape]

show queueing interface
```

További QoS kérdések

SYNERGON INFORMATION SYSTEMS PLC.

- FlexWan
  - PA
- Optical Service Module
  - Standard
  - Enhanced
- SPA Interface Processor
  - SIP-200, SIP-400
  - SIP-600

- Általában a hagyományos MQC eszközök

Sources

- Cisco Catalyst 6500 Series Switches White Papers

http://www.cisco.com/en/US/products/hw/switches/ps708/prod_white_papers_list.html

- Networkers 2004

http://www.cisco.com/networkers/nw04/presos/rst.html

- Command Reference

Kérdések & Válaszok

Balla Attila
balla.attila@synergon.hu