

CESNET Technical Report 3/2003
**QoS in Layer 2 Networks with Cisco
Catalyst 3550**

**Sven Ubik, ubik@cesnet.cz
Josef Vojtěch, xvojtecj@fel.cvut.cz**

April 20, 2003

1 Introduction

Capacity of backbone Internet circuits have increased such that most circuits are now lightly loaded or “overprovisioned” (loosely defined as being loaded up to 10% of their installed capacity). Access networks (LANs and MANs) are often loaded to higher percentage of their capacity. Utilization peaks are also higher in access networks because of lower traffic multiplexing when compared to backbone circuits.

Access networks are often based on layer 2 (link layer) infrastructure or on the combination of layer 2 and layer 3 (network layer) infrastructure. If we find that some explicit network QoS provisioning is needed in access networks to provide end-to-end QoS guarantees, such as priority or bandwidth sharing, it must be implemented with equipment available in access networks. This is usually represented by layer 2 switches sometimes enhanced with some layer 3 functionality.

In this report we provide an overview of QoS provisioning in access networks and we summarize our practical experience with Cisco Catalyst 3550 switch, which is now commonly used in newly installed access networks.

2 Layer 2 QoS

2.1 Service codepoint

A service codepoint denotes a class of service for a particular packet. IEEE 802.1Q [1] standard defines architecture for virtual bridged LANs (VLANs). A part of this architecture is the format of a tag header that can be used to carry VLAN ID and user priority, which is a service codepoint in IEEE 801.1Q networks. The tag header is inserted in an Ethernet frame after the source address field. The format of the tag header is illustrated in Fig. 1. When the first two

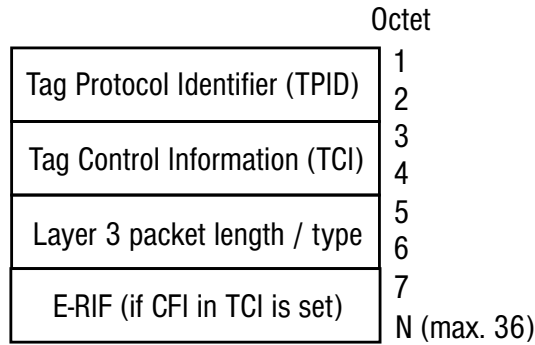


Figure 1: Tag header format

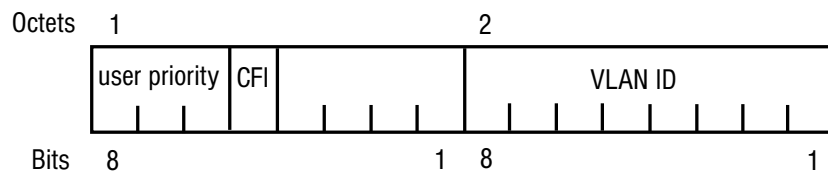


Figure 2: Tag control information (TCI) format

User priority	Acronym	Purpose
0 (default)	BE	best effort
1	BK	background
2	-	spare
3	EE	excellent effort
4	CL	controlled load
5	VI	“video” < 100 ms latency and jitter
6	VO	“voice” < 10 ms latency and jitter
7	NC	network control

Table 1: User priority values and their purpose

bytes, where the layer 3 packet length / type field is expected, is set to 0x8100, this field is considered as a Tag protocol identifier (TPID), which informs that the tag header is present in the frame. Following TPID is the Tag control information (TCI) field, which includes user priority, CFI and VLAN ID. The format of the TCI field is illustrated in Fig. 2. It is followed by the original layer 3 packet length / type field and optionally by the E-RIF field, if CFI bit is set in the TCI field. The user priority field is three bits long allowing for eight different priority values. Each priority value was originally designed for certain purpose, as shown in Table 1.

User priority traffic class	Number of available traffic classes							
	1	2	3	4	5	6	7	8
0 (default)	0	0	0	1	1	1	1	2
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1
3	0	0	0	1	1	2	2	3
4	0	1	1	2	2	3	3	4
5	0	1	1	2	3	4	4	5
6	0	1	2	3	4	5	5	6
7	0	1	2	3	4	5	6	7

Table 2: Recommended mapping of user priorities to traffic classes

2.2 Selecting service

On a Linux host, the user priority field can be set as a part of IEEE802.1Q VLAN support or using a RAW socket. Alternatively, the user priority can be set on the connecting switch. Cisco uses the term CoS (Class of Service) to denote user priority. Cisco Catalyst 2900, 3500XL and 3550 switches allow one default CoS value per input port. On Cisco Catalyst 3550 switches, you can define multiple access lists selecting certain packets and specify one CoS value for each access list. For already tagged frames, their CoS value can be trusted or it can be overridden by a default CoS value.

2.3 Service provisioning

According to IEEE 801.1Q standard, bridges must have ability to regenerate user priority. For each input port, a user priority regeneration table specifies correspondance between input and regenerater user priority. Default mapping is 1:1.

Bridges can provide more than one transmission queue for each port. Frames are assigned to transmission queues according to their user priority using a traffic class table. Queues are mapped 1:1 with traffic classes. Recommended mapping of user priorities to traffic classes for different number of available traffic classes is shown in Table 2.

2.4 Implications of layer 2 QoS provisioning

According to IEEE 802.1Q standard, packets of the same user priority for a given combination of source and destination addresses must not be reordered. We should consider this requirement when using policing on Cisco Catalyst 3550 switch. Out-of-profile packets can be dropped or marked down, that is their user priority can be changed. Due to different user priority, these packets can be put in a different queue, which can result in packet reordering. Our experiments described in section 5 confirm that reordering can happen.

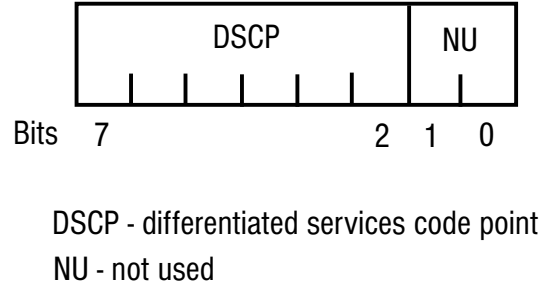


Figure 3: Differentiated services codepoint (DSCP)

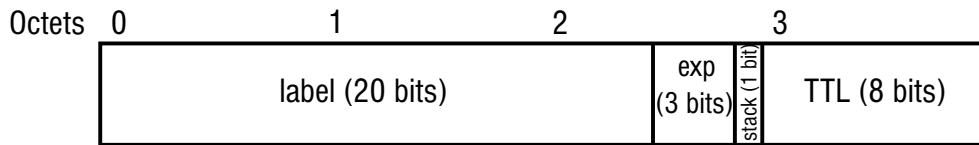


Figure 4: MPLS label with experimental bits

3 Layer 3 QoS

Numerous articles on layer 3 QoS provisioning have been published. We just provide here a review of the service codepoint storage for layer 3, which is important for interaction with layer 2 QoS.

3.1 Service codepoint

Within the context of differentiated services in IPv4 and IPv6 networks, differentiated services codepoint (DSCP) field in the packet header can be used to indicate desired service, as specified in [2] and [3]. DSCP redefines older IPv4 ToS octet and IPv6 traffic class octet. DSCP uses upper six bits of the former ToS octet, as shown in Fig. 3.

In MPLS networks, the desired service can be indicated in three experimental bits in an MPLS label stack entry. A sequence of these entries, comprising a label stack, can be stored after a link layer header and before a network layer header. A structure of one entry is illustrated in Fig. 4 [4]. Three experimental bits allow for up to eight different classes of service. MPLS also permits to infer a class of service directly from the label, to support more different classes of service.

When interfacing a pure IP network with a MPLS cloud, the DSCP field in an IP header can be copied into experimental bits and the label in a label stack entry or vice versa. Network devices within a MPLS cloud can use experimental bits and the label to differentiate between classes of service.

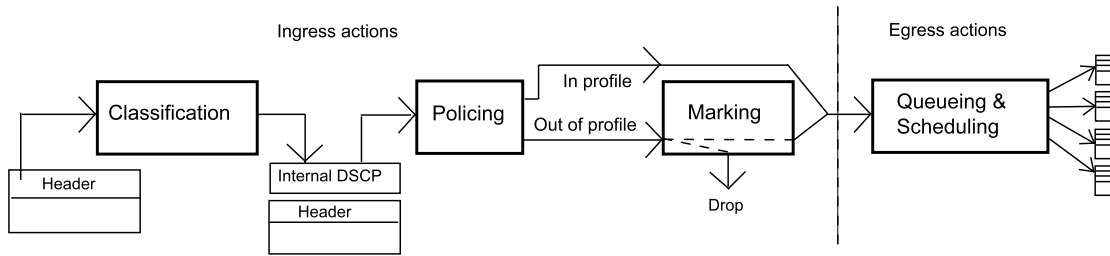


Figure 5: QoS model on Cisco Catalyst 3550

4 QoS model in Cisco Catalyst 3550 switch

When QoS support is configured on a Cisco Catalyst 3550 switch, packets are processed in several stages: classification, policing and marking are performed on input (ingress), queueing and scheduling are performed on output (egress) [5], see Fig. 5. Classification process is illustrated in more detail in Fig. 6. Double-line boxes denote various maps described in section 4.1. As a result of classification, each packet is assigned an internal DSCP, which determines further actions performed on the packet.

For both IP packets and non-IP packets the following classification actions can be performed:

- An access list can be configured and a packet matching this list is assigned a specified internal DSCP value
- User priority (called CoS in Cisco terminology) in a packet can be trusted and transformed into an internal DSCP value with the CoS to DSCP map
- All packets can be assigned a default internal DSCP value of 0

For IP packets the following additional classification actions are possible:

- DSCP value in a packet can be trusted and transformed into an internal DSCP value with the DSCP to DSCP mutation map
- IP precedence in a packet can be trusted and transformed into an internal DSCP value with the IP precedence to DSCP map

The function of policing and marking is illustrated in Fig. 7. An out-of-profile packet can be either dropped or its internal DSCP can be changed using the DSCP to policed DSCP map.

Finally, queueing and scheduling is performed at egress, as shown in Fig. 8. The internal DSCP value is converted to a CoS value using the DSCP to CoS map. This CoS value is then used to determine the number of queue to put the packet for scheduling according to the CoS to egress queue map. Four queues are available at each output port. The queues are serviced by a weighted round robin (WRR) scheduler according to the bandwidth share configured for each queue. Tail drop or weighted random early detection (WRED) is used for queue management. WRED thresholds can be configured using the DSCP to threshold map.

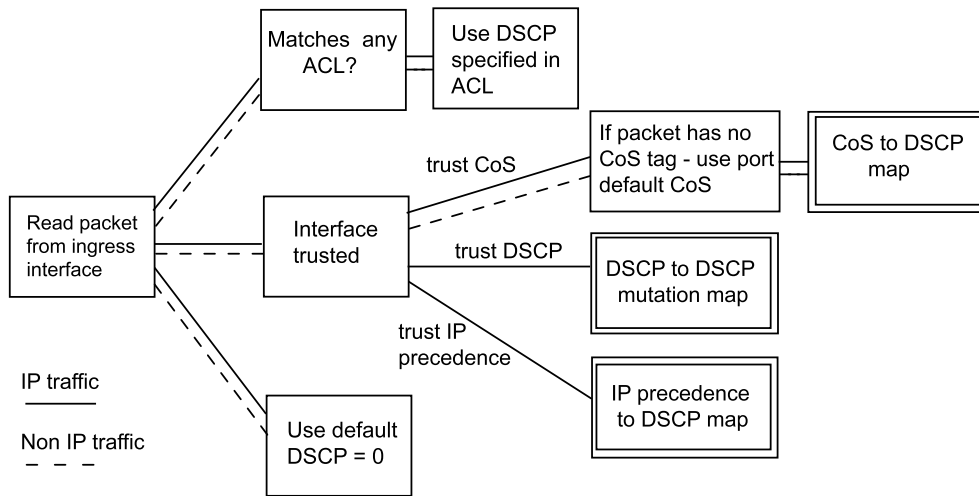


Figure 6: Classification

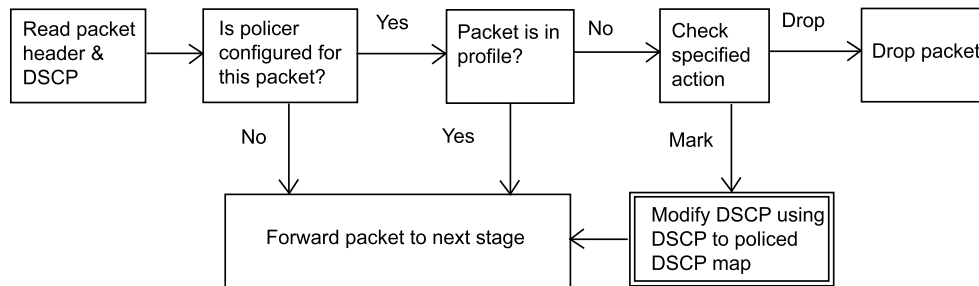


Figure 7: Policing and marking

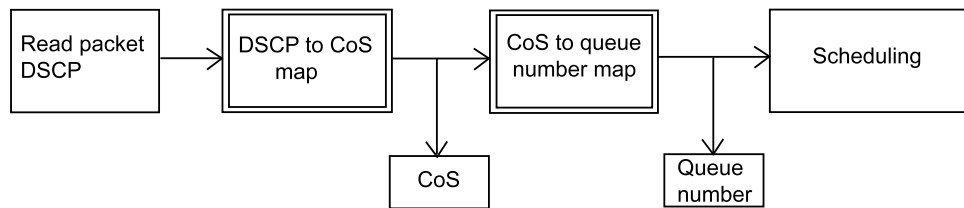


Figure 8: Queueing and scheduling

CoS value	0	1	2	3	4	5	6	7
DSCP value	0	8	16	24	32	40	48	56

Table 3: Default CoS to DSCP map

CoS value	0	1	2	3	4	5	6	7
DSCP value	10	15	20	25	30	35	40	45

Table 4: Modified CoS to DSCP map

4.1 Mapping tables

4.1.1 CoS to DSCP map

Default map content is shown in Table 3. To modify the map content, we specify new DSCP values for all CoS values, for instance:

```
mls qos map cos-dscp 10 15 20 25 30 35 40 45
```

The content of the modified map is shown in Table 4.

4.1.2 DSCP to DSCP mutation map

Default mapping is 1:1. To modify the map content, we specify new DSCP values for specified old DSCP values. Each port has its own DSCP to DSCP mutation map. For example, to trust all input DSCP values except values 1-7, which should be replaced by 0 and except values 8-13, which should be replaced by 10 on Gigabit Ethernet interface 0/1:

```
mls qos map dscp-mutation OurMap1 1 2 3 4 5 6 7 to 0
mls qos map dscp-mutation OurMap1 8 9 10 11 12 13 to 10
interface gigabitEthernet 0/1
  mls qos trust dscp
  mls qos dscp-mutation OurMap1
```

4.1.3 IP precedence to DSCP map

Default map content is the same as for the CoS to DSCP map. To modify the map content, we specify DSCP values for all IP precedence values, for example:

```
mls qos map ip-prec-dscp 10 15 20 25 30 35 40 45
```

4.1.4 DSCP to policed DSCP map

Default mapping is 1:1. To modify the map content, we specify new DSCP values for specified old DSCP values. For example, to replace values 50-57 by 0:

```
mls qos map policed-dscp 50 51 52 53 54 55 56 57 to 0
```

DSCP value	0-7	8-15	16-23	24-31	32-39	40-47	48-55	56-63
CoS value	0	1	2	3	4	5	6	7

Table 5: Default DSCP to CoS map

user priority (CoS)	queue
0, 1	0
2, 3	1
4, 5	2
6, 7	3

Table 6: Default CoS to egress queue map

4.1.5 DSCP to CoS map

Default map content is shown in Table 5. To modify the map content, we specify new CoS values for all DSCP values, for instance:

```
mls qos map dscp-cos 8 16 24 32 40 to 0
```

4.1.6 CoS to egress queue map

Default map content is shown in Table 6. Default behaviour is thus different from the recommendation in IEEE 802.1Q standard for the case of four egress queues, as shown in Table 2. To modify the map content to comply with the recommendation:

```
interface gigabitEthernet 0/1
  wrr-queue cos-map 1 1 2
  wrr-queue cos-map 2 0 3
  wrr-queue cos-map 3 4 5
  wrr-queue cos-map 4 6 7
```

4.1.7 Cisco Catalyst 2900 and 3500 XL (for comparison)

Each output port has only two queues - normal priority and high priority. Frames with user priority 0-3 are placed in the normal priority queue and frames with user priority 4-7 are placed in the high priority queue. Frames in the normal priority queue are sent only after frames in the high priority queue.

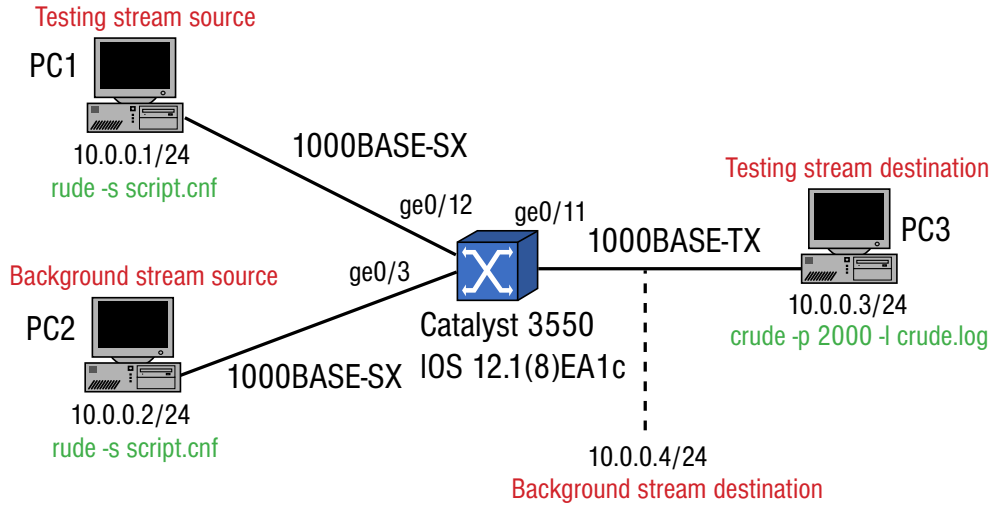


Figure 9: Experimental configuration

5 Experiments

Experimental configuration is shown in Fig. 9. PC1 sent testing streams to PC3. PC2 sent background streams to the virtual target (an address manually entered in the switch CAM table) on the same output port where PC3 was connected. Therefore, both testing streams and background streams shared capacity of the output port. All streams were generated and captured with RUDE/CRUDE [6] utilities and consisted of 1500-byte long packets. The resulting QoS characteristics were computed with the qosplot [7] utility.

Measurement 1

PC1 sent a testing stream of 300-1000 Mb/s in 50 Mb/s steps. PC2 sent a constant background stream of 500 Mb/s. Both streams shared the same egress queue. Throughput and packet loss rate of the testing stream is shown in Fig. 10. When the testing stream rate exceeded 500 Mb/s, some packets were lost. As the testing stream was more aggressive than the background stream, it got a larger share of the output link capacity.

Measurement 2

PC1 sent a testing stream of 300-1000 Mb/s in 50 Mb/s steps. PC2 sent a constant background stream of 700 Mb/s. Packets of each stream had a different DSCP value. We used the following configuration to put each stream in its own egress queue:

```
class-map Class1
  match access-group 1
policy-map Policy1
  class Class1
```

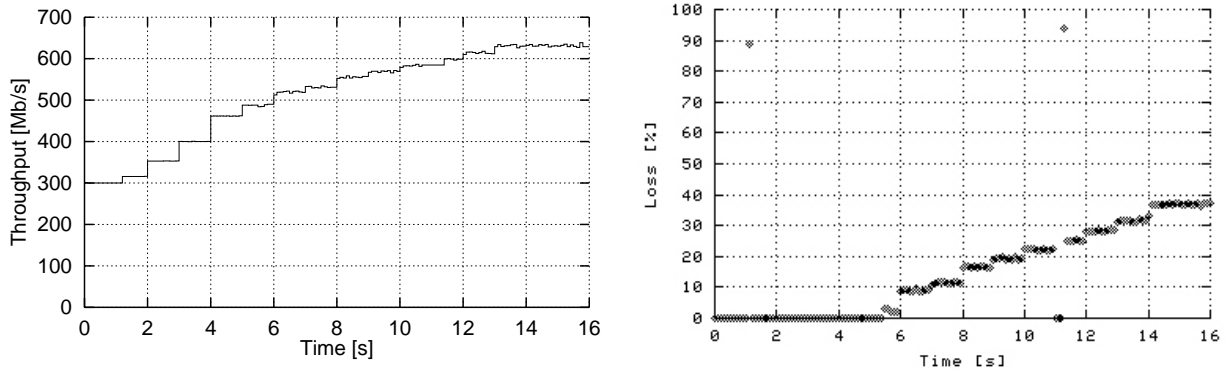


Figure 10: Measurement 1 - one shared queue

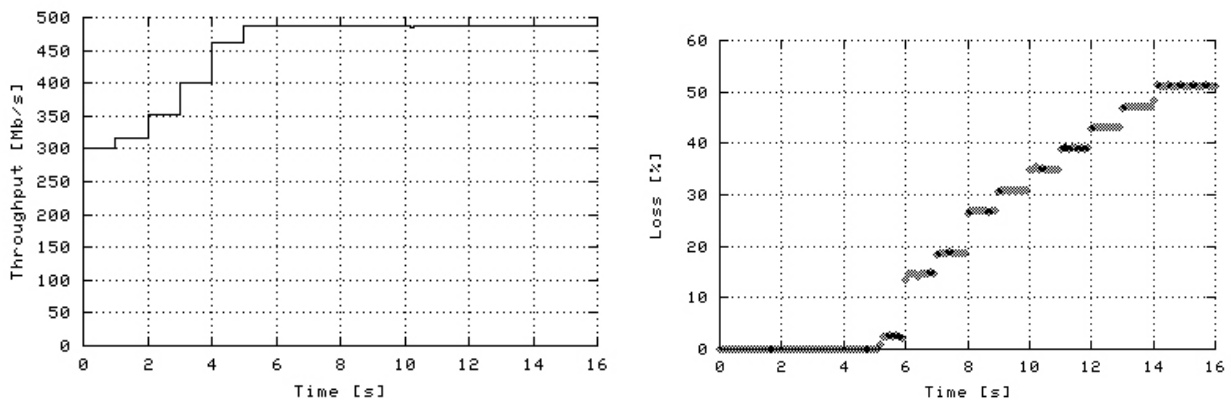


Figure 11: Measurement 2 - two queues with the equal share of the link capacity

```
trust dscp
interface gigabitEthernet 0/12
 service-policy Policy1
access-list 1 permit 10.0.0.0 0.0.0.255
```

Scheduling of egress queues was left at the default setting when each queue gets an equal share of the link capacity. Throughput and packet loss rate of the testing stream is shown in Fig. 11. The maximum throughput of the testing stream was 500 Mb/s corresponding to its share of the link capacity.

Measurement 3

The same as measurement 2, but we changed the share of the link capacity between queues to 70% and 30%. We added the following configuration commands to make this change:

```
interface gigabitEthernet 0/11
 wrr-queue bandwidth 30 70 0 0
```

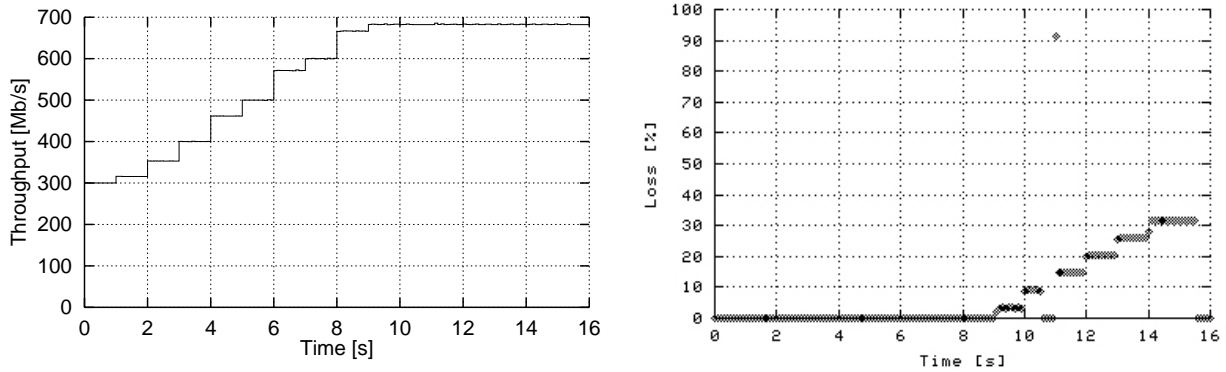


Figure 12: Measurement 3 - two queues with the share of the link capacity set to 70% and 30%

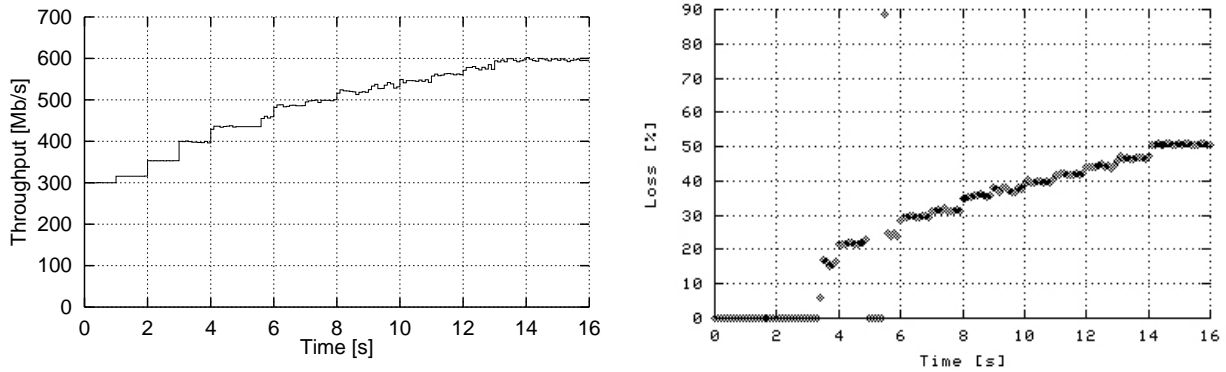


Figure 13: Measurement 4 - two queues with the equal share of the link capacity, one queue is for in-profile packets of the testing stream, the other queue is for out-of-profile packets of the testing stream and for the background stream

Throughput and packet loss rate of the testing stream is shown in Fig. 12.

Measurement 4

The same as measurement 2, but when the testing stream rate exceeds 400 Mb/s, the exceeding packets are “marked down”, that is their DSCP is changed so that they are put in the same queue with the background stream and thus compete for the share of the link capacity assigned to this queue. We added the following configuration commands to make this change:

```
policy-map Policy1
  class Class1
    police 400000000 512000000 exceed-action policed-dscp-transmit
```

Throughput and packet loss rate of the testing stream is shown in Fig. 13.

As expected, this configuration resulted in significant reordering (approx. 18%) of the testing stream packets.

6 Conclusion

Cisco Catalyst 3550 is a multilayer switch now commonly used in access networks. Individual ports can be configured as switched or routed, all commonly used routing protocols are available. However, we do not need any routing capability to implement QoS with Cisco Catalyst 3550 in an access network. All ports can remain configured as switched. The switch can convert layer 2 service codepoint (user priority, called CoS in Cisco terminology) to layer 3 service codepoint (DSCP) and vice versa, using the CoS to DSCP map and DSCP to CoS map. Queuing decision is based on output CoS, that is on the layer 2 service codepoint.

The mapping features and capacity sharing worked for us as requested with small difficulties that should be considered when configuring the switch. First, the `access-list 1 permit any` command failed to match all traffic. We had to specify the range of IP addresses as in the `access-list 1 permit 10.0.0.0 0.0.0.255` command. Second, it was not possible to set the port to trusted or override state and attach an input policy to this port at the same time. When a policy was attached, the trusted or override port state was switched off. And when this port state was set again, the policy was detached from the port.

References

- [1] “IEEE Standards for Local and Metropolitan Area Networks”, *Virtual Bridged Local Area Networks*, IEEE Standard 802.1Q-1998, <http://standards.ieee.org/getieee802>.
- [2] K. Nichols, S. Blake, F. Baker, D. Black. “Definition of the differentiated services field”, Request for Comments 2474, Internet Engineering Task Force, December 1998.
- [3] D. Grossman. “New terminology and clarifications for diffserv”, Request for Comments 3260, Internet Engineering Task Force, April 2002.
- [4] Bruce Davie, Yakov Rekhter. “MPLS - Technology and Applications”, Morgan Kaufmann Publishers, 2000.
- [5] “Catalyst 3550 Multilayer Switch Software Configuration Guide 12.1(4) - Configuring QoS”, http://www.cisco.com/en/US/products/hw/switches/ps646/products_configuration_guide_cha
- [6] RUDE/CRUDE UDP data emitter and collector, <http://rude.sourceforge.net>.
- [7] qosplot - Utility for computing and plotting network QoS characteristics, <http://www.ces.net/project/qosip>.