Application Note

QoS Priority and Rate Limit Support for the KSZ8873/8863 Family

Introduction

Latency critical applications such as Voice over IP (VoIP) and video typically need to guarantee a high quality of service (QoS) throughout the network. QoS can be supported by various priority schemes offered by the switches and routers.

This application note describes the different priority schemes supported by the KSZ8873/8863 family of switches and how they are configured.

There are three different priority schemes supported by the devices: Port-based priority, 802.1p Tag-based Priority and DiffServ-based priority. The mechanisms for each of these priority schemes differ only with respect to the ingress port. The egress port buffer scheme is common to all three priority schemes.

Common Settings for Port Based, 802.1p Tag Based and DiffServ Based Priorities

The KSZ8873/8863 family of devices offers four priority transmits queues per port. Queue 3 is the highest priority queue as priority 3 and Queue 0 is the lowest priority queue as priority 0 in 4 queue mode. The bits 0 of the port registers control 0 are used to enable splitting transmit 4 queues for egress port 1, port 2 and host port respectively. The bits 7 of the port registers control 2 are used to enable splitting transmit 2 queues for egress port 1, port 2 and host port respectively (4Q and 2Q can not be set at same time).

Priority Scheme

Priority transmit queuing allows a switch to define two priority schemes. One is “always transmit higher priority packets first” mode. The other is the weighted fair queuing (WFQ) mode. When the transmit queue is set to WFQ mode, the transmit queue will follow a scale for the four queues and the bandwidth allocation is Q3:Q2:Q1:Q0=8:4:2:1. If any queue is empty, the highest non-empty queue will get one more weighting. For example, if Q2 is empty, Q3:Q2:Q1:Q0 will become (8+1):0:2:1.

This mechanism assures that during congestion, the higher-priority data does not get delayed by lower-priority traffic. Some examples of priority level are:

- Important Voice and video packets are assigned a highest-priority level 3.
- General Voice and video packets are assigned a higher-priority level 2.
- Web traffic is assigned a lower-priority level 1.
- Back-up data traffic is assigned the lowest-priority level 0.

The Weighted Fair Queuing (WFQ) mode tries to ensure that the lower-priority packets will not be starved during congestion. WFQ is implemented in the KSZ8873/8863 three ports switches; with a host bus as the host port, by controlling the priority scheme select bit 3 in the global register 5. These related registers as shown in table 1.

All datasheets and support documentation can be found on Micrel’s web site at: www.micrel.com.

Register	Bit	Name	Description	Default
Global register 5 bit 3=0 and the port registers 175 to 186 bit7=0	3/7	Priority Scheme select	Always high priority packets first.	1
Global register 5 bit 3=1 or 0 and the port registers 175 to 186 bit7=1	3/7	Priority Scheme Select	. 4Q: Weighted Fair Queuing (WFQ) is enabled, Q3,Q2,Q1,Q0 = 8:4:2:1. 2Q: Weighted Fair Queuing (WFQ) is enabled, Q1,Q0 = 2:1.	0

Table 1. Registers are used for the priority Schemes

Notes:

1. If the *TX Multiple Queues Select Enable* bit is not enabled in port register control 0 bit0 and control 3 bit7, then only a single output queue will be present at the egress port. Hence, the priority scheme selection will have no effect, irrespective of the other settings for the ingress and egress ports.
2. The settings highlighted in “Note 1” above will be used for all Port based priorities, 802.1p based priorities and DiffServ based priorities.

Port Based Priority

Port based priority is the simplest form of QoS. Each ingress port can be individually classified as one of the priorities 0-3. All packets arriving at the ingress port will be passed to any of the four priority queues at the egress port, depending upon the configuration of the ingress port.

Each ingress port can be configured as one of the priorities 0-3 by using the Port Based Priority Classification Enable bit shown in Table 2.

For example, if the port register control 0 bit 4-3 is set to 10, all of packets from ingress port 2 will be treated as priority 2 level packets and go to priority 2 transmit queue on the egress port which has set into four priority queues.

Register	Bit	Name	Description	Default
Port registers control 0	4-3	Port based priority classification	00 = ingress packets on port n will be classified as priority 0 queue. 01 = ingress packets on port n will be classified as priority 1 queue. 10 = ingress packets on port n will be classified as priority 2 queue. 11 = ingress packets on port n will be classified as priority 3 queue.	00

Table 2. Registers are used for Port Based Priority

Note: "Diffserv", "802.1p" and port priority can be enabled at the same time. The OR'ed result of 802.1p and DSCP overwrites the port based priority.

802.1p Tag Based Priority

802.1p priority can be enabled by the 802.1p Priority Classification Enable bit in the Port Registers Control 0 bit 5.

Ethernet packets can have an optional 4-byte 802.1q VLAN tag inserted between the source address (SA) and the length/type fields. As shown in Figure 1, there is a 3-bit priority field embedded in the 4-byte tag, the 3-bit priority field is used for 802.1p priority classification.

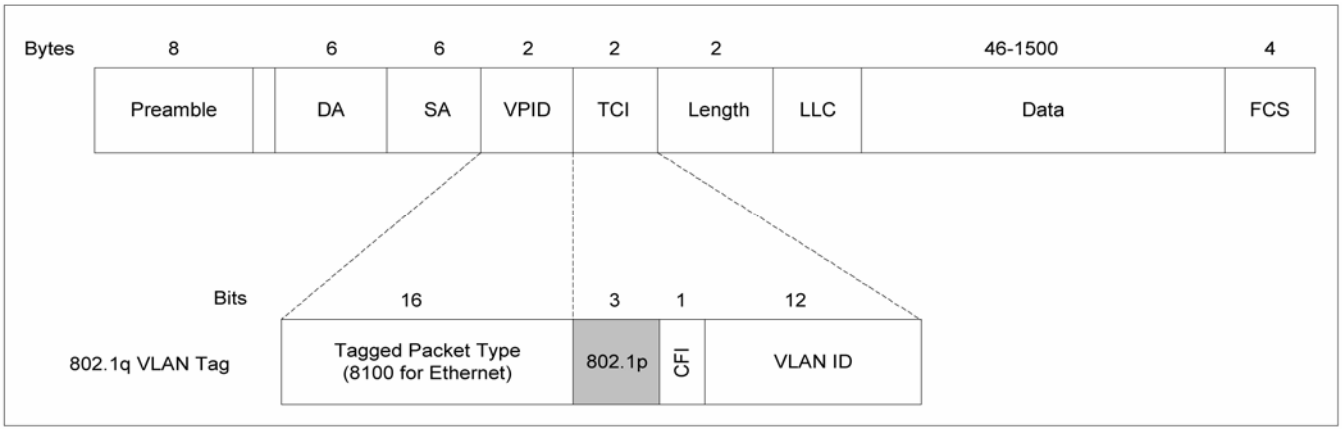


Figure 1. Ethernet Packet with 802.1q VLAN Tag

The 3-bit priority field in the VLAN tag is used to set the priority level (0-3) for each packet. The number value from 0 to 7 is configured in the switch and compared to the binary equivalent of the incoming packet's priority field in the VLAN tag. The 3-bit priority field value (0-7) can be decoded as priority level (0-3) by the global register 12 and register 13, they can be programmed by the user (See Table 3 for details). The priority field value of the incoming tagged packets will be re-classified to four priority levels based on the global register 12 and register 13 setting. The related registers are as shown in Table 3 for 802.1p based priority.

Register	Bit	Name	Description	Default
Port Register Control 0 bits 5	5	Priority Classification Enable	1 = Enable 802.1p priority classification for ingress packets on port. 0 = Disable 802.1p priority.	0
Global Register 13	7-6	Tag_0x7	IEEE 802.1p mapping. The value is used as the frame's priority when its IEEE Tag has a value of 0x7.	0X3 for priority 3
	5-4	Tag_0x6	IEEE 802.1p mapping. The value is used as the frame's priority when its IEEE Tag has a value of 0x6.	0X3 for priority 3
	3-2	Tag_0x5	IEEE 802.1p mapping. The value is used as the frame's priority when its IEEE Tag has a value of 0x5.	0X2 for priority 2
	1-0	Tag_0x4	IEEE 802.1p mapping. The value is used as the frame's priority when its IEEE Tag has a value of 0x4.	0X2 for priority 2
Global Register 12	7-6	Tag_0x3	IEEE 802.1p mapping. The value is used as the frame's priority when its IEEE Tag has a value of 0x3.	0X1 for priority 1
	5-4	Tag_0x2	IEEE 802.1p mapping. The value is used as the frame's priority when its IEEE Tag has a value of 0x2.	0X1 for priority 1
	3-2	Tag_0x1	IEEE 802.1p mapping. The value is used as the frame's priority when its IEEE Tag has a value of 0x1.	0X0 for priority 0
	1-0	Tag_0x0	IEEE 802.1p mapping. The value is used as the frame's priority when its IEEE Tag has a value of 0x0.	0X0 for priority 0
Port Register Control 1 bit 3	3	User Priority Field (Ceiling)	1 = If the packet's "priority field" is greater than the "user priority bits" in port n's VID Control register bits [15:13], replace the packet's " priority field" with the "user priority bits" in port n's VID Control register bits [15:13]. 0 = Do not compare and replace the packet's "user priority field."	0
Port Register Control 3 bits [7:5]	7-5	User Priority bits	Port n tag [7-5] is for priority field of ingress tagged packet to be compared or replaced.	000

Table 3. Registers are used for Tag Based Priority

Note: The OR'ed result of 802.1p and DSCP (see below) priority classification overrides any port priority.

Priority Re-Mapping

The KSZ8873/8863 family of devices has the ability to re-map the ingress packets 802.1p priority field by setting bit 3 of User Priority Field in the port registers control 1 and bits [7-5] of User Priority Bits in the port registers control 3. An example of the importance of priority re-mapping is shown as follows.

In the case that port 1 is connected to a PC and port 2 is connected to a VoIP router, a problem may occur if you have a 'corrupt' PC transmitting data packets containing high priority 802.1p tags. This causes the VoIP and PC (data) packets to both be tagged as high priority and hence, there is no differentiation between them. Acceptable QoS for the voice traffic can no longer be guaranteed.

Priority re-mapping is available on the KSZ8873/8863 family of ports. Each ingress port can be set as specified in the User Priority Bits in the port registers control 3. If the incoming packet's 802.1p priority field is greater than the user defined value in the User Priority Bits in the port register control 3, then the packet's priority field is replaced with the user defined value in the User Priority Bits. Priority re-mapping is enabled using the *User Priority Field (Ceiling)* bit in the port register control 1 bit 3 as shown in Table 3.

DSCP (DiffServ-based) Priority

The KSZ8873/8863 devices support DiffServ-based priority in IPv4 and IPv6 IP packets. In this note, IPv4 packets are used as an example.

The differentiated service code point (DSCP) priority operates in the Layer 3, IP protocol. The IP datagram header is embedded within the Ethernet data field (see Figure 2).

The DSCP priority bits are located inside the type of service (TOS) field, within the standard IPv4 header.

The IPv4 header is shown below in more detail. The TOS byte is the second byte located after the header length field (HLEN).

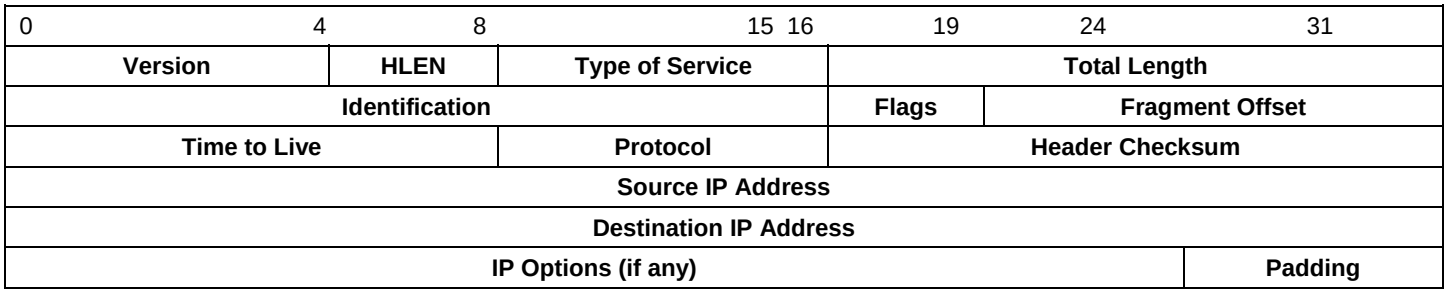
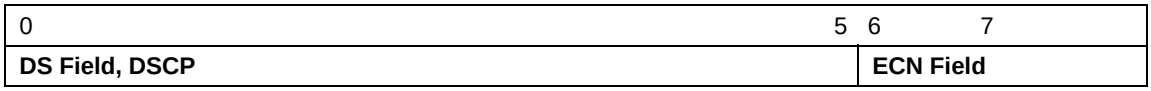


Figure 2. Format of IPv4 Datagram Header

Bits 0 to 5 of the ToS field are then taken and fully decoded in 64 separate QoS service codes as shown in Figure 3.



DSCP: Differentiated Services Code Point

ECN: Explicit Congestion Notification (Unused)

Figure 3. Differential Services (DS) Code Point within the ToS Field of an IP Datagram

The Differentiated Service Code is then compared against the corresponding bit in the *Priority Control Registers 0 to 15 (REG96-REG111)* of the KSZ8873/8863 devices (8 bits x 16 registers = 128 bits totally). The corresponding 2-bits in REG96-REG111registers stands for one code point of DSCP. 2-Bits have 4 priority levels, where 00 is priority 0, 01 is priority 1, 10 is priority 2 and 11 is priority 3. Using the 128-bits with 2-bit as a DSCP in the TOS Priority Control registers, it is possible to make 64 DSCP code points with 4 priority levels for each DSCP.

DSCP priority is enabled using the *DiffServ Priority Classification Enable* bit 6 in each Port Register Control 0. They are shown in Table 4.

Register	Bit	Name	Description	Default
Port register control 0 bit 6	6	Diffserv Priority Classification Enable	1 = Enable diffserv priority classification for ingress packets on port. 0 = Disable diffserv priority.	0

Table 4. Registers are used for DiffServ Priority

Note: The OR'ed result of 802.1p and DSCP priority classification overrides any port priority.

TOS Register	DSCP # (Dec)	Bits	Priority Level	Default
Register 96	0	[1-0] = 00, 01, 10, 11	0-3	00
	1	[3-2] = 00, 01, 10, 11	0-3	00
	2	[5-4] = 00, 01, 10, 11	0-3	00
	3	[7-6] = 00, 01, 10, 11	0-3	00
Register 97	4	[1-0] = 00, 01, 10, 11	0-3	00
	5	[3-2] = 00, 01, 10, 11	0-3	00
	6	[5-4] = 00, 01, 10, 11	0-3	00
	7	[7-6] = 00, 01, 10, 11	0-3	00
Register 98	8	[1-0] = 00, 01, 10, 11	0-3	00
	9	[3-2] = 00, 01, 10, 11	0-3	00
	11	[5-4] = 00, 01, 10, 11	0-3	00
	12	[7-6] = 00, 01, 10, 11	0-3	00
Register 99	13	[1-0] = 00, 01, 10, 11	0-3	00
	14	[3-2] = 00, 01, 10, 11	0-3	00
	15	[5-4] = 00, 01, 10, 11	0-3	00
	16	[7-6] = 00, 01, 10, 11	0-3	00
Register 100	17	[1-0] = 00, 01, 10, 11	0-3	00
	18	[3-2] = 00, 01, 10, 11	0-3	00
	19	[5-4] = 00, 01, 10, 11	0-3	00
	20	[7-6] = 00, 01, 10, 11	0-3	00
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Register 110	56	[1-0] = 00, 01, 10, 11	0-3	00
	57	[3-2] = 00, 01, 10, 11	0-3	00
	58	[5-4] = 00, 01, 10, 11	0-3	00
	59	[7-6] = 00, 01, 10, 11	0-3	00
Register 111	60	[1-0] = 00, 01, 10, 11	0-3	00
	61	[3-2] = 00, 01, 10, 11	0-3	00
	62	[5-4] = 00, 01, 10, 11	0-3	00
	63	[7-6] = 00, 01, 10, 11	0-3	00

Table 5. Registers are used for 64 DSCP Priority Level Settings

All Priority Control Registers (REG96 – REG111) are as shown in Table 5 for detail priority levels.

For example,

If DSCP=001000 (Bin) = 8 (Dec) in IP TOS field, this implies that TOS Priority Control Register 98, bits 1-0 will be examined, and the priority of those bits will set as the priority level of the packets.

Ingress Rate limit for Priority

There are 0-3 four priority levels for Port based, 802.1p based and DiffServ based priorities for the ingress packets. The register 22/38/54 bits [6-0] are for priority 0 ingress packets rate limit. The register 23/39/55 bits [6-0] are for priority 1 ingress packets rate limit. The register 24/40/56 bits [6-0] are for priority 2 ingress packets rate limit. The register 25/41/57 bits [6-0] are for priority 3 ingress packets rate limit. The rate step is 64Kbps when the rate limit is less than 1Mbps, the rate step is 1Mbps when the rate limit is from 1Mbps to 10Mbps (10BT) and 1Mbps to 100Mbps (100BT). Please refer to Data Rate Limit Table 6 below.

Data Rate Limit for ingress or egress	**100BT Register bit[6:0]**	**10BT Register bit[6:0]**
	0x01 to 0x63 for the Rate 1Mbps to 99Mbps.	0x01 to 0x09 for the rate 1Mbps to 9Mbps
	0 or 0x64 for the rate 100Mbps	0 or 0x0A for the rate 10Mbps
The rate steps below are used less than 1Mbps	Hex value for register bit[6:0]	
64 Kbps	0x65	
128 Kbps	0x66	
192 Kbps	0x67	
256 Kbps	0x68	
320 Kbps	0x69	
384 Kbps	0x6A	
448 Kbps	0x6B	
512 Kbps	0x6C	
576 Kbps	0x6D	
640 Kbps	0x6E	
704 Kbps	0x6F	
768 Kbps	0x70	
832 Kbps	0x71	
896 Kbps	0x72	
960 Kbps	0x73	

Table 6. Data Rate Limit Table

For ingress rate limiting, KSZ8873/8863 provides options to selectively choose frames from all types, multicast, broadcast, and flooded unicast frames. The KSZ8873/8863 counts the data rate from those selected type of frames. Packets are dropped at the ingress port when the data rate exceeds the specified rate limit.

Register	**Bit**	**Name**	**Description**	**Default**
Port register control 5 bit 3-2	3-2	Limit mode	Ingress Limit Mode These bits determine what kinds of frames are limited and counted against ingress rate limiting. = 00, limit and count all frames = 01, limit and count Broadcast, Multicast, and flooded unicast frames = 10, limit and count Broadcast and Multicast frames only = 11, limit and count Broadcast frames only	00

Table 7. Limit Mode Select

Egress Rate limit for Priority Queues

There are 0-3 four priority levels for Port based, 802.1p based and DiffServ based priorities for 4 queues of egress packets. On the transmit side, the data transmit rate for each priority queue at each port can be limited by setting up Egress Rate Control Registers. The register 154/158/162 bits [6-0] are for priority 0 egress packets rate limit at transmit queue Q0. The register 155/159/163 bits [6-0] are for priority 1 egress packets rate limit at transmit queue Q1. The register 156/160/164 bits [6-0] are for priority 2 egress packets rate limit at transmit queue Q2. The register 157/161/165 bits [6-0] are for priority 3 egress packets rate limit at transmit queue Q3. The rate step is same as Table 6 and description above. The port register 154/158/162 bit 7 should be set to take effect for the egress packets rate limit.

Register	Bit	Name	Description	Default
Port register 154/158/162 bit 7	7	Egress Rate Limit Flow Control Enable	1 = Enable egress rate limit flow control 0 = Disable	0

Table 8. Enable Egress Rate Limit Flow Control

Conclusion

The KSZ8873/8863 family of switches is ideal for handling Quality of Service requirements. With emerging applications such VoIP and Video Broadcasting, the network must be ready to handle different type of services in a cost effective manner. The KSZ8873/8863 also provides a fine resolution rate limiting at different priority levels for ingress and egress packets. The network should be designed with an end-to-end QoS capability for the future. As shown in this paper, Micrel's Ethernet product family of switches provides a rich set of QoS functionality to meet the needs for emerging triple play applications.

MICREL, INC. 2180 FORTUNE DRIVE SAN JOSE, CA 95131 USA
TEL +1 (408) 944-0800 FAX +1 (408) 474-1000 WEB http://www.micrel.com