BCM交换芯片原理概要

——by M.C

本文使用的寄存器图片非摘录自56620的寄存器说明文档，可能存在部分文字与图片描述不一致的地方。其匹配寄存器文档为“56620-PR101-R”(网络上没找到)

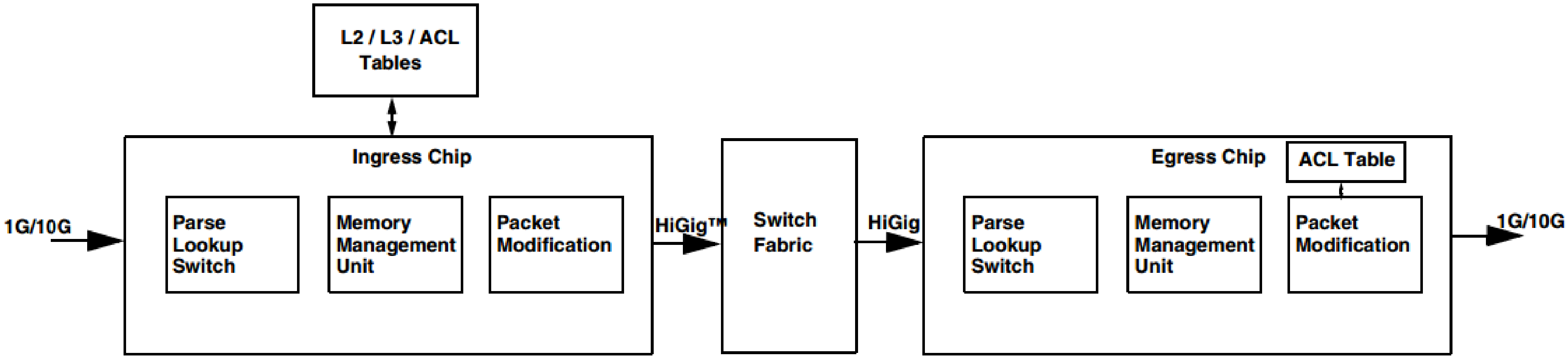
Outline

- Architectural
- ACL
- Buffer Management
- L2
- VLAN
- Link Aggregation
- Mirroring
- L3

Architectural

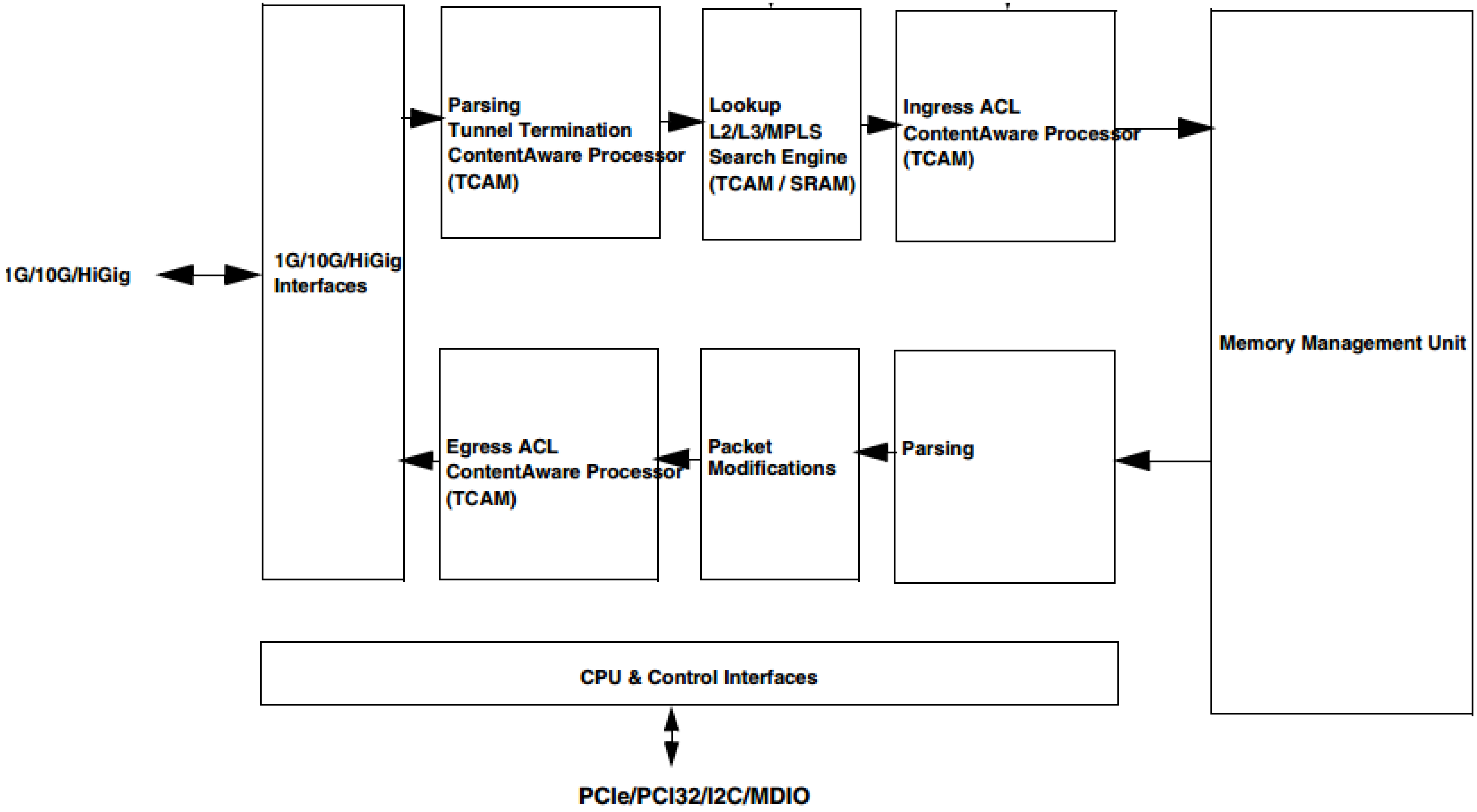
系统逻辑视图

- Ingress Chip
  - 解析报文128字节的头部(MMU Cell的最小单位)
  - 隧道终结
  - 报文头分类，以决定VRF
  - 通过VRF与报文头的信息，进行L2/L3/MPLS查找
  - 入口ACL处理；基于ACL进行计数与统计
  - 报文缓存、准入控制与调度
  - 修改报文 (如基于报文类型进行修改)
- Switch Fabric
  - 基于HiGig头部信息进行报文的交换选路
  - 多播处理
  - 支持基于服务的流量控制
- Egress Chip
  - 解析HiGig报文头
  - 根据HiGig头部信息决定出端口
  - 报文缓存、准入控制与调度
  - 修改报文
  - 出口ACL处理



芯片逻辑视图

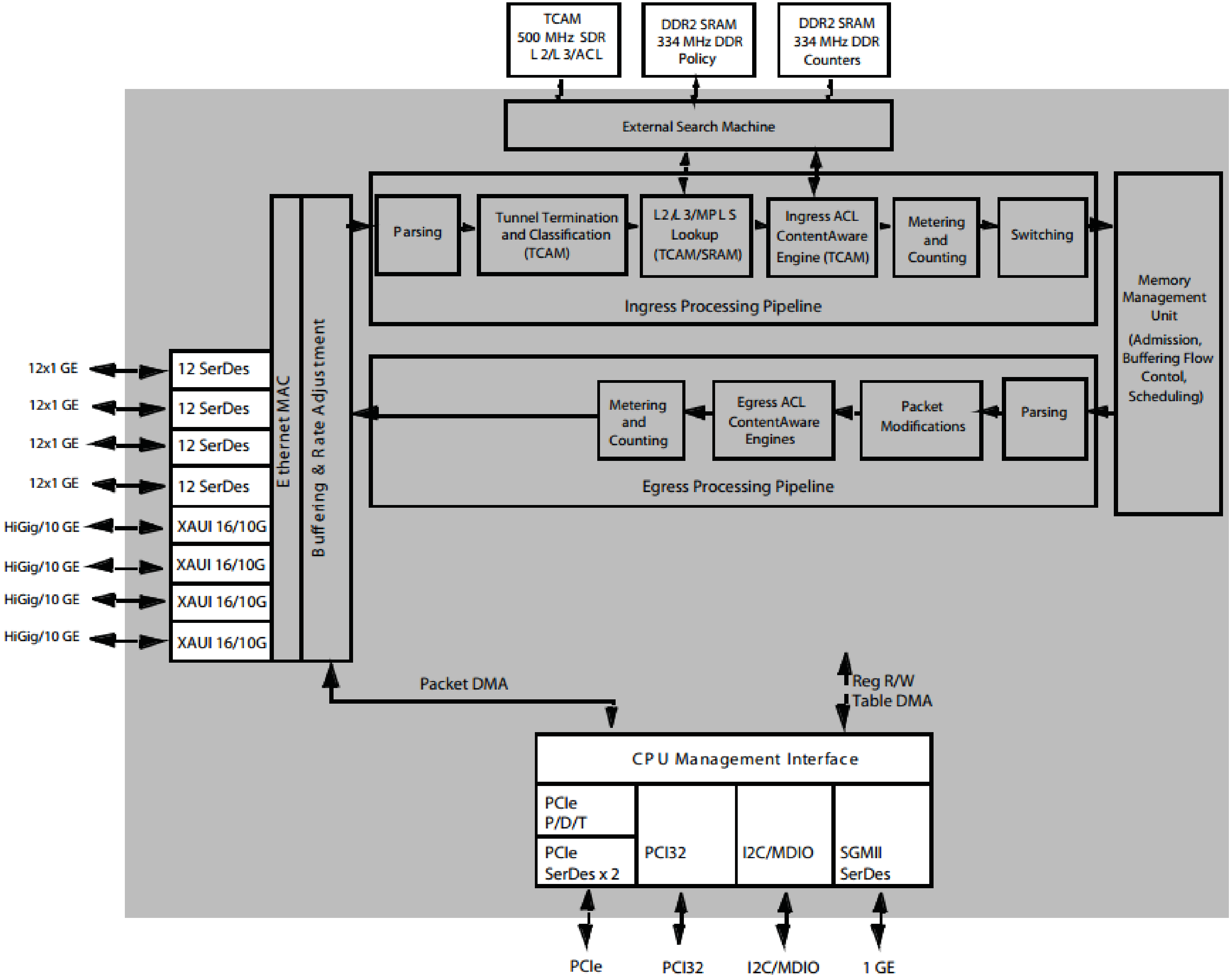
- 入端处理通道负责报文解析，隧道终结与VRF配置，L2/L3/MPLS使用隧道终结阶段获得的VRF信息进行查找与ACL处理。



TCAM：ternary content addressable memory。从CAM的基础上发展而来的。一般的CAM存储器中每个bit位的状态只有两个，“0”或“1”，而TCAM中每个bit位有三种状态，除掉“0”和“1”外，还有一个“don’t care”状态，所以称为“三态”，它是通过掩码来实现的，正是TCAM的这个第三种状态特征使其既能进行精确匹配查找，又能进行模糊匹配查找，而CAM没有第三种状态，所以只能进行精确匹配查找。TCAM 表内所有条目都可以并行访问，比如，如果你有100条ACL，TCAM能一次就能对比这100条ACL进行对比操作，过去如果有100条ACL的话，需要第一条ACL对比完后再对比第二条，然后第三条，直至N条，效率很明显没有TCAM高。

TCAM的组件FM（**特性管理器**）软件将匹配语句编译（合并）为TCAM表项，这样就可以**以帧转发速度查询TCAM**

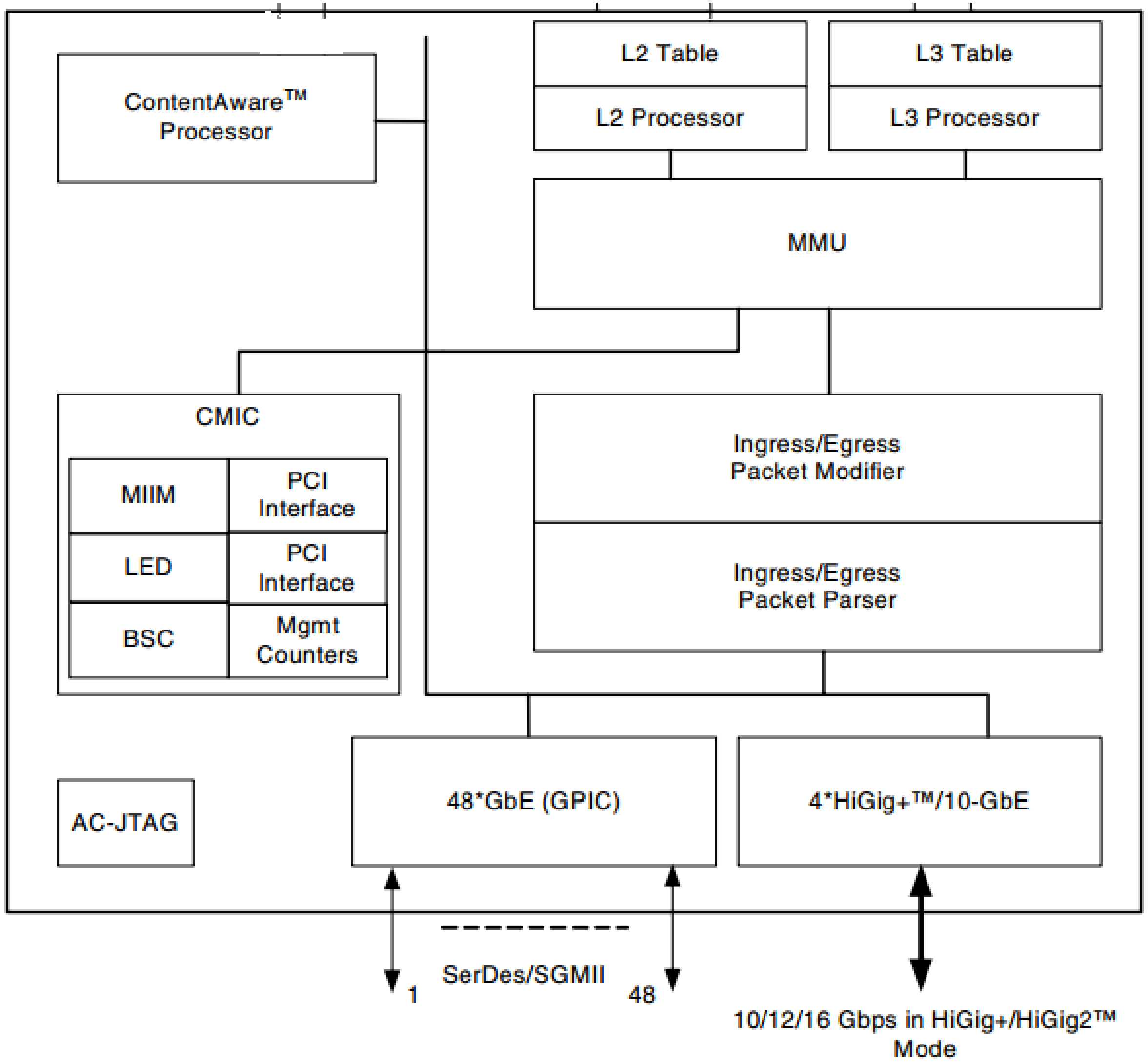
芯片框图



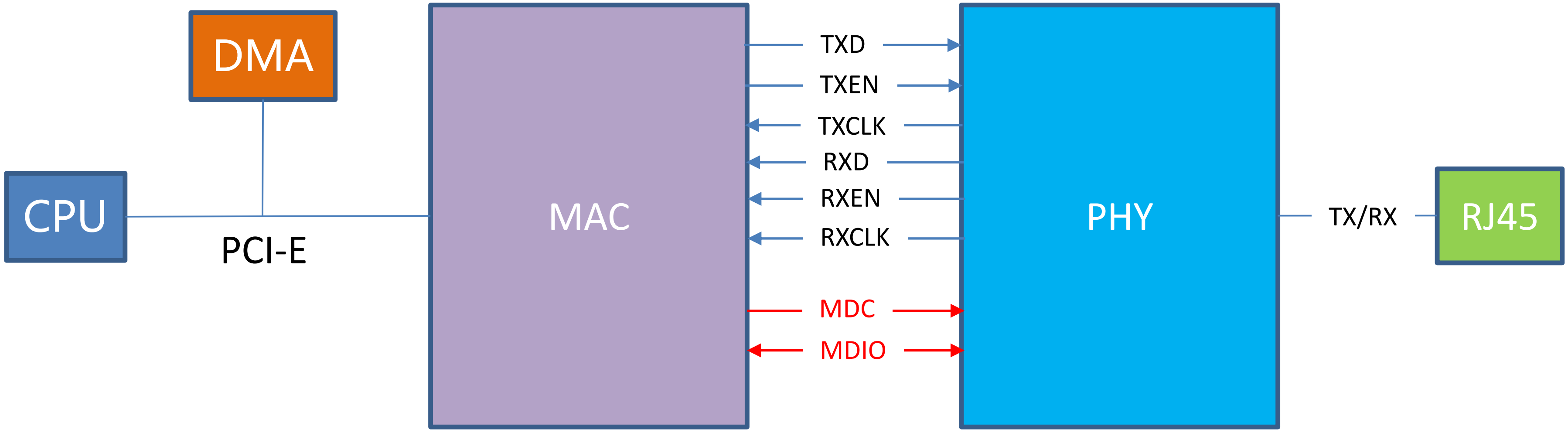
TCAM
500 MHz SDR
L2/L3/ACL
DDR2 SRAM
334 MHz DDR
Policy
DDR2 SRAM
334 MHz DDR
Counters
External Search Machine
Parsing
Tunnel Termination and Classification (TCAM)
L2/L3/MPLS Lookup (TCAM/SRAM)
Ingress ACL ContentAware Engine (TCAM)
Metering and Counting
Switching
Ingress Processing Pipeline
Memory Management Unit
(Admission, Buffering Flow Contol, Scheduling)
Metering and Counting
Egress ACL ContentAware Engines
Packet Modifications
Parsing
Egress Processing Pipeline
Ethernet MAC
Buffering & Rate Adjustment
12x1 GE
12x1 GE
12x1 GE
12x1 GE
HiGig/10 GE
HiGig/10 GE
HiGig/10 GE
HiGig/10 GE
12 SerDes
12 SerDes
12 SerDes
12 SerDes
XAUI 16/10G
XAUI 16/10G
XAUI 16/10G
XAUI 16/10G
Packet DMA
Reg R/W
Table DMA
CPU Management Interface
PCIe P/D/T
PCIe SerDes x 2
PCI32
I2C/MDIO
SGMII SerDes
PCIe
PCI32
I2C/MDIO
1 GE

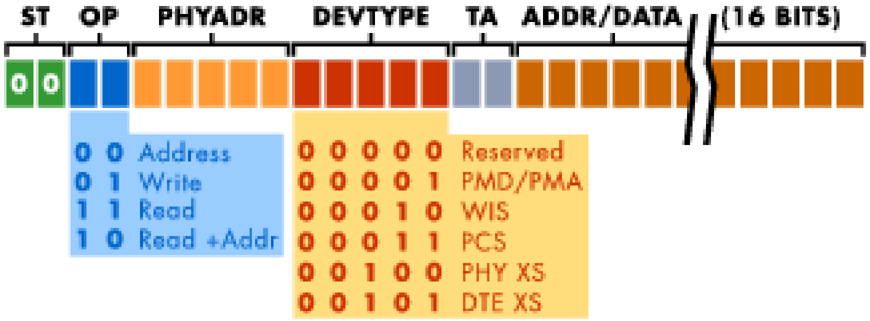
芯片框图

- GPIC：端口的配置信息驻留在GPIC(Gigabit Ethernet Port Interface Controller)里。GPIC可以配置为SGMII模式或者SerDes模式。SGMII模式可以直接连外部PHY设备；SerDes模式可以直接连光模块。
- HiGig：HiGig模式端口用于多个芯片的互连来增加整个系统的端口密度。
- CMIC：通过PCI外接CPU，实现对芯片寄存器的读写设置操作。CPU口发包的处理逻辑与端口接收报文一样(一般CPU是0口，就像0口接收到了报文一样)，处理过程也会出现报文重新被送CPU的情况。



MIIM



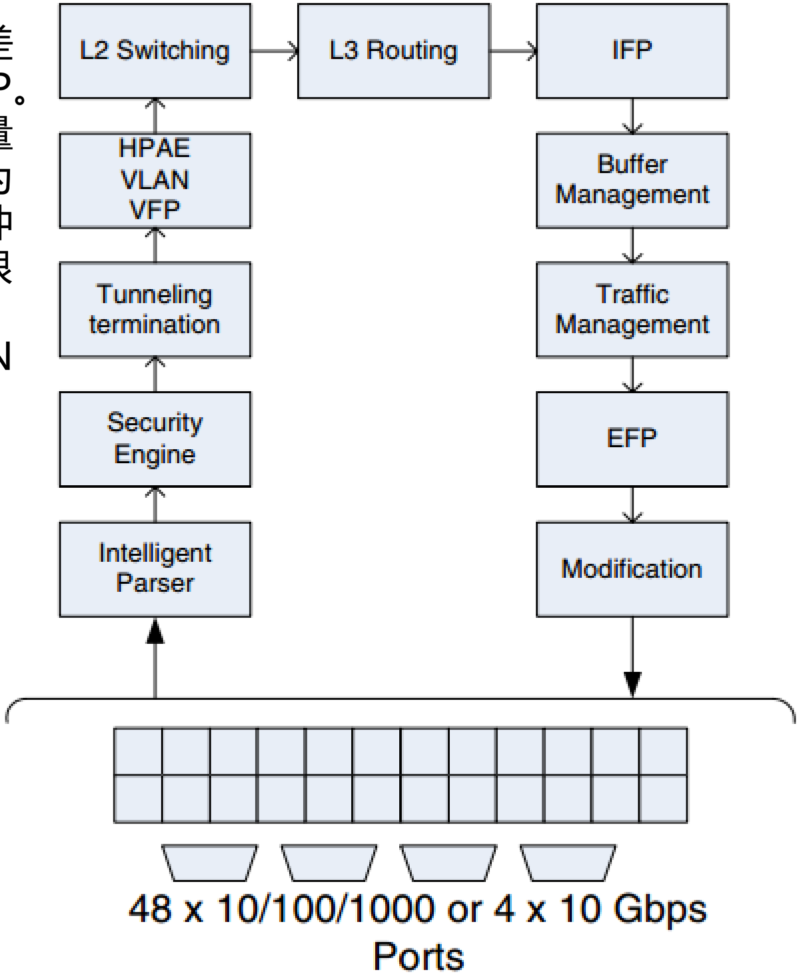


0 0	Address
0 1	Write
1 1	Read
1 0	Read +Addr

0 0 0 0 0	Reserved
0 0 0 0 1	PMD/PMA
0 0 0 1 0	WIS
0 0 0 1 1	PCS
0 0 1 0 0	PHY XS
0 0 1 0 1	DTE XS

交换架构

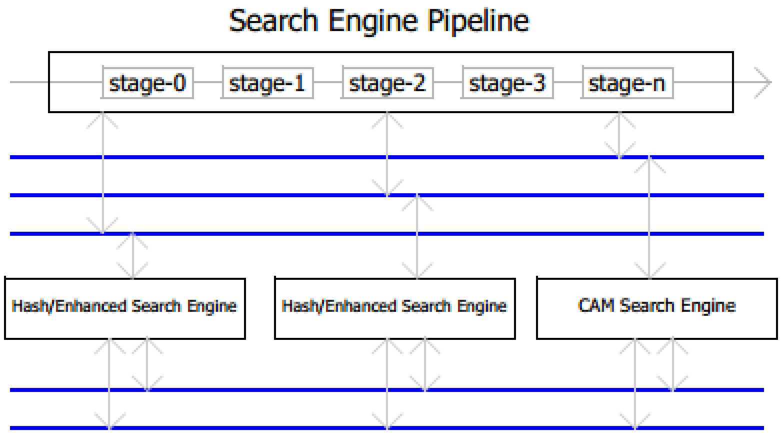
- 采用模块化、高性能的管道式报文交换处理架构。在管道上的每个模块都有各自的处理功能，并把处理的结果提供给下一模块进行处理。
  - Intelligent Parser：包括两个独立的解析器，全解析器和HiGig解析器。全解析器负责解析来自端口与CMIC的报文(面板口与CPU口)，需要的信息都可以在头128字节里获得，全解析器必须保存所有的解析信息，以备各种搜索引擎使用。HiGig解析器负责解析来自HiGig口的报文。
  - Security Engine：早期的硬件安全检测机制，防止DoS攻击。
  - L2 Switching：分配VLAN、优先级，源MAC学习，目的MAC查找转发。
  - L3 Routing：源/目的 IP查找
  - ContenAware Processing：CAP用来提供ACL、差分服务、QoS等类型的应用。图中的IFP、EFP即CAP。
  - Buffer Management：控制端口的传输行为与流量整形。每个出口关联有8个CoS队列，调度器可配置为不同的模式对这8个队列进行不同的仲裁。调试器仲裁的一个主要方面是提供最小带宽保证与最大带宽限速，通过监视每个CoS队列的计数机制实现。
  - Modification：根据搜索引擎的的结果，进行VLAN转换、隧道封装与L3路由变更。



- CAP在过去称为过滤引擎(filter processor)，所以IFP(ingress filter processor)、EFP(egress filter processor)沿用了这样的命名
- Intelligent Parser可以从逻辑上看成是硬件内部有一个微程序(硬件电路实现)负责解析报文，并把报文的各个字段(如MAC，IP等)保存到结构体的各个成员变量里，以便高效地提供给后续的搜索引擎使用。

CAP

- 芯片内部主要的搜索引擎有两种，HASH搜索引擎与CAM搜索引擎。
  - HASH搜索引擎：L2表、L3表、VLAN表、IPMC表等等
  - CAM搜索引擎：ACL表
- 每个搜索引擎都带有一定的存储空间，用来存储相应的内容表项。搜索引擎采用管道式的架构，进行每个引擎的处理。



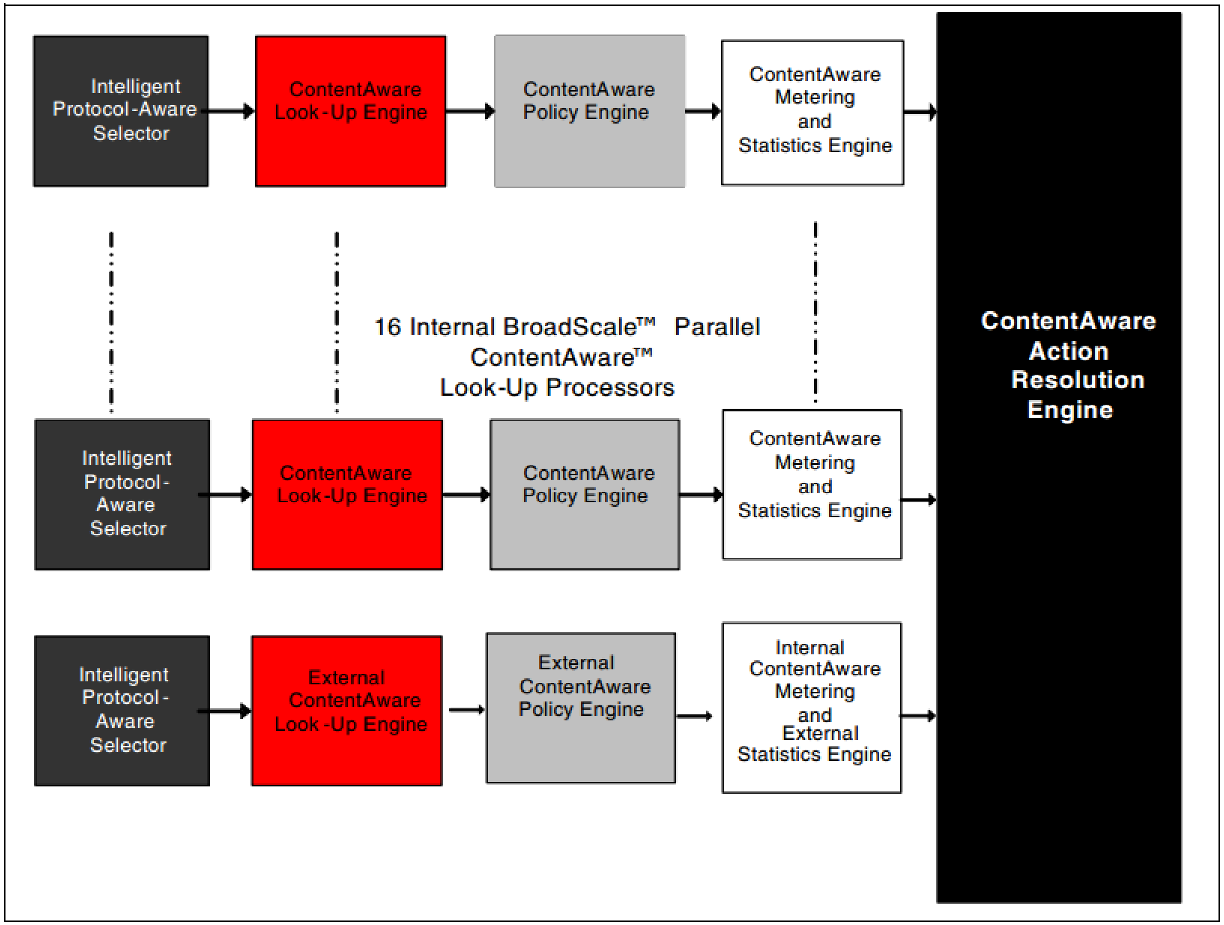
ACL

CAP

- IFP(ICAP)、EFP(ECAP)都属于CAP。IFP采用的是有16个并行内容查找处理器的CAM，CAM被分成16个slice，16个查找处理器并行进行查找，产生16个结果，如果结果的行为不是冲突的，则全部执行；如果行为是冲突的，则根据优先级选择最高优先级的行为执行。

  查找处理器的主要组成部件：

  - Intelligent protocol-aware selector：选择匹配域作为查找的关键字
  - ContentAware lookup engine：执行关键字查找并且输出完全匹配的那条表项索引。
  - ContentAware policy engine：匹配结果将执行的动作行为。根据lookup engine的索引来选择。
  - ContentAware metering and statistics engine：策略与统计收集。
  - ContentAware action resolution engine：多重匹配的处理方式。


Intelligent Protocol-Aware Selector
ContentAware Look-Up Engine
ContentAware Policy Engine
ContentAware Metering and Statistics Engine
16 Internal BroadScale™ Parallel ContentAware™ Look-Up Processors
Intelligent Protocol-Aware Selector
ContentAware Look-Up Engine
ContentAware Policy Engine
ContentAware Metering and Statistics Engine
Intelligent Protocol-Aware Selector
External ContentAware Look-Up Engine
External ContentAware Policy Engine
Internal ContentAware Metering and External Statistics Engine
ContentAware Action Resolution Engine

Intelligent protocol-aware selector

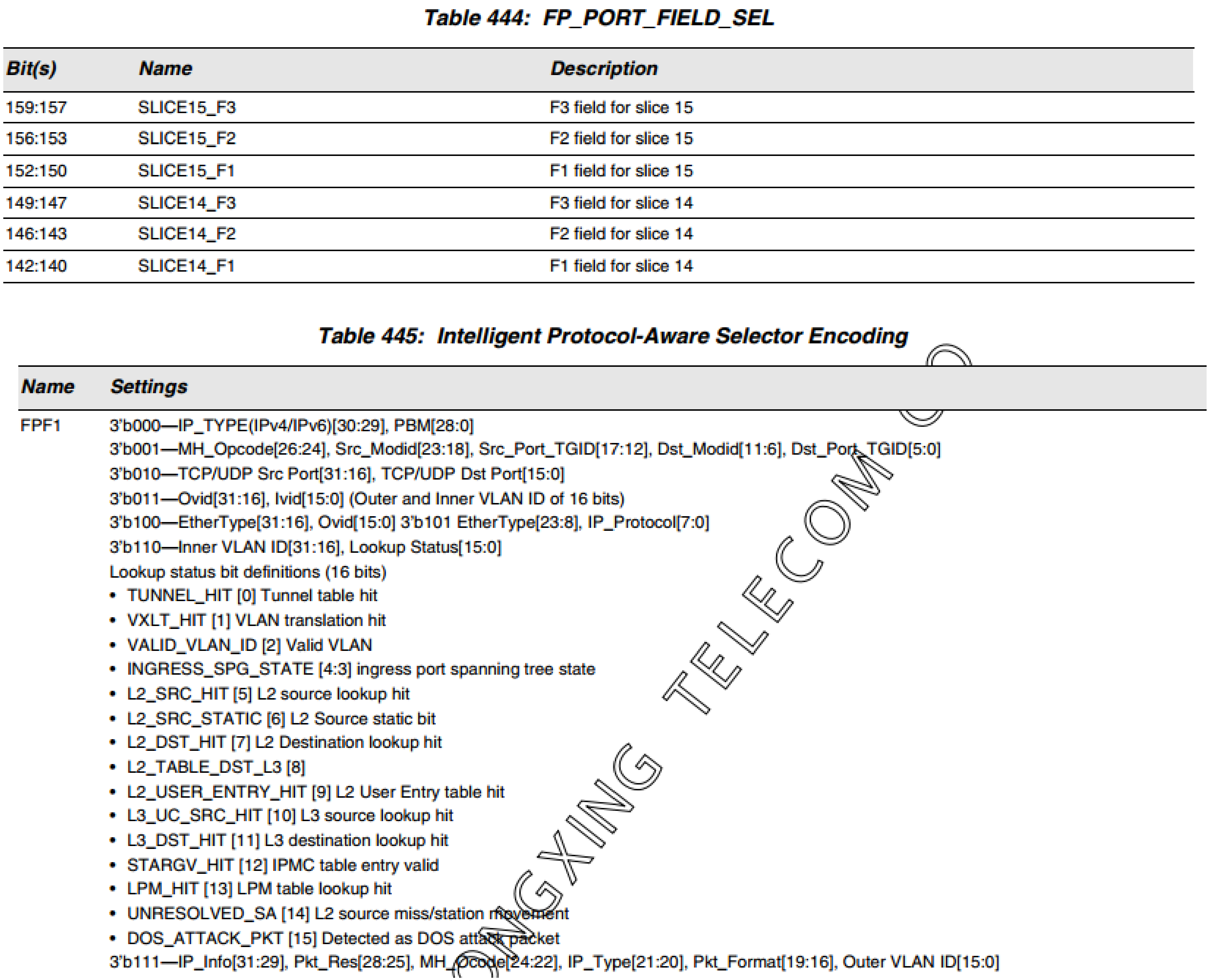


Table 444: FP_PORT_FIELD_SEL

Bit(s)	Name	Description
159:157	SLICE15_F3	F3 field for slice 15
156:153	SLICE15_F2	F2 field for slice 15
152:150	SLICE15_F1	F1 field for slice 15
149:147	SLICE14_F3	F3 field for slice 14
146:143	SLICE14_F2	F2 field for slice 14
142:140	SLICE14_F1	F1 field for slice 14

Table 445: Intelligent Protocol-Aware Selector Encoding

Name	Settings
FPF1	3’b000—IP_TYPE(IPv4/IPv6)[30:29], PBM[28:0] 3’b001—MH_Opcode[26:24], Src_Modid[23:18], Src_Port_TGID[17:12], Dst_Modid[11:6], Dst_Port_TGID[5:0] 3’b010—TCP/UDP Src Port[31:16], TCP/UDP Dst Port[15:0] 3’b011—Ovid[31:16], Ivid[15:0] (Outer and Inner VLAN ID of 16 bits) 3’b100—EtherType[31:16], Ovid[15:0] 3’b101 EtherType[23:8], IP_Protocol[7:0] 3’b110—Inner VLAN ID[31:16], Lookup Status[15:0] Lookup status bit definitions (16 bits) • TUNNEL_HIT [0] Tunnel table hit • VXLT_HIT [1] VLAN translation hit • VALID_VLAN_ID [2] Valid VLAN • INGRESS_SPG_STATE [4:3] ingress port spanning tree state • L2_SRC_HIT [5] L2 source lookup hit • L2_SRC_STATIC [6] L2 Source static bit • L2_DST_HIT [7] L2 Destination lookup hit • L2_TABLE_DST_L3 [8] • L2_USER_ENTRY_HIT [9] L2 User Entry table hit • L3_UC_SRC_HIT [10] L3 source lookup hit • L3_DST_HIT [11] L3 destination lookup hit • STARGV_HIT [12] IPMC table entry valid • LPM_HIT [13] LPM table lookup hit • UNRESOLVED_SA [14] L2 source miss/station movement • DOS_ATTACK_PKT [15] Detected as DOS attack packet 3’b111—IP_Info[31:29], Pkt_Res[28:25], MH_Ocode[24:22], IP_Type[21:20], Pkt_Format[19:16], Outer VLAN ID[15:0]

ContentAware lookup engine

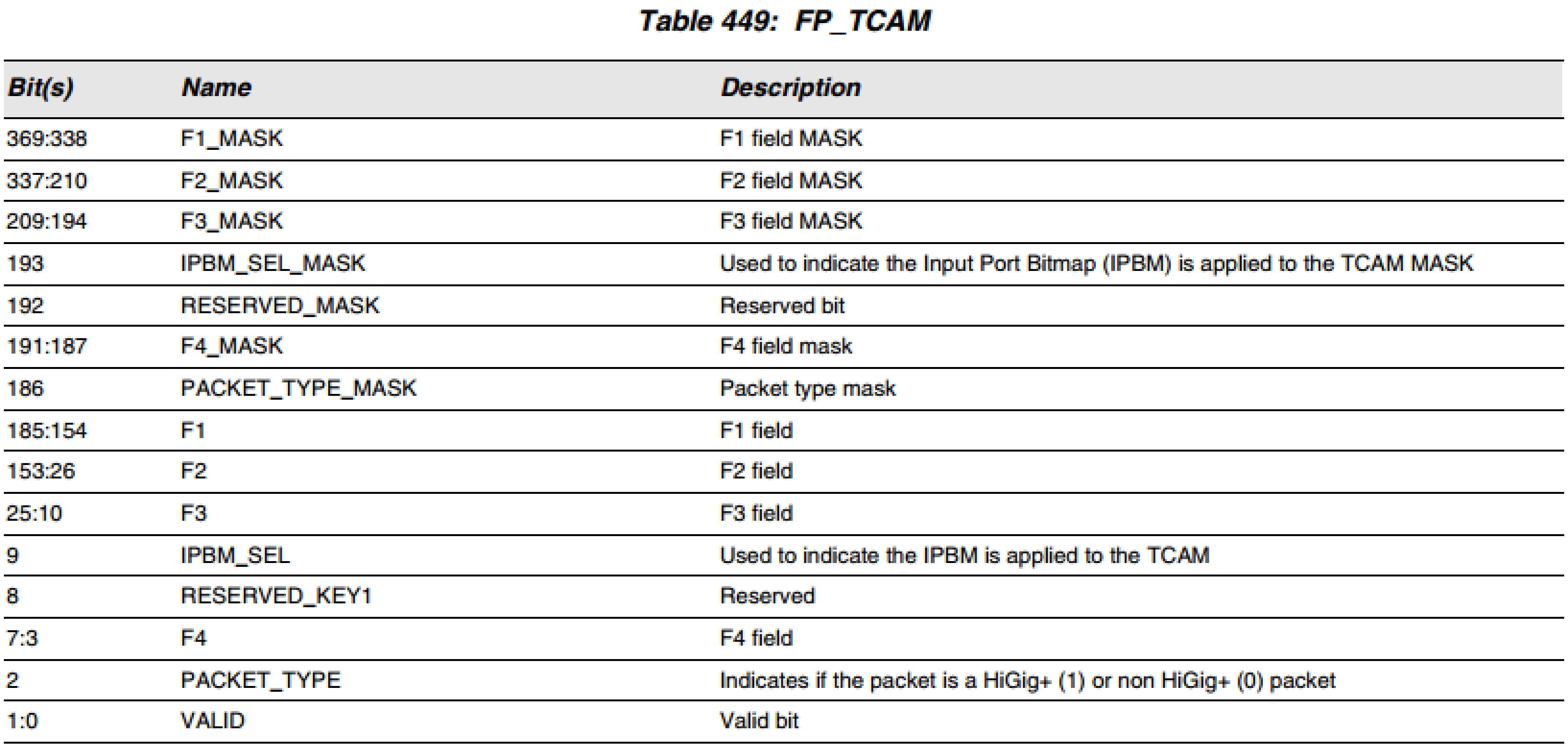


Table 449: FP_TCAM

Bit(s)	*Name*	*Description*
369:338	F1_MASK	F1 field MASK
337:210	F2_MASK	F2 field MASK
209:194	F3_MASK	F3 field MASK
193	IPBM_SEL_MASK	Used to indicate the Input Port Bitmap (IPBM) is applied to the TCAM MASK
192	RESERVED_MASK	Reserved bit
191:187	F4_MASK	F4 field mask
186	PACKET_TYPE_MASK	Packet type mask
185:154	F1	F1 field
153:26	F2	F2 field
25:10	F3	F3 field
9	IPBM_SEL	Used to indicate the IPBM is applied to the TCAM
8	RESERVED_KEY1	Reserved
7:3	F4	F4 field
2	PACKET_TYPE	Indicates if the packet is a HiGig+ (1) or non HiGig+ (0) packet
1:0	VALID	Valid bit

ContentAware policy engine

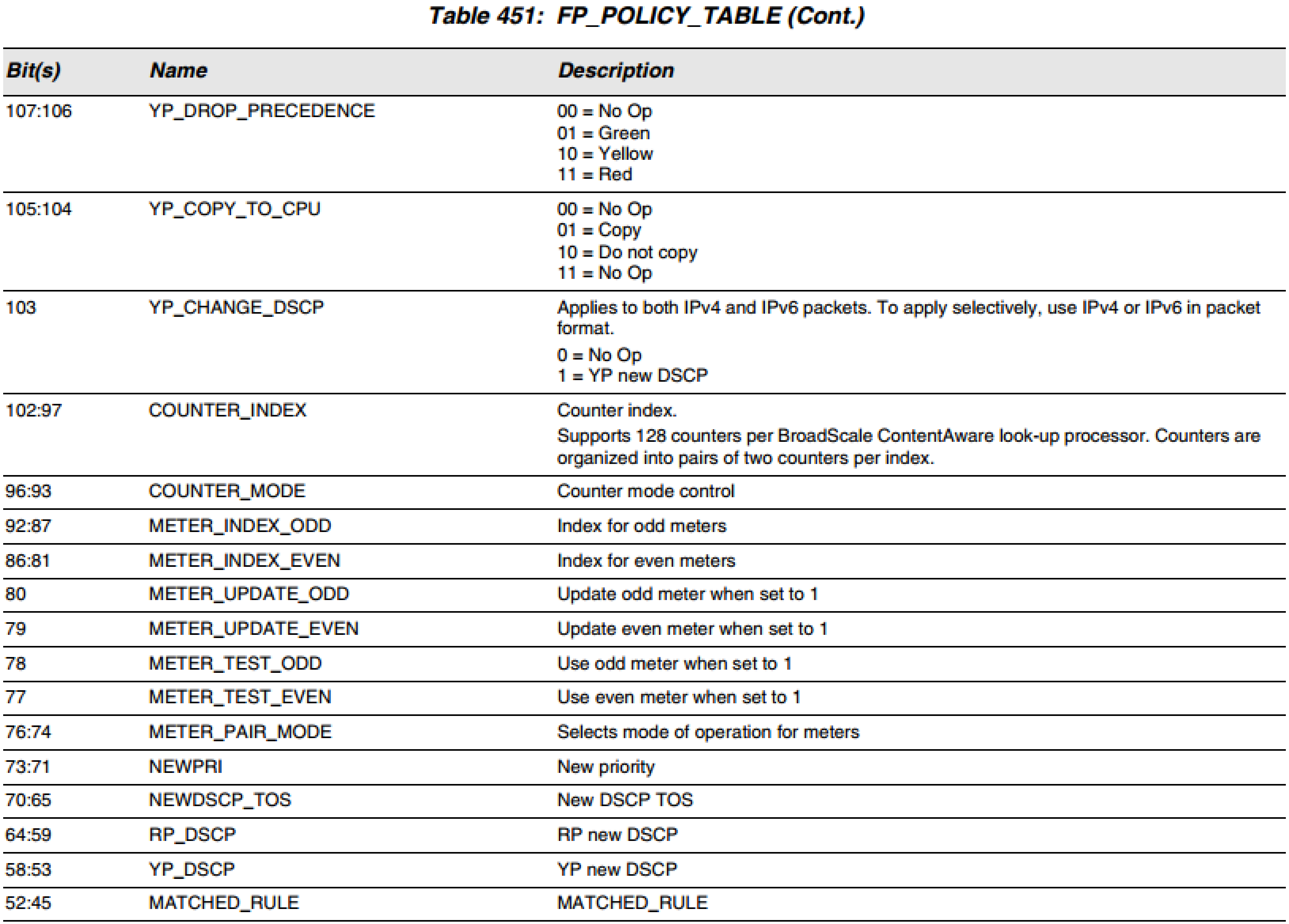
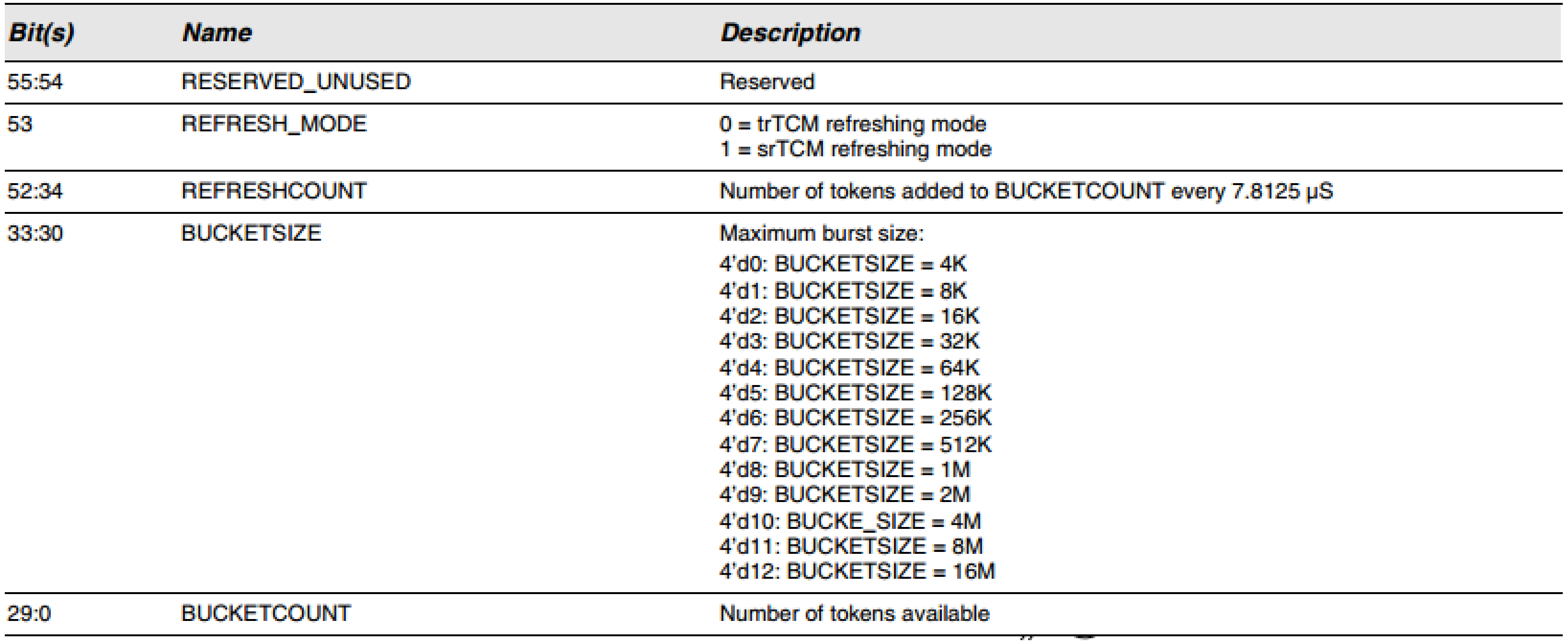


Table 451: FP_POLICY_TABLE (Cont.)

Bit(s)	Name	Description
107:106	YP_DROP_PRECEDENCE	00 = No Op 01 = Green 10 = Yellow 11 = Red
105:104	YP_COPY_TO_CPU	00 = No Op 01 = Copy 10 = Do not copy 11 = No Op
103	YP_CHANGE_DSCP	Applies to both IPv4 and IPv6 packets. To apply selectively, use IPv4 or IPv6 in packet format. 0 = No Op 1 = YP new DSCP
102:97	COUNTER_INDEX	Counter index. Supports 128 counters per BroadScale ContentAware look-up processor. Counters are organized into pairs of two counters per index.
96:93	COUNTER_MODE	Counter mode control
92:87	METER_INDEX_ODD	Index for odd meters
86:81	METER_INDEX_EVEN	Index for even meters
80	METER_UPDATE_ODD	Update odd meter when set to 1
79	METER_UPDATE_EVEN	Update even meter when set to 1
78	METER_TEST_ODD	Use odd meter when set to 1
77	METER_TEST_EVEN	Use even meter when set to 1
76:74	METER_PAIR_MODE	Selects mode of operation for meters
73:71	NEWPRI	New priority
70:65	NEWDSCP_TOS	New DSCP TOS
64:59	RP_DSCP	RP new DSCP
58:53	YP_DSCP	YP new DSCP
52:45	MATCHED_RULE	MATCHED_RULE

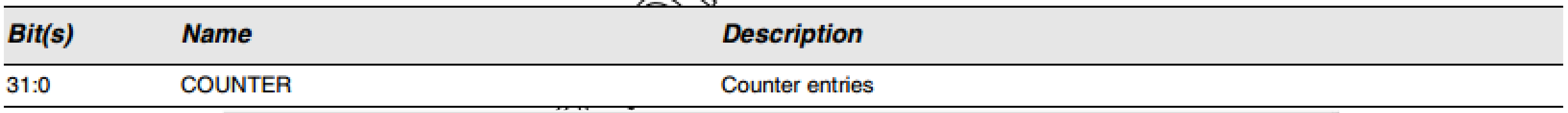
ContentAware metering and statistics engine

Table 452: FP_METER_TABLE

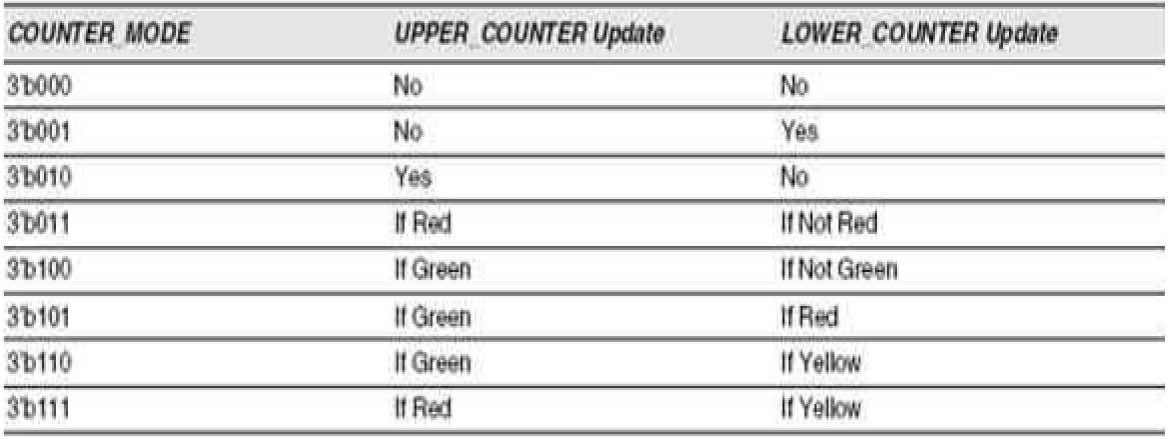


Bit(s)	Name	Description
55:54	RESERVED_UNUSED	Reserved
53	REFRESH_MODE	0 = trTCM refreshing mode 1 = srTCM refreshing mode
52:34	REFRESHCOUNT	Number of tokens added to BUCKETCOUNT every 7.8125 μS
33:30	BUCKETSIZE	Maximum burst size: 4'd0: BUCKETSIZE = 4K 4'd1: BUCKETSIZE = 8K 4'd2: BUCKETSIZE = 16K 4'd3: BUCKETSIZE = 32K 4'd4: BUCKETSIZE = 64K 4'd5: BUCKETSIZE = 128K 4'd6: BUCKETSIZE = 256K 4'd7: BUCKETSIZE = 512K 4'd8: BUCKETSIZE = 1M 4'd9: BUCKETSIZE = 2M 4'd10: BUCKE_SIZE = 4M 4'd11: BUCKETSIZE = 8M 4'd12: BUCKETSIZE = 16M
29:0	BUCKETCOUNT	Number of tokens available

Table 453: FP_COUNTER_TABLE



Bit(s)	Name	Description
31:0	COUNTER	Counter entries



COUNTER_MODE	UPPER_COUNTER Update	LOWER_COUNTER Update
3'b000	No	No
3'b001	No	Yes
3'b010	Yes	No
3'b011	If Red	If Not Red
3'b100	If Green	If Not Green
3'b101	If Green	If Red
3'b110	If Green	If Yellow
3'b111	If Red	If Yellow

Intelligent protocol-aware selector

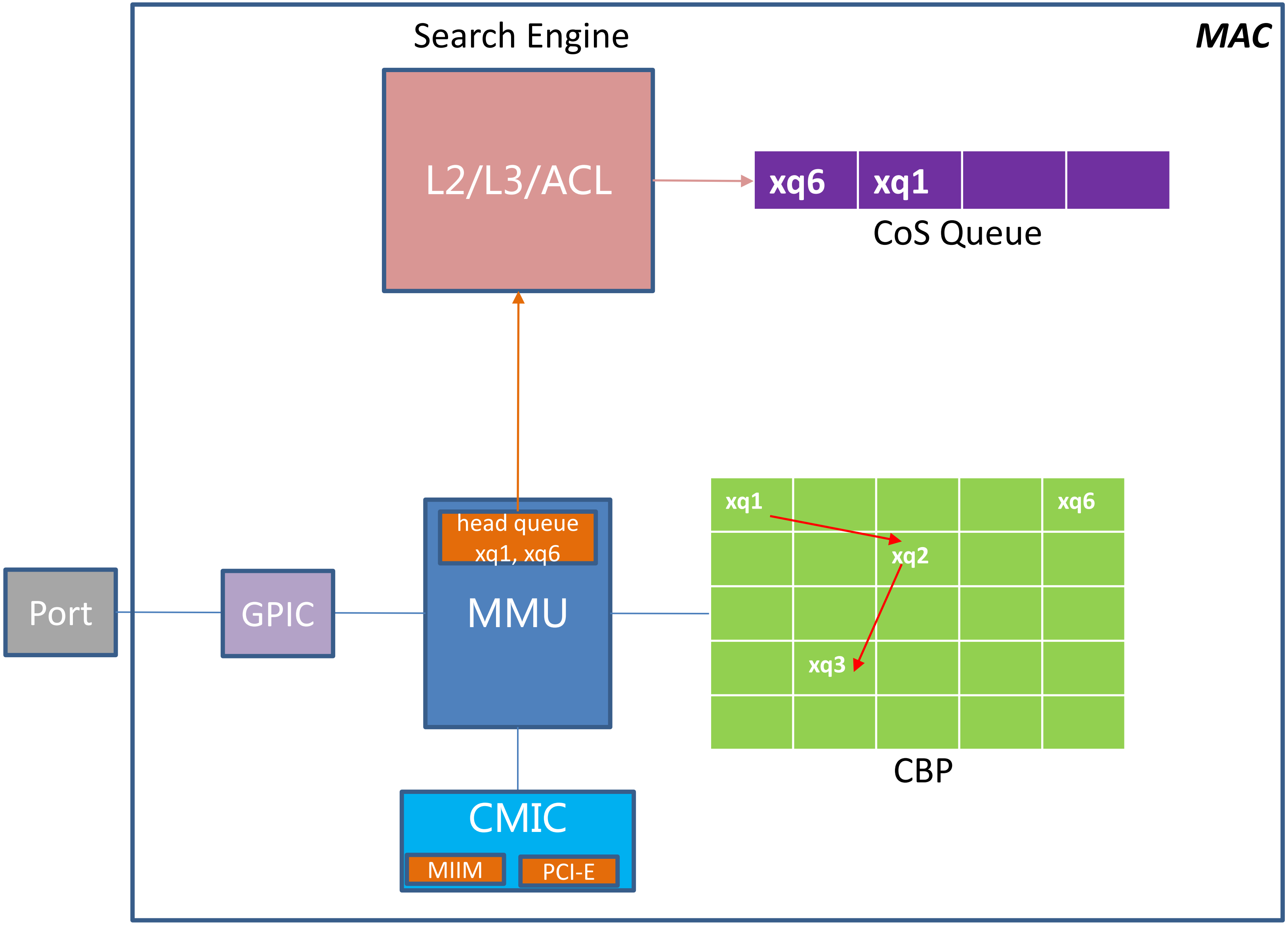
- Intelligent protocol-aware selector：FP_PORT_FILED_SEL用来选择选择Intelligent protocol-aware selector里的哪些匹配域作为解析器解析的关键字(这些关键字将作为Look-Up engine的输入)。Intelligent protocol-aware selector已经由硬件固定形成了几种不同的组合，只能从这些组合里进行选择(如果想要自定义匹配域，芯片有另外提供UDF_OFFSET table)。
- ContentAware lookup engine：lookup可以设置F域(F1,F2,F3)的值及F域的掩码，F域与F掩码的bit位宽是一致的，F掩码=1则表示关心上一步解析后输入的关键字的相应bit位与F域设置的值是否一样，如果所有关心的bit位都一样则表示匹配成功，否则匹配失败；如果F掩码=0则表示不关心相应的bit位。
- ContentAware policy engine：当lookup engine查找匹配时，其匹配项的索引值作为policy engine的索引值(lookup与policy的表项一一对应)，由policy engine设置的action来决定对报文的处理方式，如drop、permit、to_cpu等动作。
- ContentAware metering and statistics engine：由FP_METER_TABLE与FP_COUNTER_TABLE组成。Policy engine的METER_INDEX与CONTER_INDEX分别作为索引指向其配套的METER与COUNTER table。
  - METER用来测量报文的状态并使用srTCM/trTCM进行着色，把着色的结果反馈给配套的lookup engine进行处理。METER一般是两条表项配对使用，由policy engine的METER_INDEX_ODD与METER_INDEX_EVEN进行索引，分别对应srTCM的C桶与E桶或者trTCM的C桶与P桶。
  - COUNTER根据policy选择的COUNTER_MODE进行Green/Yellow/Red报文的着色个数统计。Policy engine的COUNTER_INDEX与COUNTER_INDEX+1分别指向LOWER_COUNTER与UPPER_COUNTER。

Buffer Management

MMU

- MMU主要功能是负责在正常与拥塞的传输情景下，高效地管理CBP(Cell Buffer Pool)和报文指针资源。
  - 能够以最大115Gbps的速率接收端口的报文数据流
  - 单播与组播报文能够以16Gbps的速率从出端口发送出去
  - 计算ingress/egress的容量使用情况
  - 高效地管理组播传输
  - 支持12k字节的巨帧
  - 存储与处理每个报文的12/16字节头部信息
  - 支持HiGig协议
  - 允许CPU从端口接收，或者往端口发送
  - 支持802.1p的队列优先级
  - 基于每个CoS队列的流量控制
  - 动态容量管理
  - 拥塞控制消息的处理与生成
- 缓存管理
  - 入端：监视入端口CoS队列的可用缓存使用情况。如果超过阈值，则产生流量控制消息，发往该端口队列。如果对端处理流控报文，则对端会停止发送，以缓解入端的压力。
  - 出端：监视出端口CoS队列的缓存使用情况。如果超过阈值，新到来的入口报文如果目的出口是超过阈值的端口，那么拥塞控制将会把入口报文立即丢弃。
- 报文缓存管理
  - 缓存以cell为最小单位，每个cell大小为128字节。32768个cell提供了最大4MB的缓存容量。同时提供了11k个packet pointer(XQ)，每个报文分片存储在cell里，指针(XQ)用来指示报文的第一个cell。
  - 入端与出端分别独立地统计进入芯片的报文。对于多播或者泛洪的报文，存在多个出口，入端只统计一个报文计数，所有出端也只统计一个报文计数。

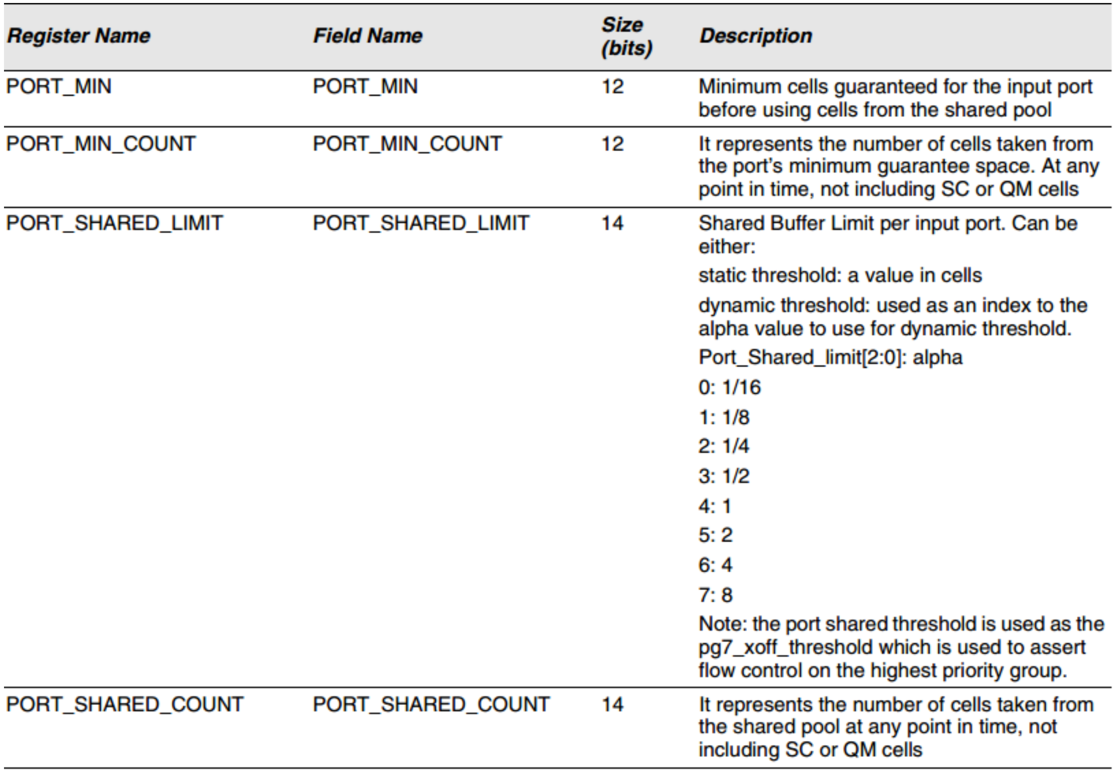
MMU转发框架


Search Engine
MAC
L2/L3/ACL
xq6
xq1
CoS Queue
head queue
xq1, xq6
MMU
Port
GPIC
xq1
xq6
xq2
xq3
CBP
CMIC
MIIM
PCI-E

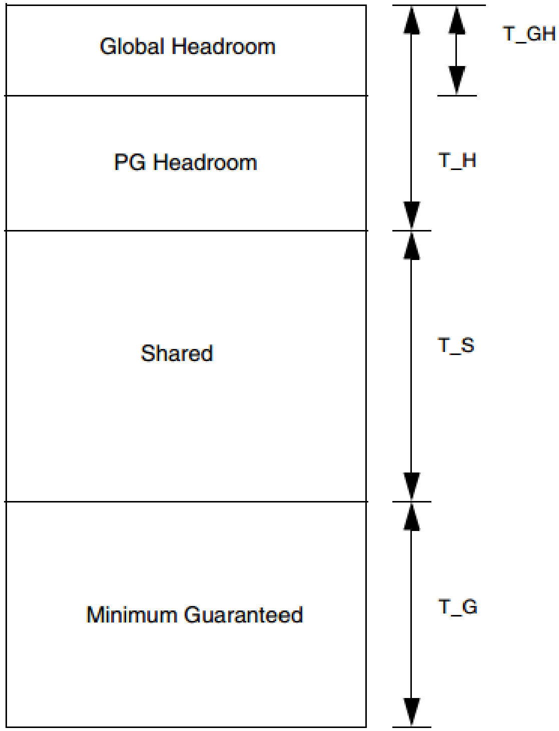
Ingress Control

- 缓存空间分为3个部分
  - Minimum Guaranteed space：为端口提供最小保证可用空间PORT_MIN。
  - Shared space：当最小保证空间不足时，为端口提供的共享缓存空间。TOTAL_SHARED_LIMIT指示总的共享缓存空间，PORT_SHARED_LIMIT指示每个端口的最大可用共享缓存空间。
  - Headromm space：当最小保证空间与共享缓存空间不足时，提供一些额外的缓存能力
    - PG Headroom：流控帧发送出去到对端停止发送报文这段时期，可能收到对端发送过来的部分报文，这分部缓存空间即用于接收这部分报文。由PG_HDRM_LIMIT指示其缓存空间。
    - Global Headroom：如果没有为每个端口单独分配PG Headroom，则使用这分部缓存空间作为所有端口的PG Headroom共享空间。如果使用此空间，则每个端口只允许保存一个报文。此空间允许存储的总报文个数由PORT_MAX_PKT_SIZE指示。
- 缓存规则：

报文接收时，优先使用Guaranteed space，然后使用Shared space，最后是Headroom。当报文发送后，则优先释放Headromm，然后是Shared space，最后是Guaranteed space。



Register Name	***Field Name***	***Size (bits)***	***Description***
PORT_MIN	PORT_MIN	12	Minimum cells guaranteed for the input port before using cells from the shared pool
PORT_MIN_COUNT	PORT_MIN_COUNT	12	It represents the number of cells taken from the port's minimum guarantee space. At any point in time, not including SC or QM cells
PORT_SHARED_LIMIT	PORT_SHARED_LIMIT	14	Shared Buffer Limit per input port. Can be either: static threshold: a value in cells dynamic threshold: used as an index to the alpha value to use for dynamic threshold. Port_Shared_limit[2:0]: alpha 0: 1/16 1: 1/8 2: 1/4 3: 1/2 4: 1 5: 2 6: 4 7: 8 Note: the port shared threshold is used as the pg7_xoff_threshold which is used to assert flow control on the highest priority group.
PORT_SHARED_COUNT	PORT_SHARED_COUNT	14	It represents the number of cells taken from the shared pool at any point in time, not including SC or QM cells



PAUSE Metering

- 每个端口都有单独的漏桶算法机制，用来监测其缓存的使用情况。PAUSE metering功能用来在入端口触发流量控制，以反压对端的流量，实现入口流量整形。
  - BUCKET_COUNT表示当前桶里的令牌个数，一开始BUCKET_COUNT是0，如果有报文到达，则按报文字节大小转换成相应的令牌个数加入到桶里。每过一个T_REFRESH(7.8125us)周期，就把REFRESHCOUNT个数的令牌从桶里取出(BUCKET_COUNT -= REFRESHCOUNT)。每个令牌的粒度大小由METER_GRANULARITY选择。
  - 当BUCKET_COUNT到达DISCARD_THD水线时，MMU将通告端口丢弃到达的报文，在通告期间可能会有一部分报文到达，这部分报文由MMU丢弃。直到BUCKET_COUNT降到RESUME_THD水线时，MMU才会开始接收报文。

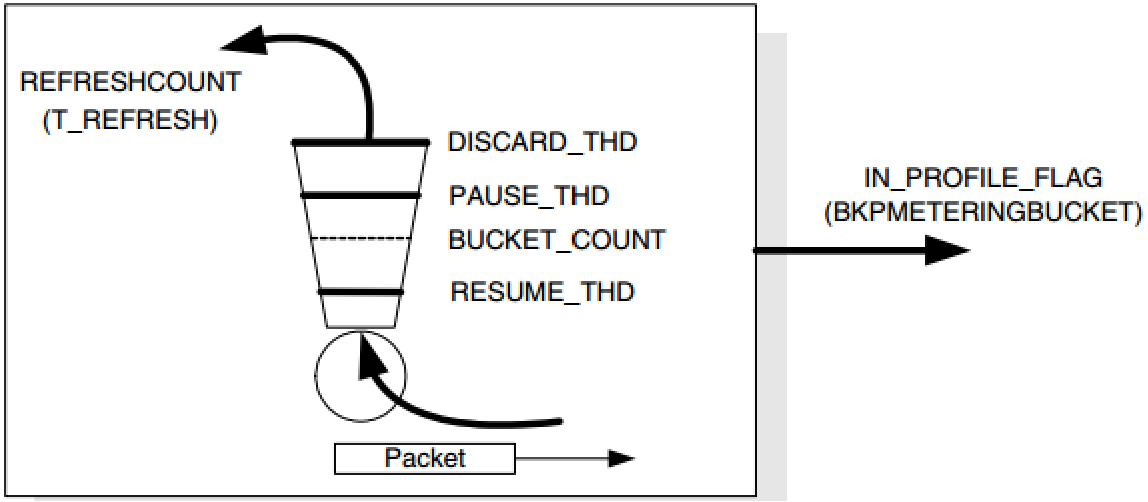


Table 272: BKPMETERINGBUCKET

Bit	Name	R/W	Description	Default
31:30	RESERVED	RO	Reserved	0x0
29	IN_PROFILE_FLAG	RO	In-profile flag, indicates the current state of the Backpressure Metering bucket 0 = In profile 1 = Out of profile (polarity backwards based upon bit name)	0
28:0	BUCKET_COUNT	RO	Pause metering bucket count	0x00000000

PAUSE Metering

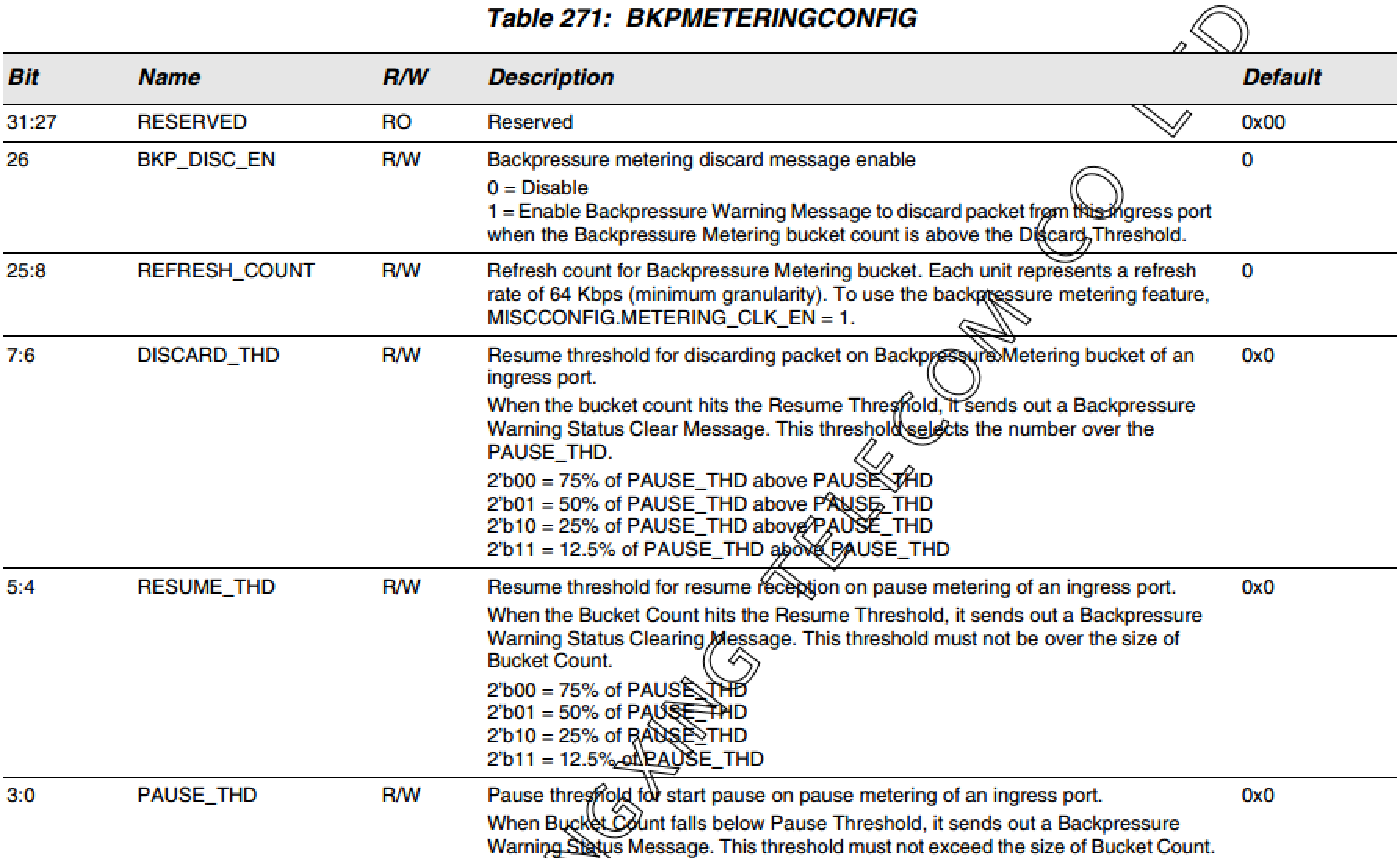


Table 271: BKPMETERINGCONFIG

Bit	*Name*	*R/W*	*Description*	*Default*
31:27	RESERVED	RO	Reserved	0x00
26	BKP_DISC_EN	R/W	Backpressure metering discard message enable 0 = Disable 1 = Enable Backpressure Warning Message to discard packet from this ingress port when the Backpressure Metering bucket count is above the Discard Threshold.	0
25:8	REFRESH_COUNT	R/W	Refresh count for Backpressure Metering bucket. Each unit represents a refresh rate of 64 Kbps (minimum granularity). To use the backpressure metering feature, MISCCONFIG.METERING_CLK_EN = 1.	0
7:6	DISCARD_THD	R/W	Resume threshold for discarding packet on Backpressure Metering bucket of an ingress port. When the bucket count hits the Resume Threshold, it sends out a Backpressure Warning Status Clear Message. This threshold selects the number over the PAUSE_THD. 2'b00 = 75% of PAUSE_THD above PAUSE_THD 2'b01 = 50% of PAUSE_THD above PAUSE_THD 2'b10 = 25% of PAUSE_THD above PAUSE_THD 2'b11 = 12.5% of PAUSE_THD above PAUSE_THD	0x0
5:4	RESUME_THD	R/W	Resume threshold for resume reception on pause metering of an ingress port. When the Bucket Count hits the Resume Threshold, it sends out a Backpressure Warning Status Clearing Message. This threshold must not be over the size of Bucket Count. 2'b00 = 75% of PAUSE_THD 2'b01 = 50% of PAUSE_THD 2'b10 = 25% of PAUSE_THD 2'b11 = 12.5% of PAUSE_THD	0x0
3:0	PAUSE_THD	R/W	Pause threshold for start pause on pause metering of an ingress port. When Bucket Count falls below Pause Threshold, it sends out a Backpressure Warning Status Message. This threshold must not exceed the size of Bucket Count.	0x0

Egress Control

- Ingress Control是牺牲吞吐达到尽量无丢包，比如流控将防止到达的报文超过阈值而丢弃，但会造成吞吐下降。Egress Control是丢弃报文尽量达到高吞吐，比如不考虑报文是否处理及时，只管设置高阈值让吞吐量变大。
- Egress Control通过设置每个端口的阈值来实现。出端口与8个CoS队列关联，每个队列都有各自的阈值，这些阈值决定了准备进入队列的报文，及哪个的目的端口为该端口的报文将会会丢弃。类似于Ingress Control，Egress Control也有两个缓存资源：
  - Minimum Guaranteed Resources
  - Shared Resouces
- 与Ingress Control不同的是当一个报文被分成多个cell时，分片cell与packet pointer只能同时属于两个资源池中的其中一个，不能跨资源池存放；同时报文发送后，cell使用量计数将会被相应地减小，而packet pointer的使用量计数却不变。

- Minimum Guaranteed Resources配置寄存器
  - OP_Queue_Config.Q_MIN
- Shared Resouces配置寄存器
  - Total shared buffer(OP_BUFFER_SHARED_LIMIT)
  - 只支持静态的Per port(OP_PORT_CONFIG.OP_SHARED_LIMIT)
  - 支持静态或动态的Per Q(OP_Queue_Config.Q_SHARED_LIMIT)

End-to-end Congestion Control

- 当出端口出现拥塞时，所有入端口接收的报文如果目的端口是该出口，则会加剧拥塞且造成丢包。通过端到端的拥塞控制机制，当出端口出现拥塞时，广播一个拥塞控制消息给所有端口，则所有端口将会在入端口丢弃目的端口是该拥塞出端口的报文。当拥塞解除时，出端口发送一个解除消息给所有端口，则入端口恢复正常的报文接收。
- 拥塞消息的通告将采用类似于RED早期检测的方法，在出端口即将出现拥塞丢包前进行消息的通告。

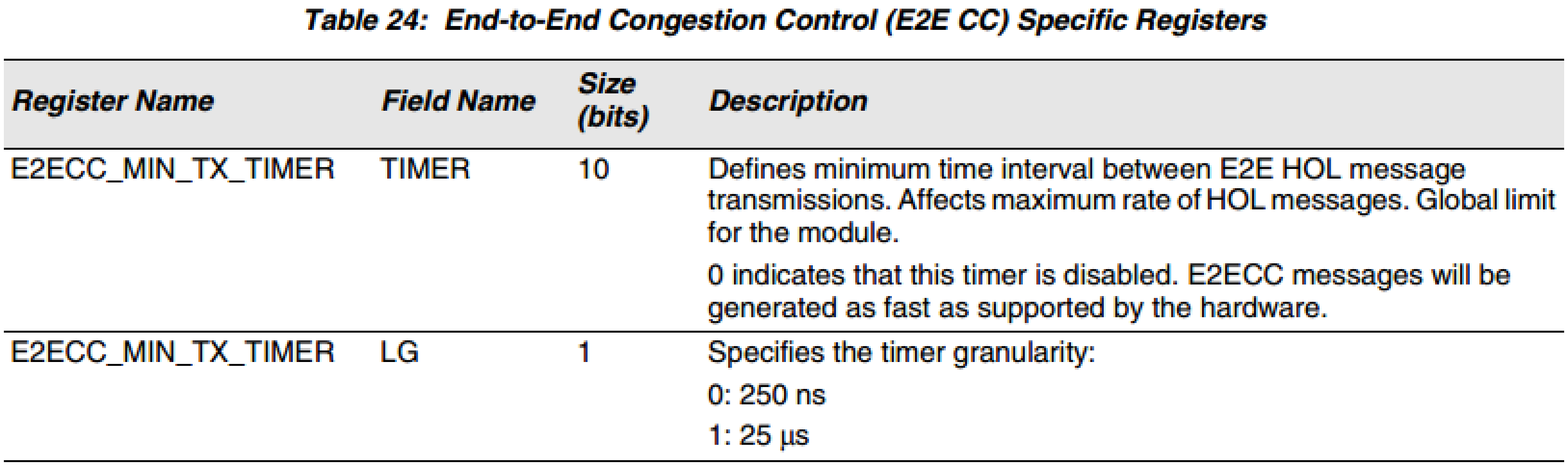


Table 24: End-to-End Congestion Control (E2E CC) Specific Registers

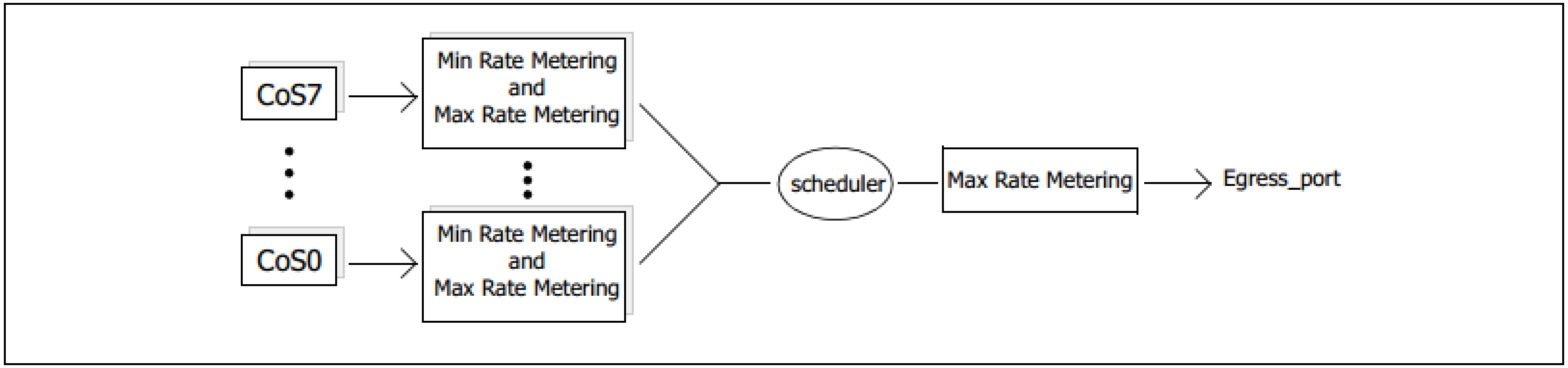
Register Name	***Field Name***	***Size (bits)***	***Description***
E2ECC_MIN_TX_TIMER	TIMER	10	Defines minimum time interval between E2E HOL message transmissions. Affects maximum rate of HOL messages. Global limit for the module. 0 indicates that this timer is disabled. E2ECC messages will be generated as fast as supported by the hardware.
E2ECC_MIN_TX_TIMER	LG	1	Specifies the timer granularity: 0: 250 ns 1: 25 µs

QoS(Traffic Management)

- 不同的应用对网络的服务质量的要求有着明显的不同，比如一些实时应用尤其是多媒体应用，如图像、语音的传输，允许偶尔的数据丢失，但对时延和抖动有严格的要求；而一般的数据传输，如FTP文件传输对时延没有太大的要求，但必须完全的可靠准确。
- 当出端口出现拥塞时，所有入端口接收的报文如果目的端口是该出口，则会加剧拥塞且造成丢包。通过端到端的拥塞控制机制，当出端口出现拥塞时，广播一个拥塞控制消息给所有端口，则所有端口将会在入端口丢弃目的端口是该拥塞出端口的报文。当拥塞解除时，出端口发送一个解除消息给所有端口，则入端口恢复正常的报文接收。
- 拥塞消息的通告将采用类似于RED早期检测的方法，在出端口即将出现拥塞丢包前进行消息的通告。
- 芯片的QoS工作机制，是由监测每个出端口CoS队列的带宽使用情况，并反馈给调度器，由调度器控制CoS队列的最大及最小带宽来实现的。

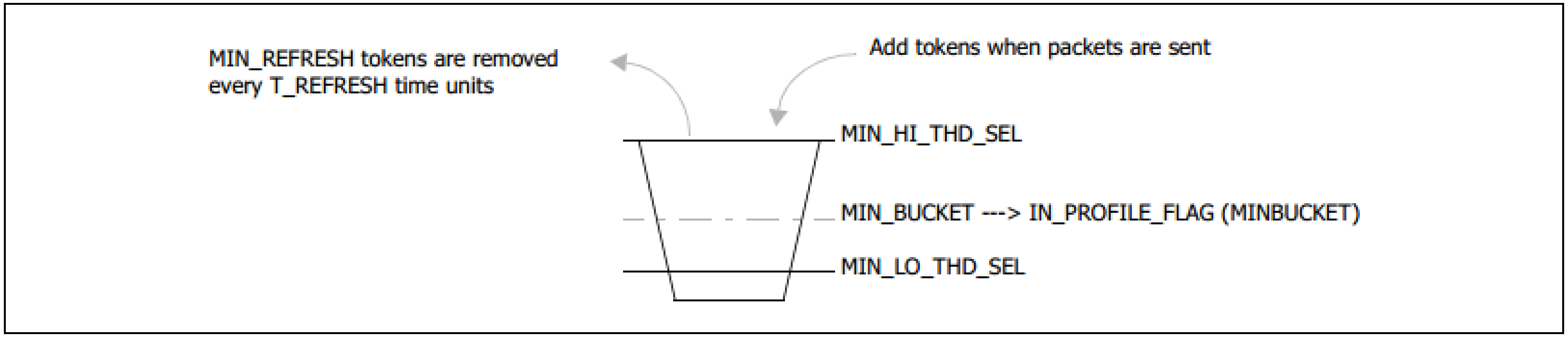
  监测机制把CoS队列分类到不同的调度组里：
  - MinNotMet：未超过最小带宽阈值的CoS队列
  - MaxNotMet：超过最小带宽阈值，但未超过最大带宽阈值的CoS队列
  - MaxExceeded：超过最大带宽阈值的CoS队列

  调度器根据调度组：
  - 如果MinNotMet组里的CoS队列非空(有报文)，且这些报文需要发送出去，则根据调度准则对CoS队列进行处理。
  - 如果MinNotMet组里的CoS队列为空(无报文)，但MaxNotMet组里的CoS队列非空(有报文)，则根据调度准则对CoS队列进行进行处理。
  - 如果MinNotMet与MaxNotMet组都是空的，或者只有MaxExceeded组非空，那么调度器不会处理任何CoS队列。


CoS7
CoS0
Min Rate Metering and Max Rate Metering
Min Rate Metering and Max Rate Metering
scheduler
Max Rate Metering
Egress_port

QoS(Minumum Bandwidth CoS Queue Metering)

- 带宽监测基于每个CoS队列单独进行。最小带宽监测是为了向每个出端口的CoS队列提供最小的带宽保证。最小带宽监测是通过漏桶机制来实现。
  - 当桶的使用超过最小带宽MIN_LO_THD_SEL，IN_PROFILE_FLAG设置为1；当桶的使用降到最小带宽MIN_LO_THD_SEL以下，则IN_PROFILE_FLAG设置为0。IN_PROFILE_FLAG=0则放入MinNotMet组，IN_PROFILE_FLAG=1则放入MaxNotMet组。
  - MIN_HI_THD_SEL用来限制桶的最高值。
  - 如果MIN_LO_THD_SEL设置得越低，则CoS队列在MinNotMet组的时间就就越少，换句话说，队列的优先处理时间就变少了(MinNotMet组最优先处理)，相当于CoS的优先级相应地变低了。


MIN_REFRESH tokens are removed every T_REFRESH time units
Add tokens when packets are sent
MIN_HI_THD_SEL
MIN_BUCKET ---> IN_PROFILE_FLAG (MINBUCKET)
MIN_LO_THD_SEL

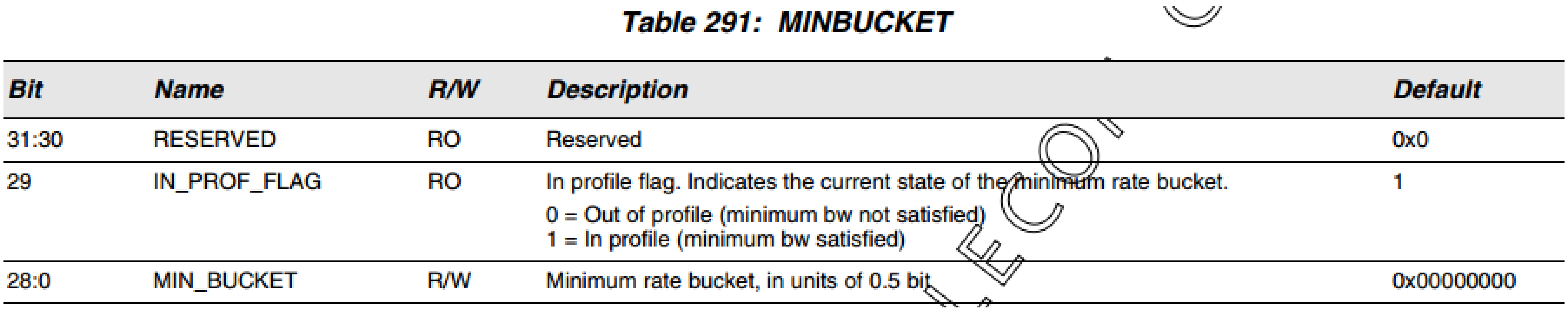


Table 291: MINBUCKET

Bit	Name	R/W	Description	Default
31:30	RESERVED	RO	Reserved	0x0
29	IN_PROF_FLAG	RO	In profile flag. Indicates the current state of the minimum rate bucket. 0 = Out of profile (minimum bw not satisfied) 1 = In profile (minimum bw satisfied)	1
28:0	MIN_BUCKET	R/W	Minimum rate bucket, in units of 0.5 bit	0x00000000

QoS(Minumum Bandwidth CoS Queue Metering)

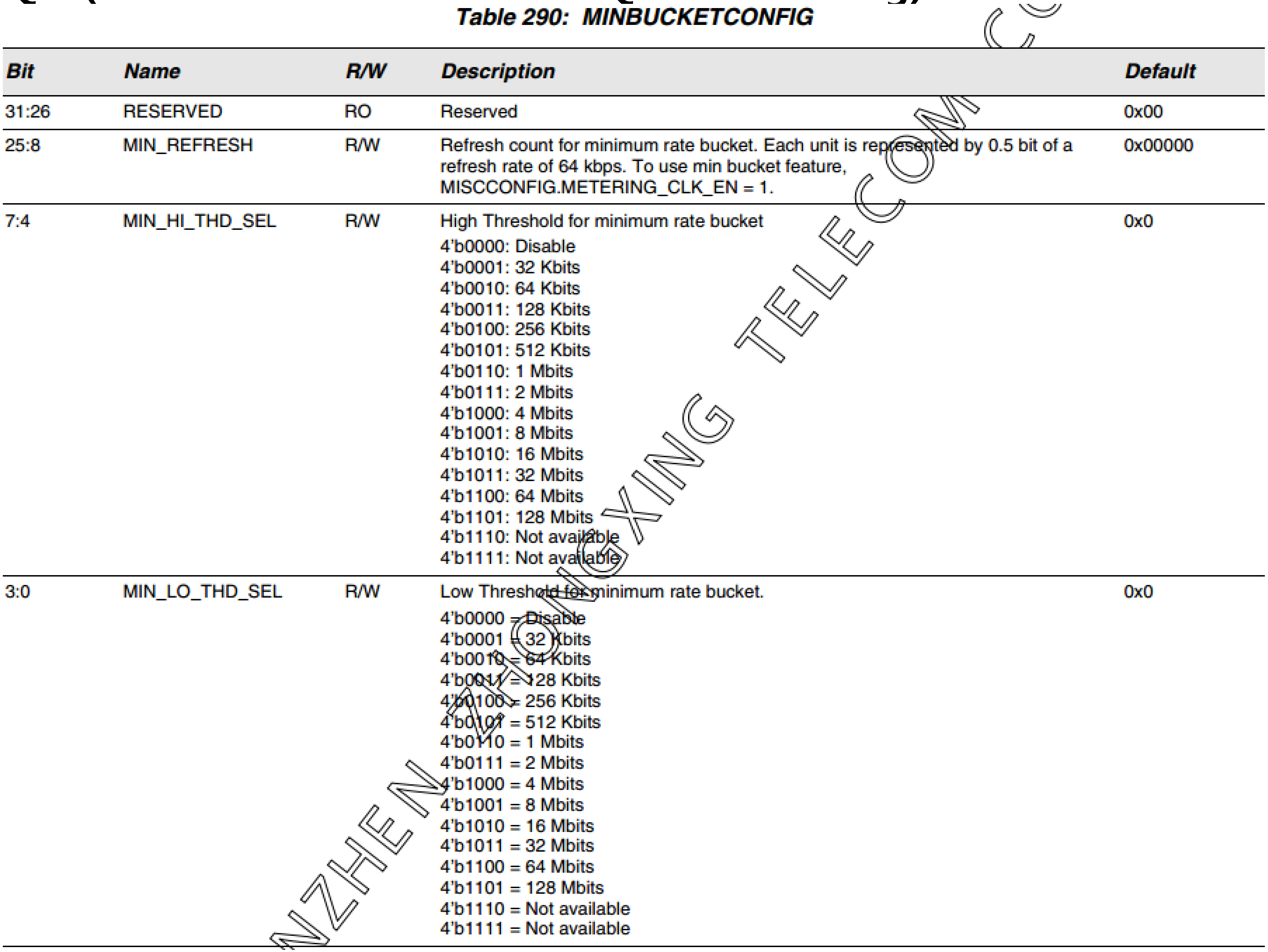
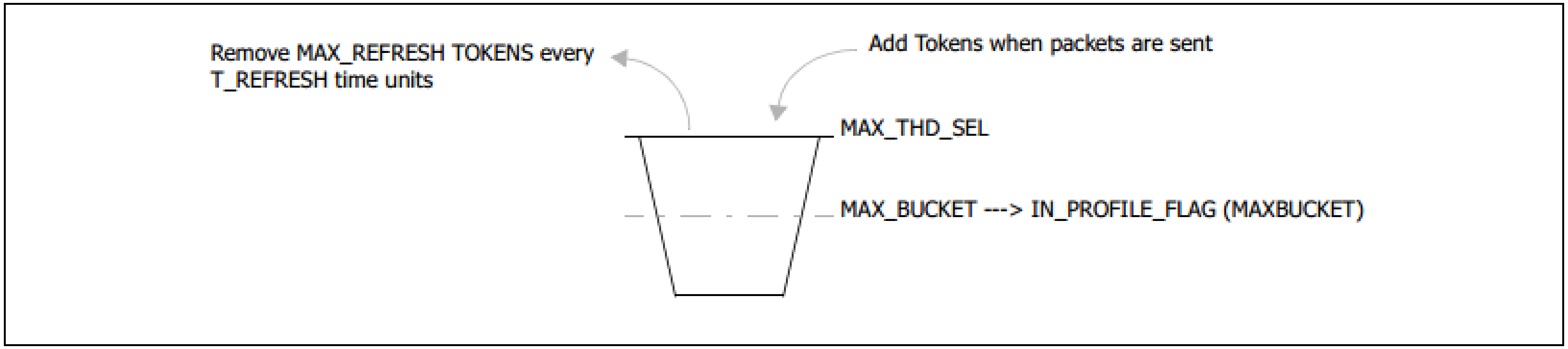


Table 290: MINBUCKETCONFIG

Bit	Name	R/W	Description	Default
31:26	RESERVED	RO	Reserved	0x00
25:8	MIN_REFRESH	R/W	Refresh count for minimum rate bucket. Each unit is represented by 0.5 bit of a refresh rate of 64 kbps. To use min bucket feature, MISCCONFIG.METERING_CLK_EN = 1.	0x00000
7:4	MIN_HI_THD_SEL	R/W	High Threshold for minimum rate bucket 4'b0000: Disable 4'b0001: 32 Kbits 4'b0010: 64 Kbits 4'b0011: 128 Kbits 4'b0100: 256 Kbits 4'b0101: 512 Kbits 4'b0110: 1 Mbits 4'b0111: 2 Mbits 4'b1000: 4 Mbits 4'b1001: 8 Mbits 4'b1010: 16 Mbits 4'b1011: 32 Mbits 4'b1100: 64 Mbits 4'b1101: 128 Mbits 4'b1110: Not available 4'b1111: Not available	0x0
3:0	MIN_LO_THD_SEL	R/W	Low Threshold for minimum rate bucket. 4'b0000 = Disable 4'b0001 = 32 Kbits 4'b0010 = 64 Kbits 4'b0011 = 128 Kbits 4'b0100 = 256 Kbits 4'b0101 = 512 Kbits 4'b0110 = 1 Mbits 4'b0111 = 2 Mbits 4'b1000 = 4 Mbits 4'b1001 = 8 Mbits 4'b1010 = 16 Mbits 4'b1011 = 32 Mbits 4'b1100 = 64 Mbits 4'b1101 = 128 Mbits 4'b1110 = Not available 4'b1111 = Not available	0x0

QoS(Maximum Bandwidth CoS Queue Metering)

- 最大带宽监测是为了控制每个出端口的CoS队列的最大带宽限制。最大带宽监测是通过漏桶机制来实现。
  - 当桶的使用超过最大带宽MAX_THD_SEL时，CoS队列被放入MaxExceeded组，在队列退出MaxExceeded组之前，队列将被停止服务。
  - 最大带宽限制将影响最大的突发流量大小。如果设置得太低，比终端填充这个队列的速率还低，可能导致无可用带宽。


Remove MAX_REFRESH TOKENS every
T_REFRESH time units
Add Tokens when packets are sent
MAX_THD_SEL
MAX_BUCKET ---> IN_PROFILE_FLAG (MAXBUCKET)

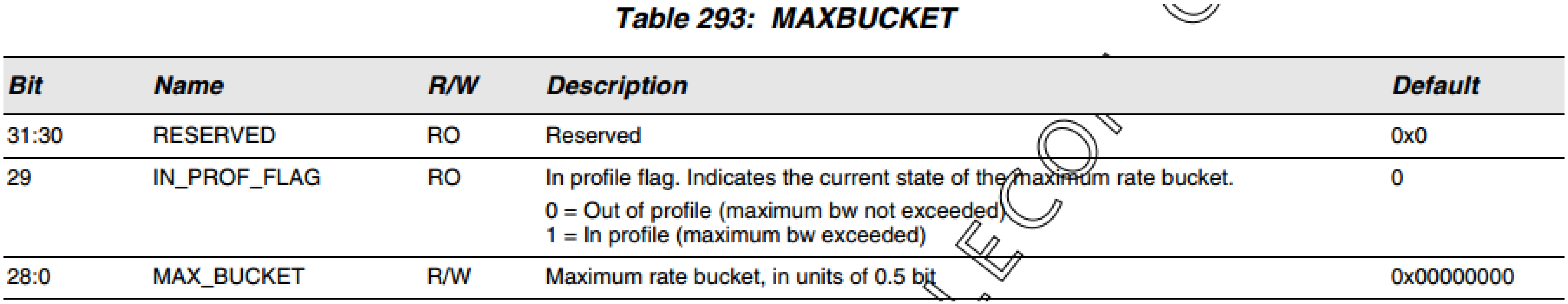


Table 293: MAXBUCKET

Bit	Name	R/W	Description	Default
31:30	RESERVED	RO	Reserved	0x0
29	IN_PROF_FLAG	RO	In profile flag. Indicates the current state of the maximum rate bucket. 0 = Out of profile (maximum bw not exceeded) 1 = In profile (maximum bw exceeded)	0
28:0	MAX_BUCKET	R/W	Maximum rate bucket, in units of 0.5 bit	0x00000000

QoS(Maximum Bandwidth CoS Queue Metering)

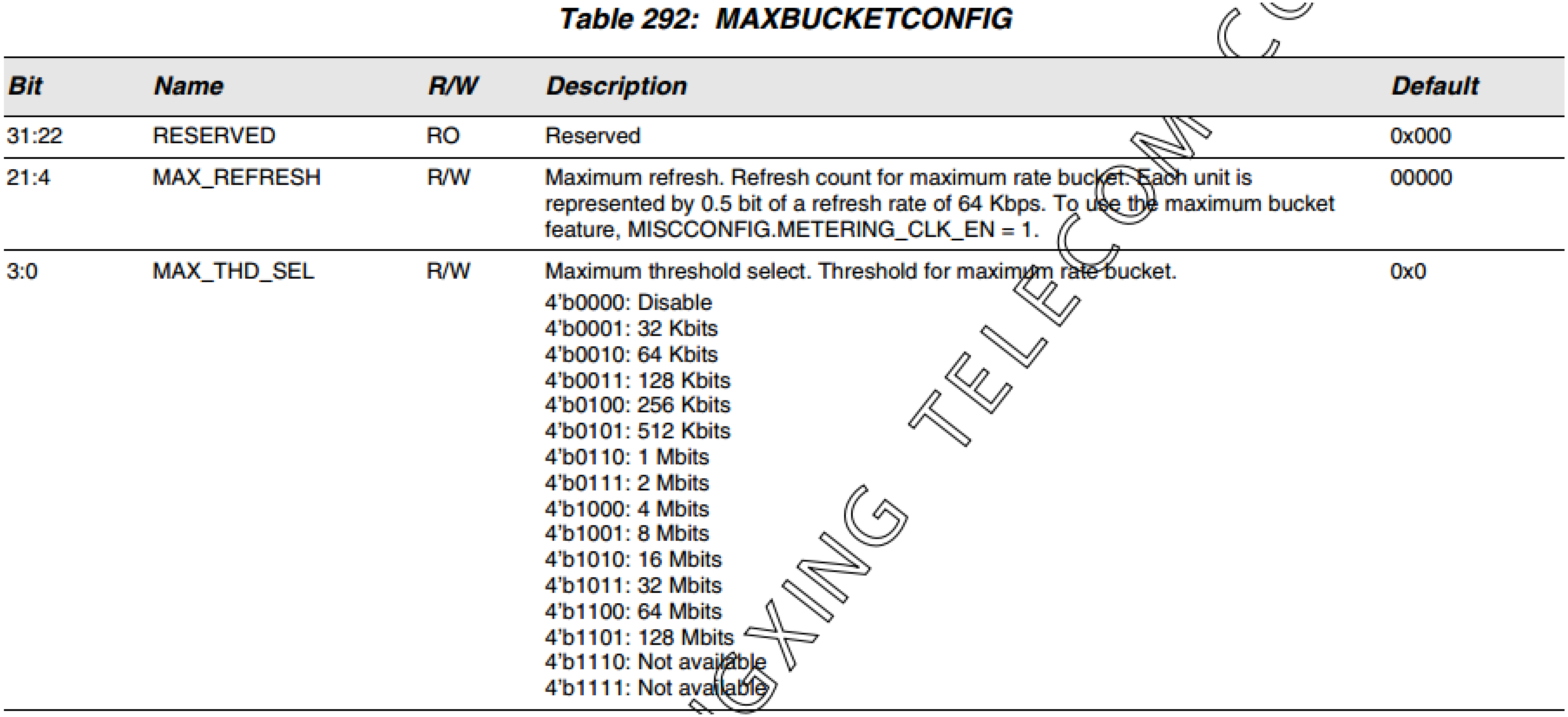


Table 292: MAXBUCKETCONFIG

Bit	Name	R/W	Description	Default
31:22	RESERVED	RO	Reserved	0x000
21:4	MAX_REFRESH	R/W	Maximum refresh. Refresh count for maximum rate bucket. Each unit is represented by 0.5 bit of a refresh rate of 64 Kbps. To use the maximum bucket feature, MISCCONFIG.METERING_CLK_EN = 1.	00000
3:0	MAX_THD_SEL	R/W	Maximum threshold select. Threshold for maximum rate bucket. 4'b0000: Disable 4'b0001: 32 Kbits 4'b0010: 64 Kbits 4'b0011: 128 Kbits 4'b0100: 256 Kbits 4'b0101: 512 Kbits 4'b0110: 1 Mbits 4'b0111: 2 Mbits 4'b1000: 4 Mbits 4'b1001: 8 Mbits 4'b1010: 16 Mbits 4'b1011: 32 Mbits 4'b1100: 64 Mbits 4'b1101: 128 Mbits 4'b1110: Not available 4'b1111: Not available	0x0

QoS(Schedule Algorithms)

- 支持四种CoS队列调度算法用来实现出口队列调度
  - Strict Priority across CoS queues
    - 严格优先级CoS队列调度，调度器完全按照队列的优先级进行报文调度出队转发，不轮询，调度次优先级队列的条件是：最高优先级的CoS队列空，或者使能关闭。
  - Round Robin across CoS queues
    - 轮询优先级CoS队列调度，调度器按照队列顺序，从由高优先级队列开始，依次调度整个队列组，循环往复，各队列调度机会均等。空CoS队列或者使能关闭的CoS队列不参与轮询，并释放占用带宽。对于报文长度是一样的流是公平的，但是如果报文长度不一样，就不公平了。
  - Weighted Round Robin（WRR） across CoS queues
    - 加权轮询优先级CoS队列调度，调度器执行RR的调度规则，周期性访问每个CoS队列。WRR为每一个CoS队列预先分配一个权值，同时为每个队列维护一个计数器，计数器初始化为默认权值。
    - 某个CoS队列被调度，待发分组长度小于等于计数器值，WRR允许发送；待发分组长度大于计数器值，WRR将当前计数器值和默认权值累加，作为新的计数器值，在下一次轮询时比较；WRR每调度一个单位报文，计数器减相应单位，直到计数器减为0，重新分配默认权值。权值越高，表明该队列获得调度的机会，包括成功的机会越大。同时，该权值又可以反映CoS队列占有带宽的百分比，从而保证最小优先级的CoS队列同样能够得到带宽，而实现成功调度。
  - Weighted Deficit Round Robin（WDRR）
    - 加权差额轮询优先级CoS队列调度。WDRR给每个待调度队列分配一个可配置的最大服务额度MTU_QUANTA，并为每个CoS队列维护一个信贷计数器Credit_Counter，初始化为MTU_QUANTA。如果一个包A从一个CoS队列成功发送，则从该队列的Credit_Counter中减去Len(A)作为新的信贷计数，即Credit_Counter=Credit_Counter-Len(A)；如果Credit_Counter < Len(A)，则该CoS队列此次轮询不能得到调度，须等待下一个轮询，WDRR将Credit_Counter和Len(A)的差值加给Credit_Counter，即Credit_Counter=Credit_Counter +（Len(A) - Credit_Counter），作为下一轮调度的新的信贷计数。

QoS(Schedule Algorithms)

- 四种CoS队列调度算法的选择

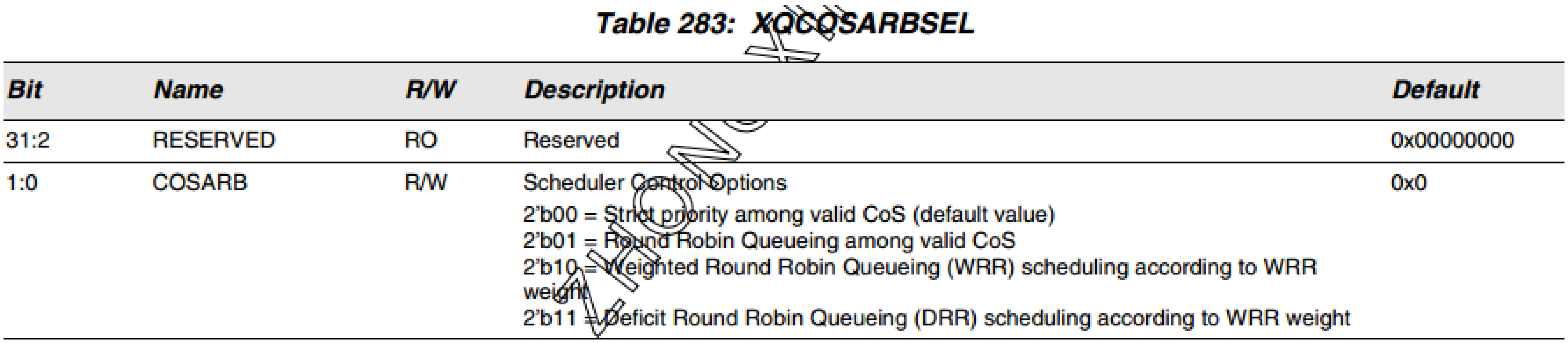


Table 283: XQCOSARBSEL

Bit	Name	R/W	Description	Default
31:2	RESERVED	RO	Reserved	0x00000000
1:0	COSARB	R/W	Scheduler Control Options 2'b00 = Strict priority among valid CoS (default value) 2'b01 = Round Robin Queueing among valid CoS 2'b10 = Weighted Round Robin Queueing (WRR) scheduling according to WRR weight 2'b11 = Deficit Round Robin Queueing (DRR) scheduling according to WRR weight	0x0

- Ip dhcp trust

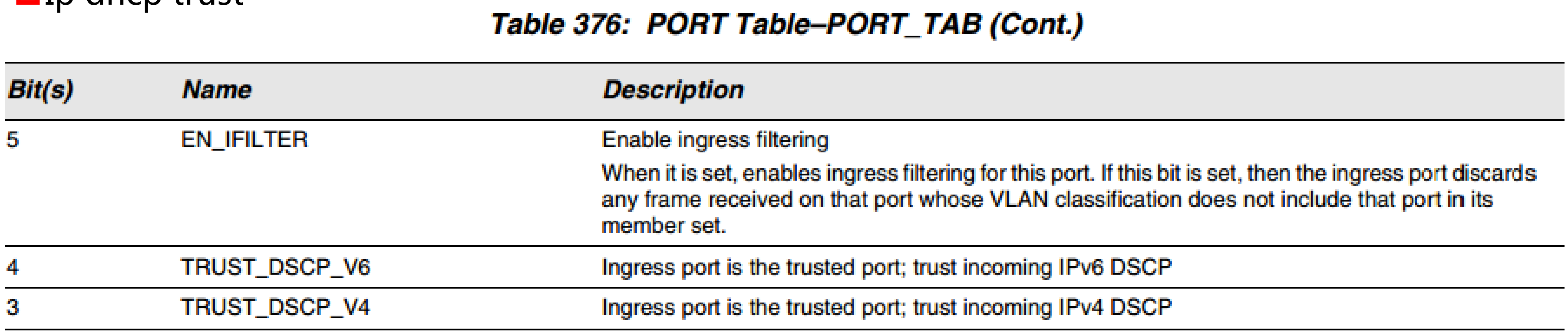


Table 376: PORT Table–PORT_TAB (Cont.)

Bit(s)	Name	Description
5	EN_IFILTER	Enable ingress filtering When it is set, enables ingress filtering for this port. If this bit is set, then the ingress port discards any frame received on that port whose VLAN classification does not include that port in its member set.
4	TRUST_DSCP_V6	Ingress port is the trusted port; trust incoming IPv6 DSCP
3	TRUST_DSCP_V4	Ingress port is the trusted port; trust incoming IPv4 DSCP

- 64个dscp优先级重映射表，由报文自带的IP头部信息里的DSCP值作为索引

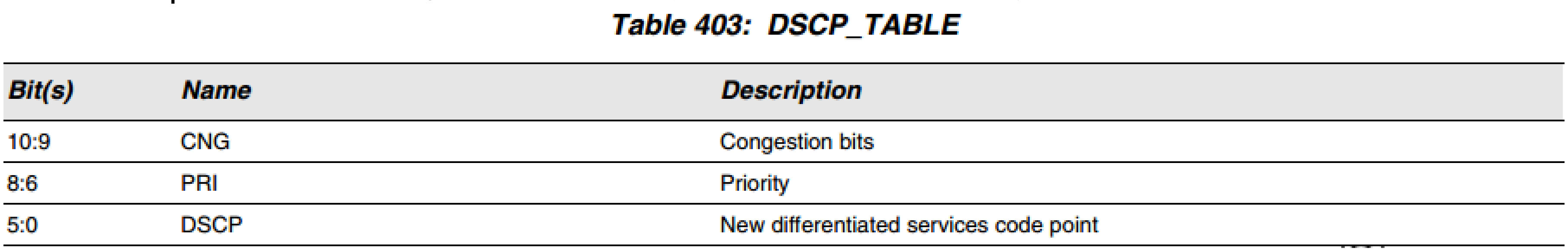


Table 403: DSCP_TABLE

Bit(s)	Name	Description
10:9	CNG	Congestion bits
8:6	PRI	Priority
5:0	DSCP	New differentiated services code point

QoS(Schedule Algorithms)

- WRR&WDRR配置

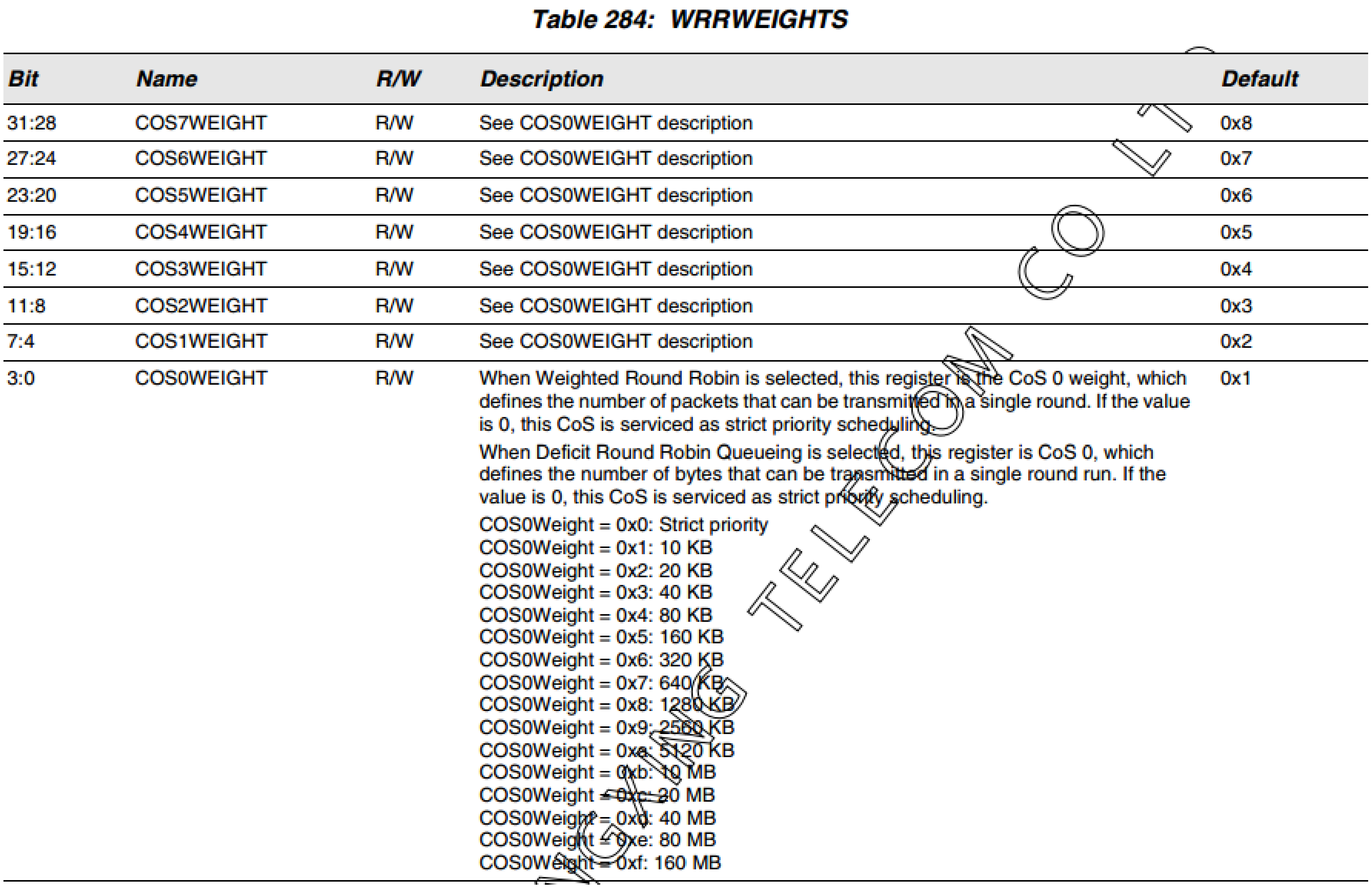


Table 284: WRRWEIGHTS

Bit	Name	R/W	Description	Default
31:28	COS7WEIGHT	R/W	See COS0WEIGHT description	0x8
27:24	COS6WEIGHT	R/W	See COS0WEIGHT description	0x7
23:20	COS5WEIGHT	R/W	See COS0WEIGHT description	0x6
19:16	COS4WEIGHT	R/W	See COS0WEIGHT description	0x5
15:12	COS3WEIGHT	R/W	See COS0WEIGHT description	0x4
11:8	COS2WEIGHT	R/W	See COS0WEIGHT description	0x3
7:4	COS1WEIGHT	R/W	See COS0WEIGHT description	0x2

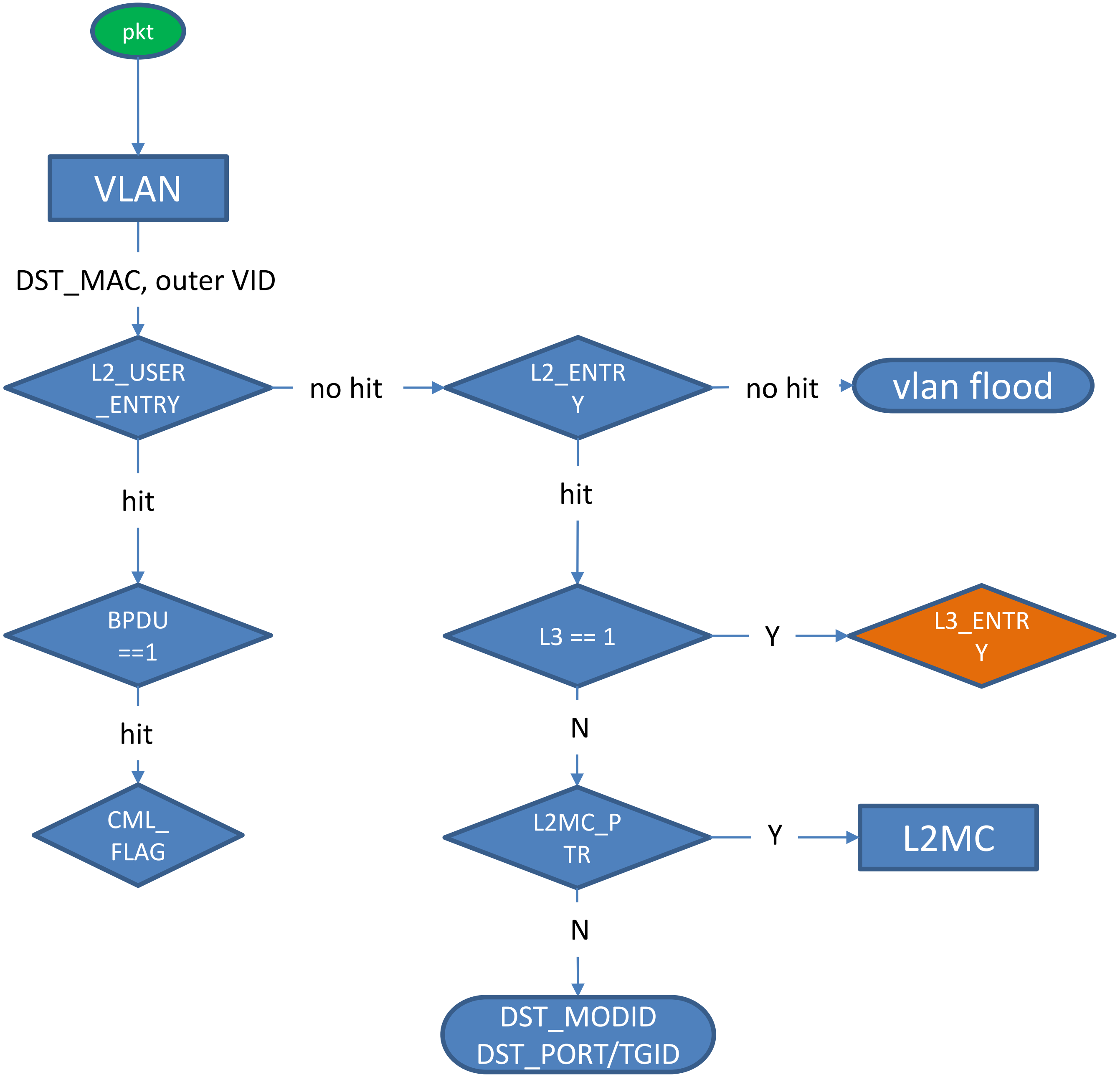
3:0	COS0WEIGHT	R/W	When Weighted Round Robin is selected, this register is the CoS 0 weight, which defines the number of packets that can be transmitted in a single round. If the value is 0, this CoS is serviced as strict priority scheduling.
			When Deficit Round Robin Queueing is selected, this register is CoS 0, which defines the number of bytes that can be transmitted in a single round run. If the value is 0, this CoS is serviced as strict priority scheduling.
			COS0Weight = 0x0: Strict priority
			COS0Weight = 0x1: 10 KB
			COS0Weight = 0x2: 20 KB
			COS0Weight = 0x3: 40 KB
			COS0Weight = 0x4: 80 KB
			COS0Weight = 0x5: 160 KB
			COS0Weight = 0x6: 320 KB
			COS0Weight = 0x7: 640 KB
			COS0Weight = 0x8: 1280 KB
			COS0Weight = 0x9: 2560 KB
			COS0Weight = 0xa: 5120 KB
			COS0Weight = 0xb: 10 MB
			COS0Weight = 0xc: 20 MB
			COS0Weight = 0xd: 40 MB
			COS0Weight = 0xe: 80 MB
			COS0Weight = 0xf: 160 MB
				0x1

L2

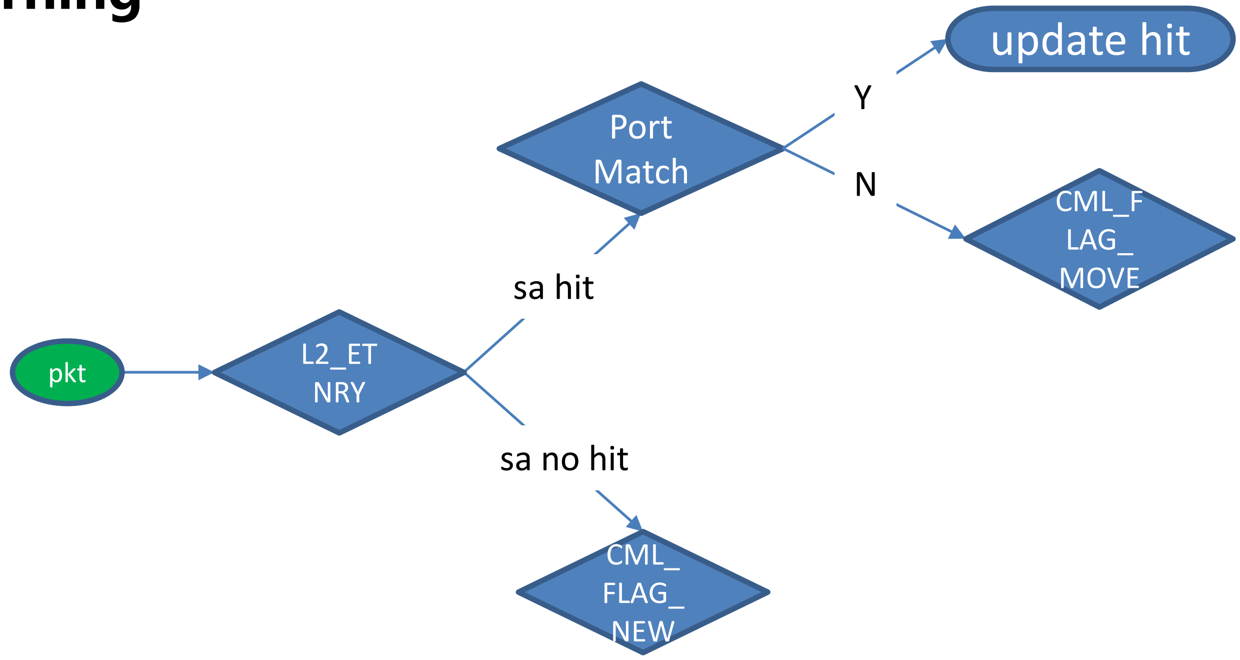
L2 Packet Flow

- 单播转发
  - ① VID分配：如果是untag报文，VLAN模块进行VID分配。如果是tag报文，根据端口的VLAN列表，判断是否丢弃。
  - ② 源MAC地址学习：根据源MAC地址与VID(IVL模式)或者MAC地址与FID(SVL模式)，查找L2_ENTRY，如果已经存在则更新hit位；如果已经存在但端口发生变化，则进行地址迁移更新；如果不存在，则进行地址的芯片自动学习、丢弃或者送CPU(CML决定)。
  - ③ 目的MAC地址查找：如果匹配L2_USER_ENTRY且BPDU=1，将报文在VLAN内泛洪、丢弃或者送CPU；如果BPDU=0，报文根据DST_MODID与DST_PORT进行转发。如果不匹配L2_USER_ENTRY，但是匹配L2_ENTRY，报文根据DST_MODID与DST_PORT/TGID进行转发；如果不匹配L2_ENTRY，报文在VLAN内泛洪。
- 多播转发：目的MAC地址查找如果匹配L2_ENTRY且l2mc_ptr指向L2MC，则转发到L2MC的指定端口。
- 广播转发：地址学习阶段后，向VLAN内泛洪。

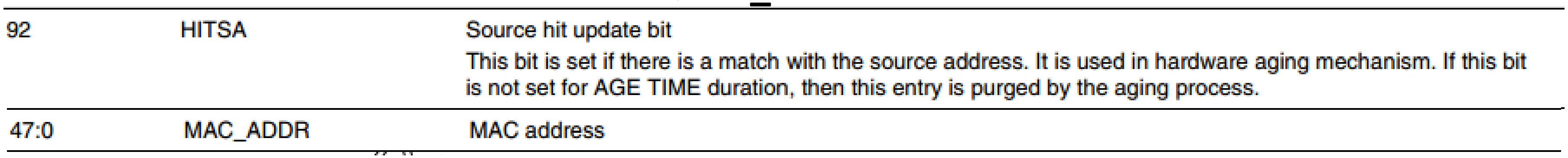
L2 Process


pkt
VLAN
DST_MAC, outer VID
L2_USER _ENTRY
no hit
L2_ENTR Y
no hit
vlan flood
hit
hit
BPDU ==1
L3 == 1
Y
L3_ENTR Y
hit
N
CML_ FLAG
L2MC_P TR
Y
L2MC
N
DST_MODID
DST_PORT/TGID

L2 Learning

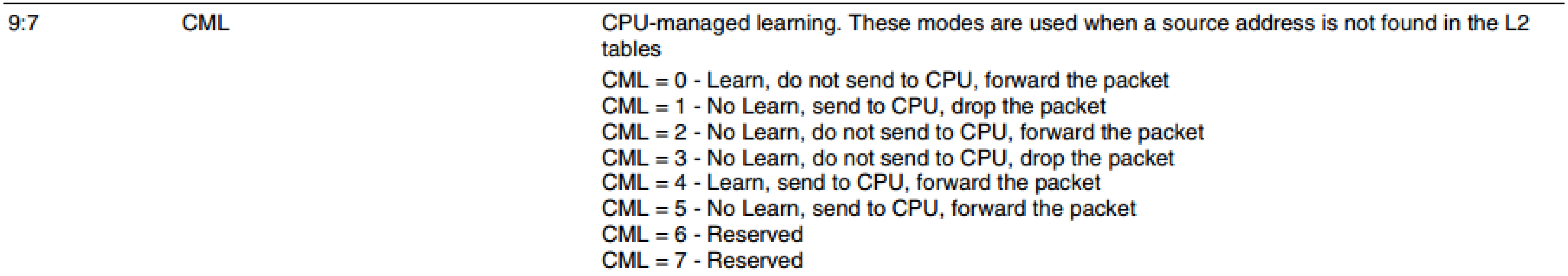


Table： L2_ENTRY



92	HITSA	Source hit update bit This bit is set if there is a match with the source address. It is used in hardware aging mechanism. If this bit is not set for AGE TIME duration, then this entry is purged by the aging process.
47:0	MAC_ADDR	MAC address

Table： PORT



9:7	CML	CPU-managed learning. These modes are used when a source address is not found in the L2 tables
		CML = 0 - Learn, do not send to CPU, forward the packet
		CML = 1 - No Learn, send to CPU, drop the packet
		CML = 2 - No Learn, do not send to CPU, forward the packet
		CML = 3 - No Learn, do not send to CPU, drop the packet
		CML = 4 - Learn, send to CPU, forward the packet
		CML = 5 - No Learn, send to CPU, forward the packet
		CML = 6 - Reserved
		CML = 7 - Reserved

L2 Aging

- 新的源MAC地址学习或者源MAC地址查找匹配时，会设置hit位。当AGE_VAL时间到期，则会清除hit位。当AGE_VAL时间再次到期时，如果hit位已经处于清除状态，则把表项删除。

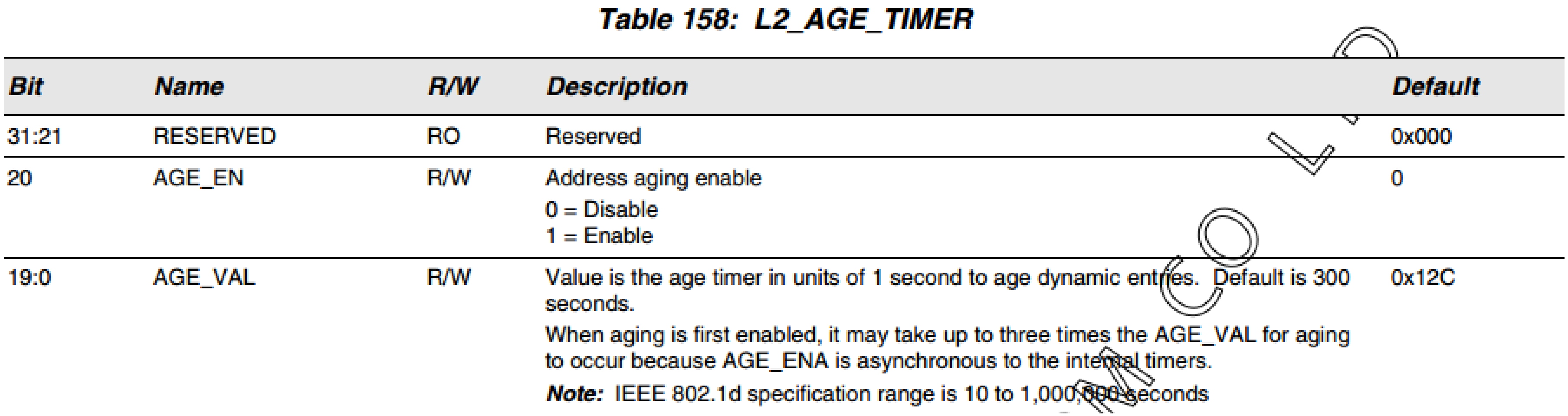


Table 158: L2_AGE_TIMER

Bit	Name	R/W	Description	Default
31:21	RESERVED	RO	Reserved	0x000
20	AGE_EN	R/W	Address aging enable 0 = Disable 1 = Enable	0
19:0	AGE_VAL	R/W	Value is the age timer in units of 1 second to age dynamic entries. Default is 300 seconds. When aging is first enabled, it may take up to three times the AGE_VAL for aging to occur because AGE_ENA is asynchronous to the internal timers. ***Note:*** IEEE 802.1d specification range is 10 to 1,000,000 seconds	0x12C

L2 Multicast

- PFM(Port Filtering Mode)
  - Mode A：所有的多播报文在VLAN内泛洪
  - Mode B：已知(目的组播MAC地址表项存在)多播报文在多播组成员端口内泛洪，未知多播报文在VLAN内泛洪
  - Mode C：已知多播报文在多播组成员端口内泛洪，未知多播报文被丢弃

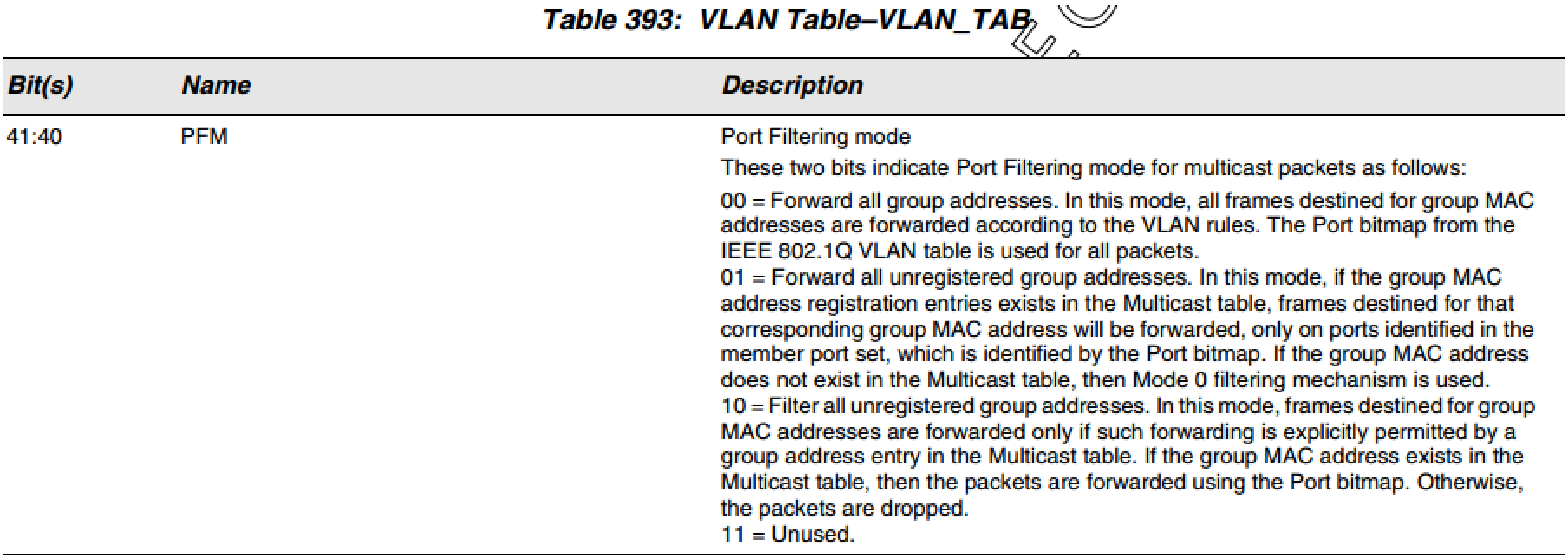


Table 393: VLAN Table–VLAN_TAB

Bit(s)	Name	Description
41:40	PFM	Port Filtering mode These two bits indicate Port Filtering mode for multicast packets as follows: 00 = Forward all group addresses. In this mode, all frames destined for group MAC addresses are forwarded according to the VLAN rules. The Port bitmap from the IEEE 802.1Q VLAN table is used for all packets. 01 = Forward all unregistered group addresses. In this mode, if the group MAC address registration entries exists in the Multicast table, frames destined for that corresponding group MAC address will be forwarded, only on ports identified in the member port set, which is identified by the Port bitmap. If the group MAC address does not exist in the Multicast table, then Mode 0 filtering mechanism is used. 10 = Filter all unregistered group addresses. In this mode, frames destined for group MAC addresses are forwarded only if such forwarding is explicitly permitted by a group address entry in the Multicast table. If the group MAC address exists in the Multicast table, then the packets are forwarded using the Port bitmap. Otherwise, the packets are dropped. 11 = Unused.

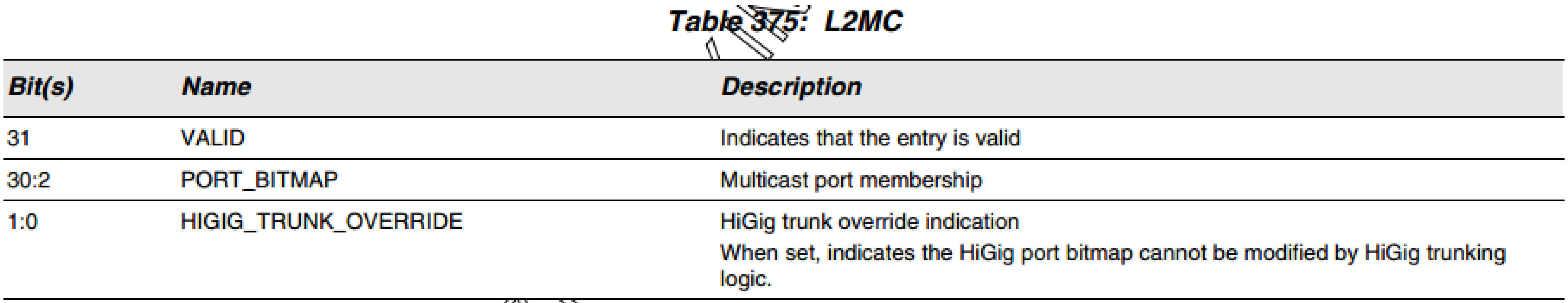


Table 375: L2MC

Bit(s)	Name	Description
31	VALID	Indicates that the entry is valid
30:2	PORT_BITMAP	Multicast port membership
1:0	HIGIG_TRUNK_OVERRIDE	HiGig trunk override indication When set, indicates the HiGig port bitmap cannot be modified by HiGig trunking logic.

L2 Port Bridge

- 通常情况下，从一个端口进来的报文，不会再从该端口出去(即使是广播泛洪)。但是有些场景下(如WAP)，我们需要这样的功能，通过设置PORT_BRIDGE，可以达到这个目的。

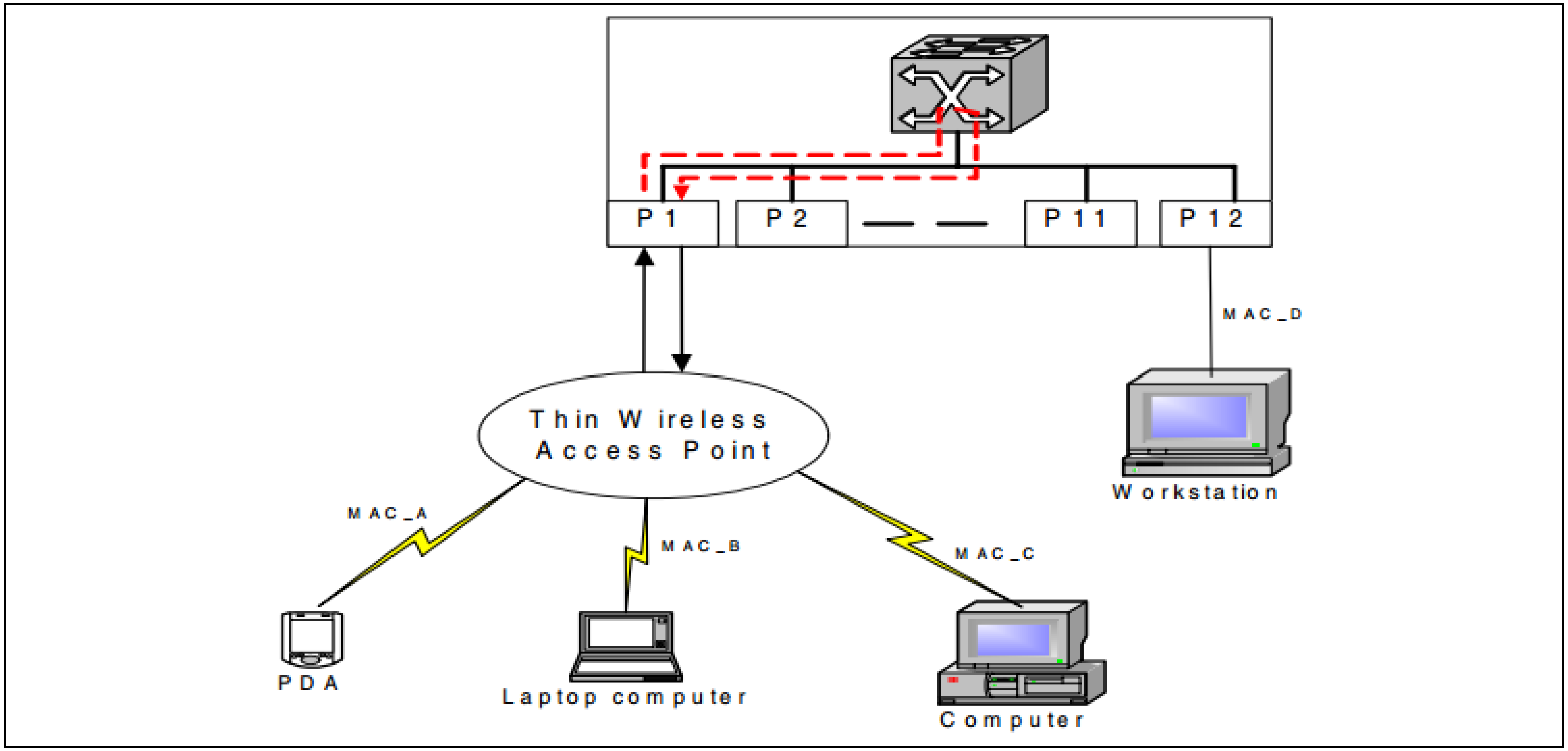


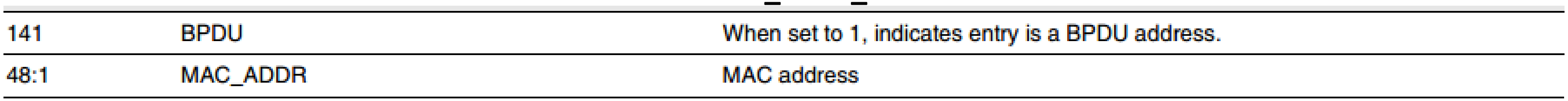
Table 376: PORT Table–PORT_TAB

Bit(s)	*Name*	*Description*
67:66	RESERVED	Reserved
65	ALLOW_SRC_MOD	Allow packets with MH.SRC_MODID is same as MY_MODID
64	IGNORE_IPMC_L3_BITMAP	Set this bit to disable L3 routing of IPMC packets on this port.
63	IGNORE_IPMC_L2_BITMAP	Set this bit to disable L2 bridging of IPMC packets on this port.
62	PORT_BRIDGE	When set to 1, the port supports L2 port bridging. The ability for a packet to be Layer 2 switched where the source destination ports are the same.

Spanning Tree

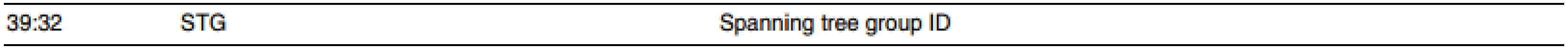
- STP使用组播地址0x0180c2000000，通过设置L2_USER_ENTRY.MAC_ADDR=0x0180c2000000，L2_USER_ENTRY.BPDU=1且PORT.CML=2即可实现STP报文送CPU。
- VLAN(STG)用来索引VLAN_STG table，VLAN_STG table共有512个entry，每个entry都注明了所有端口的STP状态。因此MSTP最大可以支持512个实例。
  - 0=Disabled：所有报文被丢弃(包括BPDU控制报文)
  - 1=Blocking：所有BPDU控制报文被送往CPU处理，其它报文被丢弃(关闭地址学习)
  - 2=Learning：所有BPDU控制报文被送往CPU处理，其它报文被丢弃(开启地址学习)
  - 3=Forwarding：所有BPDU控制报文被送往CPU处理，其它报文按正常转发

Table：L2_USER_ENTRY



141	BPDU	When set to 1, indicates entry is a BPDU address.
48:1	MAC_ADDR	MAC address

Table： VLAN



39:32	STG	Spanning tree group ID

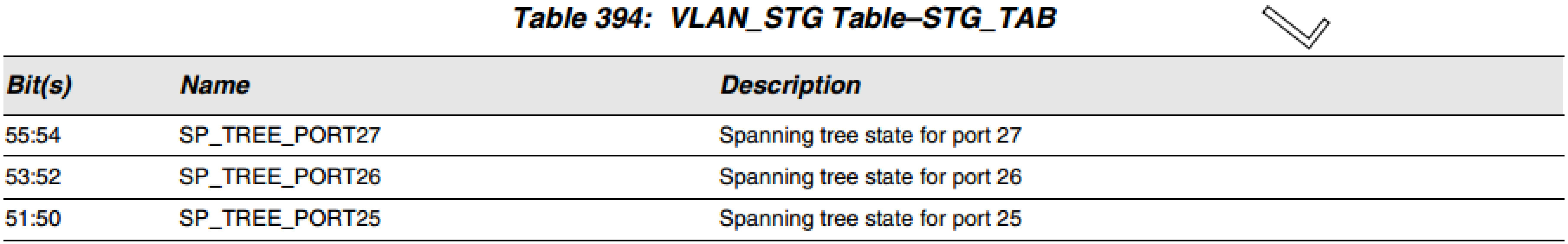


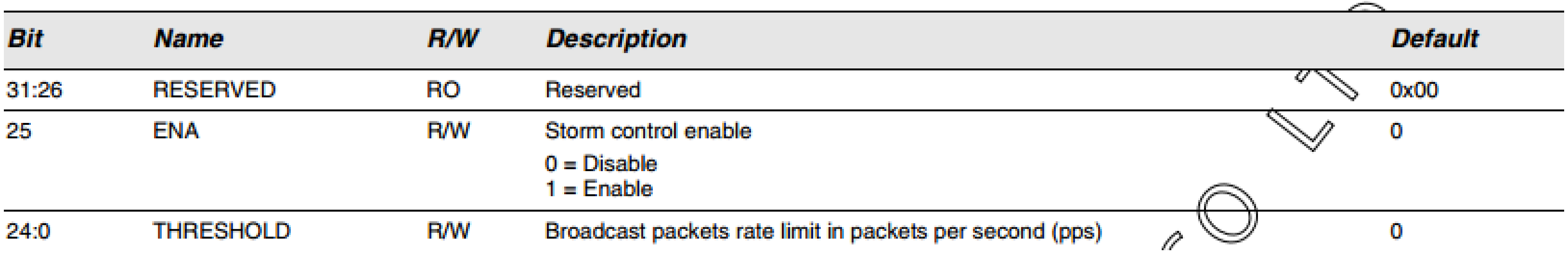
Table 394: VLAN_STG Table–STG_TAB

Bit(s)	***Name***	***Description***
55:54	SP_TREE_PORT27	Spanning tree state for port 27
53:52	SP_TREE_PORT26	Spanning tree state for port 26
51:50	SP_TREE_PORT25	Spanning tree state for port 25

Storm Control

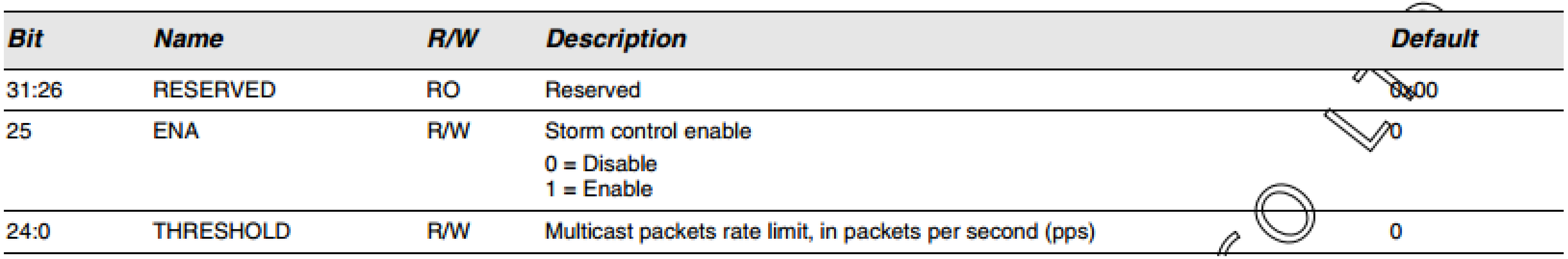
为了防止广播、组播以及未知名单播报文泛洪对网络所造成的影响，可以通过设置这些报文每秒允许的处理阈值，丢弃超过阈值的报文来保证网络的安全

Table 177: BCAST_STORM_CONTROL



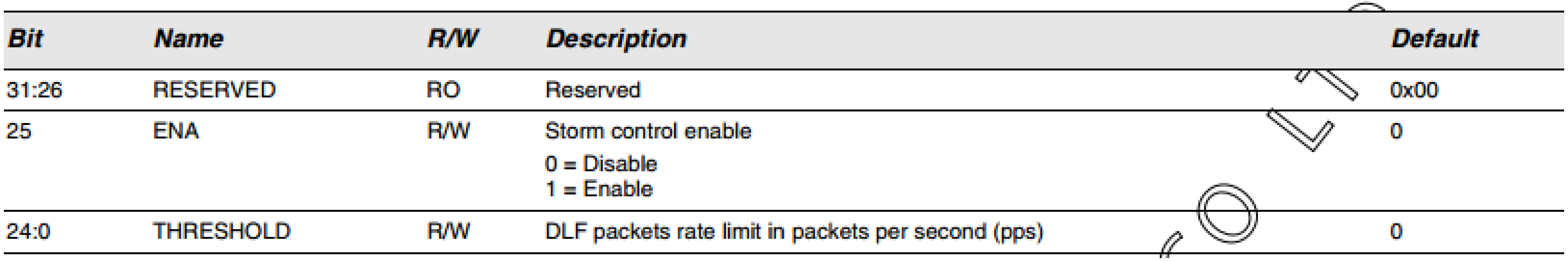
Bit	Name	R/W	Description	Default
31:26	RESERVED	RO	Reserved	0x00
25	ENA	R/W	Storm control enable 0 = Disable 1 = Enable	0
24:0	THRESHOLD	R/W	Broadcast packets rate limit in packets per second (pps)	0

Table 178: MCAST_STORM_CONTROL



Bit	Name	R/W	Description	Default
31:26	RESERVED	RO	Reserved	0x00
25	ENA	R/W	Storm control enable 0 = Disable 1 = Enable	0
24:0	THRESHOLD	R/W	Multicast packets rate limit, in packets per second (pps)	0

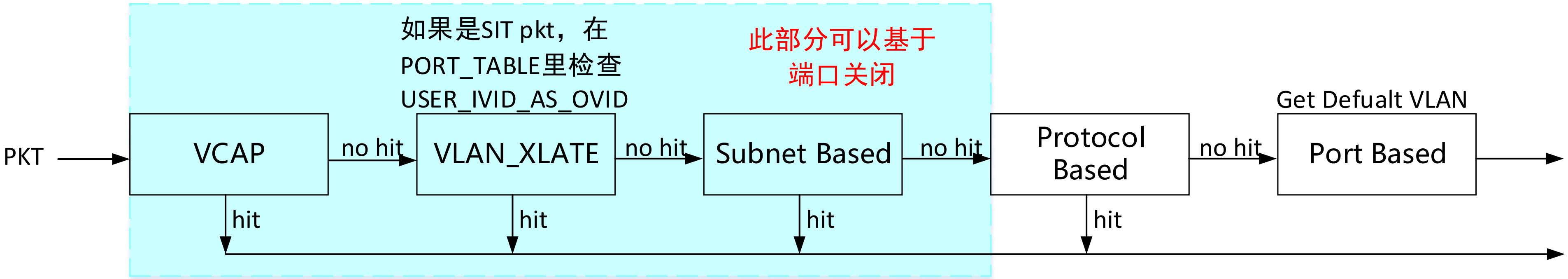
Table 179: DLF_STORM_CONTROL



Bit	Name	R/W	Description	Default
31:26	RESERVED	RO	Reserved	0x00
25	ENA	R/W	Storm control enable 0 = Disable 1 = Enable	0
24:0	THRESHOLD	R/W	DLF packets rate limit in packets per second (pps)	0

VLAN

VLAN Assign

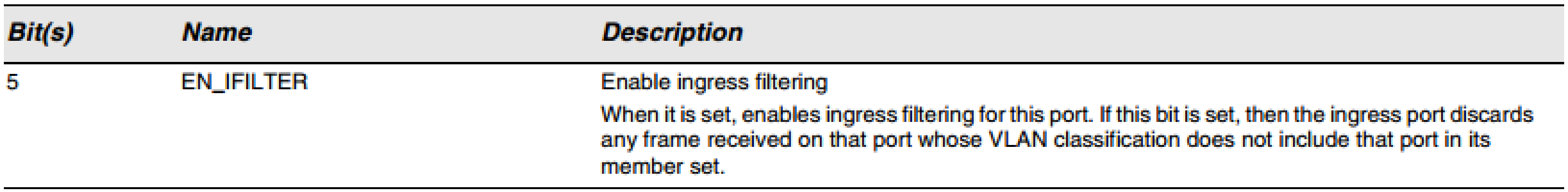


如果是SIT pkt，在PORT_TABLE里检查USER_IVID_AS_OVID

此部分可以基于端口关闭

PORT.EN_IFILTER=1，表示如果入口报文的VLAN不属于端口所分配的VLAN范围，设备将会丢弃该报文。

Table 376: PORT Table–PORT_TAB (Cont.)



Bit(s)	Name	Description
5	EN_IFILTER	Enable ingress filtering When it is set, enables ingress filtering for this port. If this bit is set, then the ingress port discards any frame received on that port whose VLAN classification does not include that port in its member set.

Flow-Based

通过VLAN Content Aware Processor（VCAP），基于报文内容的筛选条件进行VLAN ID（VID）的分配。VCAP是一个内容可识别的硬件引擎模块，分4个Slice，每个Slice有512个entry，即总共支持2K个entry。VCAP可以基于Port，Group ID，L2，L3，L4的内容进行IVID与OVID的添加、删除、替换等操作。
支持Single wide mode（512 rules x 210-bits wide），Double wide mode（256 rules x 420-bits wide）以及Quad wide mode（Double Slice）（256rules x 840-bits wide）。

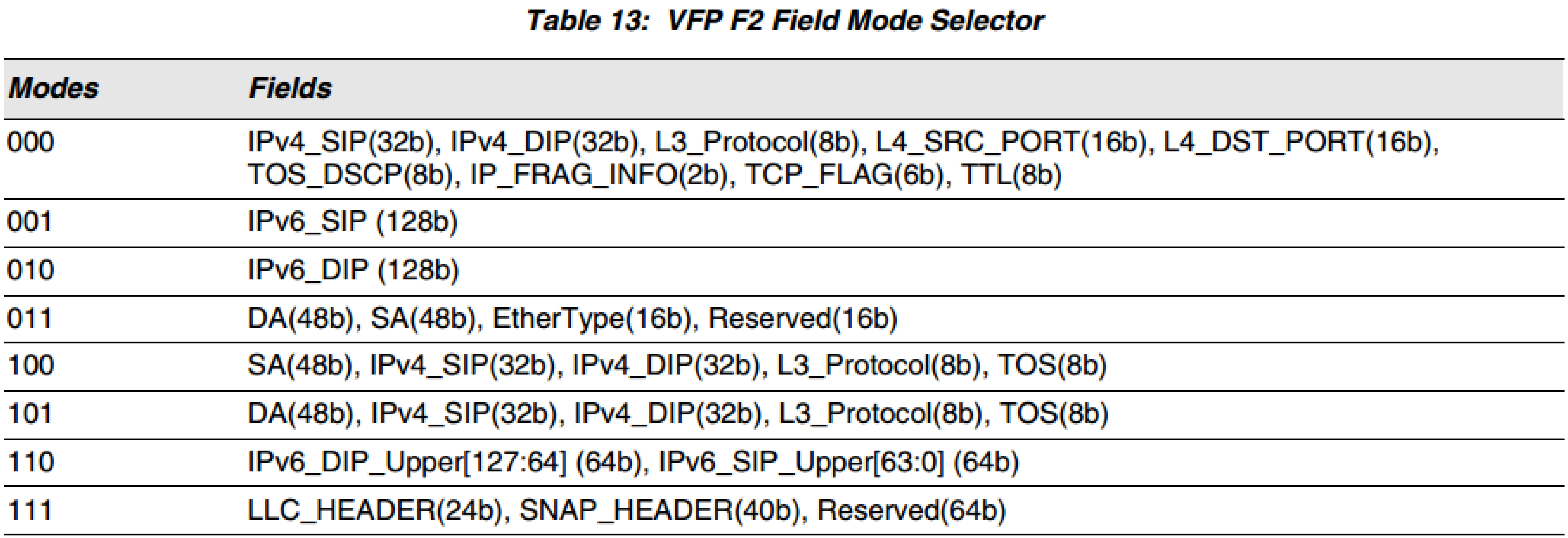


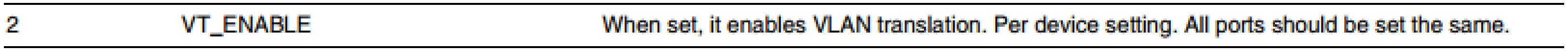
Table 13: VFP F2 Field Mode Selector

Modes	Fields
000	IPv4_SIP(32b), IPv4_DIP(32b), L3_Protocol(8b), L4_SRC_PORT(16b), L4_DST_PORT(16b), TOS_DSCP(8b), IP_FRAG_INFO(2b), TCP_FLAG(6b), TTL(8b)
001	IPv6_SIP (128b)
010	IPv6_DIP (128b)
011	DA(48b), SA(48b), EtherType(16b), Reserved(16b)
100	SA(48b), IPv4_SIP(32b), IPv4_DIP(32b), L3_Protocol(8b), TOS(8b)
101	DA(48b), IPv4_SIP(32b), IPv4_DIP(32b), L3_Protocol(8b), TOS(8b)
110	IPv6_DIP_Upper[127:64] (64b), IPv6_SIP_Upper[63:0] (64b)
111	LLC_HEADER(24b), SNAP_HEADER(40b), Reserved(64b)

VLAN Translation

- 通过VLAN Content Aware Processor（VCAP），基于报文内容的筛选条件进行VLAN ID（VID）的分配。VCAP是一个内容可识别的硬件引擎模块，分4个Slice，每个Slice有512个entry，即总共支持2K个entry。VCAP可以基于Port，Group ID，L2，L3，L4的内容进行IVID与OVID的添加、删除、替换等操作。
- 支持Single wide mode（512 rules x 210-bits wide），Double wide mode（256 rules x 420-bits wide）以及Quad wide mode（Double Slice）（256rules x 840-bits wide）。

PORT Table



2	VT_ENABLE	When set, it enables VLAN translation. Per device setting. All ports should be set the same.

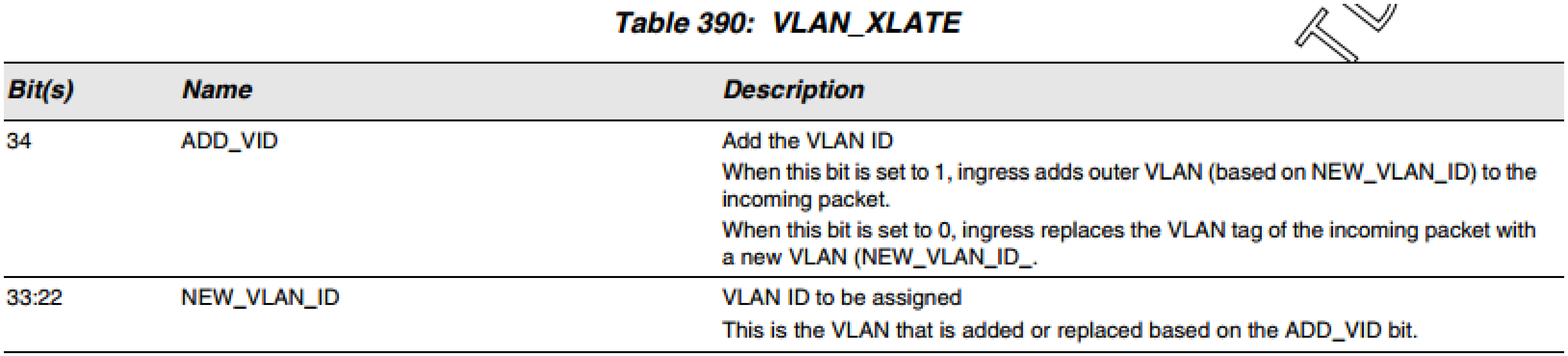


Table 390: VLAN_XLATE

Bit(s)	Name	Description
34	ADD_VID	Add the VLAN ID When this bit is set to 1, ingress adds outer VLAN (based on NEW_VLAN_ID) to the incoming packet. When this bit is set to 0, ingress replaces the VLAN tag of the incoming packet with a new VLAN (NEW_VLAN_ID_.
33:22	NEW_VLAN_ID	VLAN ID to be assigned This is the VLAN that is added or replaced based on the ADD_VID bit.

IP Subnet-Based VLAN

- VLAN_SUBNET table包含IP address和subnet mask域。通过mask可以设置灵活的IP匹配模式。
- 芯片的有255个entry的VLAN_SUBNET table及VLAN_SUBNET_DATA table。VLAN_SUBNET table一旦匹配命中，则其entry index作为VLAN_SUBNET_DATA的entry index，从VLAN_SUBNET_DATA中获取VID及PRI的分配值。

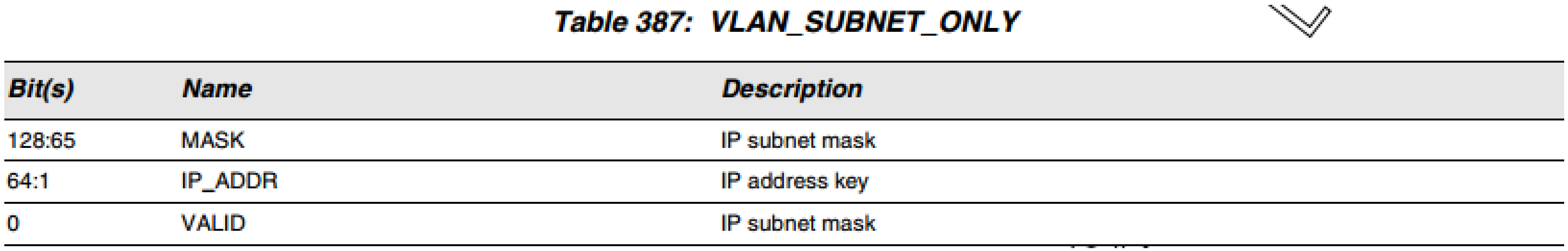
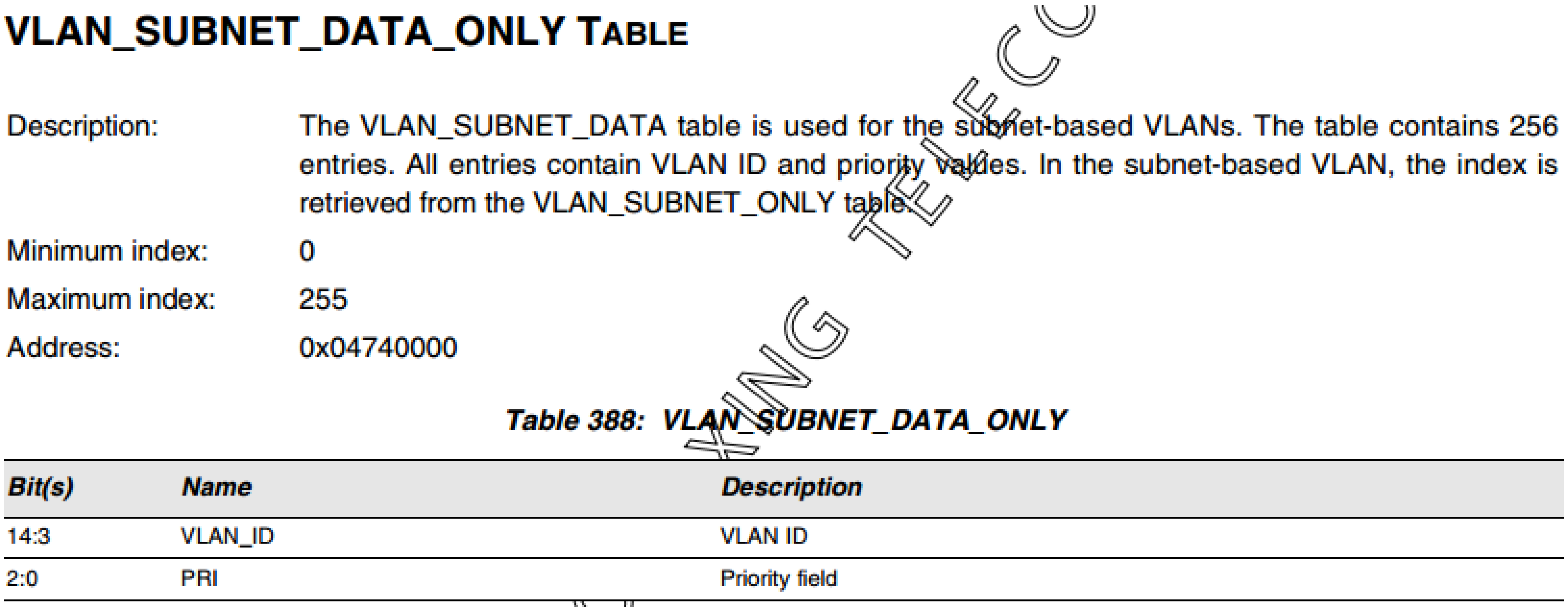


Table 387: VLAN_SUBNET_ONLY

Bit(s)	Name	Description
128:65	MASK	IP subnet mask
64:1	IP_ADDR	IP address key
0	VALID	IP subnet mask



VLAN_SUBNET_DATA_ONLY TABLE

Description: The VLAN_SUBNET_DATA table is used for the subnet-based VLANs. The table contains 256 entries. All entries contain VLAN ID and priority values. In the subnet-based VLAN, the index is retrieved from the VLAN_SUBNET_ONLY table.

Minimum index: 0

Maximum index: 255

Address: 0x04740000

Table 388: VLAN_SUBNET_DATA_ONLY

Bit(s)	Name	Description
14:3	VLAN_ID	VLAN ID
2:0	PRI	Priority field

Protocol-Based VLAN

- 基于报文以太网协议（FrameType，EtherType）类型进行VID的分配。通过对VLAN_PROTOCOL与VLAN_PROTOCOL_DATA tables的两步查找操作，如果匹配条件命中，则返回TAG_ACTION_PROFILE_PTR域，完成VID的分配。VLAN_PROTOCOL匹配命中后，
- 匹配域模式如下：
  - VLANFrameType（3 bits）. Ethernet II, IEEE 802.3 SNAP, or IEEE 802.3 LLC
  - VLANEtherType（16bits）. IPv4, IPX-RAW, IPX-LLC, and so on
  - VLANMatchUpper. Match the upper 8-bits of the 16-bit EtherType field
  - VLANMatchLower. Match the lower 8-bits of the 16-bit EtherType file
- 芯片的VLAN_PROTOCOL table一共有16个entry，每个端口有16个VLAN_PROTOCOL_DATA entry（共464个entry）。VLAN_PROTOCOL table的每个entry只设置一种匹配域类型，一旦匹配命中，则其entry index作为VLAN_PROTOCOL_DATA的entry index，从VLAN_PROTOCOL_DATA中获取VID及PRI的分配值。

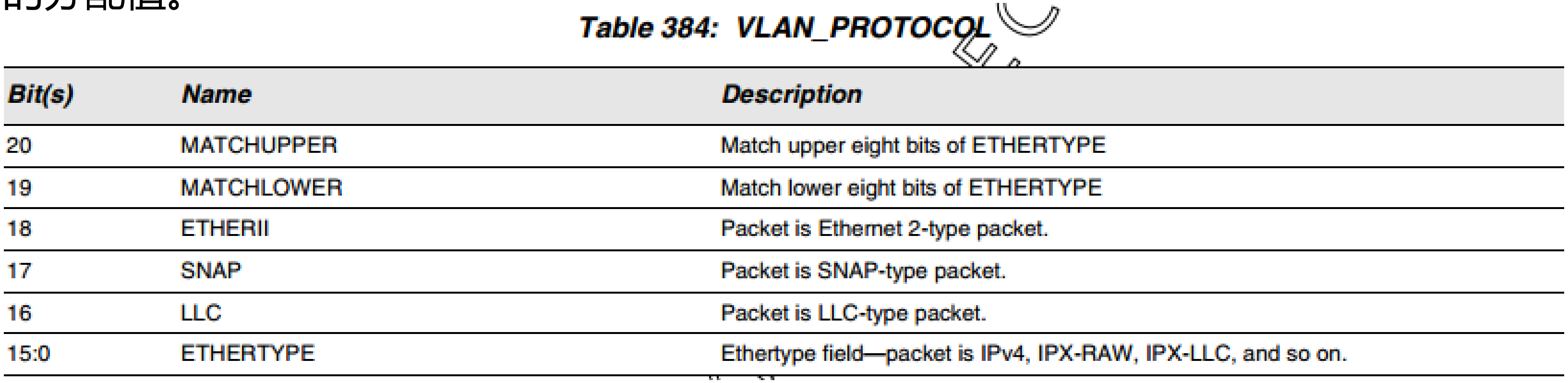


Table 384: VLAN_PROTOCOL

Bit(s)	Name	Description
20	MATCHUPPER	Match upper eight bits of ETHERTYPE
19	MATCHLOWER	Match lower eight bits of ETHERTYPE
18	ETHERII	Packet is Ethernet 2-type packet.
17	SNAP	Packet is SNAP-type packet.
16	LLC	Packet is LLC-type packet.
15:0	ETHERTYPE	Ethertype field—packet is IPv4, IPX-RAW, IPX-LLC, and so on.

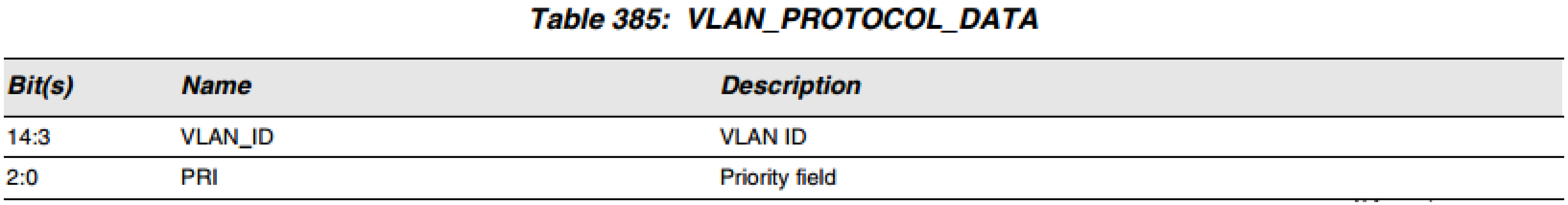


Table 385: VLAN_PROTOCOL_DATA

Bit(s)	Name	Description
14:3	VLAN_ID	VLAN ID
2:0	PRI	Priority field

Port-Based

- 默认根据PORT.PORT_VID进行VID的分配

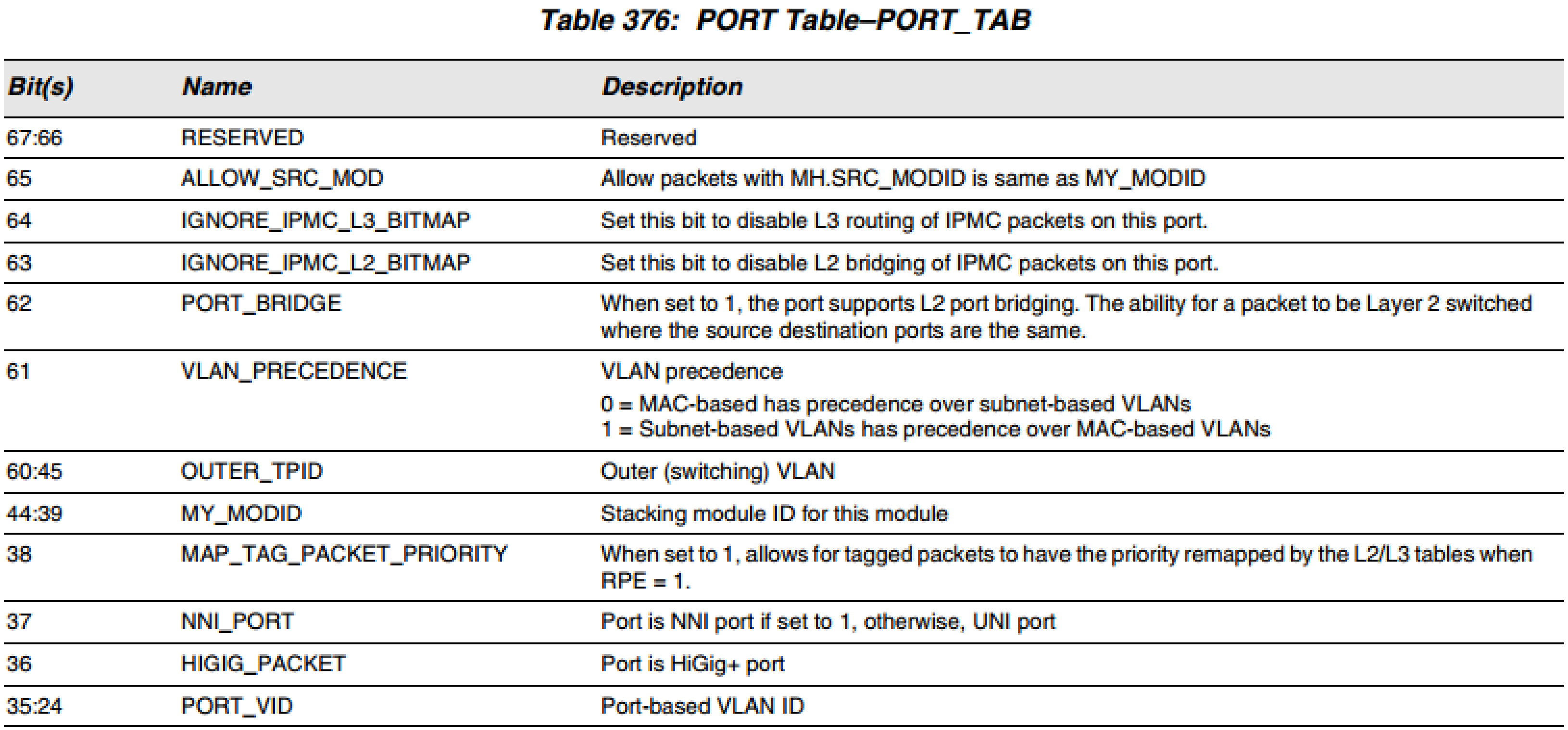


Table 376: PORT Table–PORT_TAB

Bit(s)	*Name*	*Description*
67:66	RESERVED	Reserved
65	ALLOW_SRC_MOD	Allow packets with MH.SRC_MODID is same as MY_MODID
64	IGNORE_IPMC_L3_BITMAP	Set this bit to disable L3 routing of IPMC packets on this port.
63	IGNORE_IPMC_L2_BITMAP	Set this bit to disable L2 bridging of IPMC packets on this port.
62	PORT_BRIDGE	When set to 1, the port supports L2 port bridging. The ability for a packet to be Layer 2 switched where the source destination ports are the same.
61	VLAN_PRECEDENCE	VLAN precedence 0 = MAC-based has precedence over subnet-based VLANs 1 = Subnet-based VLANs has precedence over MAC-based VLANs
60:45	OUTER_TPID	Outer (switching) VLAN
44:39	MY_MODID	Stacking module ID for this module
38	MAP_TAG_PACKET_PRIORITY	When set to 1, allows for tagged packets to have the priority remapped by the L2/L3 tables when RPE = 1.
37	NNI_PORT	Port is NNI port if set to 1, otherwise, UNI port
36	HIGIG_PACKET	Port is HiGig+ port
35:24	PORT_VID	Port-based VLAN ID

VLAN Ingress

- 入口报文的VLAN检查

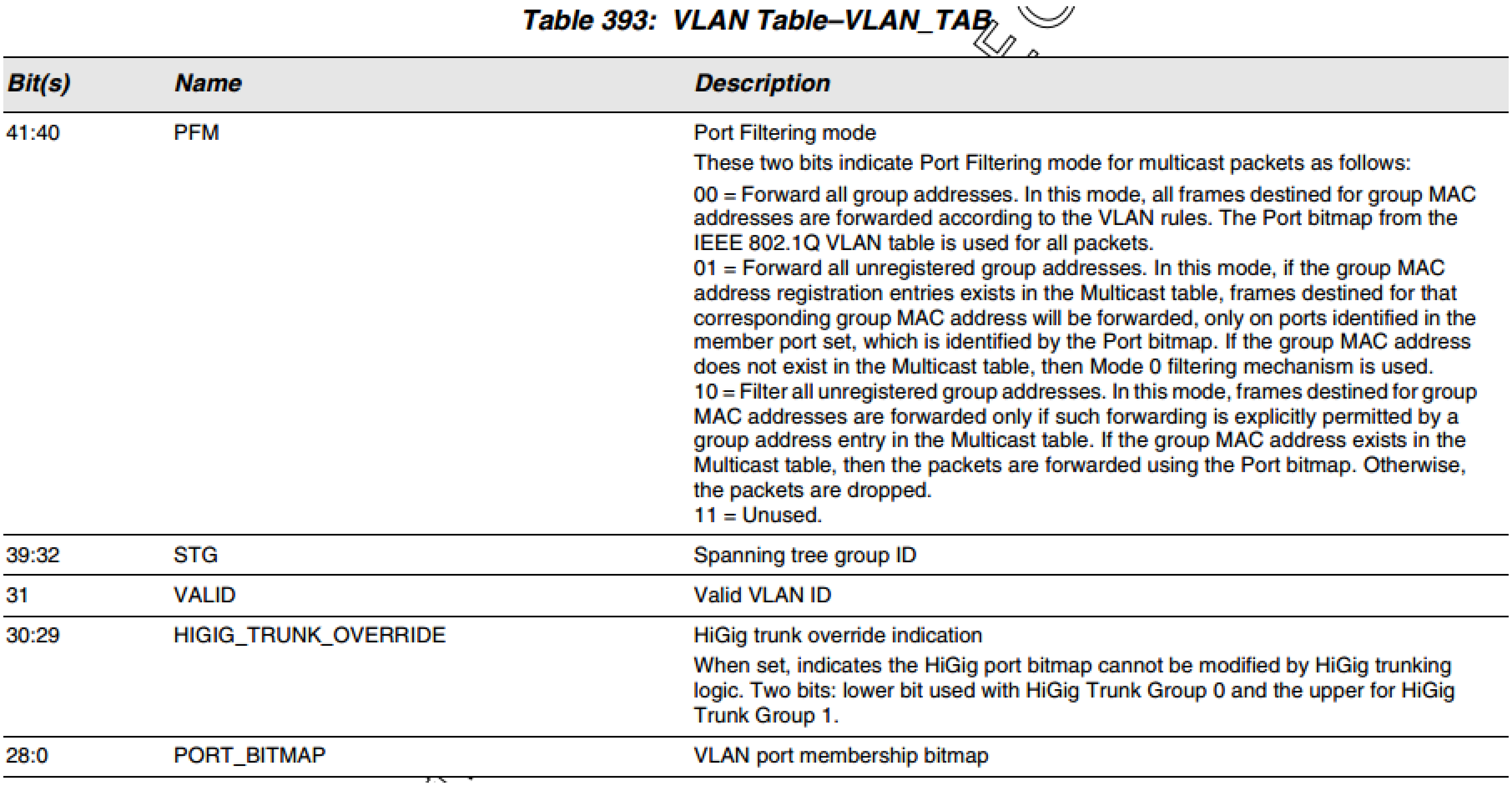


Table 393: VLAN Table–VLAN_TAB

Bit(s)	Name	Description
41:40	PFM	Port Filtering mode These two bits indicate Port Filtering mode for multicast packets as follows: 00 = Forward all group addresses. In this mode, all frames destined for group MAC addresses are forwarded according to the VLAN rules. The Port bitmap from the IEEE 802.1Q VLAN table is used for all packets. 01 = Forward all unregistered group addresses. In this mode, if the group MAC address registration entries exists in the Multicast table, frames destined for that corresponding group MAC address will be forwarded, only on ports identified in the member port set, which is identified by the Port bitmap. If the group MAC address does not exist in the Multicast table, then Mode 0 filtering mechanism is used. 10 = Filter all unregistered group addresses. In this mode, frames destined for group MAC addresses are forwarded only if such forwarding is explicitly permitted by a group address entry in the Multicast table. If the group MAC address exists in the Multicast table, then the packets are forwarded using the Port bitmap. Otherwise, the packets are dropped. 11 = Unused.
39:32	STG	Spanning tree group ID
31	VALID	Valid VLAN ID
30:29	HIGIG_TRUNK_OVERRIDE	HiGig trunk override indication When set, indicates the HiGig port bitmap cannot be modified by HiGig trunking logic. Two bits: lower bit used with HiGig Trunk Group 0 and the upper for HiGig Trunk Group 1.
28:0	PORT_BITMAP	VLAN port membership bitmap

VLAN Egress

- 决定报文出去时的VLAN状态

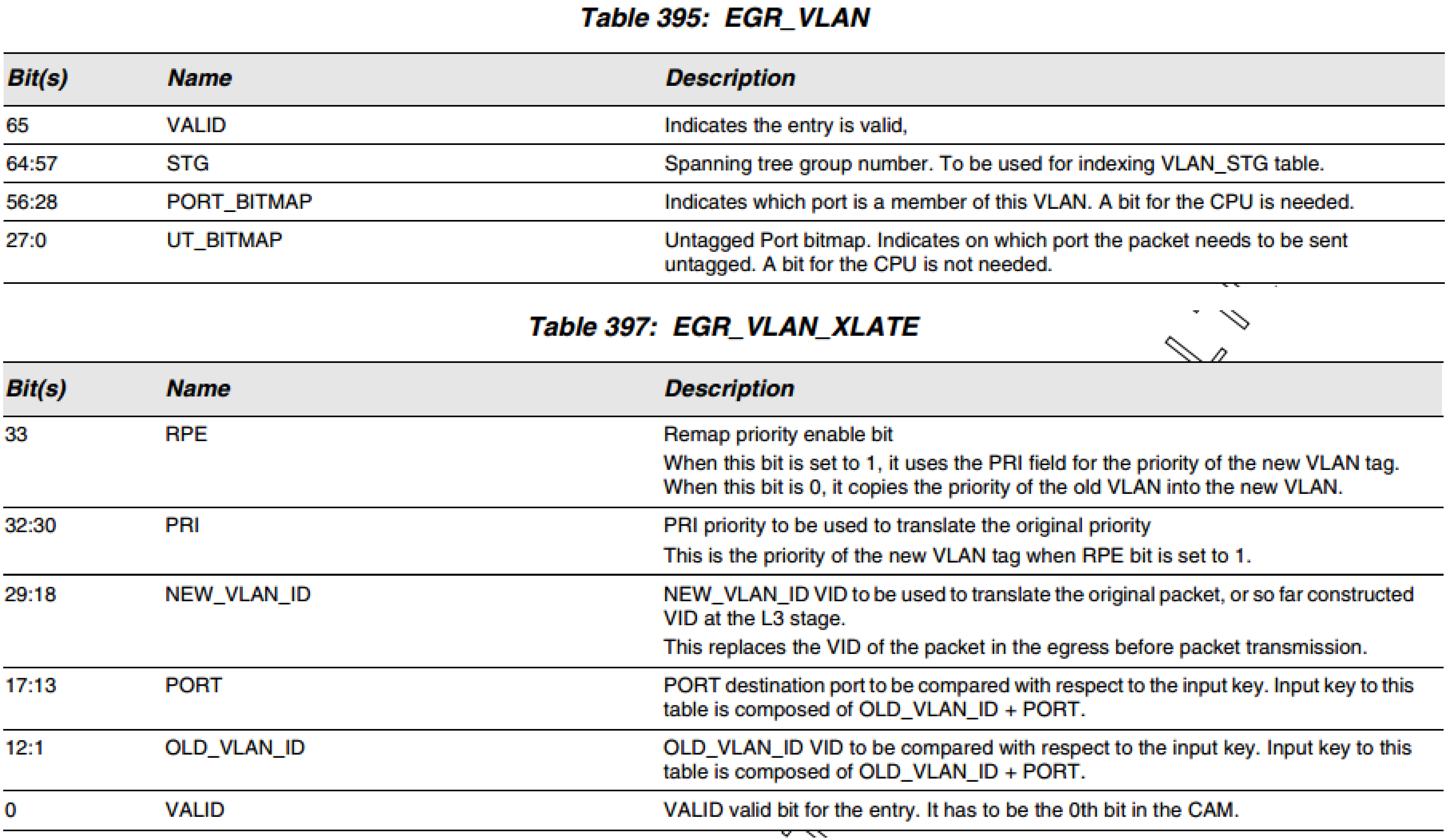


Table 395: EGR_VLAN

Bit(s)	Name	Description
65	VALID	Indicates the entry is valid,
64:57	STG	Spanning tree group number. To be used for indexing VLAN_STG table.
56:28	PORT_BITMAP	Indicates which port is a member of this VLAN. A bit for the CPU is needed.
27:0	UT_BITMAP	Untagged Port bitmap. Indicates on which port the packet needs to be sent untagged. A bit for the CPU is not needed.

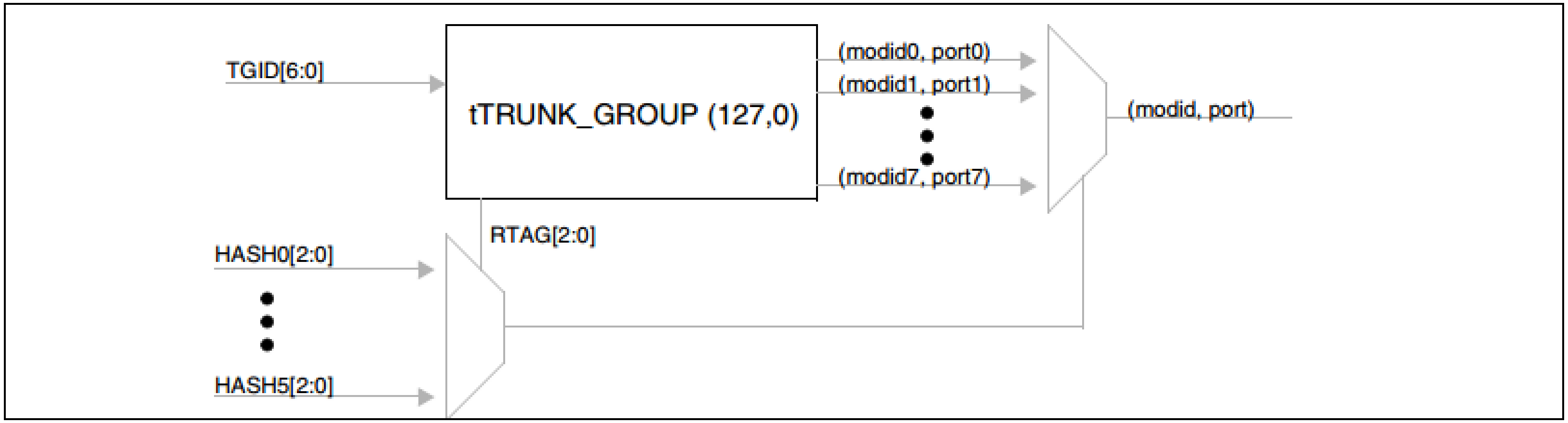
Table 397: EGR_VLAN_XLATE

Bit(s)	Name	Description
33	RPE	Remap priority enable bit When this bit is set to 1, it uses the PRI field for the priority of the new VLAN tag. When this bit is 0, it copies the priority of the old VLAN into the new VLAN.
32:30	PRI	PRI priority to be used to translate the original priority This is the priority of the new VLAN tag when RPE bit is set to 1.
29:18	NEW_VLAN_ID	NEW_VLAN_ID VID to be used to translate the original packet, or so far constructed VID at the L3 stage. This replaces the VID of the packet in the egress before packet transmission.
17:13	PORT	PORT destination port to be compared with respect to the input key. Input key to this table is composed of OLD_VLAN_ID + PORT.
12:1	OLD_VLAN_ID	OLD_VLAN_ID VID to be compared with respect to the input key. Input key to this table is composed of OLD_VLAN_ID + PORT.
0	VALID	VALID valid bit for the entry. It has to be the 0th bit in the CAM.

Link Aggregation

Link Aggregation

RTAG可选择以报文的哪个域作为hash算法的key，以最终决定出口的modid与port



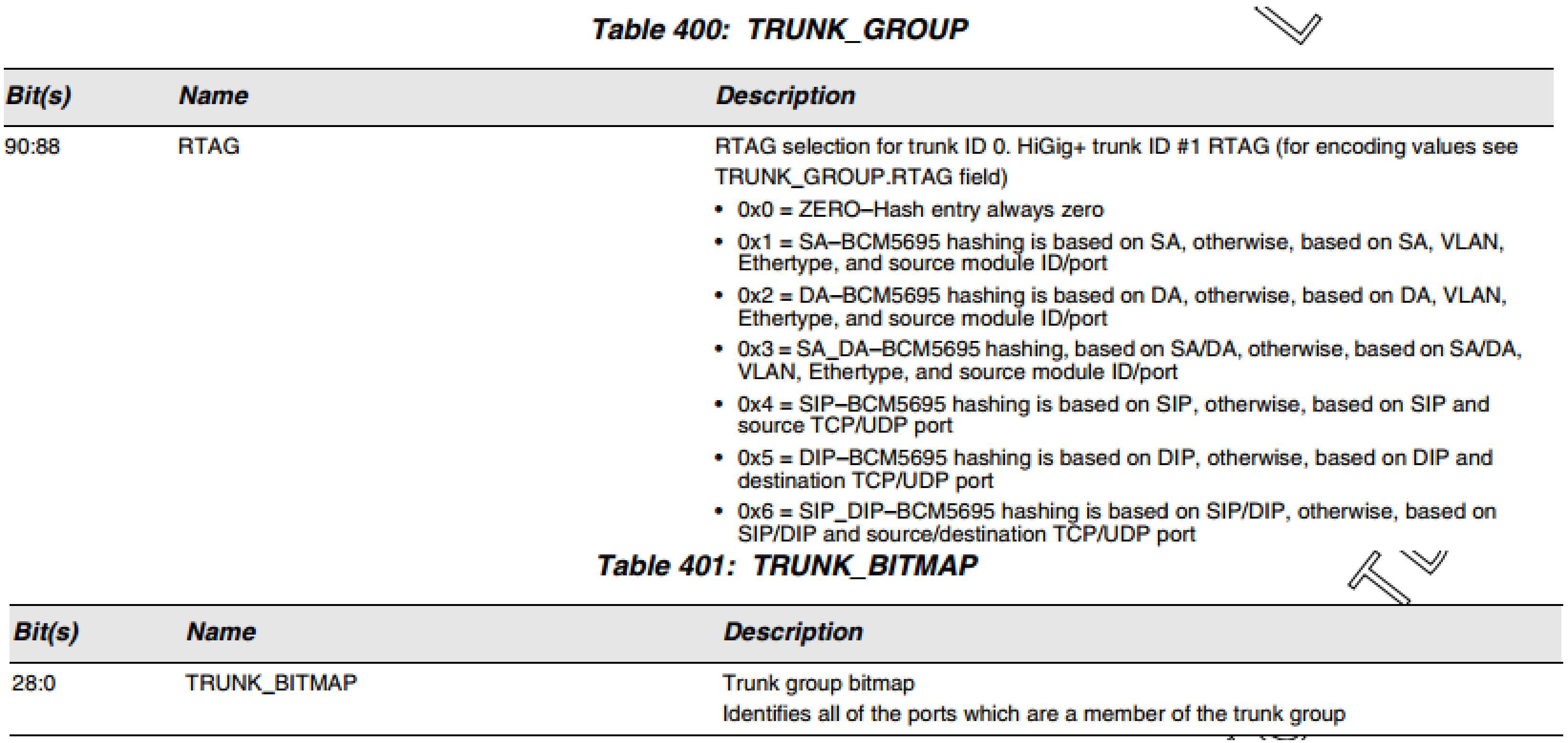


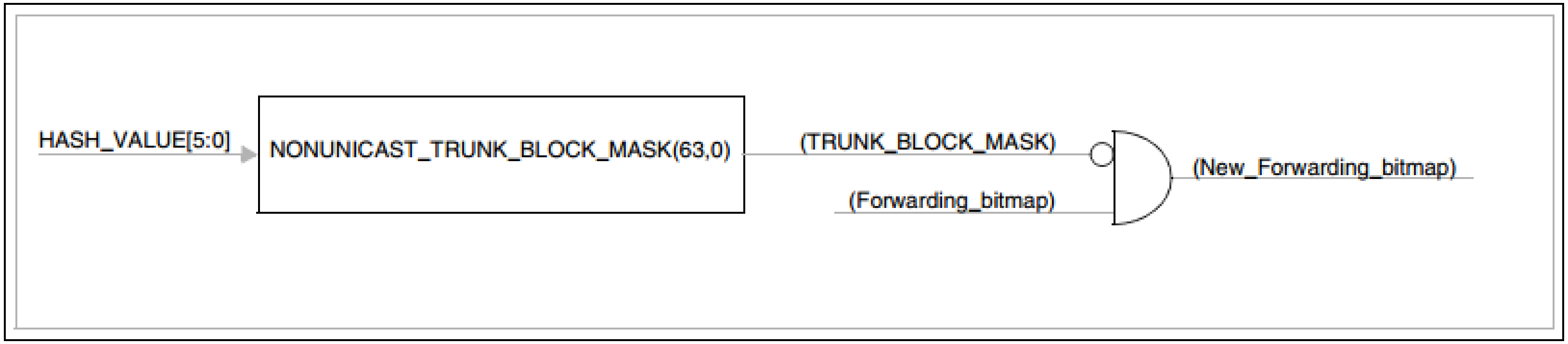
Table 400: TRUNK_GROUP

Bit(s)	*Name*	*Description*
90:88	RTAG	RTAG selection for trunk ID 0. HiGig+ trunk ID #1 RTAG (for encoding values see TRUNK_GROUP.RTAG field) • 0x0 = ZERO–Hash entry always zero • 0x1 = SA–BCM5695 hashing is based on SA, otherwise, based on SA, VLAN, Ethertype, and source module ID/port • 0x2 = DA–BCM5695 hashing is based on DA, otherwise, based on DA, VLAN, Ethertype, and source module ID/port • 0x3 = SA_DA–BCM5695 hashing, based on SA/DA, otherwise, based on SA/DA, VLAN, Ethertype, and source module ID/port • 0x4 = SIP–BCM5695 hashing is based on SIP, otherwise, based on SIP and source TCP/UDP port • 0x5 = DIP–BCM5695 hashing is based on DIP, otherwise, based on DIP and destination TCP/UDP port • 0x6 = SIP_DIP–BCM5695 hashing is based on SIP/DIP, otherwise, based on SIP/DIP and source/destination TCP/UDP port

Table 401: TRUNK_BITMAP

Bit(s)	*Name*	*Description*
28:0	TRUNK_BITMAP	Trunk group bitmap Identifies all of the ports which are a member of the trunk group

Non-Unicast Packet



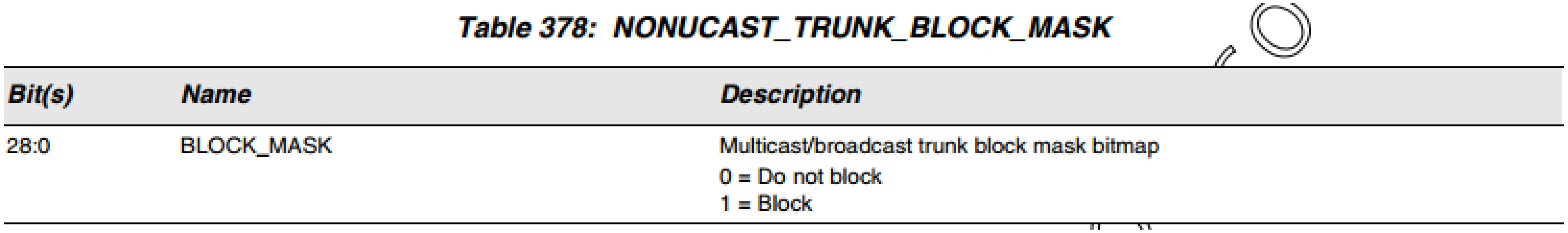


Table 378: NONUCAST_TRUNK_BLOCK_MASK

Bit(s)	Name	Description
28:0	BLOCK_MASK	Multicast/broadcast trunk block mask bitmap 0 = Do not block 1 = Block

Mirroring

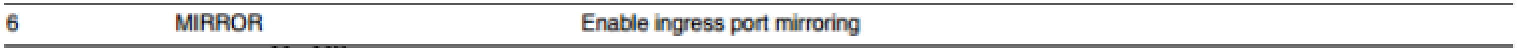
Mirroring Behavior

Mirroring支持两种模式：
Mirro only，端口只转发mirroring报文，且mirroring的报文不会进行地址学习。
Switch and mirror，端口即支持转发mirroring报文，同时也支持正常的端口报文转发。正常的报文按照正常的逻辑运转，能够进行地址学习、ACL等。

Mirroring的行为：
Ingress mirror的报文，原封不动地被发往镜像监测端口。
Egress mirror的报文，是把变化后的报文发往镜像监测端口。

Ingress Mirroring

① MIRROR_CONTROL.M_ENA=1。 /*开启端口的全局mirror功能*/
② 端口的PORT[ingress].MIRROR=1。 /*指定哪些端口需要被RX mirror*/
③ MIRROR_CONTROL.IM_MTP_INDEX /*设置指向IM_MTP_INDEX talbe的entry，IM_MTP_INDEX table包含镜像监测端口的Module ID与PORT_TGID*/
/*IM_MTP_INDEX与EM_MTP_INDEX的entry容量决定可以配置多少个RX/TX的镜像监测端口，此芯片只能支持4个*/



6	MIRROR	Enable ingress port mirroring

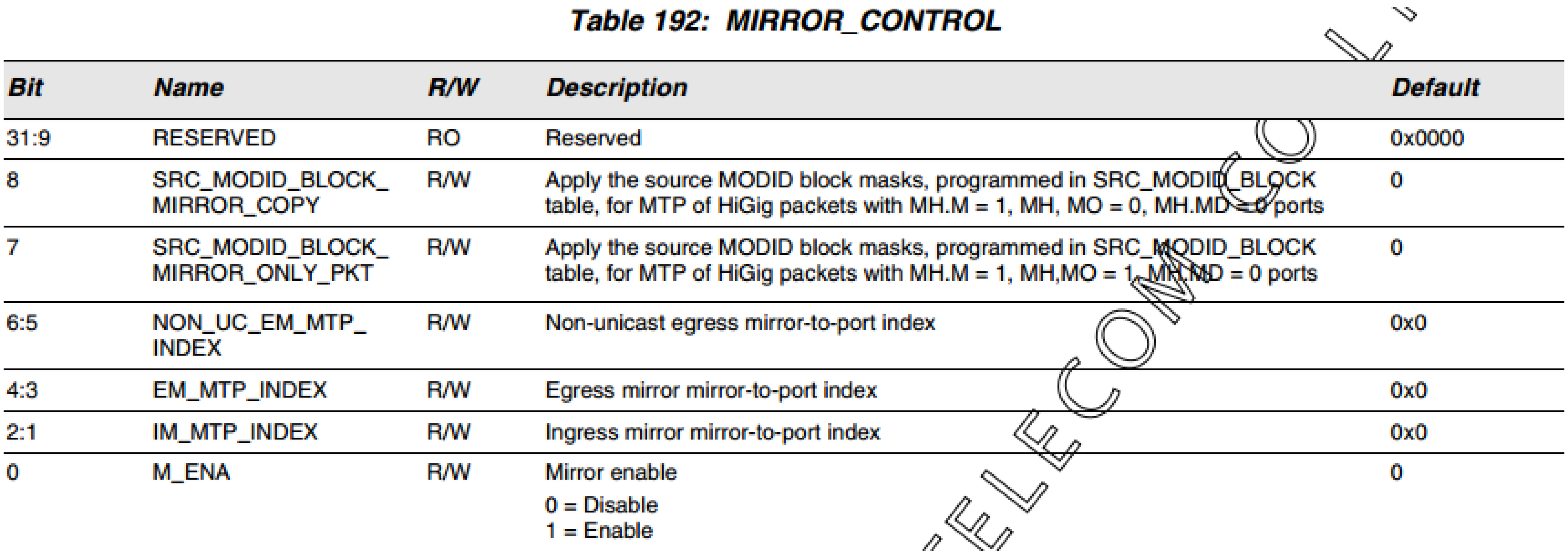


Table 192: MIRROR_CONTROL

Bit	Name	R/W	Description	Default
31:9	RESERVED	RO	Reserved	0x0000
8	SRC_MODID_BLOCK_MIRROR_COPY	R/W	Apply the source MODID block masks, programmed in SRC_MODID_BLOCK table, for MTP of HiGig packets with MH.M = 1, MH, MO = 0, MH.MD = 0 ports	0
7	SRC_MODID_BLOCK_MIRROR_ONLY_PKT	R/W	Apply the source MODID block masks, programmed in SRC_MODID_BLOCK table, for MTP of HiGig packets with MH.M = 1, MH,MO = 1, MH.MD = 0 ports	0
6:5	NON_UC_EM_MTP_INDEX	R/W	Non-unicast egress mirror-to-port index	0x0
4:3	EM_MTP_INDEX	R/W	Egress mirror mirror-to-port index	0x0
2:1	IM_MTP_INDEX	R/W	Ingress mirror mirror-to-port index	0x0
0	M_ENA	R/W	Mirror enable 0 = Disable 1 = Enable	0

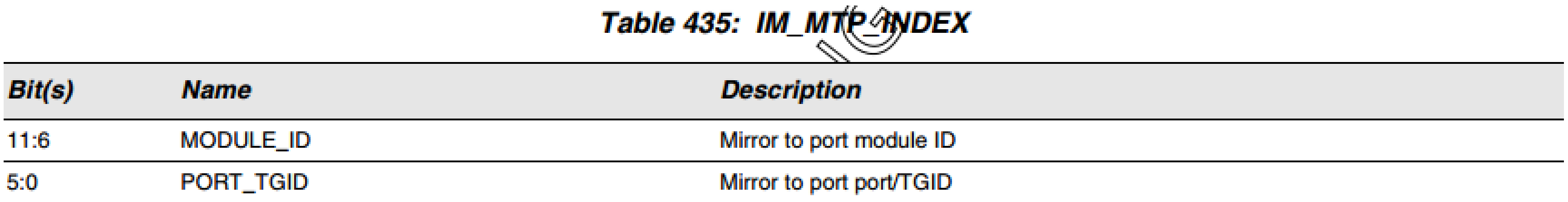


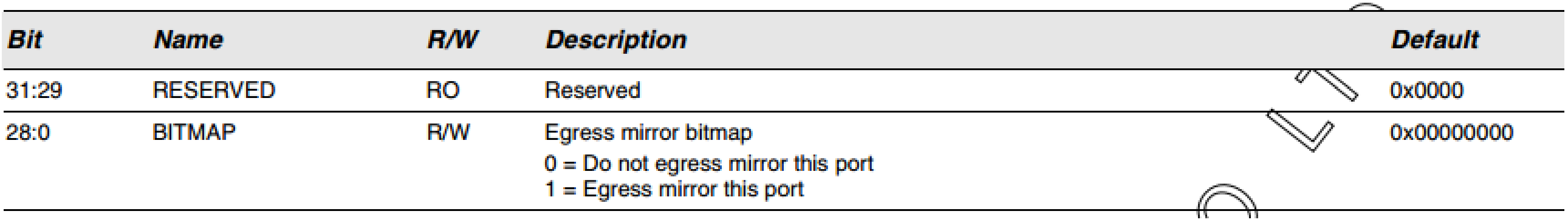
Table 435: IM_MTP_INDEX

Bit(s)	Name	Description
11:6	MODULE_ID	Mirror to port module ID
5:0	PORT_TGID	Mirror to port port/TGID

Egress Mirroring

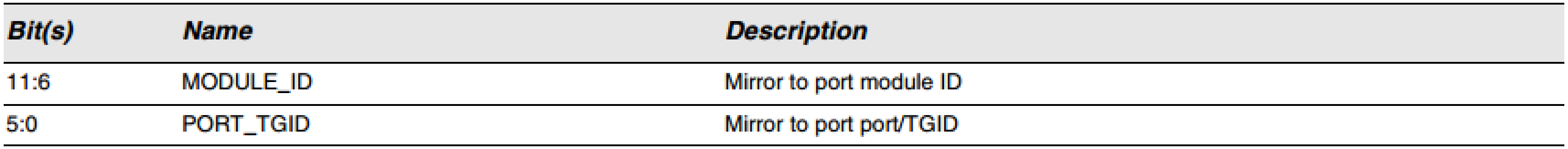
① MIRROR_CONTROL.M_ENA=1。 /*开启端口的全局mirror功能*/
② EMIRROR_CONTROL.BITMAP /*指定哪些端口需要被egress mirror*/
③ MIRROR_CONTROL.EM_MTP_INDEX /*设置指向EM_MTP_INDEX table的entry，EM_MTP_INDEX table包含镜像监测端口的Module ID与PORT_TGID*/

Table 193: EMIRROR_CONTROL



Bit	Name	R/W	Description	Default
31:29	RESERVED	RO	Reserved	0x0000
28:0	BITMAP	R/W	Egress mirror bitmap 0 = Do not egress mirror this port 1 = Egress mirror this port	0x00000000

Table 436: EM_MTP_INDEX



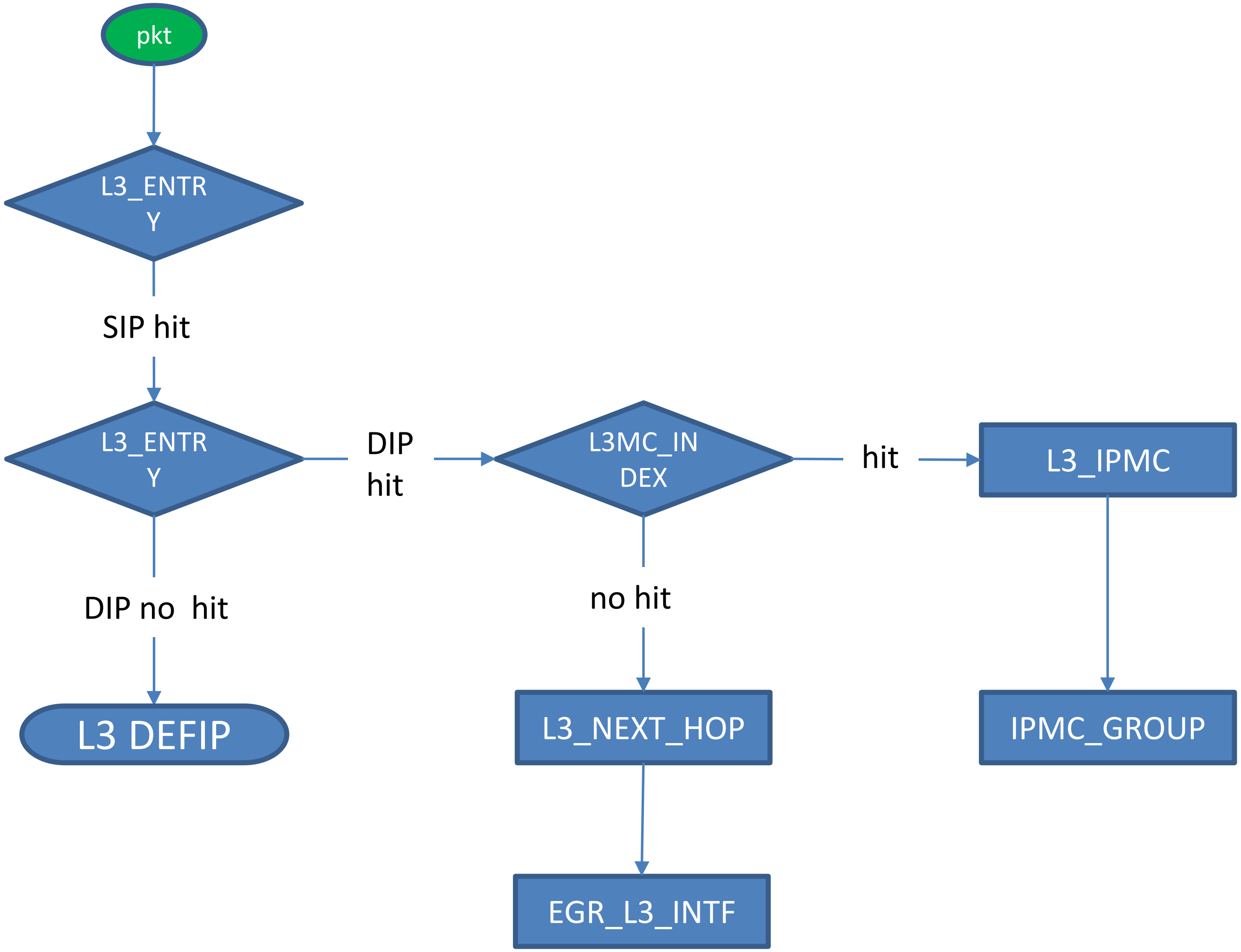
Bit(s)	Name	Description
11:6	MODULE_ID	Mirror to port module ID
5:0	PORT_TGID	Mirror to port port/TGID

L3

L3 Packet Flow

- L2进行目的MAC地址的L2_USER_ENTRY与L2_ENTRY查找时l3=1，那么报文将进入L3的处理。
  ① SIP查找：如果不匹配L3，则不进行DIP的查找处理(可选送CPU处理)；如果匹配L3，则更新L3 hit位并进行DIP的查找处理；如果匹配L3且发生端口迁移，报文将送往CPU处理，由软件负责L3表的更新，并进行DIP的查找处理。hit位可被软件用于老化处理。
  ② DIP查找：如果匹配L3，输出索引指向ING_L3_NEXT_HOP与EGR_L3_NEXT_HOP，从ING_L3_NEXT_HOP获取DST_MODID与DST_PORT/TGID，从EGR_L3_NEXT_HOP获取报文下一跳的目的MAC地址和索引指向EGR_L3_INTF，从EGR_L3_INTF获取源MAC地址与VID，获取的信息都由硬件用于自动替换报文的内容(SA,DA,VID)；如果不匹配L3，则搜索L3_DEFIP，最长匹配搜索算法用来匹配DIP的最长子网掩码，如果找到匹配，输出索引指向ING_L3_NEXT_HOP与EGR_L3_NEXT_HOP。
  对于查找命中的报文，将进行TTL减少，IP checksum与Ethernet FCS重新计算后发送。
  对于查找未命中的报文，将丢弃或者送CPU处理。

L3 Process


pkt
L3_ENTRY
SIP hit
L3_ENTRY
DIP hit
L3MC_INDEX
hit
L3_IPMC
DIP no hit
no hit
L3 DEFIP
L3_NEXT_HOP
IPMC_GROUP
EGR_L3_INTF

L3 Unicast

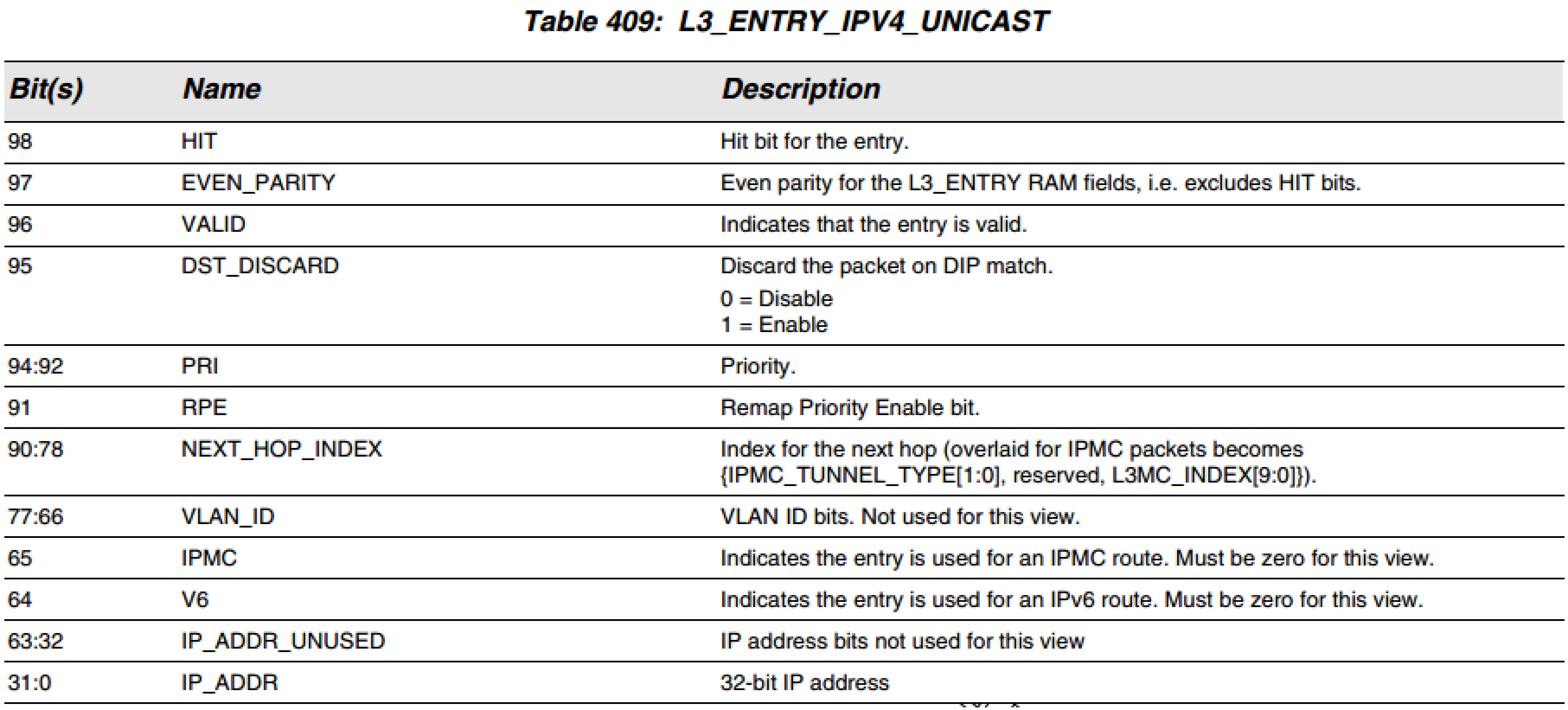
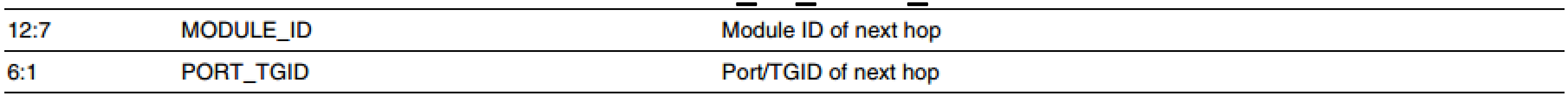


Table 409: L3_ENTRY_IPV4_UNICAST

Bit(s)	*Name*	*Description*
98	HIT	Hit bit for the entry.
97	EVEN_PARITY	Even parity for the L3_ENTRY RAM fields, i.e. excludes HIT bits.
96	VALID	Indicates that the entry is valid.
95	DST_DISCARD	Discard the packet on DIP match. 0 = Disable 1 = Enable
94:92	PRI	Priority.
91	RPE	Remap Priority Enable bit.
90:78	NEXT_HOP_INDEX	Index for the next hop (overlaid for IPMC packets becomes {IPMC_TUNNEL_TYPE[1:0], reserved, L3MC_INDEX[9:0]}).
77:66	VLAN_ID	VLAN ID bits. Not used for this view.
65	IPMC	Indicates the entry is used for an IPMC route. Must be zero for this view.
64	V6	Indicates the entry is used for an IPv6 route. Must be zero for this view.
63:32	IP_ADDR_UNUSED	IP address bits not used for this view
31:0	IP_ADDR	32-bit IP address

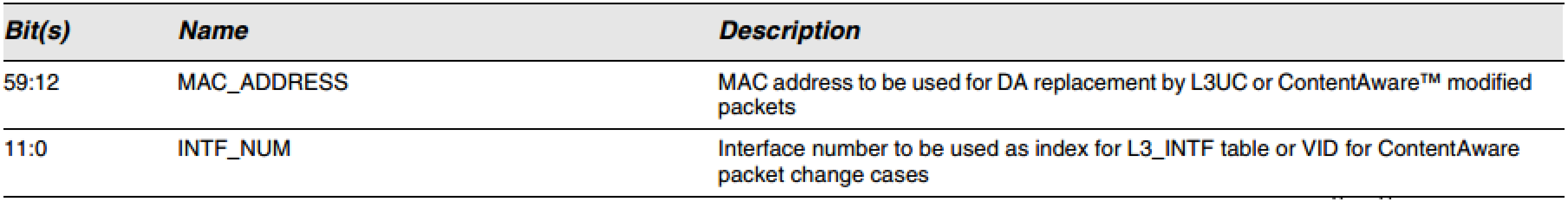
L3 Unicast

Table： ING_L3_NEXT_HOP



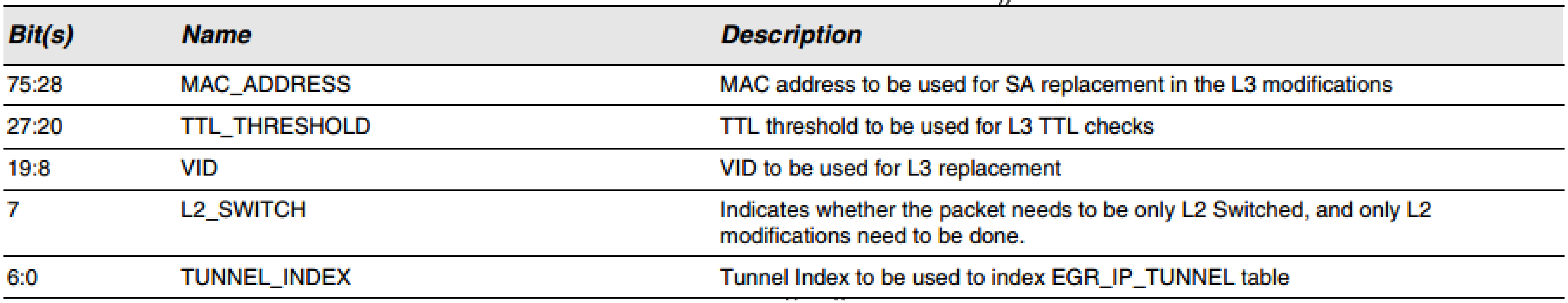
12:7	MODULE_ID	Module ID of next hop
6:1	PORT_TGID	Port/TGID of next hop

Table 421: EGR_L3_NEXT_HOP



Bit(s)	***Name***	***Description***
59:12	MAC_ADDRESS	MAC address to be used for DA replacement by L3UC or ContentAware™ modified packets
11:0	INTF_NUM	Interface number to be used as index for L3_INTF table or VID for ContentAware packet change cases

Table 422: EGR_L3_INTF



Bit(s)	***Name***	***Description***
75:28	MAC_ADDRESS	MAC address to be used for SA replacement in the L3 modifications
27:20	TTL_THRESHOLD	TTL threshold to be used for L3 TTL checks
19:8	VID	VID to be used for L3 replacement
7	L2_SWITCH	Indicates whether the packet needs to be only L2 Switched, and only L2 modifications need to be done.
6:0	TUNNEL_INDEX	Tunnel Index to be used to index EGR_IP_TUNNEL table

L3 Unicast

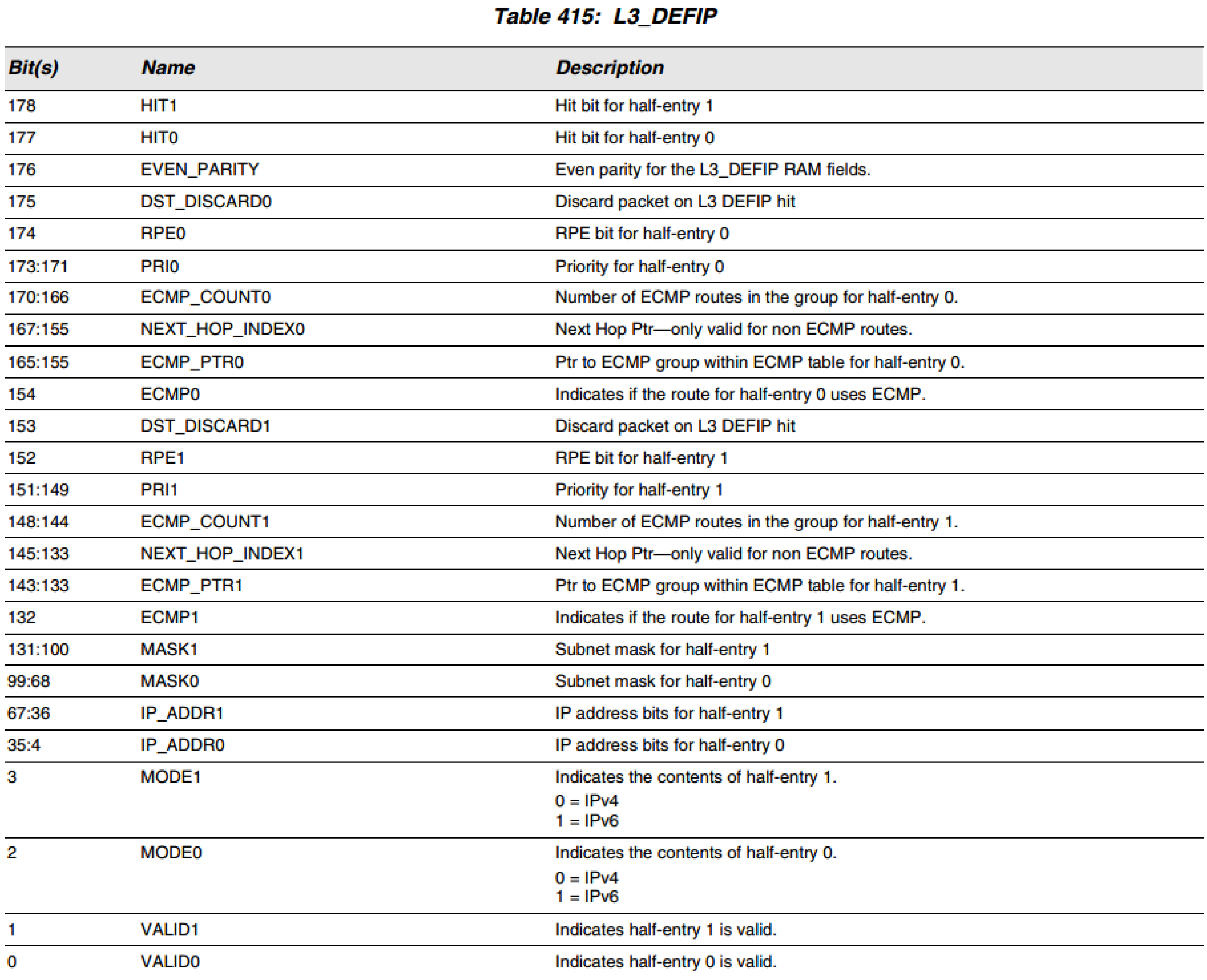


Table 415: L3_DEFIP

Bit(s)	Name	Description
178	HIT1	Hit bit for half-entry 1
177	HIT0	Hit bit for half-entry 0
176	EVEN_PARITY	Even parity for the L3_DEFIP RAM fields.
175	DST_DISCARD0	Discard packet on L3 DEFIP hit
174	RPE0	RPE bit for half-entry 0
173:171	PRI0	Priority for half-entry 0
170:166	ECMP_COUNT0	Number of ECMP routes in the group for half-entry 0.
167:155	NEXT_HOP_INDEX0	Next Hop Ptr—only valid for non ECMP routes.
165:155	ECMP_PTR0	Ptr to ECMP group within ECMP table for half-entry 0.
154	ECMP0	Indicates if the route for half-entry 0 uses ECMP.
153	DST_DISCARD1	Discard packet on L3 DEFIP hit
152	RPE1	RPE bit for half-entry 1
151:149	PRI1	Priority for half-entry 1
148:144	ECMP_COUNT1	Number of ECMP routes in the group for half-entry 1.
145:133	NEXT_HOP_INDEX1	Next Hop Ptr—only valid for non ECMP routes.
143:133	ECMP_PTR1	Ptr to ECMP group within ECMP table for half-entry 1.
132	ECMP1	Indicates if the route for half-entry 1 uses ECMP.
131:100	MASK1	Subnet mask for half-entry 1
99:68	MASK0	Subnet mask for half-entry 0
67:36	IP_ADDR1	IP address bits for half-entry 1
35:4	IP_ADDR0	IP address bits for half-entry 0
3	MODE1	Indicates the contents of half-entry 1. 0 = IPv4 1 = IPv6
2	MODE0	Indicates the contents of half-entry 0. 0 = IPv4 1 = IPv6
1	VALID1	Indicates half-entry 1 is valid.
0	VALID0	Indicates half-entry 0 is valid.

L3 Multicast

- 三层组播报文根据端口关联的VLAN数来决定需要复制多少份报文发送(1个VLAN需要1份)。L3(IPMC)索引指向L3_IPMC与IPMC_GROUP。
  - L3_IPMC指示组播报文要往哪个端口进行二层转发(L2_BITMAP)，或者三层路由(L3_BITMAP)。
  - IPMC_GROUP指示组播报文要往哪个VLAN进行转发。

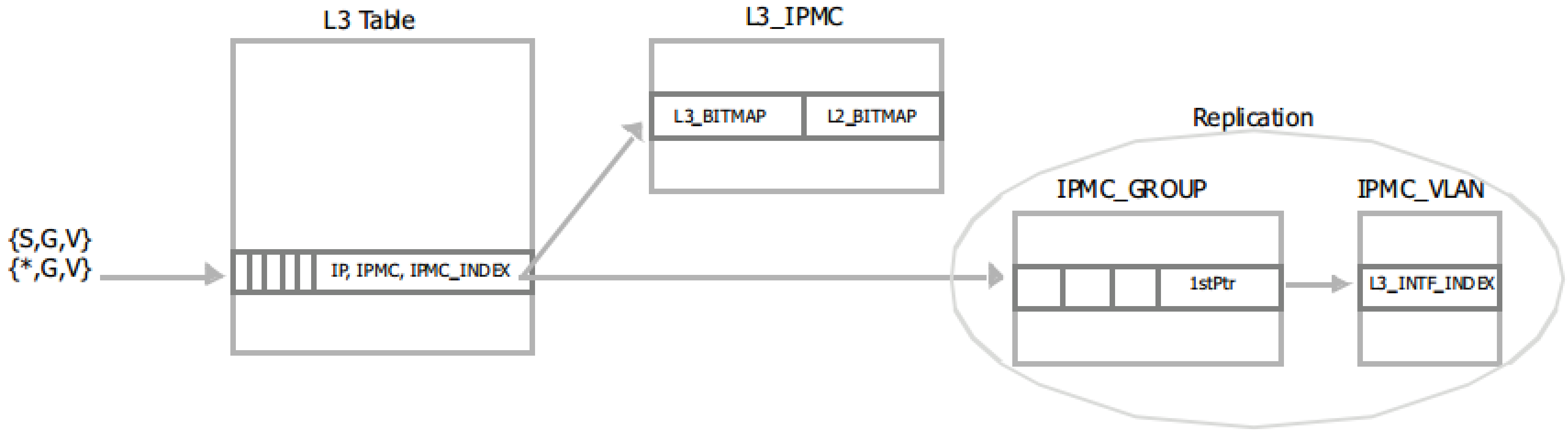
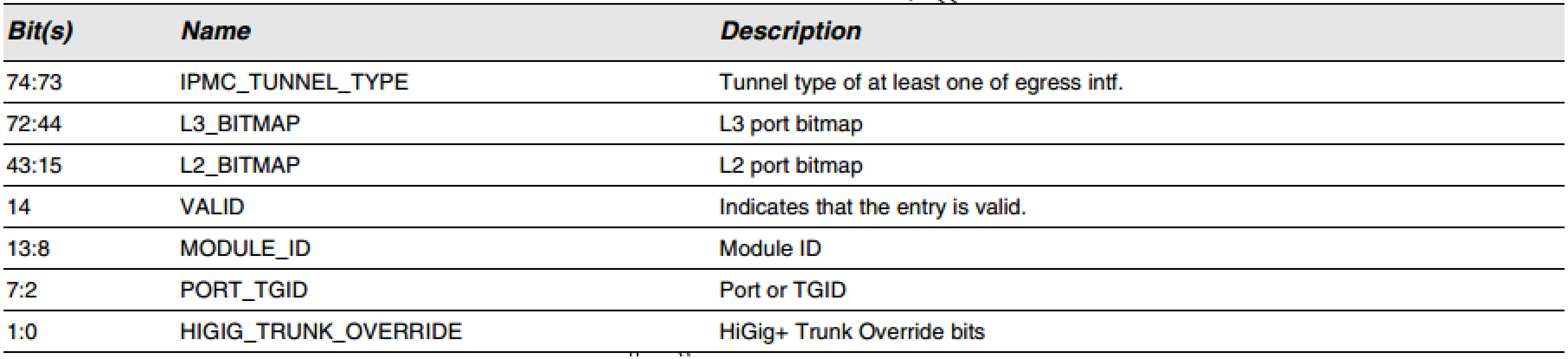


Table 414: L3_IPMC

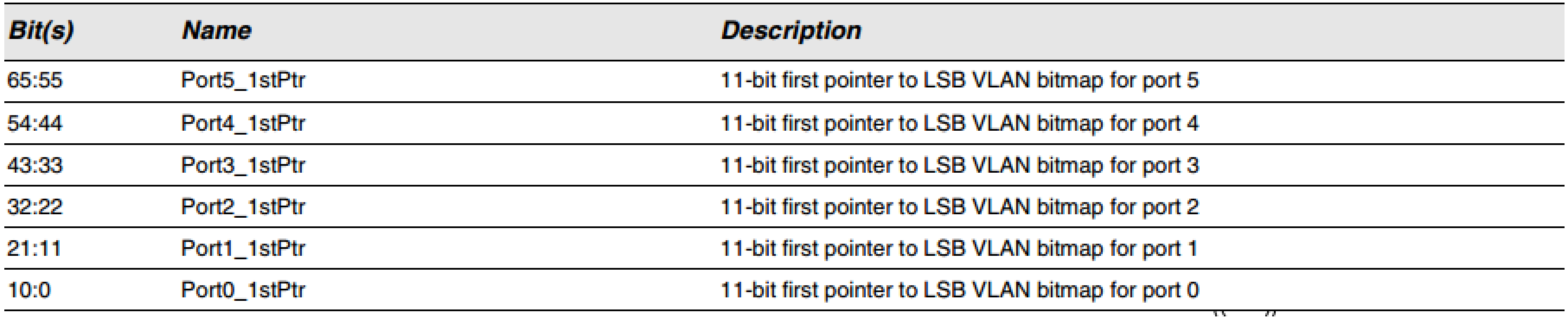


Bit(s)	***Name***	***Description***
74:73	IPMC_TUNNEL_TYPE	Tunnel type of at least one of egress intf.
72:44	L3_BITMAP	L3 port bitmap
43:15	L2_BITMAP	L2 port bitmap
14	VALID	Indicates that the entry is valid.
13:8	MODULE_ID	Module ID
7:2	PORT_TGID	Port or TGID
1:0	HIGIG_TRUNK_OVERRIDE	HiGig+ Trunk Override bits

L3 Multicast

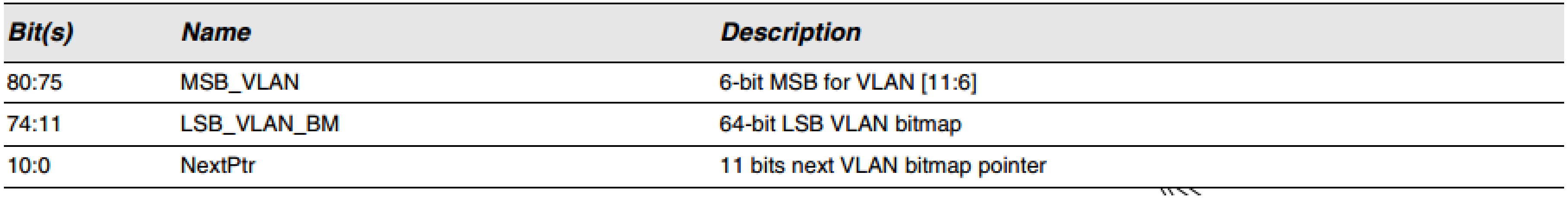
- IPMC_GROUP用来索引IPMC_VLAN，IPMC_VLAN用来指示报文应该被复制到哪个VLAN ID里进行泛洪。
- IPMC_VLAN有3个字段
  - MSB_VLAN与LSB_VLAN_BM用来组合12位的数值表示组播报文发送目的VLAN ID，VID的高6位由MSB_VLAN表示；VID的低6位由LSB_VLAN_BM的置位bit的位置表示(从左往右开始计算)。比如，MSB_VLAN=0x4，LSB_VLAN_BITMAP=0x00000080(从左往右第7位)，则最终的VID=0x000100,000111=0x107。
    如果LSB_VLAN_BITMAP有多位置位，则表示有多个VLAN，每个位分别进行计算。
  - 一条表项最多只能表示64个VLAN，由NextPrt把多条表项组成链表，来表示更多的VLAN。

Table 426: IPMC_GROUP0 Table–MMU_IPMC_GROUP_TBL0



Bit(s)	*Name*	*Description*
65:55	Port5_1stPtr	11-bit first pointer to LSB VLAN bitmap for port 5
54:44	Port4_1stPtr	11-bit first pointer to LSB VLAN bitmap for port 4
43:33	Port3_1stPtr	11-bit first pointer to LSB VLAN bitmap for port 3
32:22	Port2_1stPtr	11-bit first pointer to LSB VLAN bitmap for port 2
21:11	Port1_1stPtr	11-bit first pointer to LSB VLAN bitmap for port 1
10:0	Port0_1stPtr	11-bit first pointer to LSB VLAN bitmap for port 0

Table 434: IPMC_VLAN Table–MMU_IPMC_VLAN_TBL



Bit(s)	*Name*	*Description*
80:75	MSB_VLAN	6-bit MSB for VLAN [11:6]
74:11	LSB_VLAN_BM	64-bit LSB VLAN bitmap
10:0	NextPtr	11 bits next VLAN bitmap pointer