




商用交换芯片SDN支持 现状分析

 pica8.com  sales@pica8.com  [@pica8](https://twitter.com/pica8)

内容简介

- 商用交换芯片简介
- 商用L2/L3芯片
 - L2/L3交换处理流程
 - OpenFlow功能限制
 - 基于硬件表项TTP支持
- 可编程交换芯片
 - 芯片体系结构
 - Openflow功能支持
- Pica8 SDN产品系列
- 问题回答



商用交换芯片简介

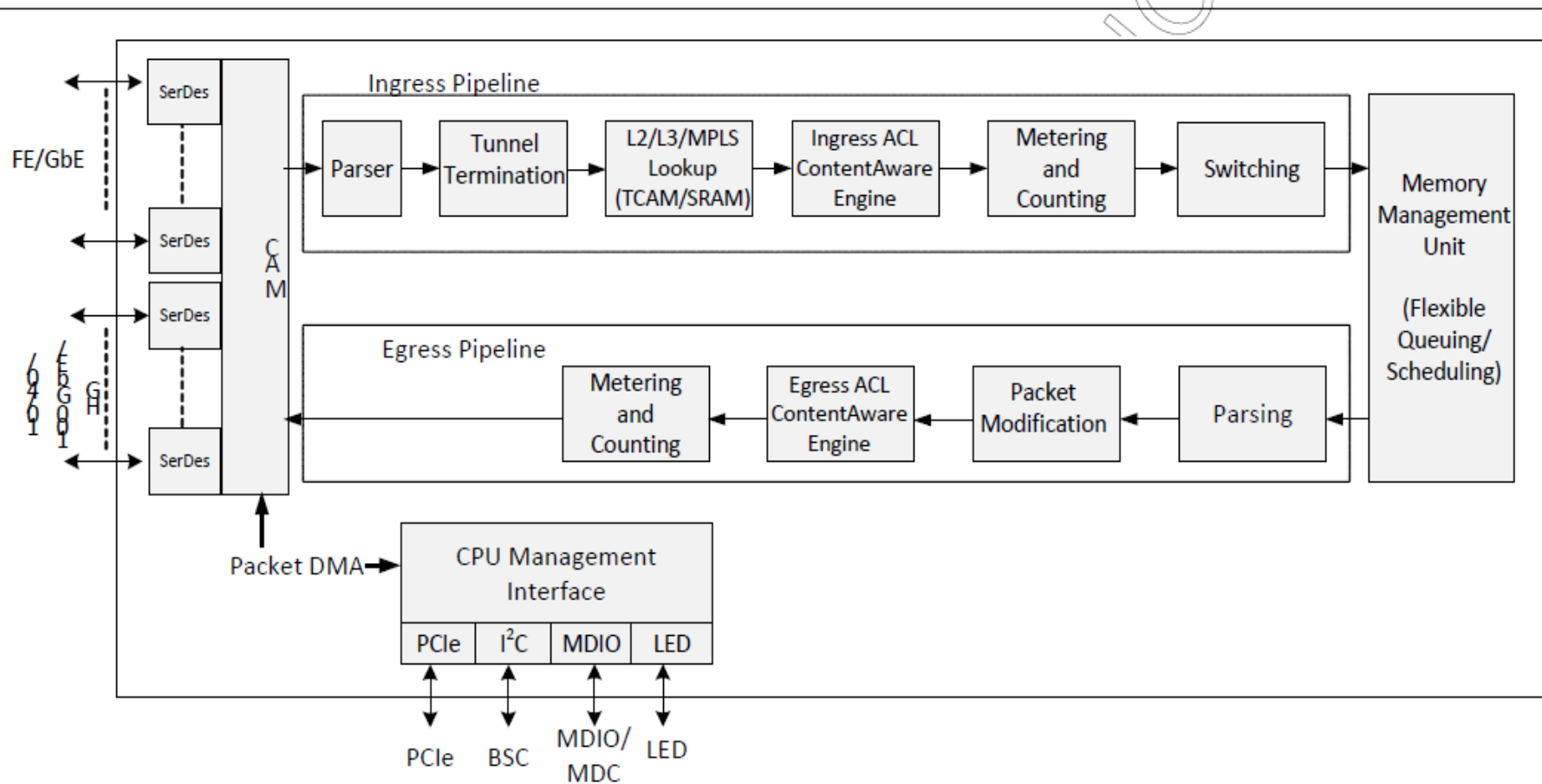
▪ 市场上主要国、内外交换芯片

- Broadcom StrataXGS Family – Trident, Trident II and Tomahawk
- Broadcom Strata DNX Family – Dune
- Cisco
- Mavell Presteria CX/DX
- Intel Fulcrum
- Cavium Xpliant, 10G, 25G, 40G, 100G
- Mellanox Spectrum, 10G, 25G, 40G, 100G
- Huawei ENP
- Centec
- MTK

L2/L3交换芯片的数据转发

- 二层转发表
 - 数据包的MAC、VLAN、端口属性
 - 地址信息学习
 - 依据地址表信息转发或Flood出去
- 三层路由表
 - IP网络路由信息
 - 路由表通常由路由协议生成
 - 按照最大匹配转发
- 数据包处理逻辑单元
 - 包头字段分析器
 - Tunnel终结处理
 - VLAN处理
 - L2转发表、L3路由表(组播)
 - ACL/Policy路由
 - Tunnel及其它包修改动作

L2/L3交换处理流程

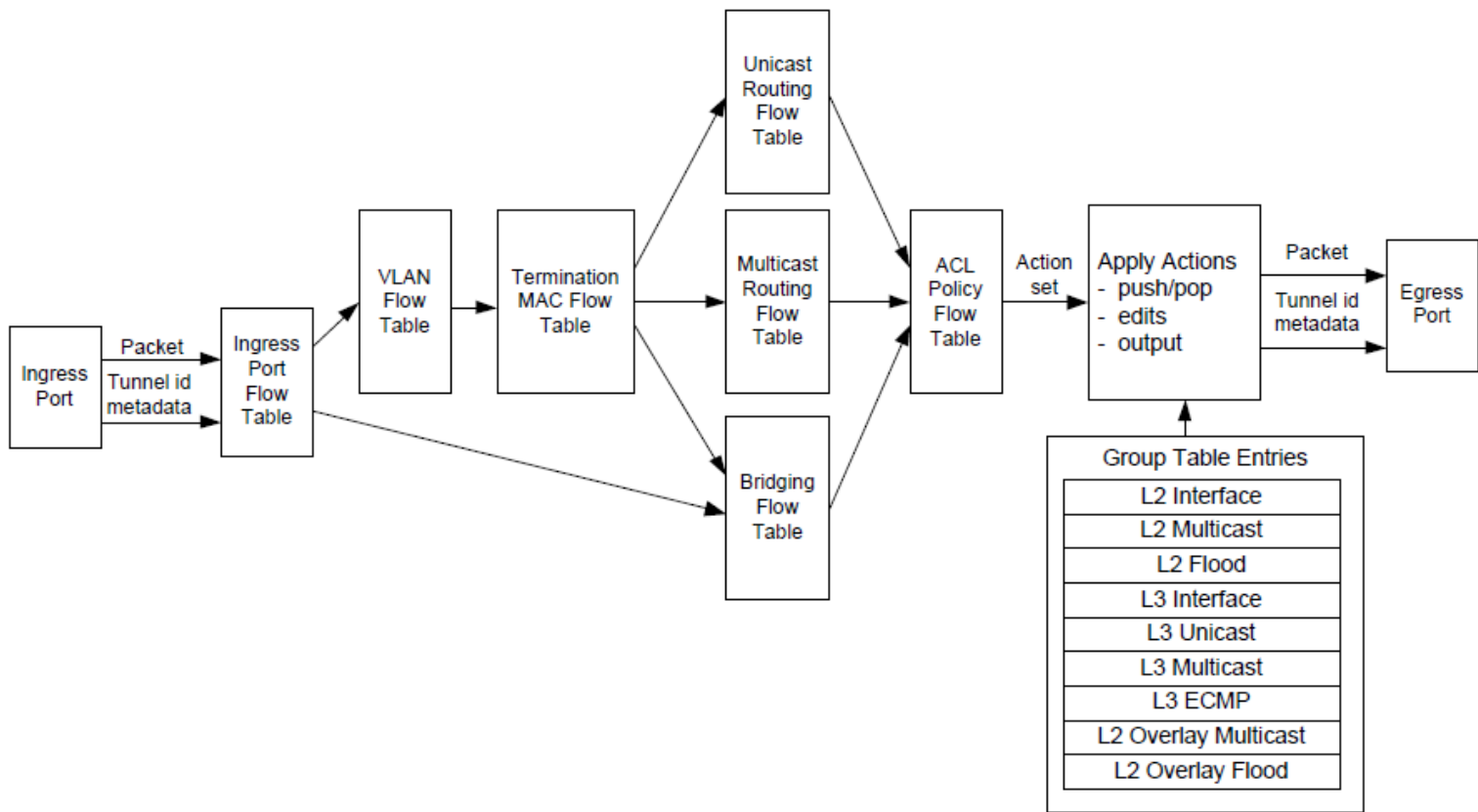


Activate Windows

OPENFLOW功能限制

- L2/L3交换芯片支持Openflow存在的问题
 - TCAM规则匹配表项较小，一般所支持的表项数目在4K-16K之间。外接TCAM成本较高，影响转发性能。
 - TCAM宽度决定了同时能够匹配的字段数目，如IPv6地址需要占用较宽的资源进行匹配
 - 硬件TCAM流表位置、数量固定，主要用于二、三层处理。
 - 各流表匹配执行的动作，无法立即生效。
 - 部分字段只读、无法修改。
 - 无法支持数据包loopback处理
 - Group动作非常受限
 - CPU端口带宽受限，影响数据包送控制器处理

基于硬件表项TTP支持



硬件表项及TTP表项映射

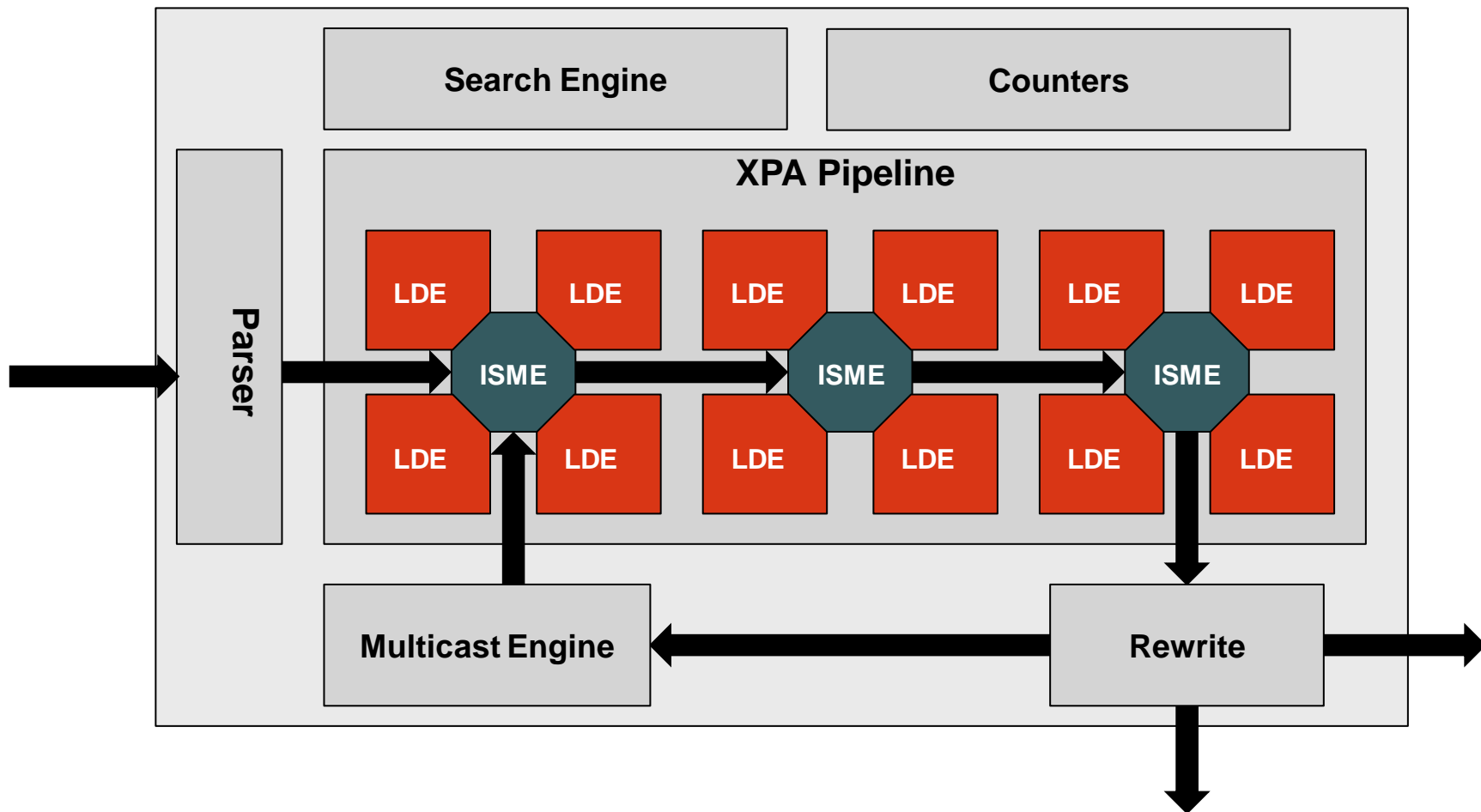
- Ingress Port Flow Table
- VLAN Flow Table
- Termination MAC Flow Table
- Bridging Flow Table
- Unicast Routing Flow Table
- Multicast Routing Flow Table
- Policy ACL Flow Table
- Group Table

TTP Group Table

- Group Table
 - L2 Interface - egress VLAN filtering and tagging
 - L2 Rewrite - modify Ethernet header
 - L3 Unicast - routing next hop and output interface
 - L2 Multicast – forward to multi L2 Interface
 - L2 Flood - flood in VLAN
 - L3 Interface - outgoing routing interface properties
 - L3 Multicast - forward to multi L3 interface
 - L3 ECMP – ECMP forwarding
 - L2 Overlay - logical ports
 - L1 Fast Failover

可编程交换芯片

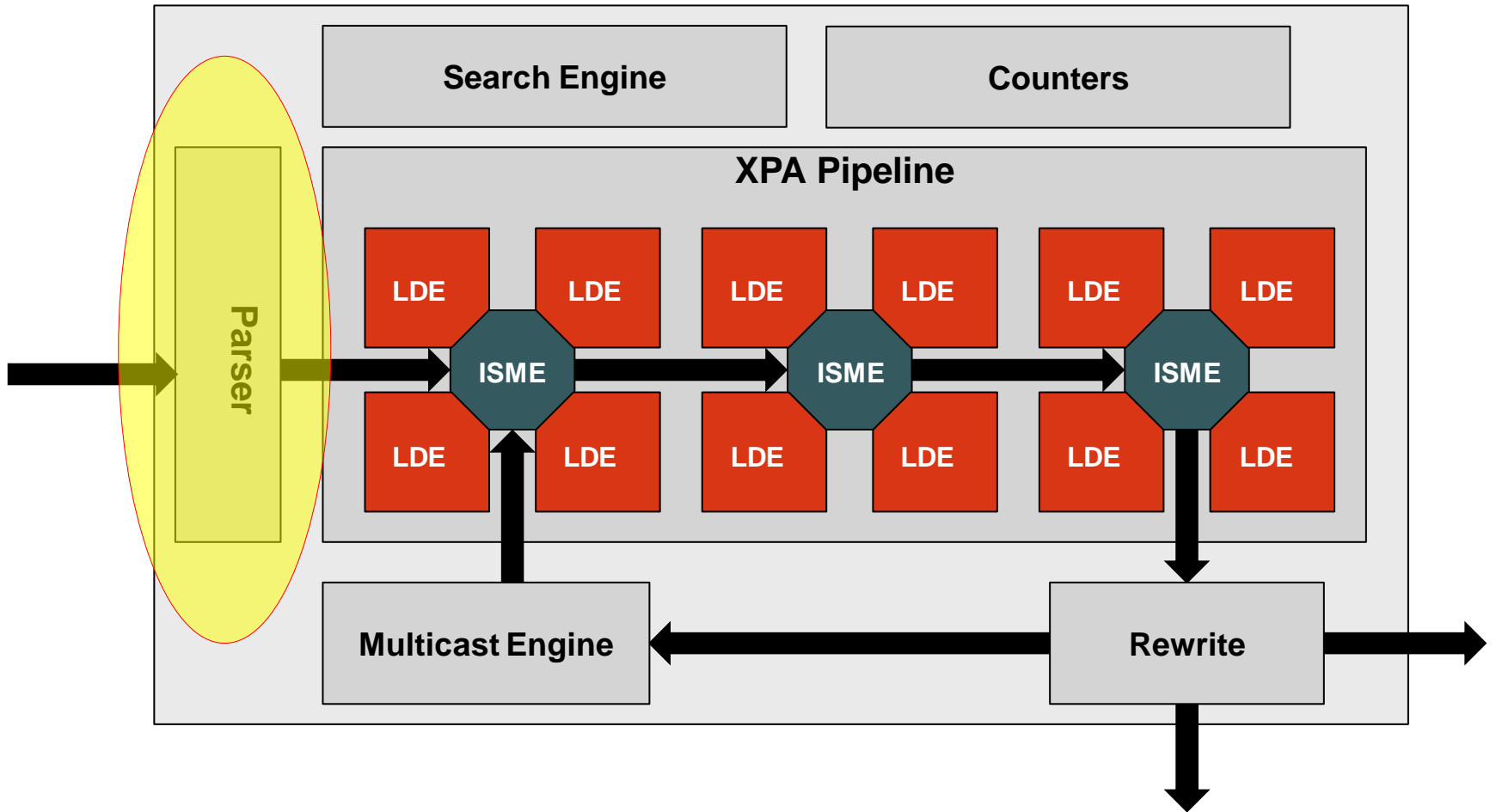
- XPliant Software Defined Engine



XPliant Software Defined Engine(C)

- **Programmable Header Parser**
 - Performs header parsing and store the parsed result in a layer stack
- **Lookup and Decision Engine – LDE**
 - Generic Lookup and Decision Engine that can be programmed to perform packet processing functionality
- **Inter Switch and Memory Element – ISME**
 - Interconnects 4 LDEs and other ISMEs
- **Header Rewrite Engine**
 - Resolves packet's destination, modifies existing packet header, and adds tunneling header
- **Multicast Replication Engine (MRE)**
 - MRE goes through a linked list of nodes and duplicates a packet per every node
- **ACM: Analytics, Counting, sampling and Policing**
 - ACM incorporates 32 Counters blocks, enabling up to 32 counting, analytics, sampling and policing events per packet, at wire speed

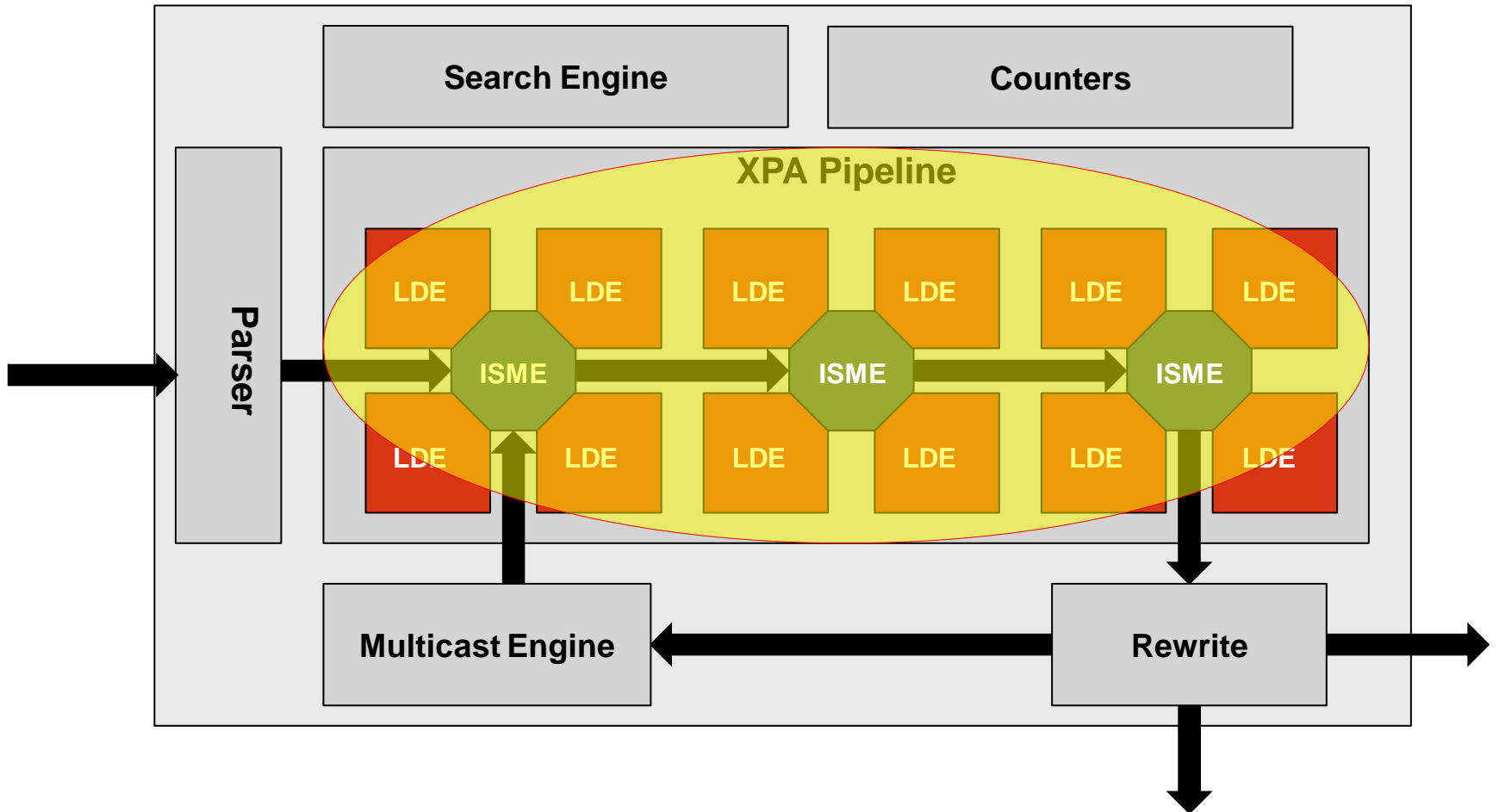
XPliant Software Defined Engine - Parser



XPliant Software Defined Engine

- Programmable Header Parser
 - Programmable Parser
 - Parsing of existing, future and proprietary header
 - Header size, 64, 128, 192 Bytes
 - Parses up to 8 layers of packet data
 - Initial VLAN assignment
 - Initial QoS assignment
 - Programmable packet hashes for ECMP
 - Validity check for well-known protocols

XPliant Software Defined Engine – LDE Pipe



XPliant Software Defined Engine

▪ Programmable LDE Pipe

- 12 Lookup and Decision Engine (LDEs)
- 3 clusters of 4 LDEs connected via Inter Switch and Memory Elements (ISMEs)
- LDEs are addressable and packet flow through LDEs is programmable

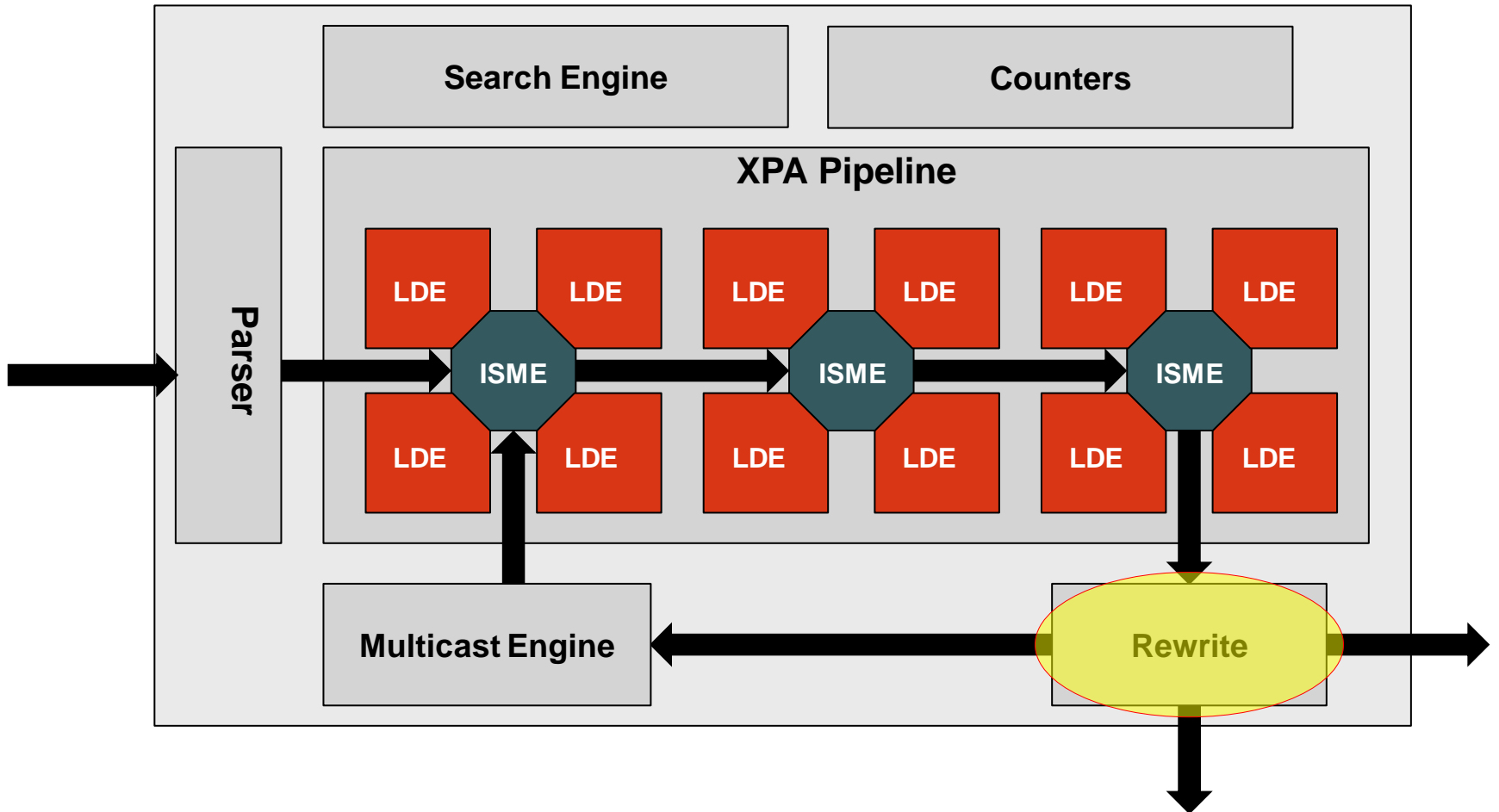
▪ Lookup and Decision Engine – LDE

- Generic Lookup and Decision Engine, programmed to perform packet processing
- Generates keys for up to 4 parallel lookups
- Programmable lookup keys and programmable lookup type, Direct Index, Hash, LPM or TCAM
- Programmed to modify any field in packet context
- Generates up to four Count, Sampling, or Policing transactions

▪ Inter Switch and Memory Element – ISME

- Interconnects 4 LDEs and other ISMEs
- Supports Unicast forwarding of tokens to LDEs and to other ISMEs

XPliant Software Defined Engine – Rewrite

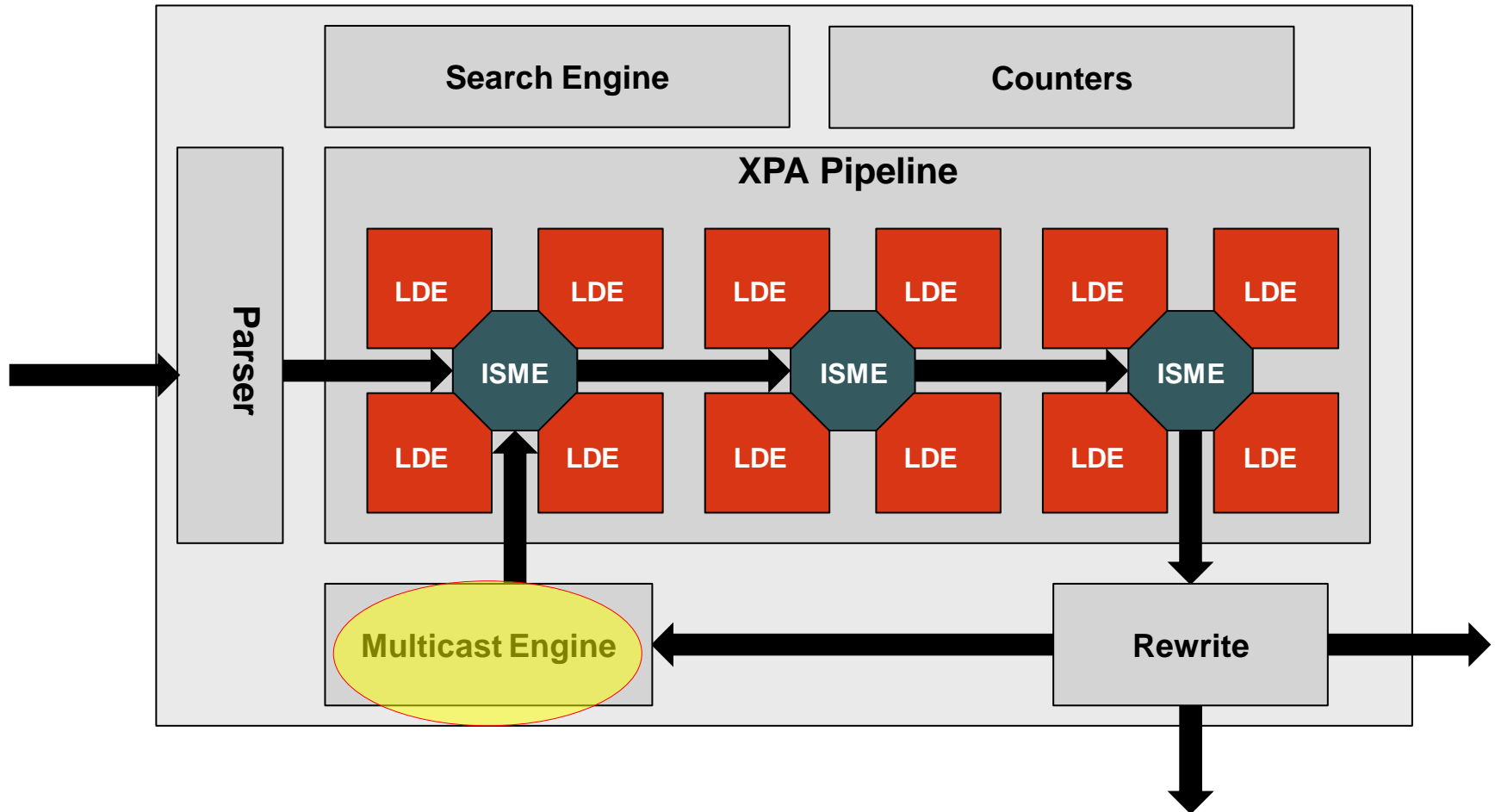


XPliant Software Defined Engine

▪ Header Rewrite Engine

- Programmable header modification
- Programmable tunnel termination
- Programmable header insertion
- IP headers and TCP headers checksum generation
- ECN marking of IP headers
- Layer 2 and LAG ECMP up to 1K paths
- Group ID based filtering
- Packet truncation for packets forwarded to CPU or to an analyzer

XPliant Software Defined Engine – Multicast

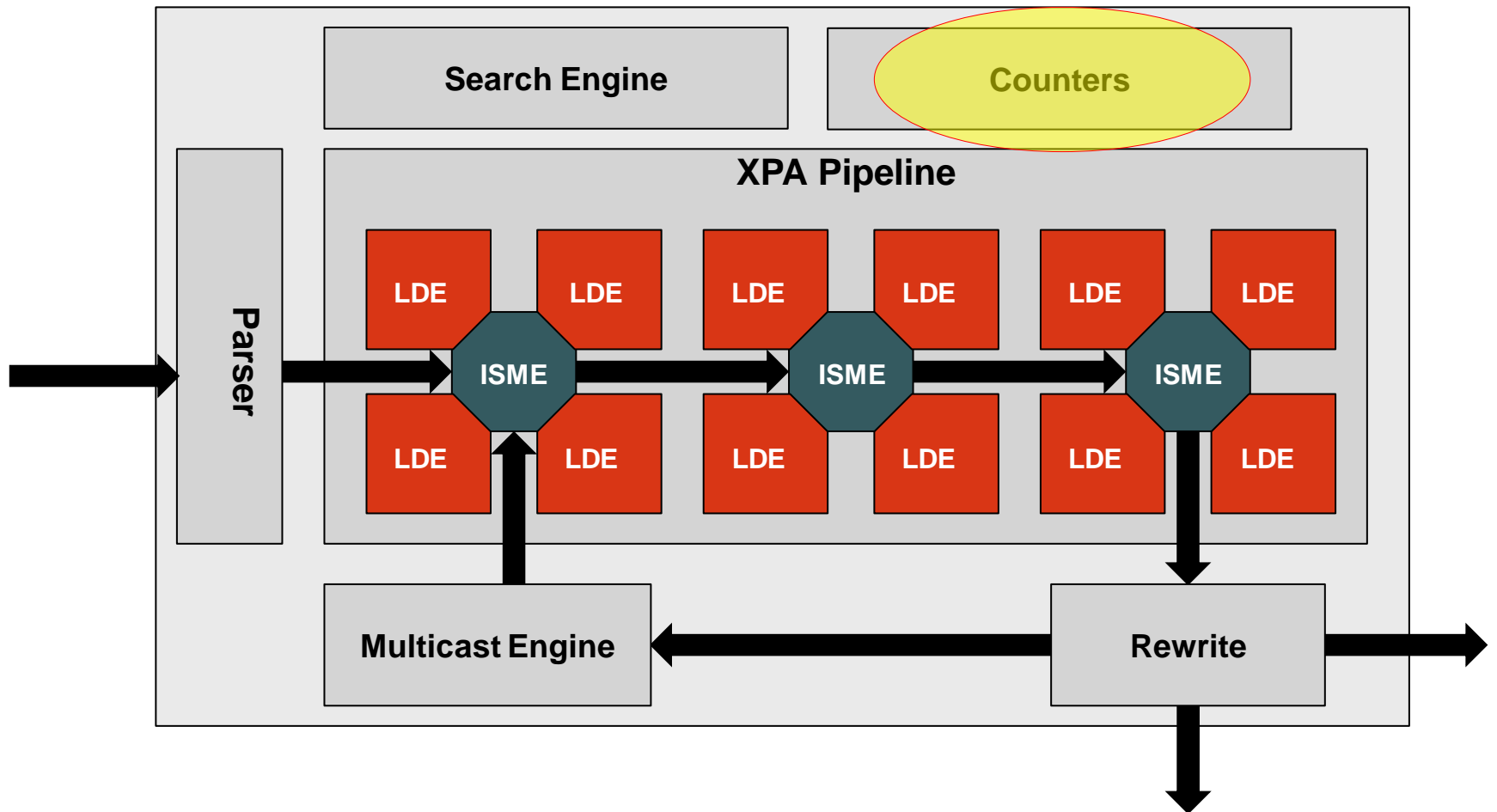


XPliant Software Defined Engine

▪ Multicast Replication Engine (MRE)

- IP multicast forwarding
- Mirroring packets to multiple analyzers
- Returning Packets to LDE Pipe after modification
- 300Mpps maximal throughput (600Mpps per device)
- Four priority based input Fifos
- Output rate shaper

XPliant Software Defined Engine – Multicast

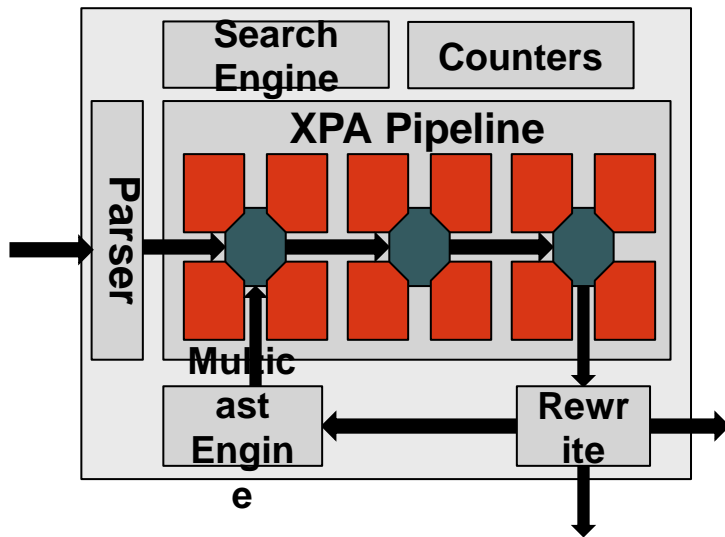
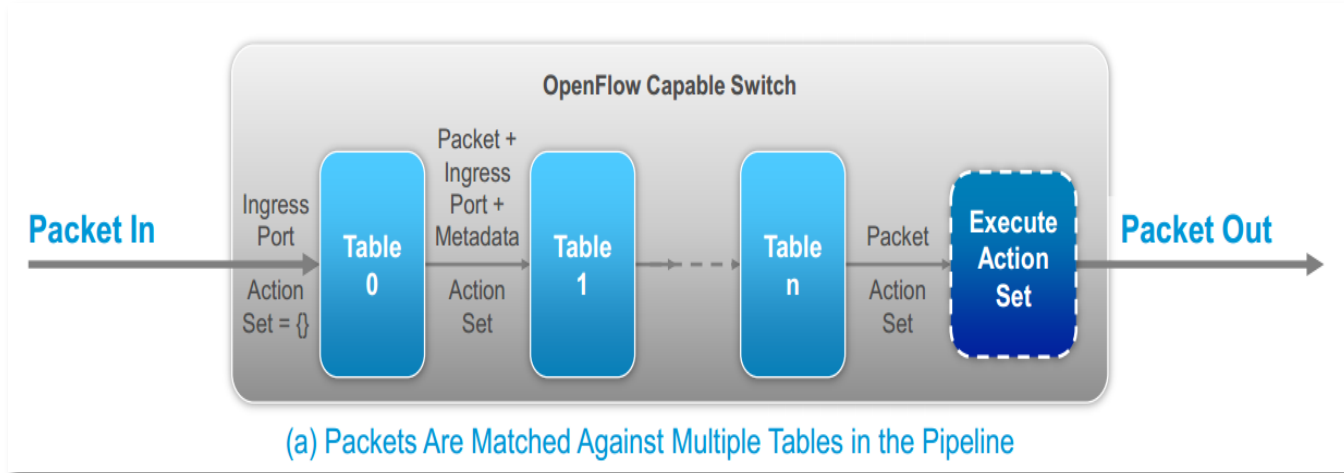


XPliant Software Defined Engine

▪ ACM: Analytics, Counting, sampling and Policing

- An LDE can trigger up to four counting events
- Each of the 32 Counters block can be configured to one of three modes: Counting, Sampling or Policing
- Policing Mode: Two Rates Three Color Marking Policers
- Sampling Mode: Event (packets) based samplers
- Counting Mode:
 - Count mode 0, 17b event counters, 8K counters per block, up to 512K Counters per device
 - Count mode 1, 34b event counters, 4K counters per block, up to 256K Counters per device
 - Count mode 2, 30b event, 38b byte counters pairs, 2K counters per block, up to 128K event, byte counter pairs per device

Openflow功能支持



- Programmable parser
- Cascade table match
- Flexible/dynamic table creation
- Fully programmable lookup key & action set
- Flexible table size allocation
- Programmable packet editor
- Flexible counters attach to any lookup/flow

MTK SDN交换芯片

- Native OpenFlow pipe and large scale of flow entries 完全针对OpenFlow设计的架构, 支持海量流表
- Native hybrid and per port/VLAN/flow pipe select 真正的混合架构和可基于端口、VLAN、流的引擎选择
- Flexible memory resource share between SDN/OF and legacy pipe 在SDN/OF引擎和传统引擎之间进行灵活的表项共享分配
- Programmable parser/flow table/packet modifier, protocol agonistic 可编程的解析器、流匹配表、数据包修改, 不局限于现有协议



PicOS及所支持白盒交换机 – 100G

硬件制造商	型号	交换芯片	流表大小
HP Enterprise	HPE6960, 32 x 100G	TOMAHAWK	TCAM 6K IPv4 128K
EdgeCore	AS7712, 32 x 100G	TOMAHAWK	TCAM 6K IPv4 128K
Dell	Z9100, 32 x 100G	TOMAHAWK	TCAM 6K IPv4 128K
Inventec	D7032Q28B, 32 x 100G	TOMAHAWK	TCAM 6K IPv4 128K
EdgeCore	AS7512, 32 x 100G	XPLIANT	TCAM 64bx512x16 256bx6x4+256bx8x4

PicOS及所支持白盒交换机 – 40G

硬件制造商	型号	交换芯片	流表大小
HP Enterprise	HPE6940 32 x 40G	TRIDENT II	TCAM - 4K IPv4 - 128K
EdgeCore	AS6712/AS6701 32 x 40G	TRIDENT II	TCAM - 4K IPv4 - 128K
PICA8	P-5401 32 x 40G	TRIDENT II	TCAM - 4K IPv4 - 128K
InterfaceMaster	N2632 32 x 40G	TRIDENT II	TCAM - 4K IPv4 - 128K

PicOS及所支持白盒交换机 – 10G

硬件制造商	型号	交换芯片	流表大小
HP Enterprise	HPE6920/1, 48x10G + 6x40G	Trident II Trident II Plus	TCAM – 4K/16K IPv4 – 16K/128K
EdgeCore	AS5712/5812-54T/X 48x10G + 6x40G	Trident II Trident II Plus	TCAM – 4K/16K IPv4 – 16K/128K
PICA8	P-5101 40x10G + 8x40G	Trident II	TCAM – 4K IPv4 – 16K
PICA8	P-3930/3922/3920 40x10G + 4x40G	Trident+	TCAM – 2K IPv4 – 16K
DELL	S4048 48x10G + 6x40G	Trident II	TCAM – 4K IPv4 – 16K

PicOS及所支持白盒交换机 – 1G

硬件制造商	型号	交换芯片	流表大小
HP Enterprise	HPE6900T/P 48/24x1G+4/2x10G	Helix4	4K
EdgeCore	AS4610T/P 48/24x1G + 4/2x10G	Helix4	4K
Penguin	Arctica 4804i 48x1G + 4x10G	Triumph 2	8K
PICA8	P-3297 48x1G + 4x10G	Triumph 2	8K



Questions & Feedback

Please contact lin.du@pica8.com