Cisco Nexus 9300 平台缓冲区和排队架构

白皮书

2014 年 11 月

目录

概述

Cisco Nexus® 9300 平台交换机是固定配置的 Cisco Nexus 9000 系列交换机。该平台以紧凑的外形提供业界领先的 1、10、40 千兆以太网端口密度和性能以及高能效。

Cisco Nexus 9300 平台交换机可以在传统的思科® NX-OS 模式或思科以应用为中心的基础设施 (ACI) 模式下运行。在思科 NX-OS 模式下运行时，Cisco Nexus 9300 平台使用全面的思科 NX-OS 软件第 2 层和第 3 层功能集以及广泛的可编程功能，以提供具有高性能、运营效率和设计灵活性的数据中心解决方案。在以思科 ACI 模式部署时，Cisco Nexus 9300 平台交换机在高速、完全自动化、由两部分构成的思科 ACI 交换矩阵架构中作为枝叶节点运行。它们为应用终端提供连接点，并为思科 ACI 租户应用执行基于策略的转发和实施。

本白皮书讨论了思科 NX-OS 模式下的 Cisco Nexus 9300 平台的缓冲区和排队架构。思科 ACI 模式和枝叶节点功能不属于本文档的讨论范围。

数据中心网络接入层的缓冲区要求

虽然数据中心网络汇聚层对深度缓冲区的需要已经由诸如 Cisco Nexus 9500 平台之类的交换机平台消除，因为此类平台提供无阻塞、低延迟的线速性能，具有 10 和 40 千兆以太网端口密度，但是出于以下几个原因，网络接入层的足够缓冲容量仍然是一个关键的网络设计原则：

- 接入交换机经常发生端口速度不匹配，因为存在多种主机连接类型。上行链路的速度通常比主机端口高。如果流量从快速端口流动到慢速端口，例如从 10 千兆以太网端口流动到 1 千兆以太网端口，则需要额外的缓冲区空间来适应端口速度差异。
- 接入层通常会设计主机端口与上行链路端口之间的超用比。
- 采用 in-cast 流量模式的应用要求接入交换机端口上有更深度的缓冲区。

Cisco Nexus 9300 平台缓冲区结构

Cisco Nexus 9300 平台交换机包括一个网络转发引擎 (NFE) 和一个应用枝叶引擎 (ALE) 或 ALE-2。当 Cisco Nexus 9300 平台交换机在思科 NX-OS 模式下运行时，NFE 执行大多数网络功能，ALE 或 ALE-2 提供额外的缓冲区空间并促进高级网络功能，例如虚拟可扩展局域网 (VXLAN) 之间的路由。

可以在 Cisco Nexus 9300 平台交换机的通用扩展模块 (GEM) 或某些类型的 Cisco Nexus 9300 平台交换机的交换机基板上找到 ALE 或 ALE-2。表 1 列出了它们配备的不同 GEM 类型和 ALE 类型。表 2 列出了不同类型的 ALE 专用集成电路 (ASIC) 及其支持的 Cisco Nexus 9300 交换机平台。

表 1. Cisco Nexus 9300 平台 GEM

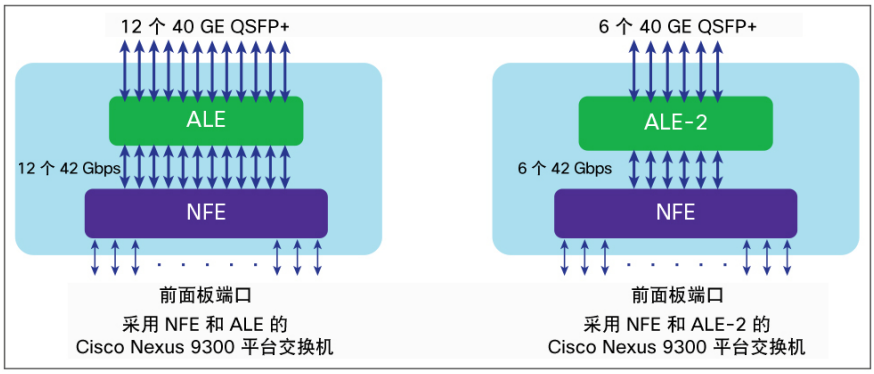
GEM 类型	ALE 类型	支持的 Cisco Nexus 9300 平台
N9K-M12PQ	ALE	所有具有一个 GEM 插槽的 Cisco Nexus 9300 平台交换机，包括 Cisco Nexus 9396PX、9396TX、93128PX 和 93128TX 交换机
N9K-M6PQ	ALE-2	所有具有一个 GEM 插槽的 Cisco Nexus 9300 平台交换机，包括 Cisco Nexus 9396PX、9396TX、93128PX 和 93128TX 交换机

表 2. ALE 类型和支持的 Cisco Nexus 9300 平台交换机

ALE 类型	缓冲区大小	支持的 Cisco Nexus 9300 平台
ALE	40 MB	具有 N9K-M12PQ 模块的 Cisco Nexus 9396PX 和 9396TX 具有 N9K-M12PQ 模块的 Cisco Nexus 93128PX 和 93128TX
ALE-2	25 MB	具有 N9K-M6PQ 模块的 Cisco Nexus 9396PX 和 9396TX 具有 N9K-M6PQ 模块的 Cisco Nexus 93128PX 和 93128TX Cisco Nexus 9372PX 和 9372TX 交换机（ALE-2 在交换机基板上） Cisco Nexus 9372PQ 交换机（ALE-2 在交换机基板上）

根据其使用的 ALE 类型，Cisco Nexus 9300 平台交换机具有图 1 中使用的两种内部架构之一。

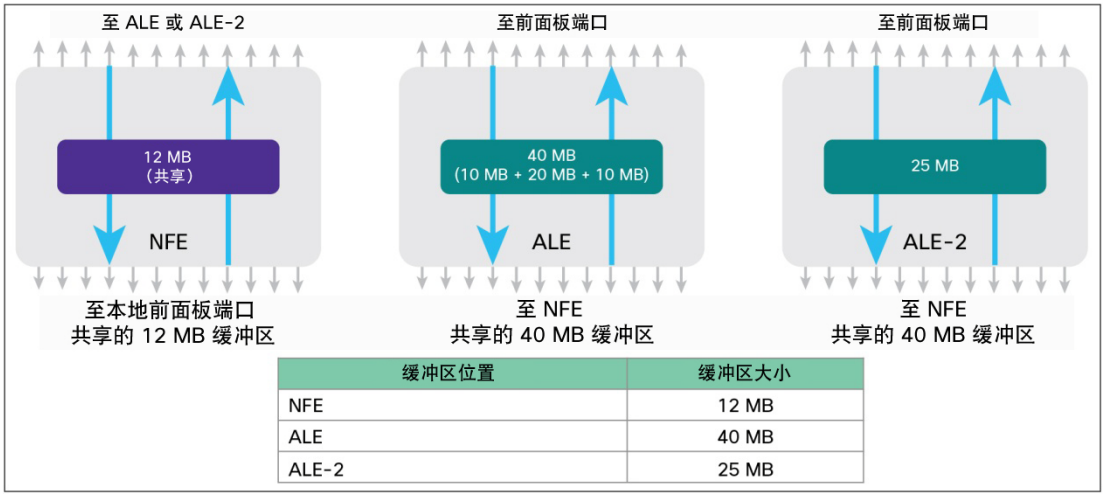
图 1. Cisco Nexus 9300 平台交换机的内部框图



NFE 和 ALE/ALE-2 提供了板载缓冲区空间。图 2 显示了 Cisco Nexus 9300 平台交换机中可能的缓冲区空间。它包括：

- NFE 上的 12 MB（由 NFE 上的所有端口共享，用于所有流量）
- ALE 上的 40 MB（分为三个区域：用于从 ALE 前面板端口到 NFE 前面板端口的流量，用于从 NFE 前面板端口到 ALE 前面板端口的流量，以及用于两个 NFE 前面板端口之间的发夹式流量）
- ALE-2 上的 12 MB（由 ALE-2 上的所有端口共享，用于所有流量）

图 2. Cisco Nexus 9300 平台交换机中的缓冲区



缓冲区位置	缓冲区大小
NFE	12 MB
ALE	40 MB
ALE-2	25 MB

根据 ALE 类型，Cisco Nexus 9300 平台交换机可以具有 52 MB 的缓冲区内存（NFE 上的 12 MB 和 ALE 上的 40 MB）或 37 MB 的缓存区内存（NFE 上的 12 MB 和 ALE-2 上的 25 MB）。

网络转发引擎上的缓冲区

NFE 上的 12 MB 缓冲区由 NFE 上的所有端口动态共享。它分为三个服务池（图 3）：

- 控制服务池
- 带外流量控制 (OOBFC) 单播服务池
- 默认服务池

控制流量使用控制服务池中的专用缓冲区资源提供服务。OOBFC 单播服务池为在 Cisco Nexus 9300 平台交换机的 ALE 上具有扩展输出队列的单播流量提供服务。

图 3. Cisco Nexus 9300 平台 NFE 缓冲区服务池

Cisco Nexus 9300 平台的思科 NX-OS 软件为用户提供了用于动态监控交换机缓冲区配置和利用率的命令行界面 (CLI) 命令。图 4 显示了交换机中的 NFE 缓冲区的监控命令的输出示例。在命令输出中：

- SP-0 是默认服务池
- SP-2 是 OOBFC 服务池
- SP-3 是控制服务池

请注意，NFE 最多支持四个缓冲区服务池。Sp-1在 Cisco Nexus 9300 平台交换机上保持未使用状态。

图 4. Cisco Nexus 9300 平台 NFE 缓冲区显示

```
n9396-1# show hardware internal buffer info pkt-stats detail

slot  1
=======


INSTANCE: 0
============


|-------------------------------------------------------------------------|
            Output Shared Service Pool Buffer Utilization (in cells)
                          SP-0      SP-1      SP-2      SP-3
|-------------------------------------------------------------------------|

Total Instant Usage           0         0         0         0
Remaining Instant Usage   29938         0     14346      6344
Peak/Max Cells Used           0         0         0         0
Switch Cell Count         29938         0     14346      6344
|-------------------------------------------------------------------------|
```

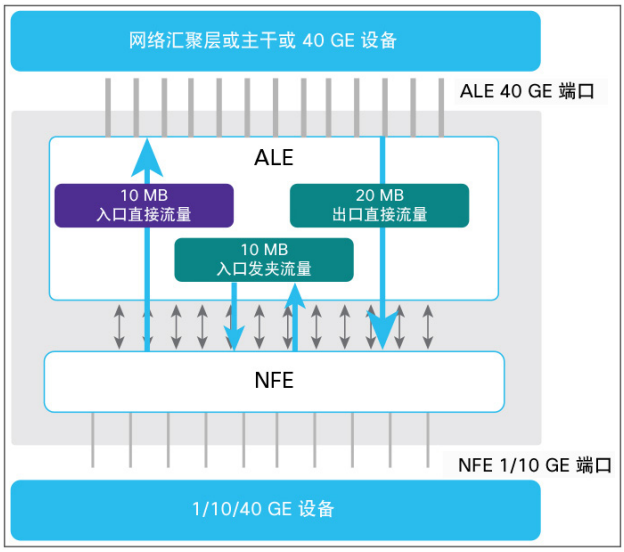
SP-0：默认服务池
SP-2：OOBFC 单播服务池
SP-3：控制服务池

应用枝叶引擎上的缓冲区

ALE 上的 40 MB 缓冲包括三个不同区域（图 5）：

- 入口直接流量的缓冲区（10 MB）：流量方向与网络有关。此处的入口意味着进入网络汇聚层或主干。因此该缓冲区适用于从 Cisco Nexus 9300 平台交换机的 ALE 40 千兆以太网端口流出的流量。
- 入口发夹流量的缓冲区（10 MB）：该缓冲区适用于在 NFE 上的两个前面板端口之间流动的流量。可以使用缓冲区提升功能将流量以发夹式传送到 ALE，以便利用 ALE 上额外的 10 MB 缓冲区空间。
- 出口直接流量的缓冲区（20 MB）：出口意味着来自网络并流向主机设备。因此该缓冲区适用于来自 ALE 40 千兆以太网端口并在 NFE 前面板端口流出的流量。

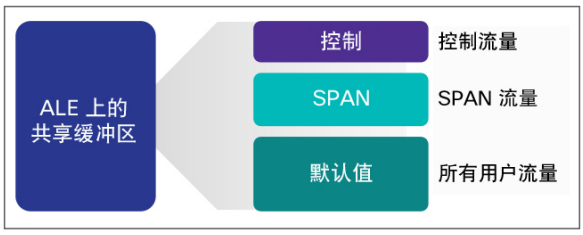
图 5. ALE 上的缓冲区区域



这三个区域中的缓冲区内存均由其在对应方向上提供服务的端口动态共享。它们分为三个服务池（图 6）：

- 控制服务池：用于所有控制平面流量
- 思科交换端口分析器 (SPAN) 服务池；用于 SPAN 流量
- 默认流量池：用于所有其他数据流量

图 6. ALE 上的缓冲区服务池



Cisco Nexus 9000 系列的思科 NX-OS 提供了用于显示 ALE 缓冲区分配和动态利用率的 CLI 命令。图 7 显示了监控命令的输出示例。

图 7. ALE 缓冲区服务池显示

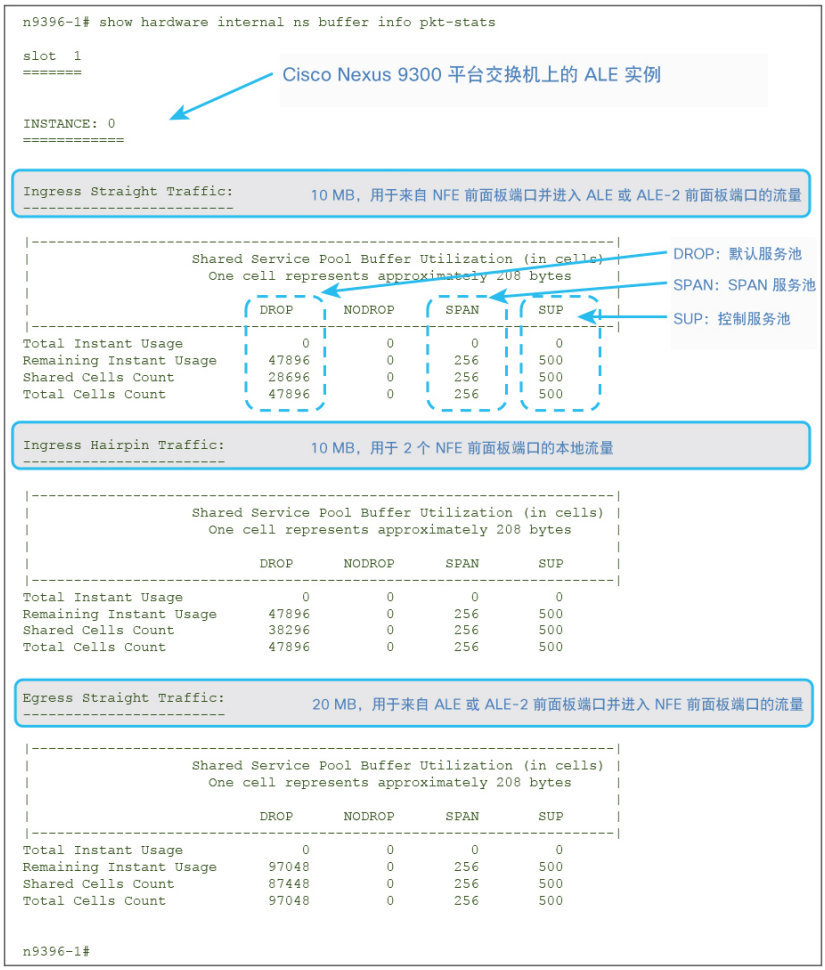


图 7 的命令输出标识了如下所述的三个 ALE 缓冲区服务池：

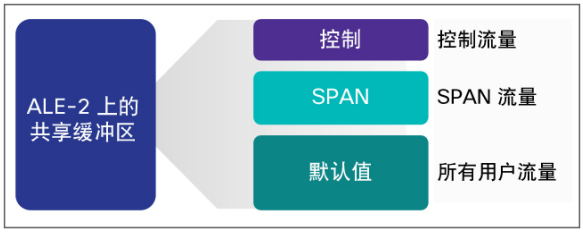
- DROP：默认服务池
- SPAN：SPAN 服务池
- SUP：控制服务池

请注意，ALE 可以支持四个服务池：DROP、NODROP、SPAN 和 SUP。当前 Cisco Nexus 9300 平台上未使用 NODROP 服务池。它可以在将来用于优先级流量控制 (PFC)。

应用枝叶引擎 2 上的缓冲区

ALE-2 具有由 ALE-2 上的所有端口动态共享的 25 MB 缓冲区内存，用于所有流量。它结合了 ALE 具有的三个区域，但是保持相同的服务池定义：控制、SPAN 和默认（图 8）。

图 8. ALE-2 上的缓冲区服务池



在缓冲区监控命令中标识 ALE-2 服务池的方式与 ALE 缓冲区服务池相同：

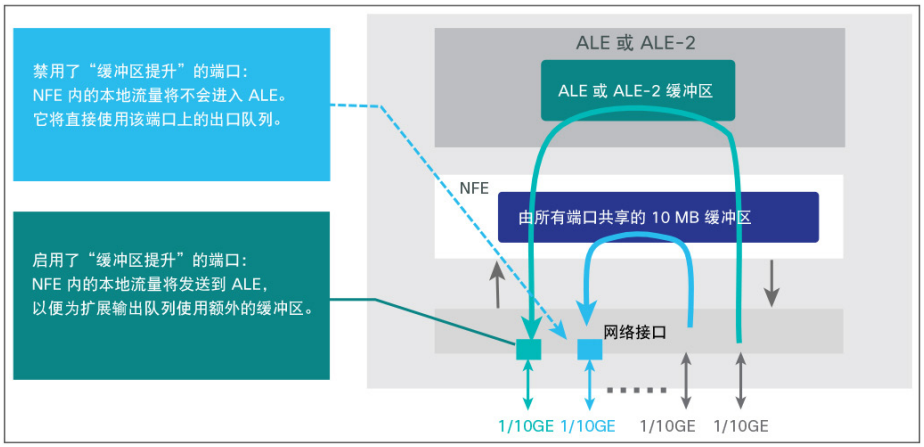
- DROP：默认服务池
- SPAN：SPAN 服务池
- SUP：控制服务池

Cisco Nexus 9300 平台上的缓冲区提升功能

与具有相同或相似端口密度的其他接入交换机平台相比，Cisco Nexus 9300 平台的一项显著优势是其较大的缓冲区大小。除了 NFE 上的 12 MB 缓冲区之外，它还具有由 ALE 提供的 40 MB 缓冲区，或由 ALE-2 提供的 25 MB 缓冲区。ALE 上的 40 MB 缓冲区中的10 MB 保留给 NFE 上的两个 1 和 10 千兆以太网前面板端口之间的本地流量使用。

如果源端口的速度高于目的端口（例如从 10 千兆以太网端口至 1 千兆以太网端口），或者本地流量是突发流量或处于 in-cast 模式，则即使对于 NFE 本地流量而言，额外的缓冲区空间也是值得拥有的。由于 NFE 执行数据包查找和转发，因此两个 NFE 前面板端口之间的本地流量无需进入 ALE 进行转发处理。但是，数据包需要发送到其 ALE 才可利用额外的 ALE 缓冲区。为此引入了缓冲区提升功能（图 9）。

图 9. Cisco Nexus 9300 平台缓冲区提升功能



在某个 NFE 前面板端口上启用缓冲区提升功能后，从另一个 NFE 前面板端口流向此端口的单播流量都会被重定向到 ALE 或 ALE-2，以便为本地流量使用额外的缓冲区空间。ALE 和 ALE-2 会将 NFE 的流量发回，以将数据包转发到出口端口。在 ALE 上，10 MB 的缓冲区空间专用于发夹式 NFE 本地流量。在 ALE-2 上，发夹式本地流量与所有其他流量共享 25 MB 的缓冲区。在 NFE 前面板端口上禁用了缓冲区提升功能之后，NFE 不将从其他本地端口到此端口的流量重定向到 ALE。相反，它直接将流量转发到此出口端口。

缓冲区提升是一个出口端口配置属性。它可以按照端口来启用或禁用。默认情况下，会在所有 NFE 1 和 10 千兆以太网前面板端口上启用它。缓冲区提升仅适用于本地单播流量。它不会更改组播流量转发。

Cisco Nexus 9300 平台出口队列和扩展输出队列

Cisco Nexus 9300 平台交换机使用简单而有效、基于类别的出口排队机制来处理链路拥塞。Cisco Nexus 9300 系列交换机使用以下类型的流量类进行排队：

- 控制流量类
- SPAN 流量类
- 用户流量类

控制流量类和 SPAN 流量类是在系统内部定义，对用户透明。网络控制平面流量，包括网络控制协议（例如开放最短路径优先 [OSPF]、边界网关协议 [BGP] 和网络时间协议 [NTP]）的流量，分类到控制类中。

SPAN 流量，包括本地 SPAN 和远程封装交换端口分析器 (ERSPAN)，分到 SPAN 类中。控制流量作为最高优先级处理，具有保留的缓冲区资源。SPAN 流量在端口上的优先级最低，会使用剩余的带宽。

四个用户流量类用于出口排队：

- c-out-q-default：出口默认队列
- C-out-q1：出口队列 1
- C-out-q2： 出口队列 2
- C-out-q3： 出口队列 3

用户可以在入口端口上定义和应用流量分类规则，以控制流量映射到四个类别的方式。可以根据 IP 差分服务代码点 (DSCP) 或优先顺序、IEEE 802.1q 服务类别 (CoS)、IP 访问控制列表 (ACL)、MAC 地址 ACL 等将流量分类。会为每个类分配服务质量 (QoS) 组编号，作为其在交换机系统中的内部标识。QoS 组编号的范围可以从 0 至 3。

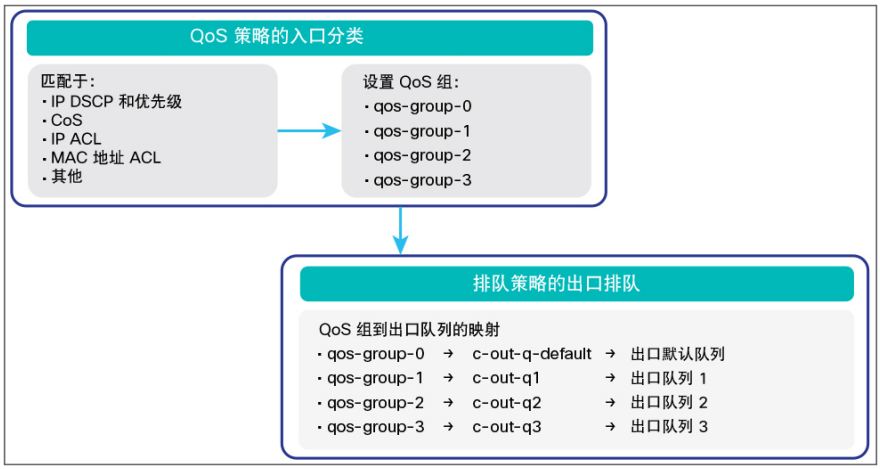
在出口端口上，QoS 组映射到流量类，如下所示：

- qos-group-0 > c-out-q-default（出口默认队列）
- qos-group-1 > c-out-q1（出口队列 1）
- qos-group-2 > c-out-q2（出口队列 2）
- qos-group-3 > c-out-q3（出口队列 3）

用户可以定义每个类的排队策略。分类到入口端口上的 QoS 组之后，将根据出口端口上为此 QoS 组定义的出口排队策略来处理流量。

图 10 显示入口流量分类和出口排队流程。

图 10. Cisco Nexus 9300 平台 QoS 分类和排队

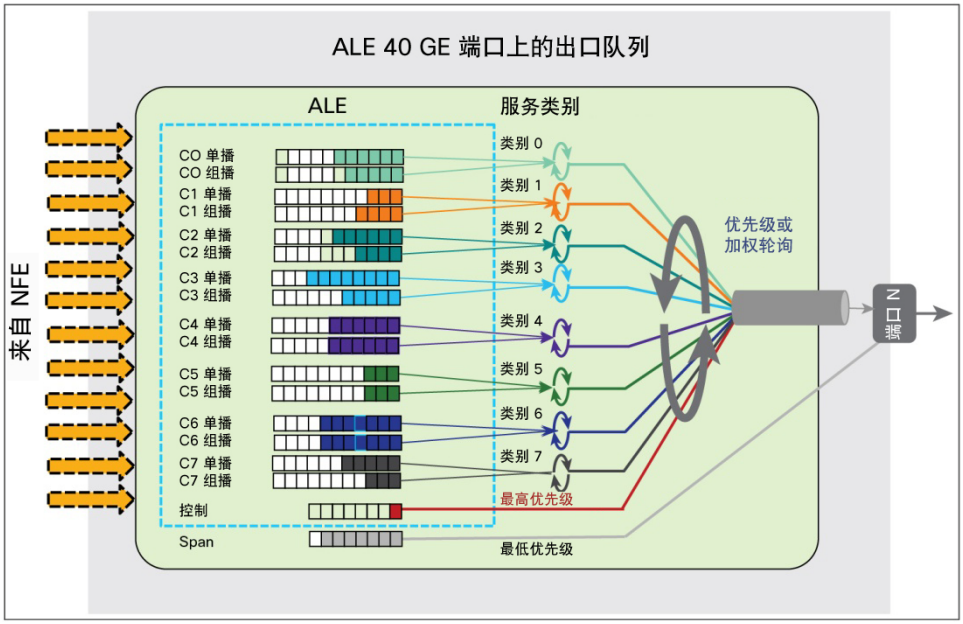


ALE 和 ALE-2 40 千兆以太网端口

图 11 描绘了由 ALE 或 ALE-2 提供的 40 千兆以太网端口的出口队列结构。队列使用六个流量类构建，包括控制流量类、SPAN 流量类和四个用户可定义的类（在内部由 QoS 组标识）。在每个用户定义的类中，存在一个单播队列和一个组播队列。因此，每个 40 千兆以太网端口具有以下出口队列：

- 一个控制流量队列
- 一个 SPAN 流量队列
- 四个单播队列
- 四个组播队列

图 11. ALE 和 ALE-2 上的 40 千兆以太网端口上的输出队列



40 千兆以太网端口上的这些出口队列使用 ALE 上的 10 MB 入口直接流量缓冲区，或者，如果端口在 ALE-2 上，则它们通过 ALE-2 与其他流量共享 25 MB 缓冲区。

NFE 前面板端口上的出口和扩展出口队列

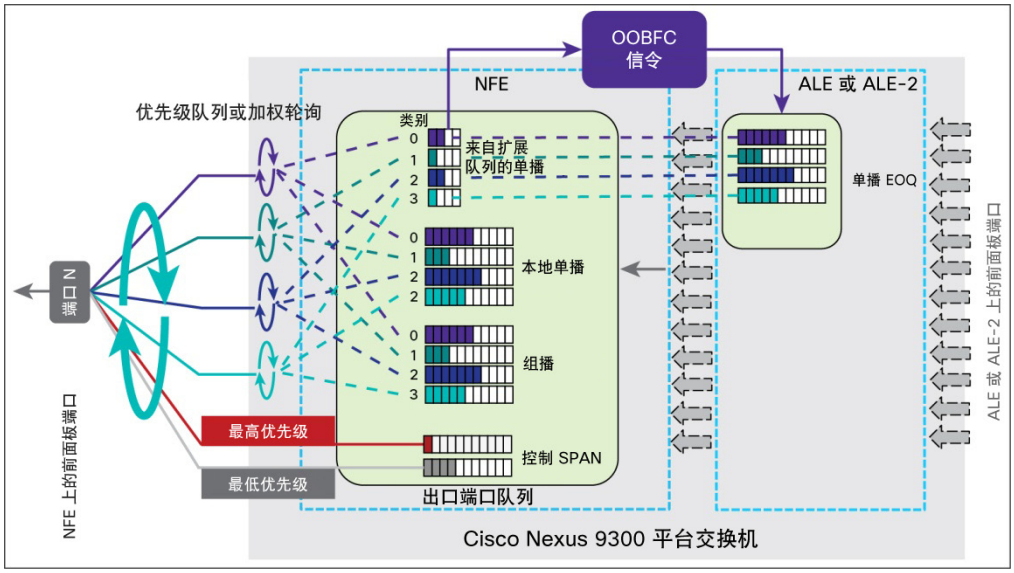
与 ALE 上的 40 千兆以太网端口相似，NFE 上的每个前面板端口具有控制流量、SPAN 流量、组播流量和单播流量的一组出口队列。此外，每个 NFE 端口有四个 OOBFC 单播队列。这些队列用于 ALE 上的单播扩展出口队列。因此，在每个 NFE 1 和 10 千兆以太网出口端口上可看到以下队列：

- 一个控制流量队列
- 一个 SPAN 流量队列
- 四个组播队列
- 四个单播队列（用于本地单播流量）
- 四个 OOBFC 单播队列（这些队列用于 OOBFC 控制的单播流量，包括迂回的 NFE 本地单播流量，以及从 ALE 40 千兆以太网端口到 NFE 前面板端口的出口直接单播流量。）

在 ALE 上，每个 NFE 出口端口有 4 个相应的单播扩展输出队列 (EoQ)。按照出口端口和单播类，NFE 使用 OOBFC 信令通道来通知 ALE 何时停止或何时恢复向 NFE 发送流量。当 ALE 被指示停止将流量发送到 NFE 时，它会使用自己的缓冲区，将相应 EoQ 中的数据包排队。因此，NFE 上的出口单播队列扩展到 ALE EoQ，以使用额外的 ALE 缓冲区资源。ALE 上可以利用 OOBFC 发出的 EoQ 信号的单播流量，包括来自 ALE 40 千兆以太网端口至 NFE 前面板端口的出口直接流量，以及在两个 NFE 端口之间流动的发夹式本地流量。

图 12 显示了 Cisco Nexus 9300 平台交换机的 NFE 前面板端口的 NFE 出口队列和 ALE EoQ。

图 12. Cisco Nexus 9300平台 NFE 前面板端口出口和扩展出口队列



Cisco Nexus 9300 平台上的加权轮询和优先级排队

Cisco Nexus 9300 系列交换机使用加权轮询 (WRR) 和优先级排队 (PQ) 机制来管理 NFE 和 ALE 上的出口队列以及扩展出口队列。

以下是是四个用户流量类的默认排队策略：

- c-out-q3
- c-out-q2
- c-out-q1
- c-out-q-default

```
n9396-1# sh policy-map type queuing default-out-policy


  Type queuing policy-maps
  ========================

  policy-map type queuing default-out-policy
    class type queuing c-out-q3
      priority level 1
    class type queuing c-out-q2
      bandwidth remaining percent 0
    class type queuing c-out-q1
      bandwidth remaining percent 0
    class type queuing c-out-q-default
      bandwidth remaining percent 100
n9396-1#
```

在 WRR 排队策略中，带宽可以按照链路带宽的百分比来定义，也可以按照剩余带宽的百分比定义。

当您使用优先级排队时，其他非优先级队列（WRR 队列）只能将带宽定义为占剩余带宽的百分比。Cisco Nexus 9300 平台交换机最多支持 3 个优先级队列。这些队列必须从策略映射配置中的类 c-out-q3 开始，然后按顺序移至 c-out-q2 和 c-out-q1。

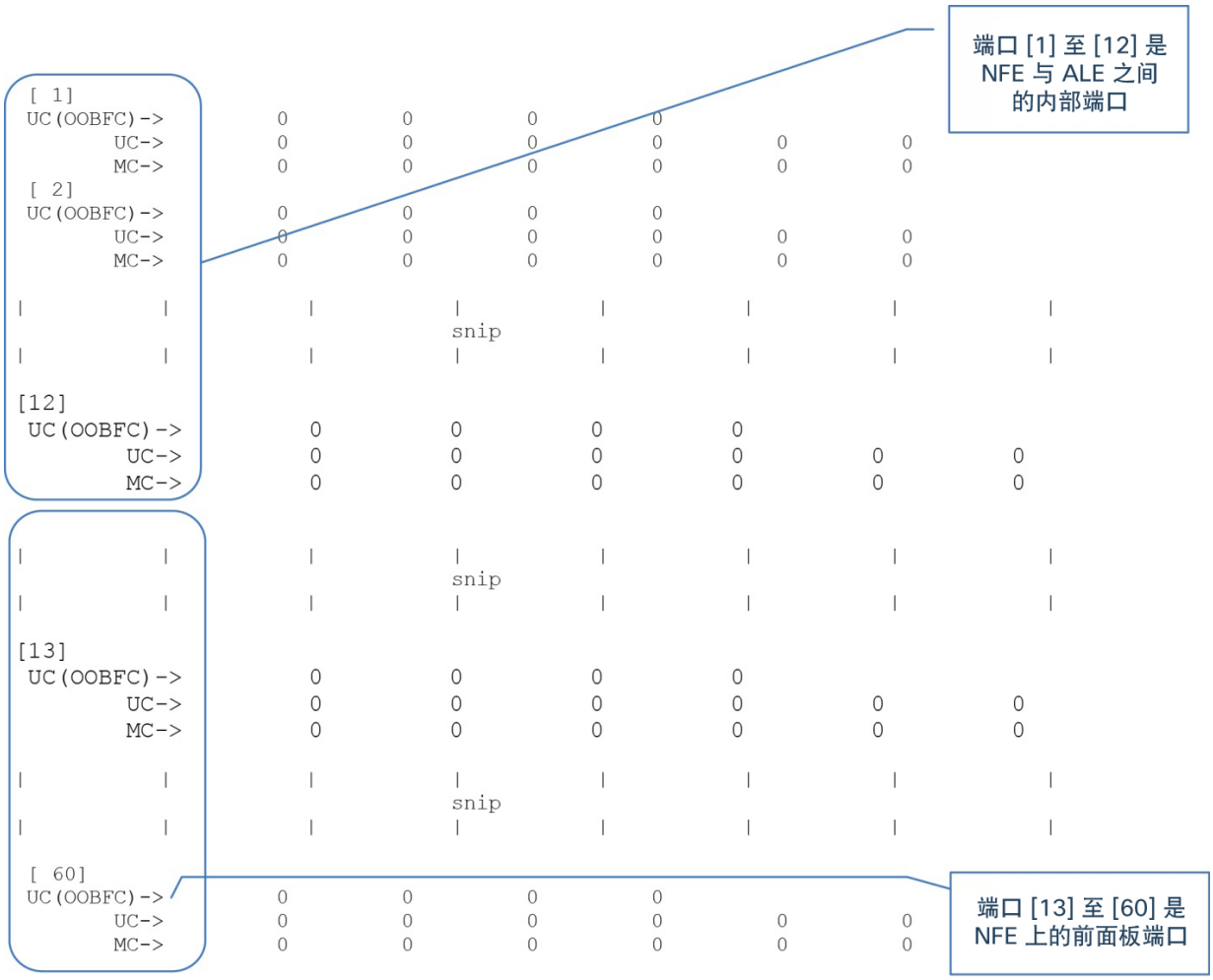
出口队列和扩展出口队列监控

NFE 上的缓冲区和队列监控

以下示例显示了 Cisco Nexus 9396PX 交换机的 NFE 上的缓冲区和队列监控结果。**show hardware internal buffer info pkt-state detail** 命令按照每个流量类和每个队列显示 NFE 上所有端口的动态缓冲区统计信息。每个端口有 6 个类：Q3、Q2、Q1、Q0、CPU 和 SPAN。类 Q3 至 Q0 具有 OOBFC 单播队列、非 OOBFC 单播队列和组播队列。类 CPU 和 SPAN 均具有单播队列和组播队列。

```
n9396-1# show hardware internal buffer info pkt-stats  detail

slot  1
=======


INSTANCE: 0
============


|--------------------------------------------------------------------|
             Output Shared Service Pool Buffer Utilization (in cells)
                           SP-0       SP-1       SP-2       SP-3
|--------------------------------------------------------------------|

Total Instant Usage             0          0          0          0
Remaining Instant Usage     29938          0      14346       6344
Peak/Max Cells Used            33          0       1531        163
Switch Cell Count           29938          0      14346       6344
|--------------------------------------------------------------------|

|------------------------------------------------------------------------|
|              Instant Buffer utilization per queue per port             |
|        Each line displays the number of cells utilized for a given     |
|                     port for each QoS queue                            |
|             One cell represents approximately 208 bytes                |
|--------------+---------+---------+---------+---------+---------+------|
|ASIC Port         Q3        Q2        Q1        Q0        CPU      SPAN   |
|--------------+---------+---------+---------+---------+---------+------|
```

每个 NFE 端口 6 个类别



注：此命令输出显示了 NFE 上所有活动端口的缓冲区统计信息，从 NFE 与 ALE 或 ALE-2 之间的内部端口开始，之后是 NFE 前面板端口。上例取自 Cisco Nexus 9396PX 交换机，该交换机在 NFE 与 ALE 之间有 12 个内部 40 千兆以太网端口，在 NFE 上有 48 个 1 和 10 千兆以太网前面板端口。因此，命令输出显示 60 个 ASIC 端口：

- 端口 1 至 12：NFE 与 ALE 之间的内部端口
- 端口 13 至 60：NFE 上的前面板端口

以上缓冲区监控命令的变化显示了各队列中缓冲区利用率的峰值。高水位监控的输出示例如下所示：

```
n9396-1# show hardware internal buffer info pkt-stats peak

slot  1
=======


INSTANCE: 0
============


|--------------------------------------------------------------------|
             Output Shared Service Pool Buffer Utilization (in cells)
                           SP-0       SP-1       SP-2       SP-3
|--------------------------------------------------------------------|

Total Instant Usage              0          0          0          0
Remaining Instant Usage      29938          0      14346       6344
Peak/Max Cells Used             33          0       1531        163
Switch Cell Count            29938          0      14346       6344
|--------------------------------------------------------------------|

|-------------------------------------------------------------------------|
|               Peak Buffer utilization per queue per port                |
|        Each line displays the number of cells utilized for a given      |
|                     port for each QoS queue                             |
|             One cell represents approximately 208 bytes                 |
|--------------+---------+---------+---------+---------+---------+--------+|
|ASIC Port          Q3        Q2        Q1        Q0        CPU       SPAN   |
|--------------+---------+---------+---------+---------+---------+--------+|

 [ 1]
 UC(OOBFC)->          0         0         0         0
        UC->          0         0         0         3        74         0
        MC->          0         0         0         1         0         0
 [ 2]
 UC(OOBFC)->          0         0         0         0
        UC->          0         0         0         1        74         0
        MC->          0         0         0         1         0         0
 [ 3]
 UC(OOBFC)->          0         0         0         0
        UC->          0         0         0         1        72         0
        MC->          0         0         0         1         1         0
 [ 4]
 UC(OOBFC)->          0         0         0         0
        UC->          0         0         0         3        73         0
        MC->          0         0         0         1         1         0
[20]
 UC(OOBFC)->          0         0         0       224
        UC->          0         0         0         0         8         0
        MC->          0         0         0         0         1         0
```

ALE 和 ALE-2 上的缓冲区和队列监控

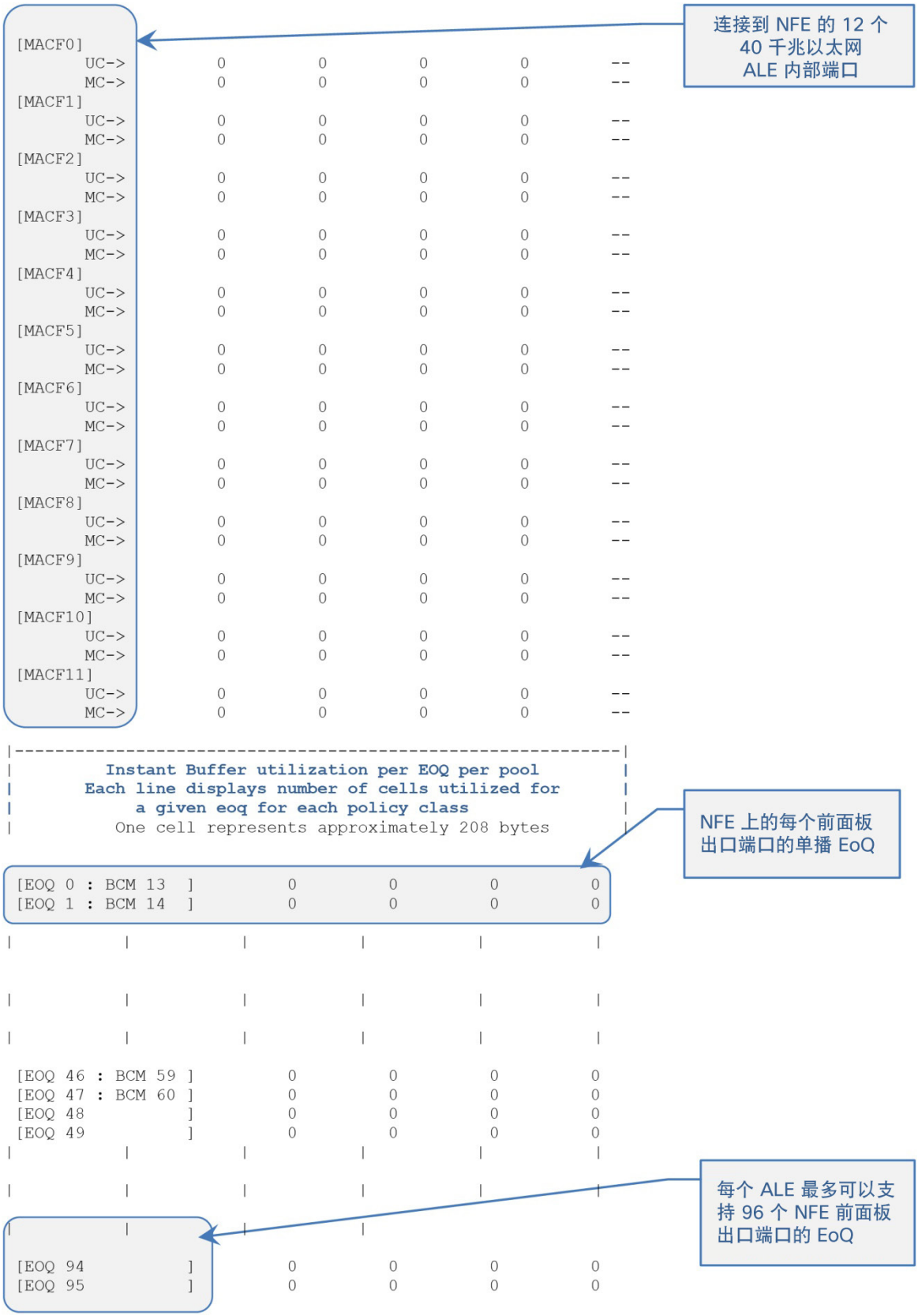
show hardware internal ns buffer info pkt-stats 命令用于监控 ALE 的缓冲区利用率和队列统计信息。

```
n9396-1# show hardware internal ns buffer info pkt-stats detail

slot  1
=======

INSTANCE: 0
============


Ingress Straight Traffic:
-------------------------

|--------------------------------------------------------------------|
|                   Shared Service Pool Buffer Utilization (in cells) |
|                     One cell represents approximately 208 bytes    |
'                                                                    '
.                                                                    .
|                            DROP       NODROP       SPAN        SUP       |
|--------------------------------------------------------------------|
Total Instant Usage               0          0          0          0
```

```
Remaining Instant Usage      47896          0       256       500
Shared Cells Count           28696          0       256       500
Total Cells Count            47896          0       256       500

|-------------------------------------------------------------------|
|          Instant Buffer utilization per port per pool             |
|     Each line displays number of cells utilized for a given       |
|                  port for each policy class                       |
|          One cell represents approximately 208 bytes              |
|-------------+---------+---------+---------+---------+---------+-|
|ASIC Port         Q0        Q1        Q2        Q3        SUP    |
|-------------+---------+---------+---------+---------+---------+-|
[MACN0]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN1]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN2]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN3]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN4]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN5]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN6]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN7]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN8]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN9]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN10]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
[MACN11]
       UC->          0         0         0         0        --
       MC->          0         0         0         0        --
```

ALE 上的 12 个 40 千兆以太网前面板端口

```
Ingress Hairpin Traffic:
------------------------

|-------------------------------------------------------------------|
|                  Shared Service Pool Buffer Utilization (in cells) |
|                    One cell represents approximately 208 bytes    |
|                                                                   |
|                            DROP      NODROP      SPAN       SUP  |
|-------------------------------------------------------------------|
Total Instant Usage              0          0         0         0
Remaining Instant Usage      47896          0       256       500
Shared Cells Count           38296          0       256       500
Total Cells Count            47896          0       256       500
```



```
Egress Straight Traffic:
------------------------

|-------------------------------------------------------------------------|
|                  Shared Service Pool Buffer Utilization (in cells)      |
|                    One cell represents approximately 208 bytes          |
|                                                                         |
|                            DROP       NODROP       SPAN       SUP      |
|-------------------------------------------------------------------------|
Total Instant Usage                0          0          0          0
Remaining Instant Usage        97048          0        256        500
Shared Cells Count             87448          0        256        500
Total Cells Count              97048          0        256        500

[MACF0]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF1]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF2]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF3]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF4]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF5]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF6]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF7]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF8]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF9]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF10]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --
[MACF11]
       UC->          0          0          0          0         --
       MC->          0          0          0          0         --

[EOQ 0 : BCM 13  ]          0          0          0          0
[EOQ 1 : BCM 14  ]          0          0          0          0
 |          |          |          |          |          |

|          |          |          |          |          |

|          |          |          |          |          |

 [EOQ 46 : BCM 59 ]         0          0          0          0
 [EOQ 47 : BCM 60 ]         0          0          0          0
 [EOQ 48          ]         0          0          0          0
 [EOQ 49          ]         0          0          0          0
|          |          |          |          |          |

|          |          |          |          |          |

|          |          |          |          |          |

 [EOQ 94          ]         0          0          0          0
 [EOQ 95          ]         0          0          0          0

n9396-1#
```

连接到 NFE 的 12 个 40 千兆以太网 ALE 内部端口

每个 NFE 前面板出口端口的单播 EoQ，最多 96 个

接口上的队列监控

```
n9396-1# sh queuing interface e1/1 summary

slot  1
=======


Egress Queuing for Ethernet1/1 [System]
------------------------------------------------------------------
QoS-Group# Bandwidth% PrioLevel                Shape
                                     Min          Max        Units
------------------------------------------------------------------
     3             -         1           -            -     -
     2             0         -           -            -     -
     1             0         -           -            -     -
     0           100         -           -            -     -
+-------------------------------------------------------------------+
|                             QOS GROUP 0                           |
+-------------------------------------------------------------------+
|                |  Unicast       | OOBFC Unicast  |  Multicast     |
+-------------------------------------------------------------------+
|        Tx Pkts |               0|      5325011301|               0|
|        Tx Byts |               0|   5954391263104|               0|
|   Dropped Pkts |               0|               0|               0|
|   Dropped Byts |               0|               0|               0|
|   Q Depth Byts |               0|               0|               0|
+-------------------------------------------------------------------+
|                             QOS GROUP 1                           |
+-------------------------------------------------------------------+
|                |  Unicast       | OOBFC Unicast  |  Multicast     |
+-------------------------------------------------------------------+
|        Tx Pkts |               0|               0|               0|
|        Tx Byts |               0|               0|               0|
|   Dropped Pkts |               0|               0|               0|
|   Dropped Byts |               0|               0|               0|
|   Q Depth Byts |               0|               0|               0|
+-------------------------------------------------------------------+
|                             QOS GROUP 2                           |
+-------------------------------------------------------------------+
|                |  Unicast       | OOBFC Unicast  |  Multicast     |
+-------------------------------------------------------------------+
|        Tx Pkts |               0|               0|               0|
|        Tx Byts |               0|               0|               0|
|   Dropped Pkts |               0|               0|               0|
|   Dropped Byts |               0|               0|               0|
|   Q Depth Byts |               0|               0|               0|
+-------------------------------------------------------------------+
|                             QOS GROUP 3                           |
+-------------------------------------------------------------------+
|                |  Unicast       | OOBFC Unicast  |  Multicast     |
+-------------------------------------------------------------------+
|        Tx Pkts |               0|               0|               0|
|        Tx Byts |               0|               0|               0|
|   Dropped Pkts |               0|               0|               0|
|   Dropped Byts |               0|               0|               0|
|   Q Depth Byts |               0|               0|               0|
+-------------------------------------------------------------------+
```

```
+-------------------------------------------------------------------+
|                      CONTROL QOS GROUP 4                          |
+-------------------------------------------------------------------+
|                |  Unicast       | OOBFC Unicast  |  Multicast     |
+-------------------------------------------------------------------+
|        Tx Pkts |            8714|               0|               0|
|        Tx Byts |         1024410|               0|               0|
|   Dropped Pkts |               0|               0|               0|
|   Dropped Byts |               0|               0|               0|
|   Q Depth Byts |               0|               0|               0|
+-------------------------------------------------------------------+
|                      SPAN QOS GROUP 5                             |
+-------------------------------------------------------------------+
|                |  Unicast       | OOBFC Unicast  |  Multicast     |
+-------------------------------------------------------------------+
|        Tx Pkts |               0|               0|               0|
|        Tx Byts |               0|               0|               0|
|   Dropped Pkts |               0|               0|               0|
|   Dropped Byts |               0|               0|               0|
|   Q Depth Byts |               0|               0|               0|
+-------------------------------------------------------------------+

Port Ingress Statistics
---------------------------------------------------
Ingress MMU Drop Pkts                             0
Ingress MMU Drop Bytes                            0


Port Egress Statistics
---------------------------------------------------
WRED Drop Pkts                                0
NS Straight EOQ(qos-group-0) Drop Pkts                 893
NS BufferBoost EOQ(qos-group-0) Drop Pkts                 0

PFC Statistics
---------------------------------------------------
TxPPP:                    0, RxPPP:                    0
---------------------------------------------------
 COS QOS Group   TxPause   TxCount         RxPause         RxCount
   0         -  Inactive         0        Inactive               0
   1         -  Inactive         0        Inactive               0
   2         -  Inactive         0        Inactive               0
   3         -  Inactive         0        Inactive               0
   4         -  Inactive         0        Inactive               0
   5         -  Inactive         0        Inactive               0
   6         -  Inactive         0        Inactive               0
   7         -  Inactive         0        Inactive               0
---------------------------------------------------

n9396-1#
```

队列限制控制

在 Cisco Nexus 9300 平台交换机上，可以按端口和类定义队列限制。它提供一种机制来防止指定端口或指定流量类占用过多的缓冲区资源，导致缓冲区无法满足其他端口或流量类的需求。队列限制还可用于在需要时向指定端口或指定流量类分配更多缓冲区空间。

Cisco Nexus 9300 系列交换机支持静态队列限制和动态队列限制。静态队列限制指定特定流量类在队列中具有的准确字节数、KB 或 MB 数量。还可以将静态限制指定为允许将数据包保留在队列中的持续时间长度，以毫秒为单位。如果在一些端口上需要对某个特定流量类进行精确的缓冲区和队列控制，则静态队列限制非常有帮助，

而动态队列限制则提供了按照端口和类队列限制灵活、动态地进行控制的方法。通过从表 3 列出的选项中选择动态队列限制因素，用户可以指定每个端口和每个类在任何给定时间上可以使用的可用缓冲区空间量。

表 3. 动态队列限制因素

动态队列限制因素	作为可用缓冲区空间百分比的队列限制
选项 0：1/128	1%
选项 1：1/64	2%
选项 2：1/32	3%
选项 3：1/16	6%
选项 4：1/8	11%
选项 5：1/4	20%
选项 6：1/2	33%
选项 7：1	50%
选项 8：2	67%
选项 9：4	80%
选项 10：8	89%

动态队列限制可实现缓冲区空间的最佳利用率，同时防止队列使用过多缓冲区资源。默认队列限制设置为选项 8，它允许每个类和每个队列最多使用 67% 可用缓冲区空间。如果某个端口或某个特定类的流量预计是突发性的，用户可以将其队列限制更改为选项 9 或 10 以利用最多 89% 的可用带宽。

ALE 和 ALE-2 上的突发配置文件和流量优先排序

ALE 和 ALE-2 突发配置文件

ALE 和 ALE-2 提供三个突发配置文件：

- 突发：优化突发
- 网状：优化网状
- 超突发：优化超突发

网状是默认突发模式。但是，如果已知通过 Cisco Nexus 9300 平台交换机的流量是突发性的，建议使用突发模式。以下全局命令可用于更改突发配置文件。命令更改不会请求系统重新启动。

```
n9396-1(config)# hardware qos ns-buffer-profile ?
  burst        Burst optimized
  mesh         Mesh optimized
  ultra-burst  Ultra burst optimized
```

CLI 命令 **show hardware qos ns-buffer-profile** 会显示交换机运行配置中的当前突发配置文件。

```
n9396-1# show hardware qos ns-buffer-profile
NS Buffer Profile: Burst optimized
n9396-1#
```

ALE 和 ALE-2 流优先排序

ALE 和 ALE-2 具有可以根据流的持续时间确定其优先顺序的内置智能。在持续时间长的流与持续时间短的突发流混合使用的情况下，ALE 和 ALE-2 可以识别并优先处理持续时间短的流。在链路拥塞而交换机必须丢弃一些数据包时，ALE 和 ALE-2 将首先丢弃持续时间长的流中的数据包，同时允许持续时间短的流通过而不会丢失数据包。

图 13 显示了 Cisco Nexus 9396PX 交换机上的流优先排序测试结果。在测试中，恒定的 10 千兆以太网流量和短期的 10 千兆以太网突发流量发送到 NFE 上的每个出口 10 千兆以太网端口。结果显示，恒定流量丢失了数据包，但是突发流量正常通过，而不会丢失数据包。

图 13. ALE 和 ALE-2 流优先排序演示

	Tx Port	Rx Port	Traffic Item	Tx Frames	Rx Frames	Frames Delta	Loss %	Tx Frame Rate	Rx Frame Rate	Tx L1 Rate (bps)	Rx L1 Rate (bps)	Rx Bytes
1	40GE-9396-2/9	10GE-9396-1/1	const-9396	388,470,164	388,469,183	981	0.000	2,349,389.151	2,349,389.651	9,999,000,225...	9,923,821,884...	197,342,3...
2	40GE-9396-2/9	10GE-9396-1/1	burst-9396	5,000	5,000	0	0.000	0.000	0.000	0.000	0.000	2,540,000
3	40GE-9396-2/10	10GE-9396-1/2	const-9396	388,470,185	388,469,842	343	0.000	2,349,388.571	2,349,389.571	9,998,997,757...	9,923,821,548...	197,342,6...
4	40GE-9396-2/10	10GE-9396-1/2	burst-9396	5,000	5,000	0	0.000	0.000	0.000	0.000	0.000	2,540,000
5	40GE-9396-2/11	10GE-9396-1/3	const-9396	388,471,352	388,470,940	412	0.000	2,349,388.707	2,349,388.707	9,998,998,335...	9,923,817,896...	197,343,2...
6	40GE-9396-2/11	10GE-9396-1/3	burst-9396	5,000	5,000	0	0.000	0.000	0.000	0.000	0.000	2,540,000

许多数据中心应用将持续时间长的流用于数据传输，同时将持续时间短的流用于状态同步或发出请求。这些持续时间短的流对于数据包丢失或延迟更敏感。通过优先处理这些持续时间短的流而非持续时间长的数据传输流，ALE 或 ALE-2 流优先排序功能可以帮助改进数据中心应用的性能。

结论

Cisco Nexus 9300 平台交换机旨在提供高性能、具成本效益的网络连接和广泛的可编程功能集，以支持现代数据中心的运营模式。该平台以紧凑的固定配置外形提供业界领先的 1、10 和 40 千兆以太网端口密度，使组织能够将数据中心网络接入层从 1 千兆以太网迁移到 10 千兆以太网，以便进行主机访问，并从 10 千兆以太网迁移到 40 千兆以太网，以便上行链接到数据中心汇聚和主干层，Cisco Nexus 9300 平台交换机交换机上的扩展缓冲容量和增强型出口排队架构可帮助确保多元化的动态网络环境中的应用性能。

相关详细信息

有关详细信息，请访问：http://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/index.html。

美洲总部
Cisco Systems, Inc.
加州圣何西

亚太地区总部
Cisco Systems (USA) Pte.Ltd.
新加坡

欧洲总部
Cisco Systems International BV
荷兰阿姆斯特丹

思科在全球设有 200 多个办事处。 地址、电话号码和传真号码均列在思科网站 www.cisco.com/go/offices 中。

美国印刷 C11-732452-00 11/14