白皮书

T640 路由节点和 TX 矩阵™ 平台：体系结构

Juniper Networks, Inc.
1194 North Mathilda Avenue
Sunnyvale, CA 94089 USA
408 745 2000 or 888 JUNIPER
www.juniper.net

文档编号：200089-001SC

介绍

Juniper 网络公司在 2002 年 4 月开始发售 T640 路由节点。T640 路由节点是 Juniper 网络公司一个路由器新家族的第一个成员，该路由器家族旨在解决供应商网络中核心路由器的部署生命周期较短的问题。在 2002 年，核心路由器的部署生命周期通常为一年半至两年，但是服务供应商要求下一代路由器能够随着他们的网络灵活地增长，并且具有五年或更长的部署生命周期。为了满足这些要求，T640 路由节点按照满足严格的带宽密度、数据包转发性能、易用性和高可用性要求进行设计。这些设计目标包括：

- 提供总计每秒 6.4 亿个数据包(pps)的*数据包转发性能*，用以支持在网络边缘提供的现有的业务，并为提供新的创收业务提供便利。
- 在半机架外形中提供 32xOC-192 ／ STM-64 或 128xOC-48 ／ STM-16 接口的*带宽密度*。这些特性将使供应商得到所需的足够的机架空间，支持以与线路可用性和用户要求相匹配的速率部署设备。
- 提供在硬件中实施的 IP *业务交付性能*，用以在高速链路上支持汇聚的差分业务类别、数据包过滤、警管、速率限制和流量监控，同时不会降低数据包转发性能。
- 通过一致的用户界面中提供*单一软件介质*，以确保在所有平台和接口的硬件中支持所有的特性。
- 提供*高可用*的特性，确保系统不会发生单点故障。这种高可用性是由热插拔和冗余硬件组件以及基本系统体系结构和 JUNOS ™ 软件的可靠性来支持的。
- 提供全面一致的*安全工具*(过滤、速率限制、跟踪、日志记录、源地址验证等)，用以支持从网络中的任何位置(核心或边缘)或接口(DS-1 到 OC-192 ／ STM-64)管理安全性。

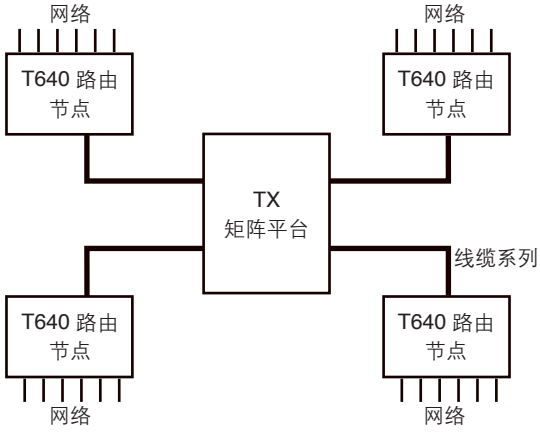
除了满足供应商对于在单机箱系统中提高带宽密度和转发性能的要求之外，Juniper 网络公司的 T640 路由节点战略的一个重要方面在于提供了一个可以轻松升级以扩展到采用多机箱路由器配置的太比特级带宽的平台。作为单一大路由器运行的多机箱路由器称为*路由矩阵*。路由矩阵使供应商能够按照预算和网络流量负荷的要求，以任意增长幅度来增加系统的规模，从而延长核心路由器设备的部署生命周期。T640 路由节点由于免除了每隔一年就更换旧设备的成本，因而大幅度节约前期设备购置成本，由于无需每隔几年就验证和安装新设备，因而节约后期运维成本，此外，该路由节点通过避免侵入式升级周期，提供更高的网络可用性。

Juniper 网络公司路由矩阵

Juniper 网络公司的路由矩阵(如图 1 所示)由三种主要部件组成：

- 四个 T640 路由节点，这些路由节点为路由矩阵提供网络接口，并执行分布式数据包转发决策。路由矩阵配置中的 T640 路由节点因为包含路由矩阵系统的网络接口卡有时也被称为线路卡机箱(LCC)。
- 一个 TX 矩阵™ 平台，该平台为路由矩阵执行路由协议，维护系统状态，并提供使单个 T640 路由节点互连的交换结构(switch fabric)的核心。TX 矩阵平台有时也被称为交换卡机箱(SCC)。
- 一组线缆，这些线缆将各机箱的数据和控制板连接为统一的路由矩阵。

图 1：Juniper 网络公司路由矩阵



路由矩阵的主要应用如下：

- 避免路由器插槽耗尽，这是目前最为紧迫的任务。发生核心路由器插槽耗尽的情况包括部署大量的边缘路由器、必须为空前增长的客户提供服务以及将路由器交叉连接来提供物理路径冗余等。Juniper 网络公司的路由矩阵通过以下途径解决这一问题：使每个路由器支持的前端插槽的数量增加四倍；允许利用后端带宽实施路由矩阵内部的交叉连接；以及提供一种占用资源规模适度的解决方案，支持在现有 POP 中的部署，并避免物理设施的升级。

- 支持部署 ATM、帧中继、话音及其他数据包业务在 IP ／ MPLS 基础设施上传输的融合式 Infranet。Infranet 在容量、时延、抖动和高可用性方面对核心路由器要求严格。Juniper 网络公司路由矩阵解决这一问题的途径是：在硬件中执行所有的数据包处理和转发；实施无阻塞的交换结构；通过面向高优先级流量的结构提供低时延；以及提供高度冗余的系统，确保不会发生单点故障。
- 整合 POP 层，这些分层中使用逻辑路由器来简化网络的物理拓扑，将大型系统分解为独立的控制和管理领域，并对核心、汇聚和边缘/分布功能进行垂直整合。尽管路由矩阵能够成功地支持这些任务，然而由于供应商永远都需要节点级的冗余性来确保网络的高可用性，因此路由矩阵*不*应被视为一种“POP in a box”解决方案。

路由矩阵的设计目标

路由矩阵通过达到以下设计目标，延长核心路由器的部署生命周期：

- 将四个 T640 路由节点和 TX 矩阵平台的数据、管理和控制板相结合，使路由矩阵看上去是一个具有统一的数据、管理和控制板的单一路由器。
- 提供一种在数据包转发性能、数据包处理特性和带宽密度方面四倍于单台 T640 容量的，作为单一路由节点运行的多机箱解决方案。
  - 一个路由矩阵包含 32 个灵活的 PIC 集中器(FPC)插槽，每个插槽支持 40-Gbps 的吞吐量(4x8 FPC ／ T640 路由节点)。
  - 一个路由矩阵提供的总吞吐量为 2.56 Tbps（4 x 640 Gbps)。
- 利用已经在世界上最大的服务供应商网络中成功部署的现有 Juniper 网络公司的技术。
  - 模块化 JUNOS 软件架构，该架构支持 Juniper 网络公司首创的单一源和版本串(release train)模型
  - T640 芯片组
  - T640 路由节点机箱组件
  - T640 路由节点物理接口卡(PIC)和灵活的 PIC 集中器(FPC)，以保护客户的现有投资

T640 路由节点数据包转发体系结构

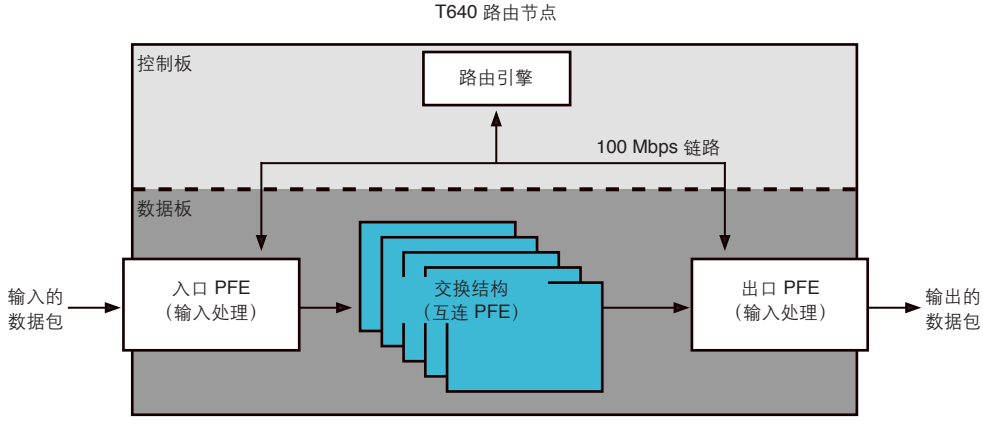
路由矩阵的数据包转发体系结构是对单独的 T640 路由节点的数据包转发体系结构的直接扩展，因此，在全面了解路由矩阵的体系结构之前，您需要首先了解 T640 路由节点的体系结构。

T640 路由节点体系结构的组件

T640 路由节点包括两种基本的体系结构组件(如图 2 所示)：

- *控制板*负责执行路由协议，维护路由表，管理控制系统接口的软件流程，以及管理用户对 T640 路由节点的访问。控制板是由在系统的路由引擎上运行的 JUNOS 软件实施。
- *数据板*负责在通过交换结构将数据包从入口接口转发到出口接口之前，在硬件中处理数据包。数据板是由 Juniper 网络公司定制的一组分布式 ASIC 实施，这些电路位于 T640 路由节点机箱中的各个电路板上。

图 2：T640 路由节点控制板和数据板体系结构

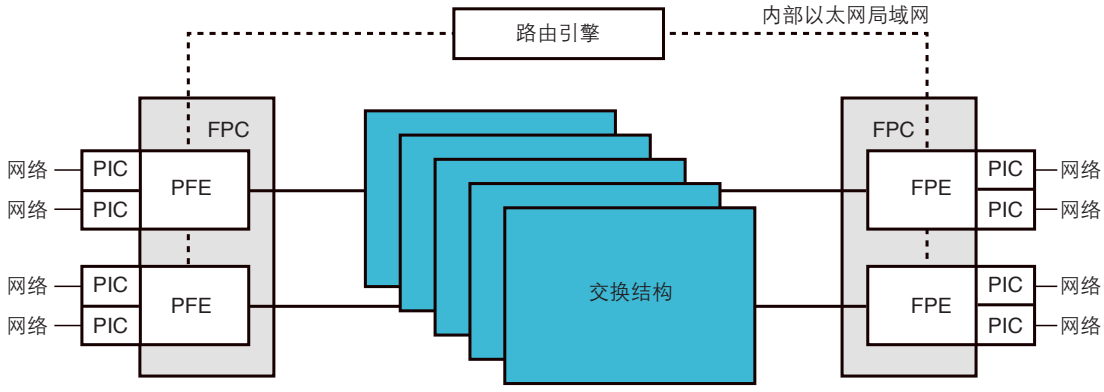


控制板和数据板独立执行它们各自的功能，同时，它们通过一条专用 100-Mbps 链路持续通信。路由引擎执行路由协议，并维护一个或多个路由表。路由引擎从路由表中获得一个主用路由表，这个表被称为*转发表*。JUNOS 内核向 T640 路由节点中的所有数据包转发引擎(PFE)复制转发表，并由 T640 路由节点做出转发决定。这种独特的设计允许更新 PFE 中的转发表，同时不妨碍数据包转发性能。

T640 路由节点的组成部件

T640 路由节点包括三种主要部件：数据包转发引擎(PFE)、交换结构以及一两个路由引擎(见图 3)。

图 3：T640 路由节点的组成部件



数据包转发引擎(PFE)

数据包转发引擎(PFE)执行第 2 层和第 3 层的数据包处理，并执行转发表的查询。 T640 路由节点中的每个 PFE 是由 Juniper 网络公司的定制 ASIC 实施，这些电路位于物理接口卡(PIC)和灵活的 PIC 集中器 (FPC)上。

每个 PFE 包含以下 ASIC 组件：

- 位于每个 PIC 上的*介质特定的 ASIC* 执行与特定的 PIC 介质类型(SONET、ATM、以太网)相关联的第 2 层功能。
- *第 2 层／第 3 层数据包处理 ASIC* 除去第 2 层数据包头，将输入的数据包分解到数据包单元进行内部处理，在数据包传输到出口网络接口之前将数据包单元重组到第 3 层的数据包中，并执行第 2 层出口数据包封装。
- T 系列*互联网处理器 ASIC* 执行转发表查询。
- *排队和内存接口 ASIC* 管理系统内存中的数据包单元的缓冲，以及出口数据包通知的排队。
- *交换机接口 ASIC* 管理穿过 T640 路由节点交换结构的数据包单元的转发。

由于每个第 3 类 FPC 支持 2 个 PFE，一个 T640 路由节点有 8 个 FPC 插槽，因此一个完全配置的 T640 路由节点包含 16 个 PFE。每个 PFE 能够支持相当于2xOC-192/STM-64 接口的带宽。

交换结构

在单独的 T640 路由节点中，交换结构在驻留于机箱中的所有 PFE 之间提供数据板连接。在多机箱路由矩阵中，交换结构在驻留于路由矩阵中的所有 PFE 之间提供数据板连接。

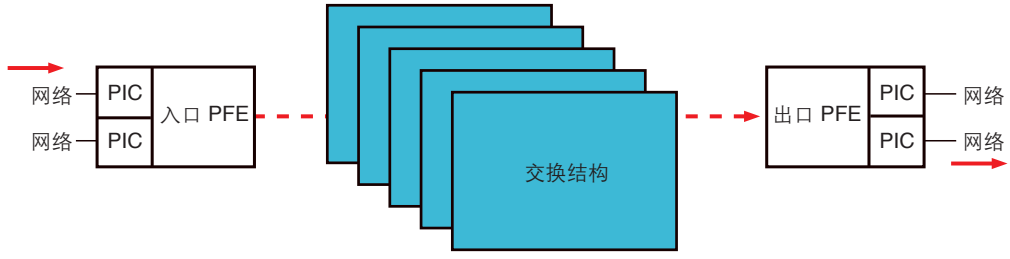
路由引擎

路由引擎执行 JUNOS 软件。该软件实施路由协议，创建路由表，获得下载到每个 PFE 中的转发表，并支持用户界面。路由引擎通过内部以太网控制路径与 T640 路由节点的其他子系统通信。

T640 路由节点数据包转发体系结构

T640 路由节点采用一种分布式体系结构进行实施(如图 4 所示)。当数据包从网络中进入一个 T640 节点的时候，入口 PFE 创建数据包通知，并将数据包分解到包含 64 字节的数据包单元中。接下来，数据包单元被写入入口内存，执行转发表查询，进行入口数据包过滤，代表数据包的数据包单元通过交换结构被转发到出口 PFE 中。当数据包单元到达出口 PFE 时，它们被写入出口内存，再次执行转发表查询，进行出口数据包过滤，代表数据包的数据包单元被重组成最初的数据包。然后，数据包在输出接口上传输到网络中。

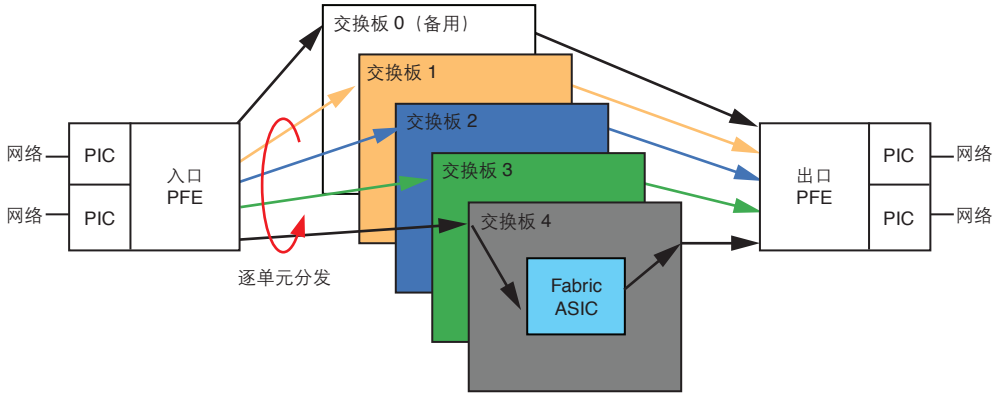
图 4：T640 数据包转发体系结构



T640 路由节点交换结构

在单独的 T640 路由节点(如图 5 所示)中，交换结构是用四个独立运行但相同的交换板实施，这四个交换板同为主用，并由另一个相同的交换板作为热备用，以提供冗余性。T640 路由节点机箱中安装的每个交换接口板(SIB)利用 Juniper 网络公司单一 16 端口 Fabric ASIC 实施其中一个交换板，该 ASIC 以纵横交换的模式运行。16 端口 Fabric ASIC 为可以驻留在单独的 T640 路由节点中的 16 PFE 提供无阻塞的连接。

图 5：T640 交换结构板



每个 PFE 连接到四个主用交换板，每个交换板负责提供一部分所需要的结构带宽。为了确保所有主用交换板之间从一个 FPC 到另一个的流量负载均衡，每个 PFE 都在四个交换板之间逐单元分发数据包单元，而不是逐数据包分发。

交换结构的特性

T640 路由节点交换结构专门设计为提供以下特性：

- 无阻塞连接
- 公平的带宽分配
- 保持数据包的顺序
- 为高优先级流量提供低时延和低抖动
- 提供分布式控制
- 提供交换板冗余性和适度降低性能

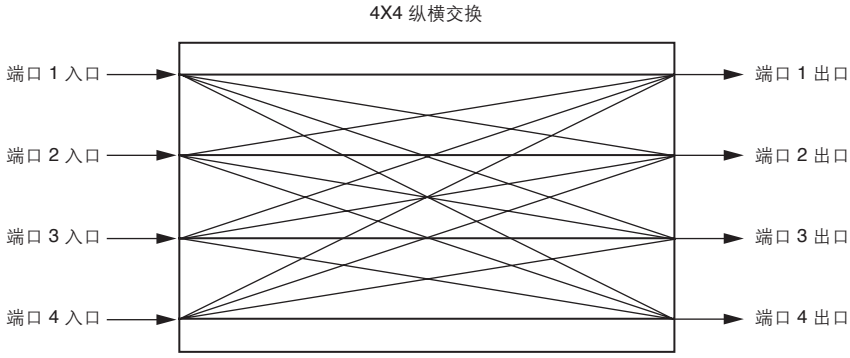
需要指出的是，多机箱路由矩阵的交换结构也设计为具有所有这些特性。

无阻塞连接

如果定向到两个不同输出端口的两个流量流从来没有发生冲突，则交换结构被视为是无阻塞的。交换结构的内部连接允许任何入口 PFE 向任何出口 PFE 发送其*合理的带宽份额*，与穿过该交换结构的其他流量流之间不会互相影响。大型 IP 网络中的通信模式可能瞬息万变，因此，交换结构无阻塞并且不需要基于端口假定流量模式是极其重要的。

图 6 显示的是无阻塞、单级、四端口纵横交换的内部拓扑。构建纵横交换的难题在于这种交换模式要求交换机内部具有 n^2 的通信路径。在下图的例子中，四端口纵横交换需要连接每个输入端口和每个输出端口的通信路径，共计 16 条。随着纵横交换所支持的端口的增加，n^2 内部通信路径的要求会变得更具挑战性，实施成本更加高昂。

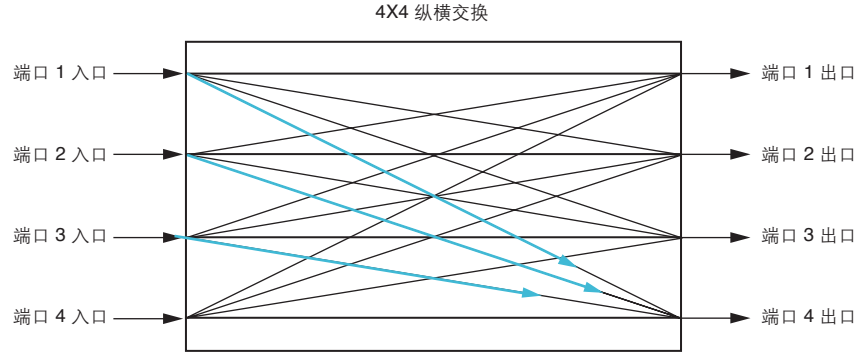
图 6：四端口(4x4)无阻塞纵横交换



公平的带宽分配

在生产网络中，不可能控制入口流量的模式以便不过量使用纵横交换的出口端口。当向出口端口传送的输入流量多于出口端口能够转发的流量时，出口端口就被过量使用。在图 7 中，入口端口 1、2 和 3 向出口端口 4 转发的总流量超过出口端口 4 的容量。

图 7：过量使用的出口交换端口



公平的带宽分配就是交换机采用一种机制，在传送到过量使用的出口端口的各入口流之间分享带宽。T640 路由节点的交换结构监控 n^2 PFE 到 PFE 流，以便每个流都公平地得到一份超量分配给出口 PFE 的可用带宽。

数据包顺序的维持

通过使用序列号和重排序缓冲器，消除了数据包单元在穿过并行的交换板传输到出口 PFE 时出现顺序混乱的潜在可能性。在这种设计中，入口 PFE 在通过交换结构转发的每个数据包单元的单元头中放置一个序列号。出口 PFE 缓冲所有预计从入口 PFE 处接收到的序列号大于下一个序列号的数据包单元。如果某个数据包单元在到达时顺序颠倒，出口 PFE 缓冲收到的数据包单元，直到正确编号的数据包单元到达，快速进行重排序缓冲。这确保数据包(以及特定数据包中的数据包单元)在穿过交换结构时不是顺序错乱。

面向高优先级流量的低时延和低抖动

像话音或视频等某些类型的流量具有时延和带宽方面的要求。T640 路由节点的交换结构设计为分配给交换结构中每个 PFE 的带宽量远远高于向 PFE 提供流量的网络接口的总带宽量。

此外，每个 PFE 向交换结构中实施优先级队列，以确保高优先级流量比来自任何其他 PFE 的低优先级流量受到更加优惠的待遇。因此，高优先级流量可确保从低时延路径通过交换结构传送到出口 PFE。

分布式控制

T640 路由节点没有连接到交换结构中所有组件的集中控制器。如果交换结构内部的任何组件发生故障，围绕该故障组件的其他组件将继续运行。分布式控制提供了更加可靠的系统设计，这是由于要使交换结构发挥功能不需要运行集中控制通道。

交换板冗余性和适度的降级

PFE 接入交换结构带宽受到“请求－许可”机制的控制。利用该机制，入口 PFE 请求出口 PFE 许可它通过交换结构传输一个数据包单元。如果在一段合理的时间后出口 PFE 没有返回许可答复，入口 PFE 就假定在用于发送请求的交换板上不可到达目的 PFE。如果一个交换板发生故障，只会丢失目前正通过交换结构传输的数据包单元，但是交换板内不发生缓冲，“请求－许可”机制确保数据包单元永远不会通过发生故障的交换板传送。

Juniper 网络公司的机制允许通过转移其周围的流量删除交换板中的故障组件，或者将使用故障交换板的所有流量交换到冗余的交换板中。如果特定交换板出现多次错误，或者如果必须交换调出该交换板，机箱管理器将进行协调，把流量转移到冗余的交换板中。在向冗余交换板迁移流量的过程中，每步都转移总体流量中的一小部分。这使系统能够保持运行，不会大幅度降低交换结构的性能。如果另一个交换板发生故障，系统仅留下三个主用交换板，交换结构将继续运行，但是容量减少。

交换结构的操作

通过交换结构传输数据包单元涉及到一个“请求－许可”协议。

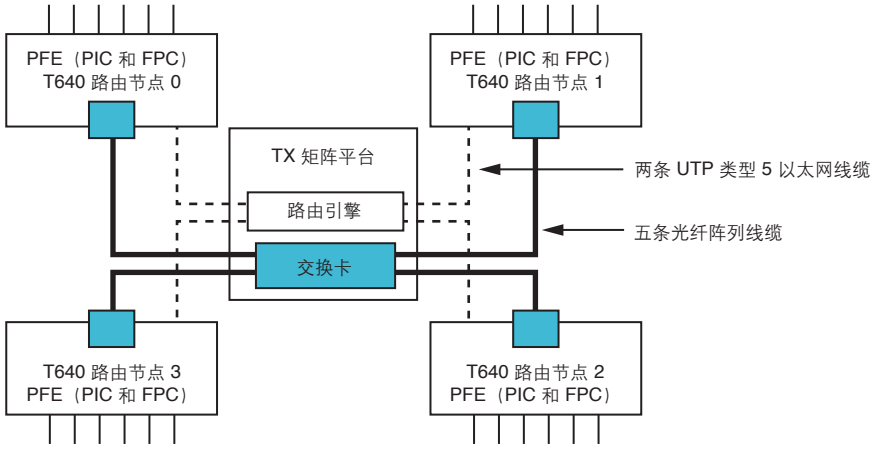
- 源 PFE 通过交换结构向目的 PFE 发送一个*请求*。针对数据包中每个单元的请求按照循环顺序通过不同的交换板传送，在交换结构的所有主用交换板之间平等地分配负载。
- 当目的 PFE 接收到请求时，它使用接收到相应请求的同一交换板向源 PFE 发送一个*许可*。
- 当源 PFE 收到许可时，它使用收到相应许可的同一交换板向目的 PFE 传输数据包单元。

如上所述，“请求－许可”机制既提供了对传送到交换结构中的数据流的控制，还提供了检测交换结构中的中断路径的途径。

路由矩阵体系结构

路由矩阵由一个 TX 矩阵平台和四个 T640 路由节点组成(如图 8 所示)。TX 矩阵充当路由矩阵的集线器，包含系统的冗余路由引擎，并作为 Clos 交换结构的第二级运行。T640 路由节点包含系统的 PIC、FPC 和分布式 PFE。

图 8：路由矩阵

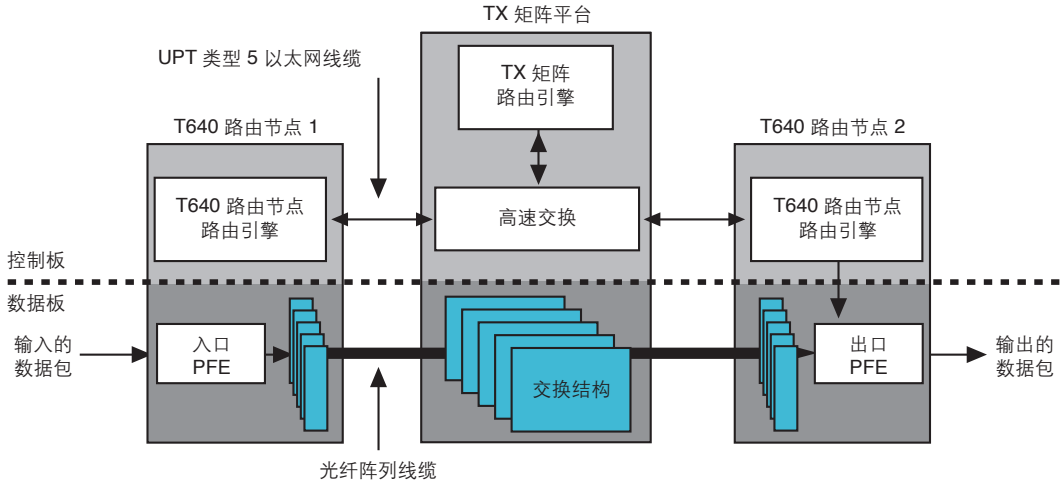


路由矩阵体系结构组件

路由矩阵与 T640 路由节点类似，也有两个基本的体系结构组件（如图 9 所示）：

- *控制板*与在单独 T640 路由节点中执行的功能相同，在 TX 矩阵平台的路由引擎上运行。
- *数据板*与在单独 T640 路由节点中执行的功能相同，由位于 T640 路由节点和 TX 矩阵平台中各个电路板上的一套 Juniper 网络公司 ASIC 实施。

图 9：路由矩阵控制板和数据板体系结构



路由矩阵的控制板和数据板单独执行各自的功能，同时，它们通过专用以太网通信通道持续通信。位于 TX 矩阵平台上的路由引擎执行路由协议，并为整个路由矩阵维护一两个路由表。TX 矩阵路由引擎从路

由表中获得一个主用路由表，这个表称为转发表。接下来，转发表通过控制板复制到驻留在路由矩阵的 T640 路由节点的 PFE 中。该体系结构允许更新位于分布式 PFE 中的转发表，而不会妨碍路由矩阵的数据包转发性能。

路由矩阵的组成部件

路由矩阵由三种基本部件组成：

- 四个 T640 路由节点
- 一个 TX 矩阵平台
- TX 矩阵线缆系统

T640 路由节点

路由矩阵中的 T640 路由节点使用其标准的 PIC、FPC 和冗余路由引擎。然而，T640 中的许多其他组件必须升级：

- 单独的 T640 路由节点中使用的五个交换接口板(SIB)更换为五个 T640 交换接口板(T640-SIB)，用以将纵横交换结构转换为多级、多机箱 Clos 交换结构。多级 Clos 交换结构的体系结构和操作将在本白皮书的后面部分讨论。
- 背面的风扇托架单位从带有五个风扇升级到带有八个风扇，以便为机箱中的新 T640-SIB 提供足够的冷却。
- 单独的 T640 路由节点中使用的两个控制板(CB)为两个 T 系列控制板(T-CB)控制板所代替，用以将单机箱控制板转换为多机箱控制板。
- T640 路由节点 FPC 上的固件可能需要升级，以便允许其控制板组件参与到路由矩阵中。这种升级不需要购买或安装新硬件，但是可能需要由 Juniper 网络公司的人员进行操作，然后，FPC 才能够在路由矩阵中使用。

TX 矩阵平台

TX 矩阵平台将四个 T640 路由节点连接为三级 Clos 网络，构成一个 2.56 太比特的 Single-headed 路由器。Single-headed 路由器由 TX 矩阵平台的路由引擎控制，单独负责运行路由协议，以及维护整个系统的状态。每个 T640 路由节点中的组件由在其各个路由引擎中运行的机箱控制流程管理，与 TX 矩阵平台

的路由引擎中运行的机箱控制流程协同配合。尽管初始实施支持 Single-headed 路由器模式，然而路由矩阵的分布式体系结构已设计为支持将来的 Multi-headed 路由器解决方案。

TX 矩阵线缆系统

TX 矩阵线缆系统实现路由矩阵中每个机箱的数据板和控制板的互连。机箱间光纤线矩阵线缆采用 VCSEL（Vertical Cavity Surface Emitting Laser）技术扩展了数据板。机箱间的电接口使用 UTP Category 5 以太网线缆扩展控制（管理）板。

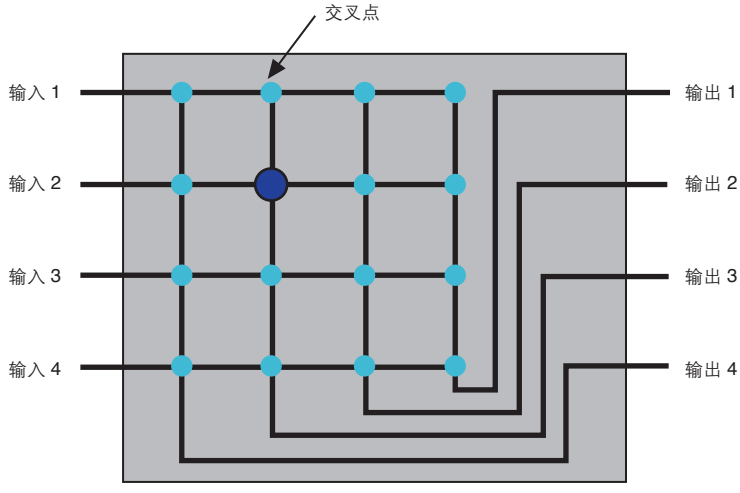
多级 Clos 网络

路由矩阵的交换结构使用多级 Clos 网络实施，而不是传统的纵横交换。在全面了解路由矩阵的交换结构的运行原理之前，您有必要首先了解纵横交换的特性和局限性，以及 Clos 网络的基本操作。

纵横交换

纵横交换有多条纵向路径，多条横向路径，以及将任意垂直路径连接到任意水平路径的*交叉点*。图 10 说明了典型的 4x4 纵横交换的内部结构。淡蓝色的点代表交叉点，深蓝色的点代表连接输入 2 和输出 3 的交叉点。

图 10：4x4 纵横交换中的交叉点



纵横交换中的每条输入路径与每条输出路径都有一个交叉点。此外，纵横交换总有可能在任何输入和任何输出之间建立一条连接路径，而不论交换机中存在哪些现有连接，从这个角度而言，纵横交换是*严格意义上的无阻塞*。最后，创建无阻塞纵横交换所需的交叉点的数量是由输入数量和输出数量的乘积决定的。由此，要创建无阻塞的交换结构，一个 n x n 的纵横交换需要 n^2 个交叉点。

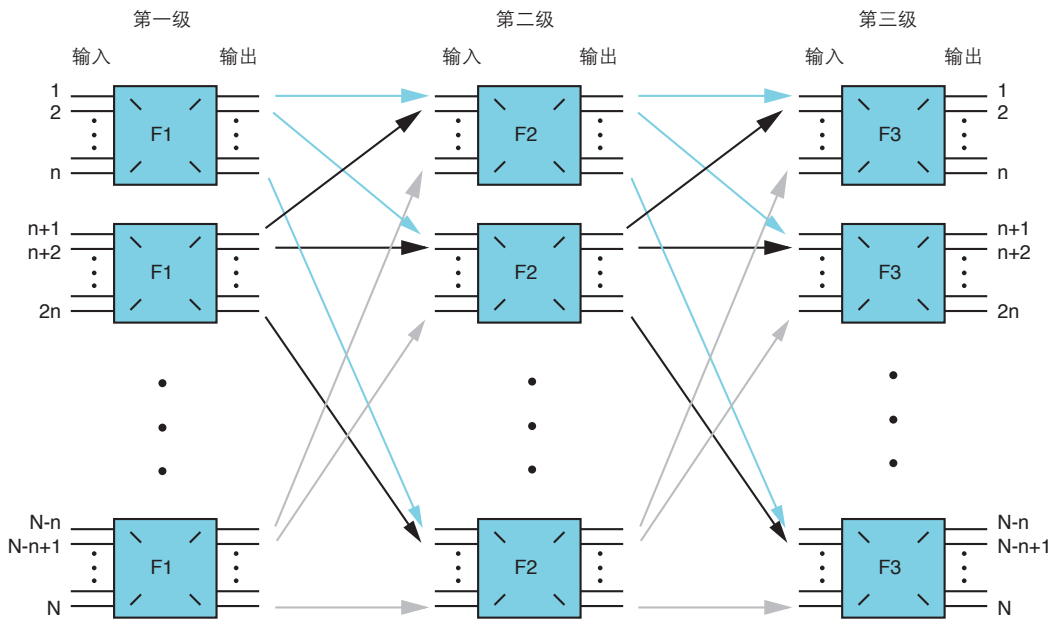
根据 n^2 的规则，有 10 条输入路径和 10 条输出路径的纵横交换需要 100 个交叉点，而有 100 条输入路径和 100 条输出路径的纵横交换需要 10,000 个交叉点。由于交叉点数量和最终成本以及交换机的大小随着交换机的容量迅速增长，工程师已开发出新的交换体系结构，这种体系结构与传统的纵横交换相比需要的交叉点较少，但仍支持相同数量的输入路径和输出路径。

多级 Clos 网络

1953 年 3 月，Charles Clos 在《Bell System Technical Journal》上发表了著名的论文《A Study of Nonblocking Switching Networks》。Clos 在论文种提出了创新的交换体系结构，并精确地证明这种体系结构支持构建严格意义上的无阻塞交换，其中包含的交叉点比纵横交换的要少，但是容量相同。例如，Clos 表明，严格的无阻塞 100 x 100 纵横交换需要 10,000 个交叉点，而 Clos 体系结构仅需要 5,700 个交叉点，更大容量的交换甚至会减少更多的交叉点。在 Clos 的论文发表后的 50 多年里，交换技术不断发展，但是 Clos 提出的体系结构在大容量交换结构的开发中继续发挥着重要作用。

图 11 说明典型的三级 Clos 交换结构的拓扑。蓝色方块代表不连续的 n x n 单级纵横交换。三级拓扑由多行单级纵横交换组成，排成三列，将总计 N 个输入连接到 N 个输出中。

图 11：典型的三级 Clos 网络



Clos 网络第一级的每个纵横交换向交换结构中提供总体输入数量(进入 Clos 结构的总计 N 个输入)的一部分(每纵横交换 n 个输入)。在 Clos 网络中，第一级中每个纵横交换的输出连接到第二级中的纵横交换的输入。第二级中的每个纵横交换的输出连接到第三级中的纵横交换的输入。Clos 拓扑第三级中的每个纵横交换提供来自交换结构的输出总体数量(来自 Clos 结构的总计 N 个输出)的一部分(每纵横交换 n 个输出)。

为了通过 Clos 拓扑建立从一个输入接口到一个输出接口的路径，系统决定哪个第一级纵横交换必须连接到哪个第三级纵横交换，然后确定 Clos 拓扑的第二级中具有未使用路径的纵横交换。如果根据 Clos 描述的条件构建多级交换结构，将会永远至少存在一个允许建立通信路径的第二级纵横交换。

Clos 网络的优势

在设计大型交换结构时，实施 Clos 网络有多种优势：

- Clos 网络通过用较小的交换结构作为基础构建组件，简化大型交换结构的构建。
- Clos 网络大幅度减少构建无阻塞交换结构所需要的交叉点。这可降低大型交换结构的成本，同时由

于可出现故障的交叉点数量减少，从而增强交换结构的可靠性。

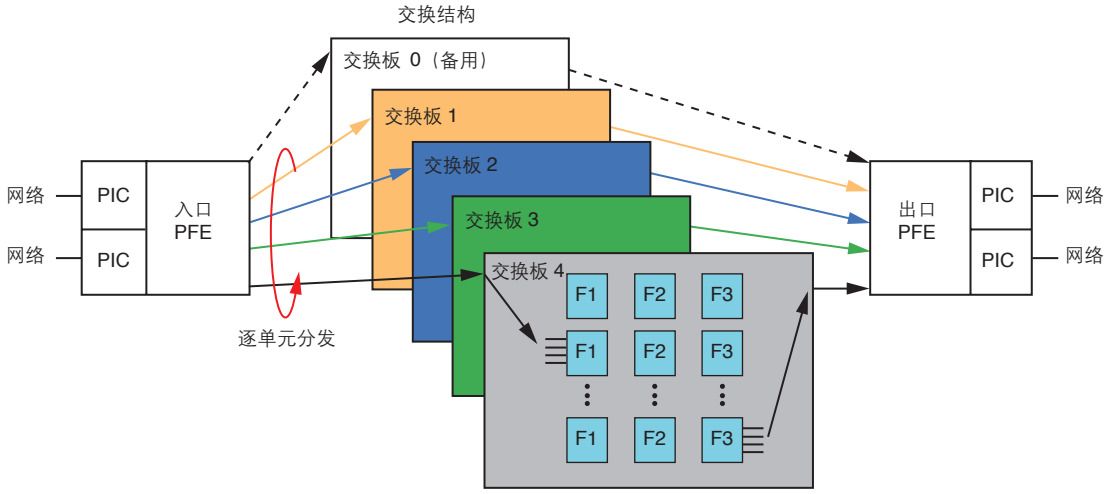
- Clos 网络对故障的抵御能力强，这是因为一个第二级纵横交换可通过使用冗余路径绕过另一个第二级纵横交换内的故障。
- Clos 网络中的每个纵横交换完全独立于该结构中的其他纵横交换。这意味着，Clos 网络不需要复杂的集中控制器来协调单个纵横交换的行为。
- Clos 网络具有高度可扩展性，支持构建极大规模的交换结构。Clos 网络的中间级别的交换可由完整的三级 Clos 网络代替。这可支持构建具有五级、七级或九级的巨大交换结构，在极大数量的输入和输出之间建立严格的无阻塞连接。

路由矩阵 Clos 交换结构的实施

路由矩阵实施 64x64 三级 Clos 网络，可提供多达 64 个 PFE 的连接性。每个 T640 路由节点最多包含 8 个 FPC，每个 FPC 最多支持 2 个 PFE，在一个 T640 路由节点中总计可支持 16 个 PFE。由于路由矩阵设计为实现多达 4 个 T640 路由节点的互连，因此，完全组装的路由矩阵包含 64 个 PFE。

路由矩阵与单独的 T640 路由节点类似，也是由四个独立运行但相同的交换板实施，这四个交换板同为主用，并由另外一个同样的交换板作为热备用，以提供交换结构的冗余性(见图 12)。

图 12：路由矩阵交换板



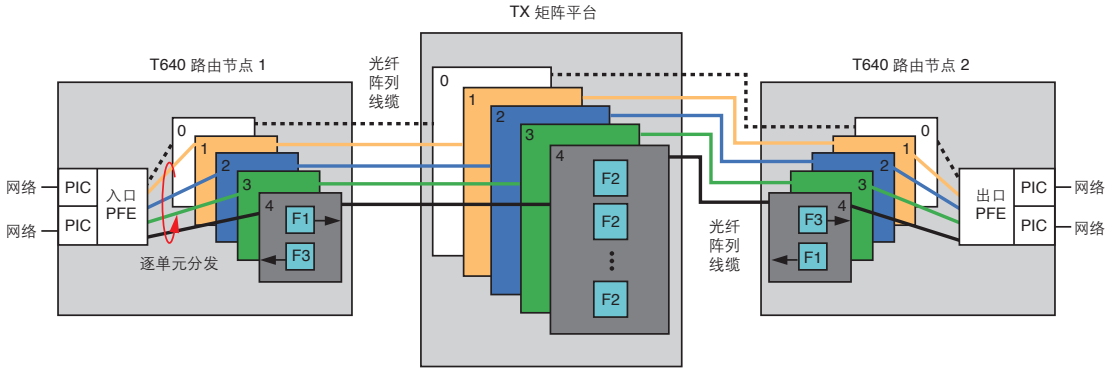
每个交换板提供利用 Juniper 网络公司 16x16 Fabric ASIC 构建的完整的三级 Clos 网络。路由矩阵中的每个 PFE 连接到四个主用交换板，而每个交换板负责提供所需的一部分结构带宽。为了确保特定数据包的单元在所有主用交换板之间的负载均衡，每个 PFE 在四个主用交换板中逐单元分发数据包单元，而非逐数据包分发。

路由矩阵的多级 Clos 结构设计为提供单独 T640 路由节点的单级纵横交换结构支持的所有特性：

- 无阻塞连接
- 公平的带宽分配
- 数据包顺序的维护
- 面向高优先级流量的低时延和低抖动
- 分布式控制
- 交换板冗余性和适度降低性能

图 13 说明三级 Clos 交换结构如何在路由矩阵的 T640 路由节点和 TX 矩阵平台中分布。

图 13：路由矩阵 Clos 交换结构的实施



T640 路由节点

路由矩阵中的 T640 路由节点同时实施 Clos 交换结构的第一级和第三级。T640 路由节点机箱中的每个 T640-SIB 都可作为交换板，包含实施第一级和第三级 Clos 功能的 Fabric ASIC。

- 当作为入口 T640 路由节点运行时(例如图 13 中的 T640 路由节点 1)，入口 PFE 处理在其网络接口上收到的数据包，并识别每个数据包的出口 PFE。然后，入口 PFE 在逐单元的基础上，在所有主用交换板中均衡地分发数据包单元。每个交换板中的第一级 Fabric ASIC 在匹配的交换板的第二级 Fabric ASIC 上分发数据包单元，相匹配的交换板位于 TX 矩阵平台中。
- 当作为出口 T640 路由节点运行时(例如图 13 中的 T640 路由节点 2)，第三级 Fabric ASIC 收集由入口 PFE 穿过 TX 矩阵平台中的所有第二级 Fabric ASIC 分发的数据包单元，并将这些单元传送到出口 PFE。出口 PFE 接收来自交换结构的数据包单元，将它们重组为最初的数据包，并在出口网络接口上转发该数据包。

TX 矩阵平台

TX 矩阵平台作为路由矩阵的交换核心。TX 矩阵平台包含五个 SIB 卡，这些卡使用机箱间的光纤阵列线缆连接到每个 T640 路由节点中的 T640-SIB 卡。TX 矩阵平台中的每个 TX-SIB 充当交换板，提供入口和出口 T640 路由节点之间的连接，并提供 640 Gbps 的交换容量。当作为 Clos 网络的第二级运行时，每个 TX-SIB 中的 Fabric ASIC 可接收来自任何入口 PFE 的数据包单元，并将这些单元交还给任何出口 PFE。

光纤阵列线缆

机箱间的光纤阵列线缆将 Clos 结构的分布式交换级连接到统一的大型交换系统中。每个 T640 路由节点由一个包括五列光纤的线缆组连接到 TX 矩阵平台，一根线缆对应一个主用并行交换板，有一根线缆支持冗余交换板。一个完全组装的路由矩阵包含四个 T640 路由节点，共需 20 个光纤阵列线缆来实现数据板的互连。

将特定 T640 路由节点连接到 TX 矩阵平台的机箱间光纤阵列线缆每根可长达 100 米。一个包括五根线缆的线缆组中所有光纤阵列线缆的长度必须相同。然而，用于将不同的 T640 路由节点连接到 TX 矩阵平台的线缆组不一定需要等长。例如，在部署的一个路由矩阵中，T640 路由节点 1 使用五根 10 米的线缆连接到 TX 矩阵平台，T640 节点 2 使用五根 100 米的线缆连接到 TX 矩阵平台。

总结

为了满足供应商对下一代核心路由器平台具有五年或更长的部署生命周期的要求，Juniper 网络公司认识到，T640 路由节点需要采取一种升级战略，以便作为路由矩阵配置的一部分来运行。T640 路由节点的设计宗旨是，交换结构可以轻松修改，以支持路由矩阵配置，同时仍然允许运营商保护其在 PIC、FPC、路由引擎和电源中的现有投资。现在，供应商可通过对现有 T640 路由节点进行简单升级，然后将它们连接到 TX 矩阵平台，从而延长他们的核心路由器的部署生命周期。TX 矩阵平台的上市表明 Juniper 网络公司致力于提供下一代核心路由器体系结构，以保护供应商对 Juniper 网络公司设备的投资，并且不需要进行“叉式”升级。

参考资料

Clos, C. “A Study of Non- blocking Switching Networks.” *Bell System Technical Journal*, March 1953, pp. 406-424.

Hwang, F. “A Survey of Nonblocking Multicast Three-Stage Clos Networks.” *IEEE Communications Magazine*, October 2003.

Jajszczyk, A. “Nonblocking, Repackable, and Rearrangeable Clos Networks: Fifty Years of the Theory Evolution.” *IEEE Communications Magazine*, October 2003.

Walker, M., and Broadcast, P. “Multistage Distribution Switching Systems, Clos and Beyond.” *http:// www. broadcastpapers. com/sigdis/Philips3StageRouter0 1 .htm*

www.juniper.net
www.cn.juniper.net

北京代表处
北京市东城区东长安街 1 号
东方经贸城西三办公楼 15 层 1508 室
邮政编码： 100738
电　　话： 8610-6528 8800
传　　真： 8610-8518 2626

上海代表处
上海市淮海中路 333 号
瑞安广场 1102-1104 室
邮政编码： 200021
电　　话： 8621-6141 5000
传　　真： 8621-6141 5090

广州代表处
广州市天河区体育东路 118 号
财富广场西塔 15 楼 101 室
邮政编码： 510620
电　　话： 8620-3886 0668
传　　真： 8620-3886 0638