目 录

交换网

1.1 交换网发展史简介

交换网技术是现代通信中最重要的技术之一，它实现信息从发送端到接收端传输。两点之间的信息传输，最简单的方式是用一条通信线路直接相连，实现点对点的通信。当有多个终端要相互通信时，如果仍采用这种点对点的方式，则需要在任意两个终端间互联。这种点对点的互联方式，会随着终端数的增加，所需的互联线急剧增加。当有 N 个终端需要互联时，一共需要 N×（N-1）/2 条互联线。例如，当有 100 个终端时，需要 4950 条互联线。为了解决大量互联线的问题，人们发明了一种终端间互联的设备，每个终端只需一根线连接到这个中心设备，通过它的自动连接功能，能够实现任意终端间的相互通信，它就是交换机（Switch）。它将互联线的数量从 N×（N-1）/2 条减少到 N 条，极大地降低了线路成本。

路由器是 IP 网络中的核心设备，其交换单元（Switch Fabric Unit）是决定路由器性能的最核心单元。通常在设计一款新路由器时，首先需要确定的就是所采用的交换网技术。在路由器 20 多年的发展史中，交换技术一步步推动路由器向更大容量，更高性能发展。

交换技术的发展，经历了共享总线式（Shared Bus Switch）、共享内存式（Shared Memory Switch）和交叉矩阵式（Crossbar Switch）三代。与此同时，路由器的发展也经历与之对应的共享总线式路由器（Shared Bus Router）、共享内存式路由器（Shared Memory Router）和交叉矩阵式路由器（Crossbar Router）三代。随着网络流量的剧增，核心路由的容量要求越来越高，路由器开始由单机路由器向集群路由器的方向发展，对应的交换网技术从单级交换网（Single Stage Switch）向多级交换网（Multi-Stage Switch）方向发展。

说明

- Fabric 本意是指织物的经纬构造，在交换网技术常用来指交换单元、交换芯片，在本文介绍中并不区分 Swtiching、Switching Fabric 和 Fabric，都是指路由器中交换单元。
- 本文以 NE80E/NE40E 的交换网原理为例进行介绍。CX600-8 和 NE40E-8 的原理类似；CX600-16 和 NE80E 的原理类似；CX600-X3/ME60-X3 和 NE40E-X3 的原理类似；CX600-X8/ME60-X8 和 NE40E-X8 的原理类似；CX600-X16/ME60-X16 和 NE40E-X16 的原理类似。

以下从如下几个方面介绍路由器的交换网：

- 交换网相关术语
- 交换网的分类

- 各种交换网技术介绍
- 华为路由器交换网介绍
- 交换网发展趋势
- FAQ

1.2 交换网相关术语

为了让您更好地理解交换网技术，首先介绍一些与交换网相关的术语和指标计算方法。

1.2.1 背板容量（Backplane Capacity）

背板是路由器内部各单元互联的重要部件。背板容量是路由器背板上业务槽位到交换单元的数据总线带宽的总和，它通常大于依据路由器吞吐量和实际性能测试所得到的容量。背板容量体现厂家的工程设计水平和该路由器未来容量提升能力，通常情况下，无法直接测试其容量。

如果说路由器拥有 400G 背板，是指其背板能够支持每槽位 400G 的业务带宽。一开始，可能现网只部署了 100G、200G 的业务单板，但当技术成熟后，推出 400G 业务单板时，客户现网使用的路由器不需要更换背板，就能升级到每槽位 400G 业务。实际设计中，为了实现 400G 业务单板的接入能力，背板容量需要远远大于 400G。

说明

背板就好比高速公路主干道，来来往往的车辆都需要在上面行驶，因此需要具备较好的扩展性。提前规划好的 8 个车行道暂时只有 4 个在用，没关系。随着交通业务的发展，预留的 4 个车行道将慢慢派上用场。

业界的背板实现中，都采用高速串行总线（俗称 Serdes，用一对表示互补的物理连线实现数据发送或接收），实现数据的高速互联。根据设计的不同， Serdes 的速率也不同，比如 2.5Gbps、3.125Gbps、6.25Gbps、12.5 Gbps。

说明

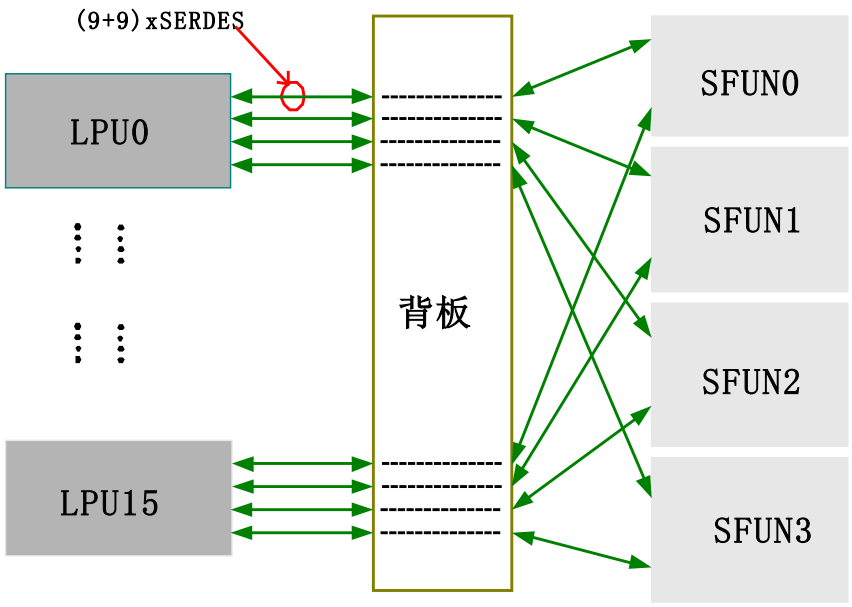
受工程实现的限制，路由器背板上的物理线路连接不能无限的增加，主要还得要依赖信号速率的提升。

整机背板容量=所有 LPU 槽位与所有 SFU 槽位之间互联的 Serdes 数量×每路 Serdes 的速率

如图 1-1 所示，该路由器背板有 16 个业务板（LPU）槽位，4 个交换网单元（SFU）槽位。每个业务板和每个交换网单元有 18 对 Serdes 相连接（其中 9 对用于数据接收，9 对数据发送），Serdes 的速率为 6.25Gbps。

其背板容量= [2×（9×4×16）]×6.25G bps = 7.2T bps。其中，2 表示“接收+发送”双向容量，9 表示每个 LPU 到每个 SFU 的 Serdes 数量， 4 表示交换网单元 SFU 的数量，16 表示业务板 LPU 的数量，6.25G 表示每路 Serdes 的速率。

图1-1 背板容量/交换容量示意图



说明

背板不仅实现业务槽位与交换槽位之间的数据链路互联，还为各种控制通道的信号提供互联，同时还将电源模块提供的电源提供给各业务槽位。

Serdes（Serializer/Deserializer，发音 sir-dees）：本意是指将多个并行信号在发送端用串行器（Serializer）转化成串行信号传输到接收端，接收端再使用解串器（Deserializer）将串行信号恢复成并行信号的传输技术，一般使用一对互补的差分信号进行串行信号的传输。后来 Serdes 泛指使用高速差分线进行数据传输的技术。

1.2.2 交换网容量(Switching Capacity)

交换网容量是指路由器交换网单元能够处理的最大容量。对于无阻塞的交换网，它等于交换网所有端口的容量之和。数据从业务端口进入路由器后，在路由器内部进行协议转换，会添加一些额外的信息（Overhead，也称开销）用于内部数据处理，因此交换网所处理的流量会大于路由器业务端口的流量。交换网容量和用户所用带宽没有直接的换算关系，它是综合用户带宽（Traffic Bandwidth）、路由器内部开销（Overhead）和交换网加速比（Speedup Factor）三者的一个综合性能参数，体现交换网的性能。

对于多平面的交换网，交换网的容量等于各个交换网平面的交换容量之和。

说明

如果把背板比作高速公路主干道，那么交换网就好比高速公路上的收费站。它对过往车辆拦截、收费、放行的处理速度；在车辆拥塞时，有效、及时的分流能力，决定了交通运行的质量。

整机交换网容量=交换网端口总数×每路 Serdes 的速率×Serdes 编码效率

如图 1-1 所示，其交换网容量= [2×（9×4×16）]×6.25G bps×0.8 = 5.76T bps。其中，0.8 为 Serdes 编码效率。

1.2.3 加速比（Speedup Factor）

加速比（用 S 表示）是一个衡量交换网性能的重要指标，其计算公式为：

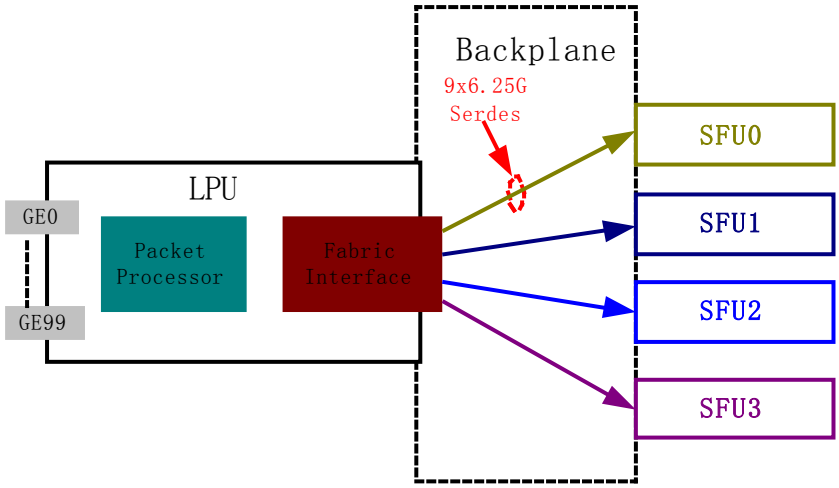
S = （交换网接口带宽）/ (物理接口带宽) =（交换网 Serdes 数量×Serdes 速度×Serdes 编码效率）/(物理端口数量×物理端口速度)

加速比大的交换网，因其内部阻塞或出口阻塞的概率小，性能一般会更好，对系统的多播性能和 QoS 支持会更好。

如图 1-2 所示，路由器有 1 个 LPU（上面有 100 个 1GE 端口），4 个 SFU。每个 LPU 和 SFU 之间通过 9 对 6.25G 的高速 Serdes 互联，高速串行总线上的编码效率为 80%，那么加速比 S 为：

S = （6.25G×9×4×0.8）/(1G×100) = 1.8

图1-2 交换网加速比示意图



说明

并非加速比越大，交换性能就越好。有时加速比较大，是因为交换算法做得太简单，需要用较大的加速比来平衡。

1.2.4 交换网的备份方式(Backup)

交换网在路由器中处于核心的位置，所有业务板均通过背板与交换网单元连接，实现业务交换。如果交换网出现了故障，就会导致与之相连的所有业务板业务中断。因此在进行系统设计时，为了提高路由器的可靠性，需要考虑交换网的冗余保护功能，备份设计是常用方式之一。常见的备份方式有主备保护方式（Master-Slave）和负载分担保护（Load Balanced）方式。

主备保护方式是指，主用交换网进行正常的流量交换，备用交换网不进行流量交换。一旦系统检测到主用交换网出现故障，会将该故障交换网降为备用交换网，同时原备用交换网升级为主用交换网，继续进行业务交换。主备保护方式也称 N：M 备份方

式，其中 N 为处于主用状态的交换网单元数量，M 为处于备份状态的交换网单元数量。主备切换的过程也就是常说的主备倒换。

负载分担保护方式是指，系统中的所有交换网都同时工作，当有任意一个出现故障时，系统自动将它的流量平均分配给其他交换网，保证系统性能不受影响。负载分担方式也称 N+M 备份方式，当有 M 个交换网单元故障后，N 个交换网单元仍然能够满足系统性能要求。M 是最大允许的故障交换单元，当故障单元数超过 M 时，系统的交换性能会下降。

1.2.5 交换网吞吐量（Fabric Throughput）

交换网吞吐率是衡量交换网每秒处理的报文的能力，对于定长交换的交换网（Cell-based）是指每秒处理的定长信元数，对于变长交换的交换网（Packet-based）是每秒处理的数据包数。交换网的吞吐率要大于路由器的转发能力，单位为 Mpps (Million Packets per second)或 Mcps (Million Cells per second)。

1.2.6 交换网延时（Fabric Latency）

交换网的延时是指从向交换网申请数据交换到目的端口输出数据之间的时间间隔。交换网在系统中是一个公用部件，各个接口板竞争使用交换网的资源，因此延时包括数据包等待进入交换网的时间和交换网转发数据时间之和。受交换网调度算法的影响，交换网的延时与系统的流量相关。交换网的调度算法会影响到系统带宽的分配，因此延时对系统的 QoS 有影响。交换网的延时越小，体现交换网性能越好。

1.2.7 交换网扩展性（Fabric Scalability）

交换网的端口数，一般以 N×M 表示，N 为输入端口数，M 为输出端口数。一般来说，在路由器的交换网中，N=M。对交换网的性能扩展，一般从如下方面考虑，端口速率（Port rate）、系统容量（System capacity）和业务支持能力（Service scalability）。

1.2.8 单播、组播、反压

单播(unicast)

从一台服务器送出的每个数据包只能传输给一个客户端,这种传输方式称为单播。在交换网中，交换网将单播流从一个端口交换到指定的另一个端口，完成数据包的交换。

组播(multicast)

组播技术允许路由器一次将数据包复制到多个通道上。采用组播方式，单台服务器能同时对大量的客户端连续发送数据流，而服务器只需要发送一个信息包，所有发出请求的客户端共享同一信息包。信息可以发送到任意地址的客户端，达到减少网络上传输的信息包总量，提升网络利用率，降低传输成本的目的。

交换网芯片内部进行多播复制的方式通常也叫作空间多播，交换网根据多播组 ID，将一个数据包从一个交换网端口复制到多个交换网端口。

反压（Backpressure）

反压是一种流控（Flow control）方式，即常说的单向流控，主要用于缓解拥塞持续恶化，从而减轻，直至解除拥塞。当 A、B 两个端口通信时，如果 A 端口发现自身接收缓冲区拥塞，A 端口会把一个特殊的数据帧（即反压帧）发送给 B 端口。B 端口收到反压帧后会停止向 A 端口发送数据，直到 A 端口接受缓冲区无阻塞。作为系统内部的的公用部件，交换网也面临数据拥塞的可能，根据交换网实现的不同，内部会有多种反压机制，同时对交换网外部的单元也有一定的反压机制。

说明

反压不能防止拥塞，它只是对拥塞的一种响应机制。换句话说，出现反压时，系统实际上已经出现了拥塞。反压是用于缓解拥塞持续恶化，从而减轻直至解除拥塞。正如，发烧是人体对病毒入侵的一种响应机制，人体出现发烧症状时，病毒实际早已入侵。发烧只能通过高温的方式来抑制病毒生存，缓解病情恶化，协助身体康复。

1.2.9 交换网性能评估参数

综上所述，路由器交换网结构（Switch Fabric）的设计需要考虑如下因素：

- 吞吐量（Throughput）
- 延时（Latency）
- 端口数、端口速率、业务类型（Scalability (Port size、Port Rate、Service type)）
- 加速比（Speedup Factor）
- 成本（Cost）
- 多级交换支持能力（Multi-stage Switching）
- QoS 业务扩展性（QoS Scalability）

1.3 交换网的分类

同一种交换网，因为分类标准不一样，会有多种名称和归类，在讨论各种交换网的区别时，需要明确交换网的分类标准。常见的分类方式为：根据上送交换网的报文类型分类、根据缓存所处交换网的位置分类、根据数据交换次数分类、根据数据通过交换网的方式分类。

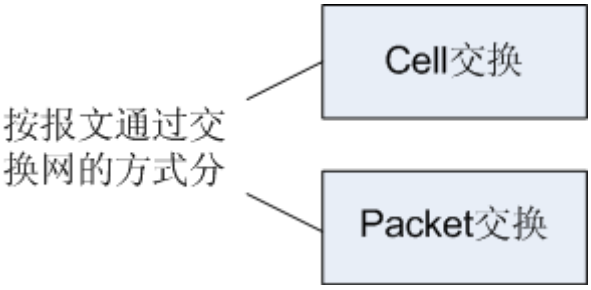
1.3.1 根据上送交换网的报文类型分类

根据数据报文上送交换网的类型，可以分为信元交换（Cell Switch，即定长交换）和包交换（Packet Switch，即变长数据交换）。

信元交换采用和 ATM 交换类似的方式，数据在进入交换网之前被切割成定长的信元，加上一定的信元头信息，送到交换网中进行交换。信元长度是固定的，交换网可以通过高速的硬件单元对信元进行处理，因此在路由器交换网中得到了广泛的应用。当信元到达出端口时，再将信元重组成数据包。

包交换是指直接使用变长的报文进行交换，没有报文分片的操作。一般用于共享内存式的交换，直接使用报文的头，或者添加一定的报文头进入交换网交换，其优点是不需要报文分片和重组，在以太网交换机中得到比较广泛的应用。

图1-3 根据上送交换网的报文类型分类图



1.3.2 根据缓存所处交换网的位置分类

交换网是路由器中的公共部件，各个线路板竞争使用，通过交换网的仲裁算法进行调度。数据在交换网的输入端口或输出端口进行缓存，以等待被调度。交换网根据信元缓存和交换单元的相对位置的不同，可分为 OQ、IQ、CIOQ 三类：

- 输出排队(Output Queuing，OQ)

  信元缓存全部放在交换矩阵之后，信元被送到输出端口进行排队，等待输出。输出排队型交换网，要求输出端口的缓存能够处理所有输入同时流向同一输出的突发情况，对缓存的访问带宽有很高的要求。

- 输入排队(Input Queuing，IQ)

  信元缓存全部放在交换矩阵之前，一旦数据被调度，直接穿过交换网从输出端口输出，输入端口排队有可能因为输入端口始终得不到调度引起端口拥塞，也可能因为前面低优先级的数据得不到调度，引起线头阻塞 HOL(Head of Line)。

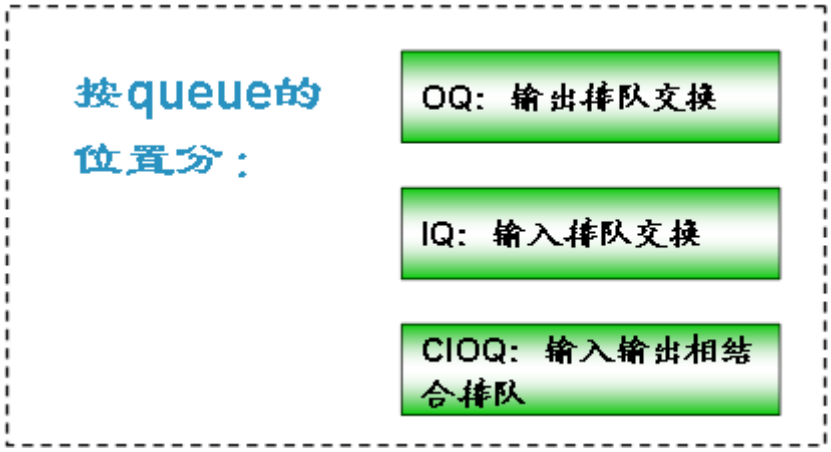
说明

虚拟输出排队（Virtual Output Queuing）：将目的输出端口不同的数据放在不同的输入队列中缓存，因此发往不同输出端口的信元间不存在 HOL 阻塞。虚拟输出排队并非一种新的缓存方式，它是对输入排队的改进。

- 输入输出相结合排队(Combined Input and Output Queuing, CIOQ)

  信元缓存部分放在交换矩阵之前，部分放在交换矩阵之后。输入输出相结合的排队方式能够有效地解决对输出缓存带宽要求较高和输入线头阻塞的问题，是交换网中常用的一种方式。

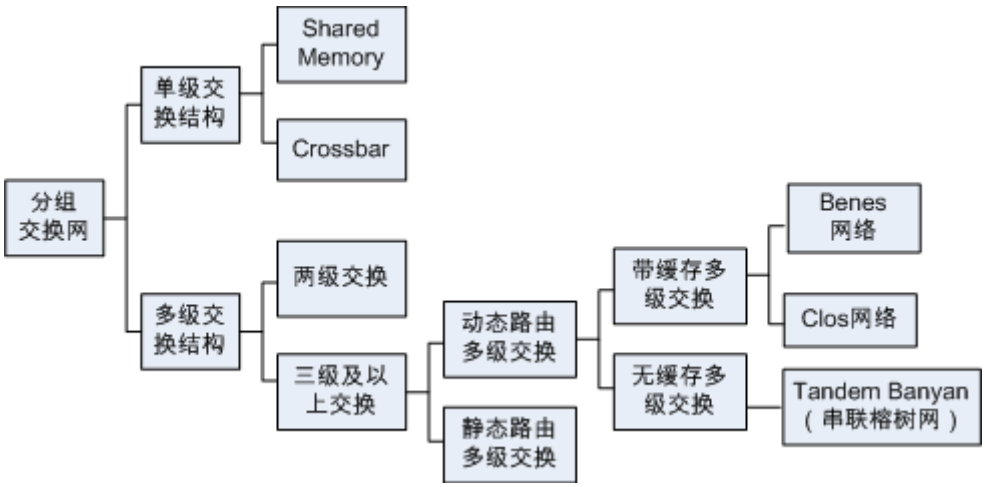
图1-4 根据缓存所处交换网的位置分类图



1.3.3 根据数据交换次数分类

根据数据在交换网中经历的交换次数，交换网可以分为单级交换网和多级交换网。其中多级交换网根据实现方式的不同，又可分为不同类型，如图 1-5 所示。

图1-5 根据数据交换次数分类图



1.3.4 根据数据通过交换网的方式分类

根据数据通过交换网的方式不同，可以将交换网分成直通式交换网（Cut through switching）和存储转发式交换网（Store and Forward switching）。

直通式交换网，并不等输入端口进来的数据帧（可以是定长的信元 cell，也可以是变长的数据包 packet）接收完成，就开始向输出端口发送数据。理论上，直通式交换网具有较快的转发速率，较低的转发延迟。

存储转发式交换网，将整个数据帧接收完成，进行必要的数据校验之后，才会向输出端口发送数据。存储转发式交换网具有很好的容错性。

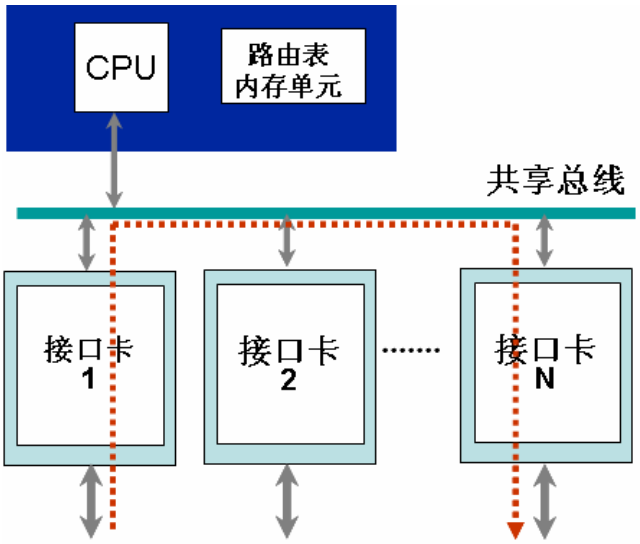
直通式交换网在端口速率适配、容错性等方面具有缺陷，因此，存储转发式交换网应用得比较广泛。

1.4 交换网发展历程与趋势

1.4.1 共享总线式交换网（第 1 代交换网）

共享总线式的交换结构是第一个成功应用的交换结构，它由计算机工业中的共享式总线演变而来。所有的输入、输出端口都连接到一条共享使用的总线上，通过一定的仲裁申请机制，在相同时刻，只允许有一对输入/输出端口利用该总线进行通信。即由输入/输出对总线的使用权提出申请，由一个中央仲裁器负责对总线的使用权进行分配，防止冲突发生。共享总线式交换网结构如图 1-6 所示。

图1-6 共享总线式交换网



在共享总线系统中，无阻塞的交换意味着所有端口的带宽之和必须小于共享总线的带宽，也就是说系统的交换性能受限于共享总线的容量，同时，系统的性能还受限于中央仲裁器（CPU）的处理速度。在图 1-6 中，当接口卡 1 和接口卡 N 通信时，独占背板上的总线，此时接口卡 2 不能和其他接口卡通信，路由器的性能受限于共享总线的性能。

第一代路由器基本都采用共享总线式交换网，例如华为 NE16E。

1.4.2 共享内存式交换网（第 2 代交换网）

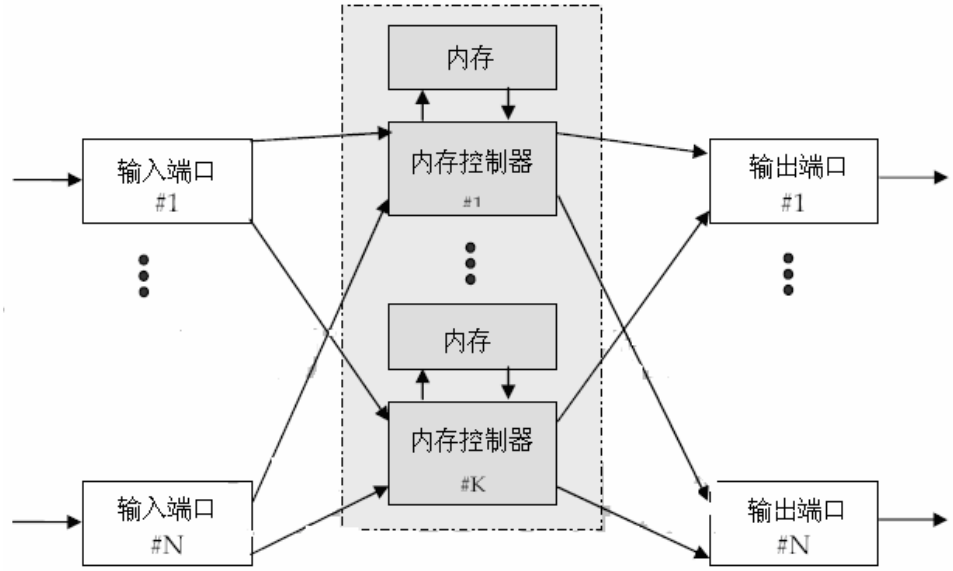
随着用户对接口带宽的需求不断增加，共享总线式交换网已无法应对新需求。首先，共享总线不能避免内部冲突；其次，共享总线的负载效应使得高速总线的设计难度太大。在 20 世纪 90 年代，一种新的基于共享内存的交换结构开始出现。

共享内存式的交换结构基于大容量的共享内存，每块内存区域由单独的内存控制器负责存取访问。内存控制器将所控制的内存区域按块分配给不同的端口，将输入端口的

数据按照一定的规则存放到内存中，然后通知输出端口将对应内存区域中的数据取走，完成一次数据交换。在共享内存式交换结构中，一次完整的数据交换包括一次内存写操作和一次内存读操作。

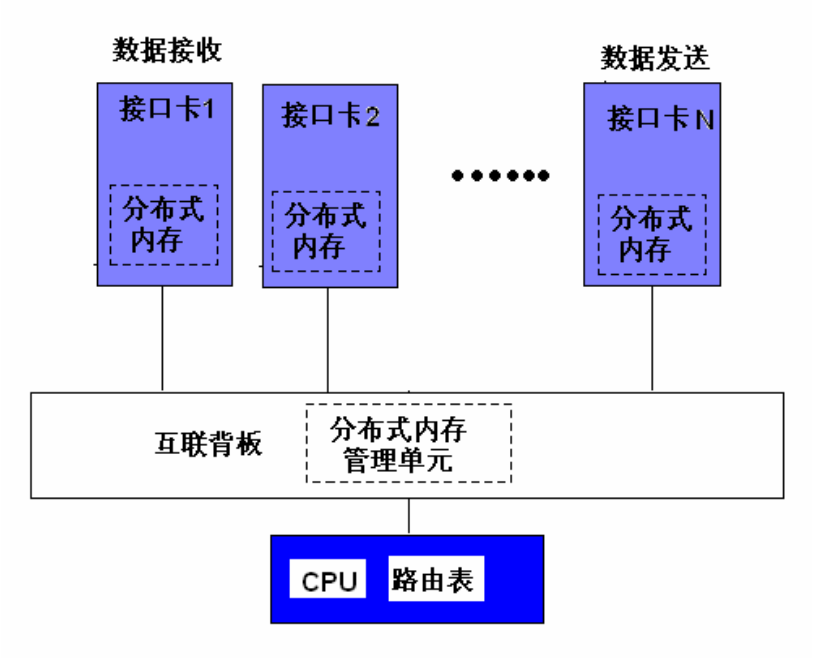
图 1-7 是一个典型的共享内存式交换结构，系统一共有 K 个内存控制器，每个内存控制器单独控制一个内存区域。每个内存控制器与每个输入/输出端口相连。从输入端口接收的数据，在内存控制器的控制下写入共享内存中，再根据输出端口的状态，将数据从共享内存中读出，发送到输出端口，完成一次数据交换。

图1-7 共享内存式交换网



典型的共享内存式路由器结构如图 1-8 所示，数据从接口卡 1 进入路由器后，首先在分布式内存管理单元的控制下写入到接口卡 1 的分布式内存中，经过路由查表将数据读出，直接发送到接口卡 N 上。

图1-8 共享内存式路由器



对于一个无阻塞的共享内存式交换网，要求内存写入的带宽大于所有输入端口带宽之和，共享内存读出的带宽大于所有输出端口的带宽之和。

1.4.3 交叉矩阵式交换网（第 3 代交换网）

交叉矩阵式交换网（Crossbar，也称纵横式交换网，矩阵式交换网）是目前主流的交换网技术之一，在业界的核心/业务路由器上有着广泛的应用。

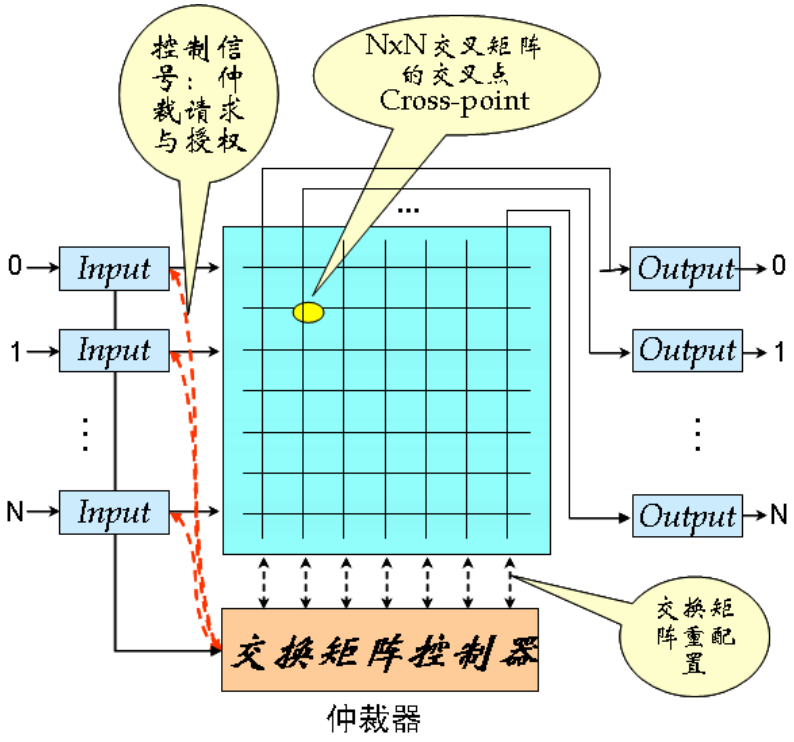
单级 Crossbar 交换网

单纯的 Crossbar 是一种简单的空分交换开关，可将 N 个输入端口与 N 个输出端口任意互连。Crossbar 交换矩阵控制器的作用是根据输入队列的状态，决定每个调度周期输入端口和输出端口间的连接关系，仲裁机构仲裁输入端口对输出端口的访问，根据仲裁结果，控制器打开或关闭交叉点开关。仲裁器是 Crossbar 交换结构的核心，Crossbar 结构的性能主要取决于仲裁器的速度。Crossbar 还可以实现将一个输入端口和多个输出端口相连，因而很容易实现组播。

Crossbar 交换网优点：同其他交换结构相比，Crossbar 可在一个信元周期内并行传送 N 个信元，因而有较高的吞吐量。外加电路实现简单，因此在交换网中得到广泛应用。

Crossbar 交换网的缺点：Crossbar 交换网随着端口数量的增加，仲裁器的复杂性随之增加（N≦64 时，Crossbar 是比较好的选择；当 N 增大时，仲裁器的复杂性随 N^2的数量级增加）。

图1-9 单级 Crossbar 交换网

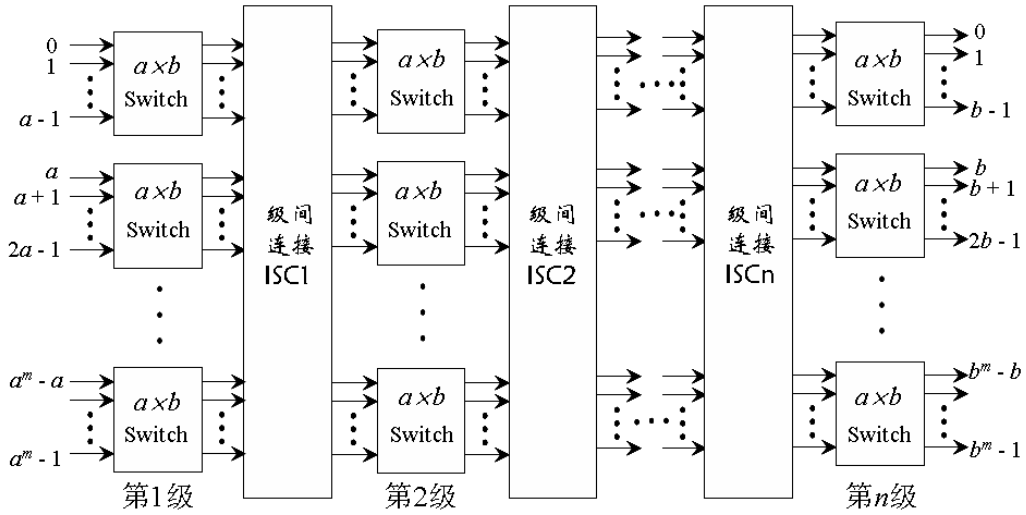


Crossbar 交换网，一般用于构建单级交换网。目前在单框交换中用得比较多，比如华为的 NE80E 交换网。

多级 Crossbar 交换网

多级交换结构是由多个单级交换单元互联起来的，每个交换单元都有一整套输入和输出，与普通矩阵交换类似，提供输入输出的连接。通过互联多个小的交换单元，就可以制造一个大型可扩展的交换结构。图 1-10 是一个通用多级交换网（具有 A^m 个输入端口，B^m 个输出端口）。

图1-10 多级交换网络



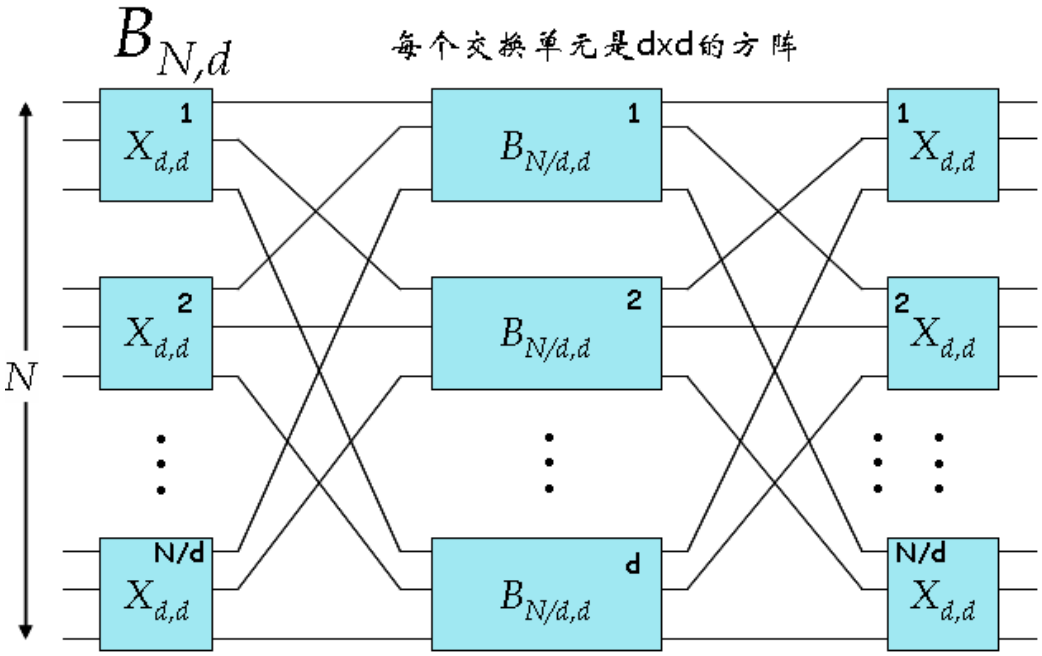
虽然单级结构的设计相对简单，成本也比较低，但是其不能满足下一代 Internet 扩展的需要。多级结构在操作上比较复杂，但是可以扩展到成百上千个端口，这对路由器的多框设计是必须的。

不同的单级交换单元和不同的级联方式，构成了多样的多级交换网络。最常见的是 Benes 交换网和 Clos 交换网，这两种交换网都是由其发明人的名字命名得来的。

- Benes 交换网

  Benes 网络由贝尔实验室科学家 Benes 于 1964 年提出。在这种构架中，每个小的交换单元都是 N×N 的方阵，Benes 网络在任意输入端口/输出端口之间提供 N/d 条可能的通路；提供较好的冗余度，中间交换级可以做到不停机维护。但 Benes 网络不能保证信元的先后顺序，因此需要额外的报文顺序整合措施。常见的三级 Benes 网络，$N=d^2$，第一级做信元分发，第二、第三级根据目的端口号做信元路由。

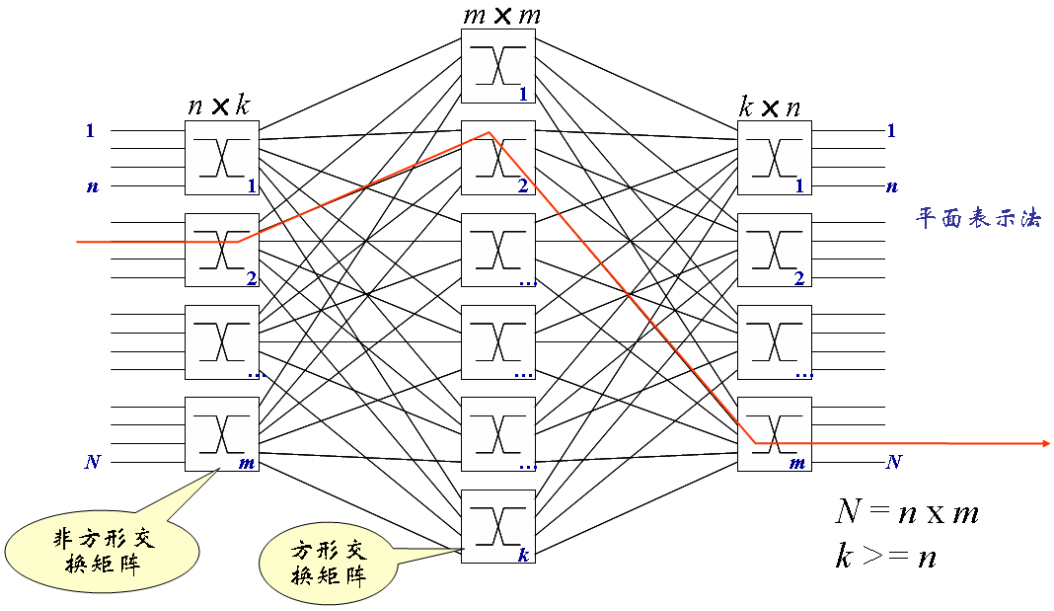
图1-11 Benes 多级交换网络



- Clos 交换网

  Clos 交换网络是有 Charles Clos 在 1953 年提出。Charles Clo 采用数学的方法证明了 Clos 网络一些非常重要的特性，即该交换网严格的无阻塞条件（Strictly nonblocking）和可重配置无阻塞条件(Rearrangeably nonblocking)。和 Benes 网络不同，Clos 第二级采用方形交换单元，而第一级、第三级采用非方形交换单元，即交换单元输入端口数和输出端口数可以不同。实现相同容量的交换网时，采用 Clos 结果可以减少交叉点，从而使网络具有更好的扩展性。例如实现 100×100 无阻塞的 Crossbar 需要 10000 个交叉点，而 Clos 只需要 5700 个交叉点，当网络规模更大时，Clos 能减少更多的交叉点。如今，Clos 结构仍然有很强大的生命力。

图1-12 Clos 多级交换网络



Clos 网络在华为的 NE5000E 上有应用。

1.4.4 交换网发展趋势

随着网络业务对物理端口容量需求的不断提高，路由器对交换网端口也提出了很高的要求，目前的发展方向主要是：

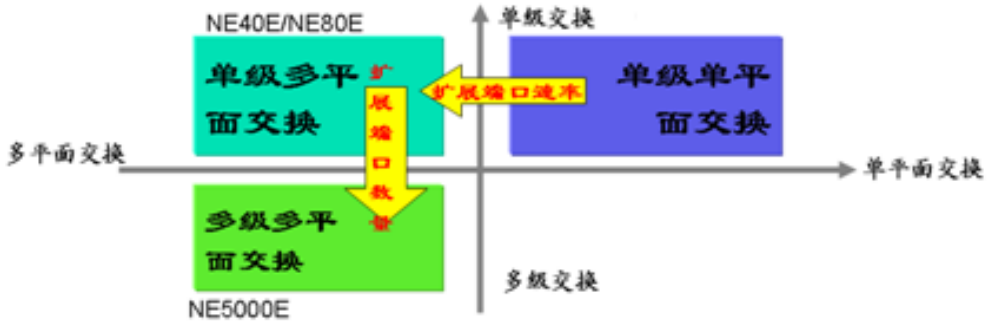
- 单端口速率提高：3.125G 提高到 6.25G，6.25G 提高到 10G，10G 提升到更高。
- 端口数量扩展：16 扩展到 64，……成倍数扩展。
- 各种交换网技术融合在一起共同发展。

  例如，业界把共享缓存交换网和 Crossbar 交换网融合在一起，开发出很多内部带缓存功能的 Crossbar 交换网。

说明

无内部缓存的 Crossbar 网络主要包含 Tandem Banyan（串联榕树）。虽然无内部缓存 Crossbar 网络的实现需要较少的逻辑，但如要取得像有缓存 Crossbar 网络一样的性能，需要大量的片内/片间高速互联，实现起来比较困难。

图1-13 交换网络发展趋势图



1.5 华为路由器交换网

华为路由器的交换网由专门的交换网板 SFU（Switch Fabric Unit）来充当，交换网板间为 N+M 备份。数据在上送交换网之前，都需要经过 LPU（Line Processor Unit）上的交换网接入处理器 FIC（Fabric Interface Controller）进行格式和接口转化。

1.5.1 NE80E/NE40E 交换网

NE80E/NE40E 采用单级多平面的 Crossbar 构架，其中 Crossbar 在交换网单元 SFU 上实现，交换网接入处理器在 LPU 板上实现。交换网一共有 8 个平面，分布在四个 SFU 上。下面分步说明数据报文是如何经过交换网的。

1. IP 数据包从 LPU 的物理接口输入，经过 LPU 上的 FIC（Fabric Interface Controller）时，数据包被分片处理，被分解成 Cell，经过缓存、调度后进入 SFU（Switch Fabric Unit）上的 Crossbar 交换单元；

说明

每个 FIC 和所有的交换平面相连，保证 Cell 单元能够均匀地被分配到各个交换平面上进行交换，这样不仅实现交换平面的负载分担，更有利于系统的容错处理。

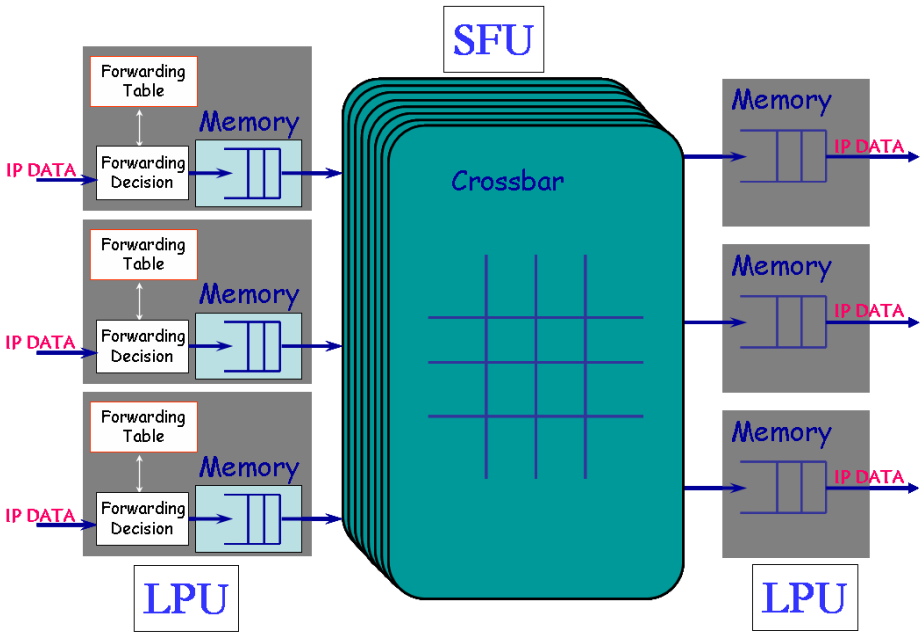
2. Cell 到达 Crossbar 后，Crossbar 将报文调度到交换网的目的端口，发送到 LPU 的 FIC 上，完成 Cell 的交换功能；
3. Cell 到达出端口的 FIC 后，FIC 将 Cell 重组成 IP 数据包，然后发送到 LPU 的目的端口，完成数据报文在路由器中的单级交换。

说明

NE80E 和 NE40E 上，FIC 是放置在 LPU 上的一颗芯片。

NE40E-8 只有 2 个独立的 SFU 板，另外 2 个 SFU 单元分别在 2 个 SRU（Switch & Routing Unit）上。因此，设备共有 4 个 SFU 单元，实现交换平面 3+1 备份。

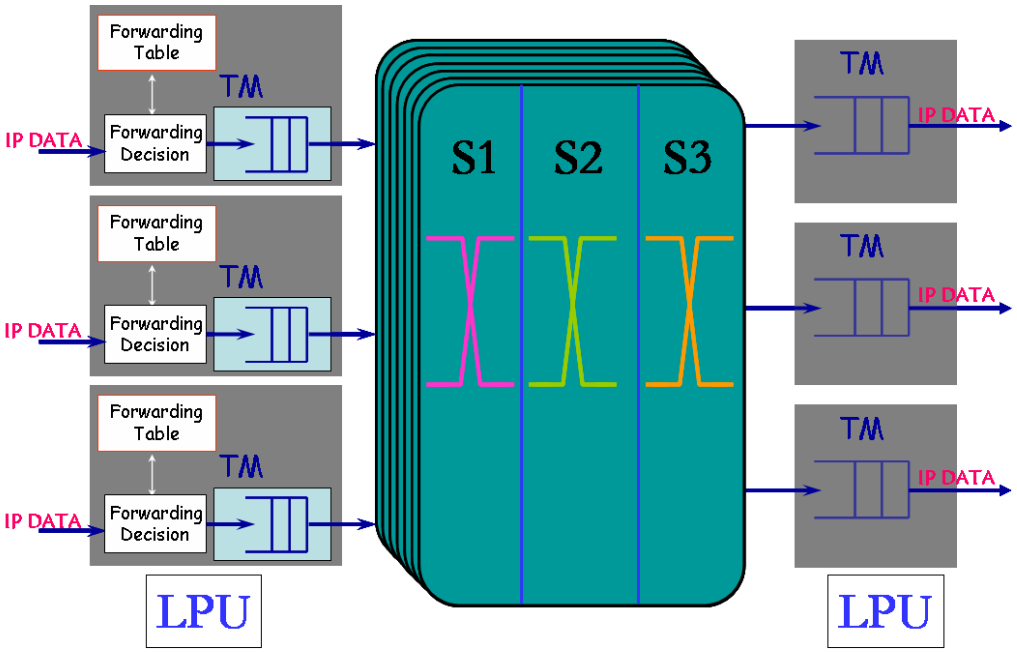
图1-14 NE40E/NE80E 交换网结构



1.5.2 NE5000E 交换网

NE5000E 的交换网采用内部带有缓存能力的 CIOQ（Combined Input and Output Queuing，输入输出相结合排队）交换网，在 NE5000E 单框和 NE5000E CCC-0 系统中，采用了单级多平面的交换网架构。在 NE5000E 的多框集群系统 CCC-1、CCC-2 中，采用了多级多平面的交换网构架，如图 1-15 所示。一个三级交换网，各级交换单元分别命名为 S1、S2、S3，下面介绍数据报文的交换过程。

图1-15 NE5000E 交换网结构



1. 数据包从线卡框 A（CLC A）LPU 的物理接口输入，经过 LPU 上的数据处理单元处理，然后发送到 TM（Traffic Manager），数据包在 TM 上被分片，数据被分解成 Cell，经过缓存、队列调度后进入第 1 级交换网 S1（即 CLC A 的 SFU）；

   说明

   每一个 TM 都和交换平面有一个或多个连接，这样就能保证 Cell 单元能够均匀地被分配到各个交换平面上。

2. Cell 到达 S1 后，交换网（fabric）根据目的端口将报文均匀分配到 S2（即 CCC 的 SFU），S2 进行选框交换，交换到目的线卡框 B 的 S3（CLC B 的 SFU），S3 进行选板交换，把报文交换到目的 LPU 上；

   说明

   单级交换网和多级交换网原理类似，只是多级交换网更容易组建大容量的交换网，提升系统交换性能。

3. Cell 到达目 LPU 的 TM 后，TM 将 Cell 重组成 IP 数据包，然后发送到 LPU 的目的端口，完成数据报文在路由器中的三级交换。

1.6 FAQ

1.6.1 如何从交换网容量计算其所支持的业务板接口容量？

一般来说，路由器的背板容量大于交换网容量，交换网容量大于业务板接口容量。交换网交换容量和业务板接口容量的比值为加速比。

为了降低成本及提高系统的可扩展性，一般将背板容量做得足够大，方便后续业务板扩容。交换网单元的容量一般只做到满足目前线路板的规格，并且能够考虑将来的升

级即可。比如配置了 2 个 SFUD 和 2 个 SRU 的 NE40E-8，目前交换网的容量为：4（交换网单元 SFU 数）×8（每个交换网单元交换网端口数）×4（每个交换网端口的 Serdes 数）×3.2G（每路 Serdes 的速率）×8B/10B（Serdes 编码效率）= 327.68G，分配到每个 LPU 的容量为 327.68 / 8 = 40.96G，如果按照加速比为 2 来评估，NE40E-8 的交换网容量每槽位最多只能支持 20G 的 LPU。

同样，对于配置了 4 个 SFUG 的 NE80E 来说，交换网的总的容量为：4（交换网单元 SFU 数）×16（每个交换网单元交换网端口数）×8（每个交换网端口的 Serdes 数）×3.2G（每路 Serdes 的速率）×8B/10B（Serdes 编码效率）=1.31T，分配到每个 LPU 的容量为 1.31T / 16 = 81.92G，如果按照加速比为 2 来评估，NE80E 的交换网，每槽位最多只能支持 40G 的 LPU。

说明

容量计算一般都按照收发双向之和来计算，即通常所说的双向容量。比如一个普通的 GE 端口，其收发速率各为 1Gbps，计算双向容量为 2Gbps，计算单向容量为 1Gbps。在本文中，如无特殊说明，都按单向容量进行描述。

如果是基于单向来计算容量，那么背板容量、SFU 交换容量、LPU 接口容量都需要是单向数据；如果是基于双向来计算容量，则都取双向数据，即度量标准要统一。

1.6.2 NE40E/NE80E/NE5000E 的交换网是否可以混插？

从外形来看，NE80E 和 NE5000E 交换网基本一致，但是通过前面的介绍，NE40E、NE80E 和 NE5000E 的交换网构架不一样，3 者的交换网板 SFU 不能混插。

1.6.3 NE40E/NE80E/NE5000E 的交换网能否不满配使用？

NE40E-8/NE80E/NE5000E 交换网单元都采用 3+1 备份设计，系统配置的 4 个交换网单元采用负载分担的方式同时工作，当其中有 1 个交换网单元出现故障或者被拔出时，剩余 3 个交换网单元能够支撑系统线速转发；当有 2 个或 2 个以上的交换网单元出现故障或被拔出时，系统虽能工作，但是转发性能会下降，因此现网中要避免此类情况发生。

NE40E-X8 交换网单元采用 2+1 备份设计，3 个交换网单元采用负载分担的方式同时工作。

1.6.4 NE40E/NE80E/NE5000E 的交换网是否支持热插拔？

NE40E/NE80E/NE5000E 的交换网单元在软硬件设计上都支持热插拔。为了提高系统在单板插拔时的可靠性，减少瞬间丢包的可能，交换网单元进行了 OFFLINE 设计，即在拔出交换网单元前请手动按下交换网单元面板上的 OFFLINE 按钮，通知系统启动必要的可靠性保护操作，等 OFFLINE 灯亮之后，方可拔出交换网单元。

1.6.5 NE40E-X3 有交换网单元吗？

NE40E-X3 采用 Full-Mesh 结构，没有单独的交换网单元。各业务接口板/子卡间通过高速总线网状互联。受网状互联设计的限制，Full-Mesh 结构的路由器一般所能支持的接口卡数量非常有限。

1.7 修订记录

版本	发布日期	修改记录
V1.0	2013-09-12	首次发布
V2.0	2014-08-26	″其交换网容量=［2×（9×4×16）］×6.25G bps×0.8 = 5.76G bps″,将 5.76Gbps 更正为 5.76Tbps。