

Internet 骨干路由器及发展中的 Internet 设计

Juniper 网络公司，爱立信公司，1998 年 9 月 16 日

引言.....	2
高速增长 Internet 对 ISP 的影响.....	2
给 ISP 带来更大增长空间的网络解决方案.....	3
路由系统的基本要素.....	3
Internet 骨干路由器的关键属性.....	5
M40 Internet 骨干路由器系统结构.....	5
路由软件：JUNOS Internet 软件.....	6
包查询：Internet 处理器 ASIC.....	6
路由寻址.....	7
可编程性.....	7
性能保证.....	7
原子更新.....	8
业务的可视性.....	9
交换结构：分布式缓冲器管理器 ASIC，I/O 管理器 ASIC 及共享内存.....	9
分布式缓冲器管理器 ASIC 和共享内存.....	10
I/O 管理器 ASIC.....	10

线路接口卡.....	11
数据包如何通过 M40 的包转发引擎.....	12
ISP 网络中的 M40 Internet 骨干路由器.....	13
故障条件下稳固的可靠性.....	13
M40 系统配置.....	14
结论.....	15

引言

伴随着我们迈入 21 世纪的步伐，Internet 正以惊人的速度发展。无论从任何一个角度衡量，如主机的数量、用户数、业务量、链路数、单条链路的带宽，或是服务提供商（ISP）网络的增长率，其增长都是惊人的。

以前，ISP 在 Internet 核心网络中只采用一般的普通设备。随着 Internet 的迅猛发展，一个专门为解决 Internet 骨干网运营商所面临的特殊问题的网络设备市场随之出现。这类新的产品设备不仅被要求能够集合带宽及业务的吞吐量，而且还要具备丰富的软件功能及控制特性。

为能同时实现提供增加的带宽及丰富的软件功能，新型路由系统的设计者必须在设计中采用以前只有在交换机中才具有的转发功能。但是，相对于为使用固定长度查询的固定长度数据包提供线速性能的直接交换，为应用最长匹配查询的不定长度数据包提供线速性能，其实现要复杂得多。特别需要说明的是，这种数据包的处理已不能通过基于微处理器或微处理器辅助的方式来实现，而必需使用一种基于随 Internet 环境变化而改变其路由软件的 ASIC 方式来完成。

这种新型的 Internet 骨干路由器所面临的系统上的挑战也是非常复杂的。它们不仅要能适应现有的 OC - 3 及 OC - 12 速率骨干网和相关的 Intra - POP 结构，而且还要能够支持具有 OC - 12 速率和吉位以太网 Intra - POP 结构的 OC - 48 速率骨干网络。另外，新的路由系统必须要加速 Internet 从“最努力”的服务向最基础的可靠性服务的方向转变。Internet 用户希望能够象使用 POTS 网络那样，在任何需要通信的时候，听到“拨号音”，便可得到高质量的服务。

Juniper 网络公司提供的 Internet 骨干路由器 M40 是目前世界上最完备的路由系统。它是第一个将 Internet 可伸缩性，Internet 控制及无可比拟的转发性能集于一身的系统。它不仅集成了只有交换机才具有的转发功能，而且还兼顾路由器软件的灵活性，可控制性，以及与在意外条件下系统的稳定性和可靠性。M40 可被应用于以光网为基础的 Internet 系统，它帮助 ISP 实现将 Internet 从“最努力”系统过渡到基础的可靠通信系统。

高速增长对 Internet 对 ISP 的影响

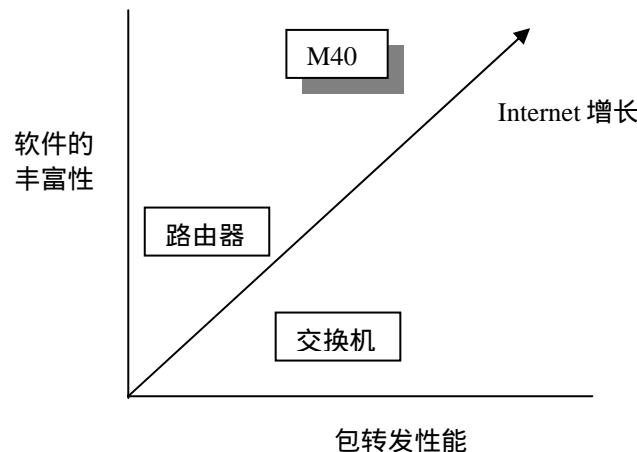
Internet 的高速增长对 ISP 如何为其客户的服务进行配置与管理都产生了深刻的影响。当 ISP 努力地扩展其网络的同时，他们会不断的碰到一些急需解决的问题：

- 在迅速提高业务量并满足客户需求的同时，增强网络可靠性的挑战（7×24 服务）。
- 在优化有效带宽的使用率及维护网络的可靠性方面，对流量工程工具的需求。
- 如何与不断发展的 Internet 技术保持一致的挑战：新的协议、新的路由硬件、新的交换硬件及光纤设备。
- 如何按照用户的需求及线路的可行性寻找足够的空间及适当的供电系统来安装设备。
- 如何在满足稳定性需求的前提下提供不同的业务以区别于其他竞争对手。

任何希望进入 ISP 市场的解决方案都需要面对所有这些挑战。如果一个新的系统不能够增强网络的性能并提供基本的可靠性，这个方案将注定失败。

给 ISP 带来更大增长空间的网络解决方案

ISP 需要那些既能提供丰富的路由软件功能又能提供高速包转发性能的解决方案。但对于 ISP 来说，很难找到一个能够同时满足两方面需求的单一系统，因为设备提供商未能开发出一个能同时满足两方面需求的令人满意的系统。过去，如果一个 ISP 希望能够在复杂的网络设计中使用丰富的软件功能以增强对网络的控制，他必须对路由器进行配置并忍受其较慢的



转发性能。

图 1：ISP 进退两难的局面

为解决这些约束，一些 ISP 决定使用“覆盖”的解决方法，即在网络的边缘使用 IP 路由器以提供软件的丰富性，而在网络的核心部分使用 ATM 交换机，以提供较高的速度。在以 OC - 12 速率传输的网络中，ISP 检测了一些有效的工具，将他们以复杂的方式组合在一起，达到了预期的效果。但是这种覆盖型的网络同时也带来了他们自身一些管理上的问题：

- 网络的物理拓扑结构与逻辑拓扑结构不匹配
- ATM 信元税 (cell-tax) 导致不能够对所提供的带宽充分利用
- PVC 的全闭合网导致的“ n^2 ”比例问题
- 覆盖型网络解决方案需要协调两个不相关的网络的管理：ATM 网络和覆盖的 IP 网络

尽管有这些局限，许多 ISP 还是决定使用这种覆盖型的解决方案，因为这是他们能够建立高带宽网络的唯一选择。但是，如果能有一种专为简化高带宽网络设计，并且比覆盖型网络解决方案更便宜、简单的系统，则无疑会成为 ISP 们共同的选择。新一代的 Internet 骨干路由器

将提供这一性能，它将使网络设计人员不再为仅仅为了满足基本的速率控制的需求，而在不同层配置备份设备而烦恼。在我们讨论新一代 Internet 骨干路由器所必须具备的性能之前，让我们先回顾一下一个路由系统所需具备的基本要素。

路由系统的基本要素

所有的路由器必须完成两个基本任务：路由及包转发。路由的过程主要是收集网络拓扑信息并建立转发表。包转发则负责依据转发表中所包含的信息将数据包从路由器的一个输入接口复制到适当的输出接口。

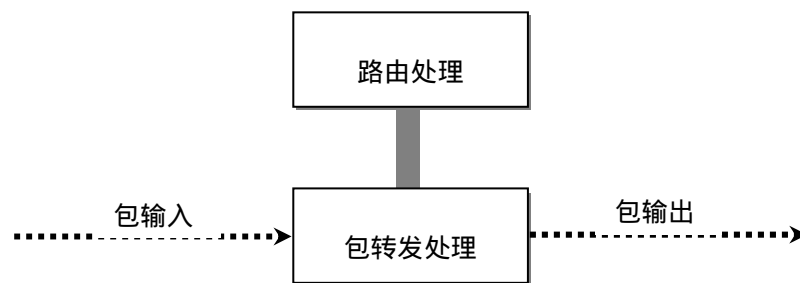
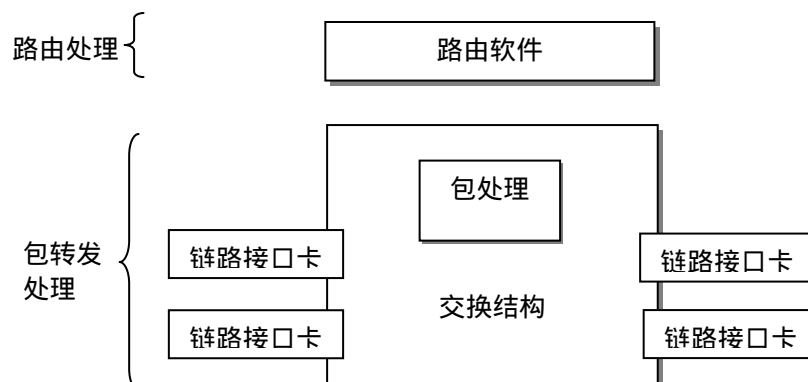


图 2：路由与包转发处理

任何一个路由系统都需要四个基本要素以实现路由及包转发的处理：路由软件，包处理，交换结构和线路接口卡。对于任何为工作在 Internet 骨干网上而设计的系统，所有四个要素的性能必须要一样强大，因为一个高性能的路由器的性能只与这四个要素中性能最差的一个



相一致。

包处理的过程可分配至每个线路卡执行，在路由器的中心处理

器部分执行，或以一种混合的方式来完成

图 3：路由系统中的基本要素

路由软件作为系统的一部分，负责执行路由功能。它负责维护对等关系，运行路由协议，建立路由表并建立转发表，以供系统的包转发部分访问。软件同时也为系统提供控制功能，包括：流量工程，用户接口，策略及网络管理。

每个进入系统的数据包都需要进行一系列与数据包长度无关的处理进程，这个过程与路由器的结构无关。进入系统时，数据包的封装必须被拆除，然后进行最长匹配路由查询，接下来，数据包需在输出端口排队并进行输出封装。如何实现最长匹配查询及相关的高速数据包处理，是设计高性能路由系统所面临的最艰巨的挑战。

交换结构提供数据包在路由器接口卡间移动的结构。设计人员已为此工作了数十年，其结果也被生产厂商很好的理解。已有许多芯片可被路由器厂商用于设计交换结构。这些方案包括纵横交换，榕树网络，Clos 网络，理想正移网络及其它。

线路卡用于端接不同物理媒体类型的线路，提供诸如 DS - 3，ATM，SONET，帧中继及 PPP 的第一层及第二层的有关技术。有关线路卡设计的有关技术也同样被生产厂商所很好的掌握。线路卡可依据那些规定了物理接口类型、光特性、电平等技术指标的主流标准进行生产制造。

Internet 骨干路由器的关键属性

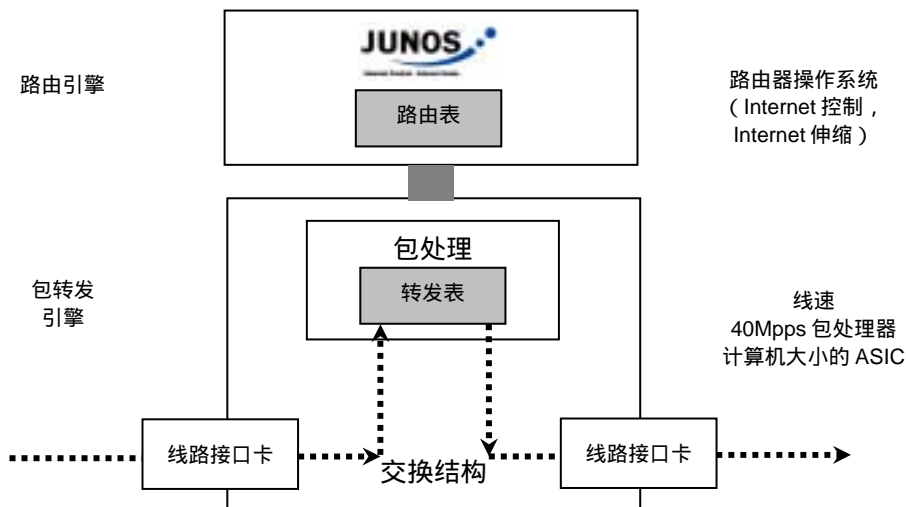
下一代的 Internet 骨干路由器必须被设计成为在光纤网络结构上提供 Internet 可伸缩性，Internet 控制，及无法比拟的性能的系统。Internet 骨干路由器所必备的主要特征包括：

- 由 Internet 专家编写，并在大型 ISP 网络上进行了成功的互操作测试的可靠、完整的路由软件。

- 支持大型 ISP 网络高效的带宽利用率，提供崭新的具有复杂控制功能的流量工程特性。
- 数据包处理需在与数据包长度及系统设置无关的情况下，以线速完成输入拆封装，路由寻找，排队及输出封装的处理过程。
- 交换结构的设计留有较大裕量，提供 40Gbps (8×OC - 48) 的有效汇聚带宽，以支持 OC - 48 速率的骨干网传输。
- 提供多种支持线速性能的接口类型。
- 机箱的端口密度最低应为每机框英寸一个插槽。
- 机械性能，适用性及管理等方面的性能，使系统在大型 ISP 网络的核心部分易于配置。
- 维护整个网络稳定性的能力，并且可以在不影响网络其它部分的情况下适应变化的环境。

M40 Internet 骨干路由器系统结构

M40 系统的基本结构 - 完全独立的路由功能和包转发功能 - 是通过将系统设计成为两个独立的组成部分：路由引擎和包转发引擎而完成的。这种分离设计的优势在于：即使路由功能极不稳定，也不会影响包转发引擎的性能。同样，即便通过非常大的业务量也不会影响路由引擎维持对等关系及计算路由表的能力。这两种功能的分离设计，使得一个系统可以同时提供



高效的转发性能和高可靠性的操作系统。

图 4：M40 系统的体系结构

下面的部分，我们将集中讨论 Internet 骨干路由器 M40 是如何为路由器结构中的关键要素，如：路由软件，基于 ASIC 的包处理及查询，交换结构和线路卡，提供领先的解决方案的。通过提供一个具备同样强有力的四个关键要素的系统，M40 为 Internet 骨干网提供了一个无可比拟的，极佳的解决方案。

路由软件：JUNOS Internet 软件

为满足 ISP 苛刻的要求，Juniper 网络公司从最底层开始开发了属于自己的一套软件系统 - JUNOS。该软件系统具有如下特点：

- 运行在保护内存下的模块化结构的软件，为系统提供了高可靠性及可伸缩性。
- 行业级的路由协议工具 - BGP，IS - IS，OSPF，路由反射，联合，组等，以满足 ISP 对控制及管理网络的需求。
- 非常灵活的策略定义语言简化了对上千条路由的路由策略管理。
- 通过使用多协议标记交换 (MPLS) 技术的流量工程，可最大限度地利用珍贵的网络资源，并且为提供不同种类的业务，包括一些新的业务提供基础。
- 用户界面提供多用户访问等级，配置更改控制，支持 ASCII 文件，并且具有恢复到以前设置版本的功能。为使软件设置错误的发生机率降至最低，它还具有将多步设置合并成一步完成的功能。
- 系统安全性通过对用户接口上的安全命令解释程序 (SSH) 访问来实现，TCP/MD5 用于 BGP 对话期，结构化安全装置用于抵抗拒绝服务的袭击等。

Juniper 网络公司充分地意识到路由软件对 ISP 们控制及管理他们的网络是何等重要，因此，JUNOS 已在世界上最大的 ISP 网络中经过测试及证明。

若您希望对 JUNOS 软件系统，包括流量控制的特点及优势，请参考 Juniper 网络公司的白皮书：优化路由软件，促进 Internet 可靠增长。

包查询：Internet 处理器 ASIC

包处理技术是 Juniper 网络公司在 Internet 骨干路由器技术上从根本上领先的一个部分。所有的路由查询都通过一个大小类似于计算机的微处理器 ASIC 来完成，但这块 ASIC 的功能及复杂程度要明显超过那些用于其他通信设备上的 ASIC。Juniper 网络的设计队伍是由许多硅谷中设计高速计算机及 ASIC 的精英们组成的，其最新设计的 ASIC 只能在极先进的计算机中才能发现。Juniper 网络的 ASIC 设计水平在通信设备领域中处于领先地位。

Internet 处理器 ASIC 是实现 M40 系统的先进性能的核心，其可编程性及路由与转发的一致性是最基本的技术先进点。Internet 处理器包含了超过六百五十万个晶体管和超过一百万个门电路。作为比较，新型的英特尔奔腾 II 处理器有七百五十万个晶体管。通过 Internet 处理器 ASIC 与 JUNOS 软件的组合，使 M40 在 Internet 骨干网中成功地架起了路由器与交换机性能之间的桥梁。

路由查询

Internet 处理器 ASIC 可以提供每秒四千万条路由的最长匹配查询，这个速度是目前应用在 Internet 上的，基于微处理器查询速度的一百倍。一块 Internet 处理器 ASIC 即可为 8×OC-48 系统的线速查询提供足够的处理能力。

另外，Internet 处理器 ASIC 可被配置成为带有每报头报表最长匹配查询。每报头报表提供 IP 转发表中被转发至各报头的字节及包的数量统计。ISP 可通过这些统计了解在其网络中流通的业务量的分布情况。

可编程性

Internet 处理器既可以作为一个普通的查找引擎，又可以作为一个完全可编程的查找引擎。在最初的版本中，它支持 IPv4（包括 IP 多点传送）及 MPLS。由于 Internet 处理器的可编程性，它可以方便的通过对包转发引擎重新编程，为路由引擎开发新的路由软件以及为 Internet 处理器传送新的转发表，从而可支持如 IPv6 及帧中继等其它协议。

Internet 处理器的可编程性能，使得现在的 JUNOS 软件可用一直持续发展并保持对硬件的完全支持。由于总会有一些意想不到的路由被开发和处理，因此，Internet 处理器 ASIC 的可编程功能使得 Juniper 网络公司可以通过现有硬件的基础上，利用软件的新功能以适应将来的考验。

运行性能保障

Internet 处理器 ASIC 使性能保障这一关键特性变得易于实现。性能保障通过隔离路由与转发，进而保护路由和转发功能的运行，以达到保障系统可靠性的目的。当系统出现故障时，通常会发生以下三个事件：

- 因为一定数量路由发生变化，因此路由处理进程将会感到压力。路由压力的出现是显然的，因为每个路由器都需要维持对等关系，传输并处理路由更新信息，计算一个新的最短路径树，将策略加到整个路由选择进程上，并修改包转发引擎使用的转发表。
- 包转发进程将会感到压力，这是因为路由器仍被连接到当前的网络链路上，突然发现他们将被迫工作在峰值状态或比平时工作水平高出很多的状态。
- 链路及路由器将被迫传输因发生故障而产生的额外业务，这样，整个网络的可靠性将到达一个临界的状态。如果这些网络单元不能够承载增加的业务及并对故障进行处理，则相关的简单的本地故障将以级联的方式扩散到整个服务提供商的网络。

服务提供商所面临的挑战是，传统的路由器结构在故障条件下不能够提供性能保障，从而无法为系统提供极为坚实的可靠性。Juniper 网络公司确信 M40 系统可为这一棘手的设计目的提供行业中最好的解决方案。它通过以下方法以提供可靠的性能保障：

- M40 提供了一种新型的分离结构，它将路由引擎和包转发引擎功能完全分开。这种设计使 M40 系统中的各个功能部分相互独立，因此，当系统的某一个部分感到压力时，不会影响到系统其他部分的性能。

- 确保 Internet 处理器 ASIC 的查询性能永远不会被降低。与查询的长度及路由表的大小无关，Internet 处理器 ASIC 可以提供 40Mpps 的查询速率。这一每秒 40M 条查询的性能，是在使用 80,000 条路由，80,000 不相关的目的地址，并不影响当前 Internet 状态的情况下实现的。
- 通过支持革命性的转发表原子更新的概念，允许包转发引擎可在转发表进行高速更新的同时，继续维持其高速转发的性能。

原子更新

Internet 处理器 ASIC 支持对其中心转发表进行原子更新 (图 5)。在 M40 系统中，路由表中包含了通过路由协议与邻居交换来的路由及静态设置产生的路由信息。转发表通过路由表生成，它包含了一个带有输出端口信息的 IP 报头 (或 MPLS 标记) 的索引。包转发利用转发表中的信息作出转发决定，而不是利用路由表中所包含的信息。

通常，修改一条特定的路由只会影响转发表中数据结构的一小部分。这意味着路由引擎只需在空闲内存中更新二叉树的一部分，然后将指针指向新的二叉树，更新便随之发生了。因为转发信息不必要被发布至多个线路卡，因此，在转发表被修改的时候，Internet 处理器并不要求将转发表锁定。在进行路由查询时使用新树还是旧树，只依赖于位置信息是否在指针被改变之前被读取。这一设计的优点在于：转发表在任何时候都保持其连续性，这就意味着，即使在路由不稳定的情况下，包转发引擎可使它的转发表继续接收更新，同时高速做出转发决定。

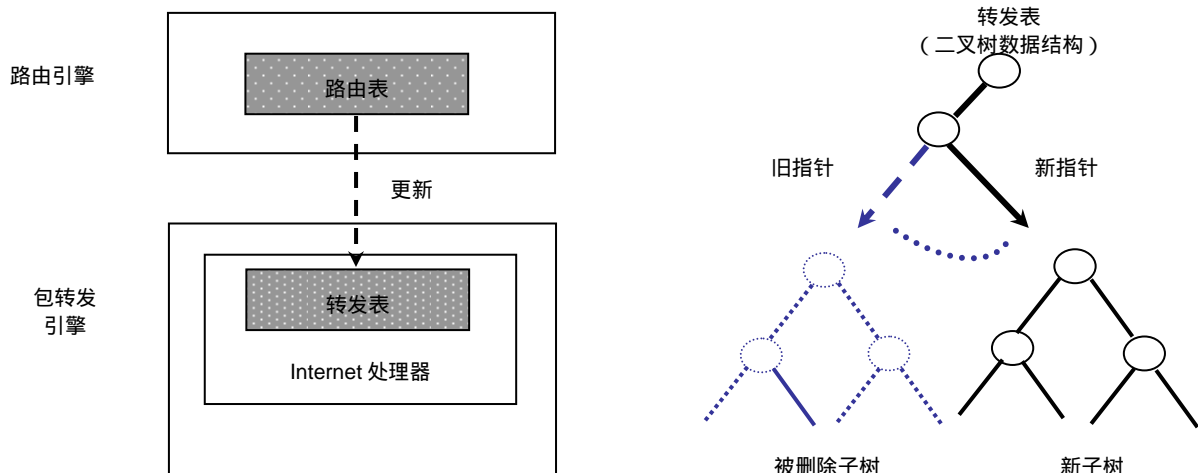


图 5：原子更新

其它高性能路由器的结构一般在其各自的线路接口卡上进行包查询。这就意味着，由于路由改变而使得转发表必须被修改时，该更新必须被发送至每个独立的线路卡。在路由更新及发送的过程中，由于中心表被修改，因此，转发表必须被锁定以保持路由的一致性。结果，包将需要进行排队，直到转发表更新完毕并被解锁。在转发表被解锁后，包才能继续转发。由此，我们可以很清楚的看到，在由于一些原因使得路由高速改变，通过路由器的业务动态增加的情况下，锁定转发表将降低系统的性能。

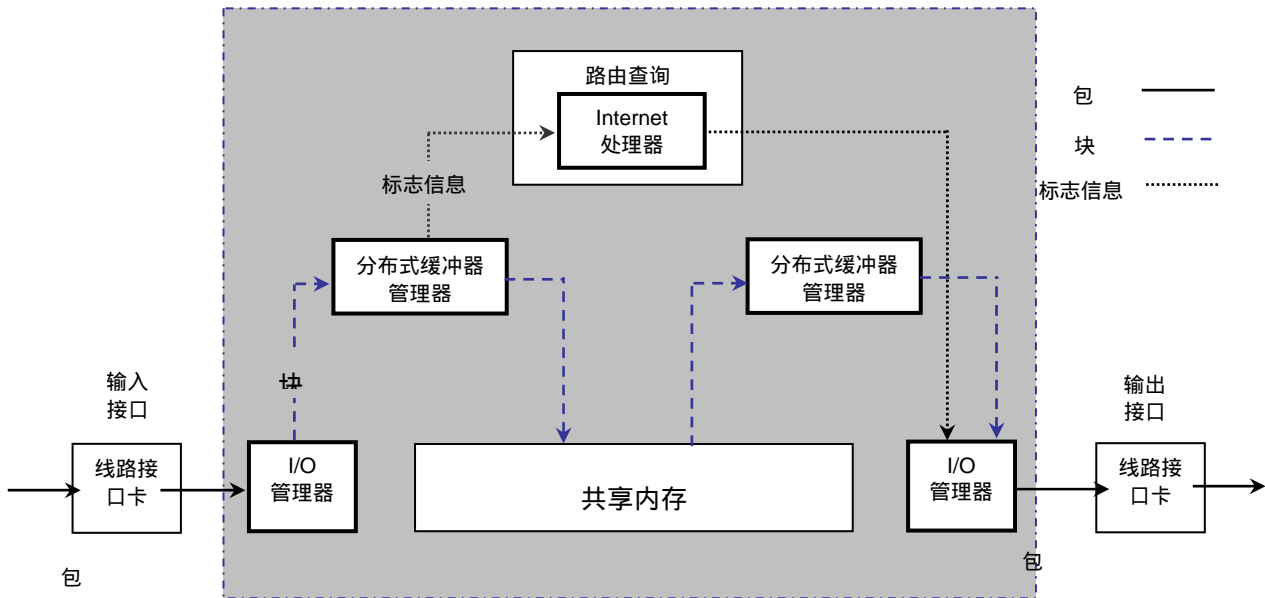
业务的可视性

因为骨干网路由器将以 40Mpps 的速率转发业务，ISP 需要了解业务在其网络中的流向趋势。例如，一个新的应用可能会影响整个网络的性能或使大量的业务在网络的不同部分移动。Internet 处理器的业务采样功能可使 ISP 对业务进行采样 - - 如 1,000 个数据包取 1 个采样。

业务采样通过向系统的其它部分发送一个带有标记的标志信息来完成。这个进程的发生不会影响 Internet 处理器的查找性能。标志信息是 Juniper 网络公司定义的一个数据结构，它包含了一个数据包被存储到共享内存后对其进行处理所需的全部信息。基于这些包含在标志信息中的信息，数据包可在共享内存中被重新获得，并将其转发至一个在路由引擎中执行的用户进程，这个过程并不影响包转发引擎的性能。

交换结构：分布式缓冲器管理器 ASIC，I/O 管理器 ASIC 及共享内存

M40 系统通过共享内存系统，提供一个有保留的，速率至少为 40Gbps 的交换结构。除了 Internet 处理器 ASIC 外，交换结构还包括分布式缓冲器管理器 ASIC，I/O 管理器 ASIC 和共享内存系统。每个包都将被分成 64 个字节的数据块结构，以利于在共享内存中存储。同



时，一个描述数据包报头的标志信息被转发至 Internet 处理器 ASIC，用来进行路由查询。

图 6：M40 包转发引擎

从厂家的观点来看，共享内存互联的限制是在技术上设计和实现的难点。但是，这并不意味着厂方将不去使用基于此种结构的交换结构。对于这样大的一个系统：8×OC - 48，或更大的系统，通过共享内存结构可提供一种可靠的途径，并使系统获得许多优势。

分布式缓冲器管理器 ASIC 和共享内存

M40 系统中所包含的两个分布式缓冲器管理器 ASIC，每个都包含多于一千零五十万个晶体管和一百七十万个门电路。类似于 Internet 处理器 ASIC，分布式缓冲器处理器 ASIC 同样提供完全的灵活性，它也可以通过 Juniper 网络公司的编程来分析第三层的报头，并根据 Internet 现在及将来的需求，产生路由查询的关键字。

分布式缓冲器管理器 ASIC 管理 Internet 骨干路由器 M40 的共享内存系统。每个 FPC (灵活的 PIC 集中器) 提供 128M 的共享内存和四个物理接口集中器 (PIC) 插槽。分布式缓冲器管理器将每个 FPC 上的内存作为系统共享内存池的一部分。

分布式缓冲器管理器 ASIC 连同共享内存结构, 为 M40 交换结构提供以下性能特点:

- 由于采用了分布式缓冲器管理器 ASIC, Juniper 网络公司开发出一个提供简单的单级缓冲器的系统, 即, 只从共享内存中读写一次。相对于需要提供多输入排队以防止线路阻塞的纵横交换式结构来说, 共享内存结构是完全没有阻塞的。
- M40 共享内存的互联容量是被过盈设计的, 使它可以轻松地支持 8 个线速为 OC - 48 接口的全双工操作。这个设计使得 M40 可将所有的排队及丢弃算法在输出接口完成, 因此, M40 不会在输入接口便将包丢弃。
- 除了单级缓存可使内存的使用效率最大化以外, 每个 FPC 提供 128M 的共享包缓冲内存, 因此, 支持大容量带宽 - 延迟设计。大容量的带宽 - 延迟设计允许末端站端的 TCP 会话期“使所有通道充满数据”, 并且获得更好的性能。
- 单级缓存共享内存的使用提供了一个低等待时间系统的结构。但是, 应该强调的是, 等待时间主要是针对一个 LAN 系统, 而对于一个 WAN 系统, 带宽 - 延迟设计便成为一个主要的因素了。
- 存储在 SDRAM 共享内存中数据的完整性是通过使用纠错码 (ECC) 内存来保证的。ECC 内存保证随机比特错误不影响数据的完整性。
- 因为系统结构使路由引擎与交换结构完全独立, 路由引擎并不消耗交换带宽。

I/O 管理器 ASIC

每个 I/O 管理器 ASIC 含有超过一千万个晶体管及一百万个门电路。I/O 管理器 ASIC 需要对不同的输入及输出数据包提供不同的功能。因此, 对于 ASIC 来说, 其在包转发处理过程中的主要职责将被分为三个独立的运行模块: 输入处理器, 内存访问接口, 输出处理器。

- 输入处理器主要作用于输入包。它负责从线路卡接收数据，解析第二层的报头，提供包 - 块的分解。
- 内存访问接口即作用于输入包，又作用于输出包。它与分布式缓冲器管理器联系并对特定 FPC 缓冲内存进行操作。对于输入包，其负责响应由分布式缓冲器管理器 ASIC 发起的包写入请求。对于输出包，I/O 管理器 ASIC 负责响应由分布式缓冲器管理器 ASIC 发起的包读出请求。
- 输出处理器作用于输出包。它负责从分布式缓冲器处理器 ASIC 接收包标志信息，提供服务等级选择，输出标志信息排队，加权循环排队服务及随机早期检测 (RED/WRED) ；管理静态计数器；提供块 - 包重组；提供第二层的输出封装。

类似于 Internet 处理器 ASIC 和分布式缓冲器管理器 ASIC，I/O 管理器 ASIC 也以可通过其极高的可编程性能提供最大的灵活性。它可通过 Juniper 网络公司的编程以提供对第二层的解析和对 PPP，帧中继和 MPLS 的封装。至于支持服务等级，I/O 管理器 ASIC 提供多种选项，用以分配排队及输出队列服务标志信息，同时也包括控制丢弃文档及包丢弃进程。最后，I/O 管理器在支持 IP 多点传送业务中的一次写入，多次读出的有效转发需求中作为一个完整的角色。

线路接口卡

M40 的线路卡通过基于媒体特性的 ASIC 来实现。例如，Juniper 网络公司在了一块高度集成的 ASIC 上集成了全部的 SONET/SDH 处理进程。对于其它厂家的系统，SONET/SDH 处理进程通常是通过一系列不同的组成部分来实现，而不是通过一块高度集成的 ASIC 来实现所有的处理功能。

M40 系统的线路卡 ASIC 可提供更高的端口密度，更好的性能，更低的功耗及增强的可靠性。超级 POP 要求路由器能够支持大量的不同的接口类型和高端口密度，这样，路由器才能适应 ISP 需求的不断增长和改变。M40 系统则可以在 ISP 不断发展的任何时期，适应大范围的超级 POP 环境。

线路卡的不同种类和插槽的数量，增强了 M40 系统在超级 POP 环境中的配置灵活性。一个全配置的 M40 系统可提供 32 个插槽 - 行业领先的，每机架英寸一个插槽的端口密度 - 为线路卡的安装提供了混合及匹配安装的灵活性。因为交换结构被过盈设计，因此，所有线路卡对任何包尺寸都可工作在线速率上。1998 年，线路接口卡的类型包括：

- OC-48 IP over SONET/SDH
- OC-12 IP over SONET/SDH
- OC-3 IP over SONET/SDH
- OC-12 IP over ATM
- OC-3 IP over ATM
- 带有内建 DSU 的 DS - 3
- 千兆以太网

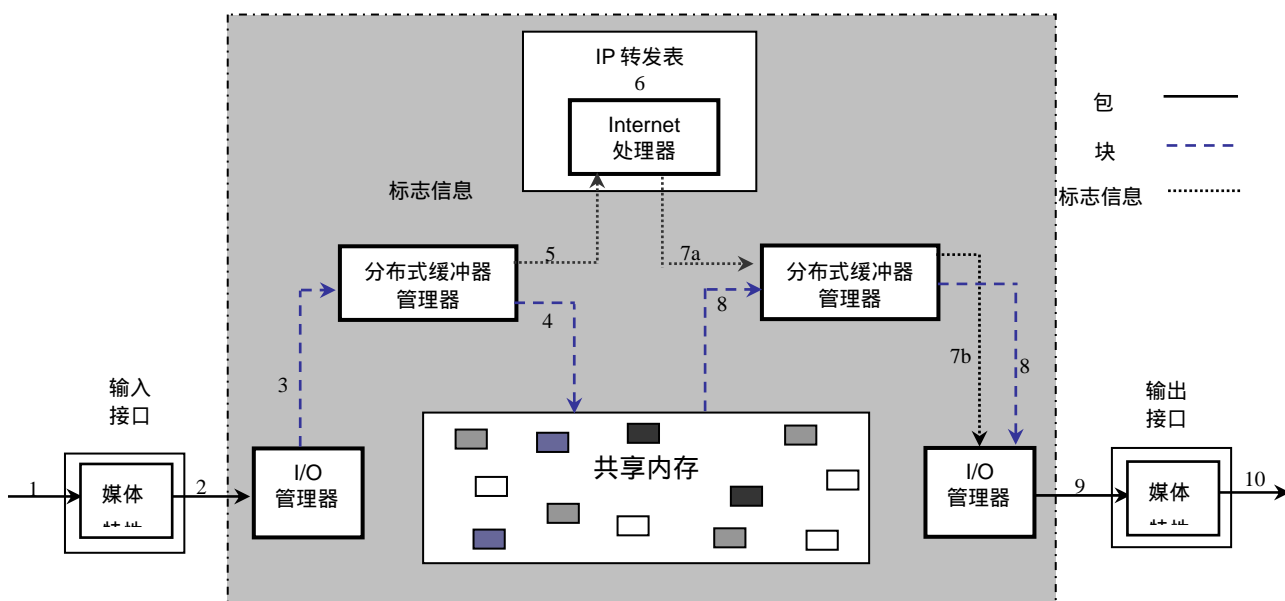
线路接口卡可在任意插槽按需要被混合安装及匹配。(OC - 48 线路卡除外，它将占用 4 个 PIC 插槽)

线路接口卡类型	每个 M40 系统的最大端口数	每个 7 英尺机架的最大端口数
OC - 48 SONET/SDH	8	16
OC - 12 SONET/SDH	32	64
OC - 12 ATM	32	64
OC - 3 SONET/SDH	128	256
OC - 3 ATM	64	128
DS - 3	128	256
千兆以太网	32	64

表 1 : M40 端口密度

数据包是如何通过 M40 包转发引擎的

首先，一个包到达 M40 系统的输入接口（步骤 1）。该端口的媒体特性 ASIC 提供所有媒体特性细节，包括从 SONET 帧，HDLC 帧中移去有效载荷，校验和。JUNIPER 网络共



计并开发了三种媒体特性 ASIC：第一种是为 SONET，第二种为 ATM，第三种为 DS3 接口。

图 7：M40 系统中的包转发过程

接下来，一个串行的字节数据流从媒体特性阶段传送到 I/O 管理器 ASIC（步骤 2）。I/O 管理器决定此帧是否为 IPv4 或 MPLS，并识别第三层包的开始。I/O 管理器同时也在保标志信息内设置一个可能被用于不同服务的标志。最后，I/O 管理器将包分割成为 64 字节的块，并将每个块传递给分布式缓冲器管理器 ASIC（步骤 3）。这些块的大小是为了有效的在共享内存中存储并从新获得而定义的，与 53 字节的 ATM 信元无关。分布式缓冲器管理器 ASIC 将这些块平均的以循环的方式分布到共享内存中去（步骤 4）。

在与将每个块分布到共享内存中去的同时，分布式缓冲器管理器 ASIC 从其收到的块中提取出路由查询关键字，并形成标志信息包。标志信息包是一个由 Juniper 网络定义的数据结构，其包含了对存储在共享内存中的包进行处理所需的全部信息。对于一个单点传送的 IPv4 包，分布式缓冲器管理器 ASIC 决定其输入端口，目标 IP 地址，源 IP 地址，协议值及源与目的地的 TCP/UDP 端口数。对于 MPLS 帧，分布式缓冲器管理器 ASIC 从输入端口提取路由查询关键字及 MPLS 标志值。在收集完这些信息之后，分布式缓冲器管理器将标志信息转发给 Internet 处理器 ASIC（步骤 5），通过它对该包作出转发决定。

Internet 处理器 ASIC 进行路由查询。对于 IPv4 包，它将对目的报头在转发表中使用最长匹配查询。对于 MPLS 帧，Internet 处理器将在 MPLS 转发表中提供精确匹配查询。通过查询，Internet 处理器将含有转发决定的标志信息传送给分布式缓冲器管理器 ASIC (步骤 7a)。分布式缓冲器管理器将标志信息转发给输出端口的 I/O 管理器 ASIC。对于 IP 多点传送帧，Internet 处理器将标志信息转发给每个输出端口。

随后，在输出接口上的 I/O 管理器 ASIC 负责管理包排队。包本身并不进行排队，而是由标志信息代替包进行排队，真正的包则仍以块的形式存储在共享内存中。对于 IP 多点传送包的特殊情况，每个输出接口的 I/O 管理器将独立地对包标志信息进行排队。

对于每个输出端口都有四个队列，每个队列都通过设置共享物理连接的一部分带宽。在输出接口上的 I/O 管理器可在决定一个包的排队时将一些因素，如 IP 优先权比特值，输入接口的使用率，目的地址等，按 RED/WRED 算法考虑进去。

当包标志信息达到其队列的头部并准备好进行传输时，I/O 管理器通过分布式缓冲器管理器产生一个请求，从共享内存中将该包的块读出 (步骤 8)。I/O 管理器将这些块重组为包，并将帧结构转发至输出接口的媒体特性 ASIC (步骤 9)。

最后，输出接口上的媒体特性 ASIC 通过提供一些必要的媒体特性操作，如 PPP - over - SONET 编码和 HDLC 组帧，在 SONET 帧中放置比特位，定义 SONET 帧的有效载荷的开始，并在光纤上将比特位串行化 (步骤 10)。由此，包离开包转发引擎转向下一跳，沿着路径向目的地行进。

Internet 骨干网路由器 M40 在 ISP 网络中的应用

除了具备作为 OC - 48 速率的 Internet 骨干网路由器所能提供的基本性能和良好的特性外，M40 系统还能够满足 ISP 提出的其它要求：在故障条件下极为牢固的可靠性和在超级 POP 环境中简易的可配置性。

故障条件下稳固的可靠性

任何一个 ISP 最关心的事情莫过于整个网络的可靠性，而不是某个单个系统的可靠性。即便 M40 系统在软件和硬件方面都被设计得极为可靠，ISP 的基本观点是使任何网络故障局部化。ISP 们不希望他们的路由器运行在这样的一种状态：当一个局部故障发生时，在网络中引发级联效应，产生附加的故障。

为了提供一个适当的稳定等级，ISP 通常对其网络进行适当的设计并预测网络的不可能性，使得客户永远不被隔离而导致完全失去服务。为实现这一目的，ISP 配置备份网络，安装大量的备份链路，在通常的工作条件下，以大约 50% 的负荷量运行。网络的这种能够调节一些故障并同时能继续对所有业务进行交换的能力，在设计中是很很难实现的。ISP 因为以下几点原因而使用这种保守的方案：

- 如果一个路由器或一条链路出现故障，网络的其他部分仍有有效的负荷量以承载附加的业务量。在例外的条件下，网络的大部分继续以 50% 的负荷量运行，但是一些特定的网络单元因业务量的增加，可能需要在接近 100% 的负荷量下运行。
- 保守设计可为 ISP 在高速增长及快速发展的时期提供一些摆动空间。如果 ISP 不能在网络拓扑上提供适当的负荷容量，它可能需要接受一些额外的风险，并使吞吐量降低 60% 或 70%，直到能提供有效的传输容量。

这种传统实现方式的挑战在于，现存的路由器通常在低负荷的条件下运行良好，但是，一旦业务量动态的跳跃至 100% 时，它们便不能提供稳定的性能。M40 互联网骨干路由器相对于其它系统，在极限压力下提供牢固的可靠性能方面是独一无二的：

- M40 互联网骨干路由器对于路由处理和包转发都提供了足够的空间。在意外条件下，路由引擎继续接收和发送路由更新，提供路由计算，维护对等关系，对接口失效进行处理等。同样，包转发引擎继续以 40Mpps 的速率对包进行交换，而与包的大小及系统负荷无关。

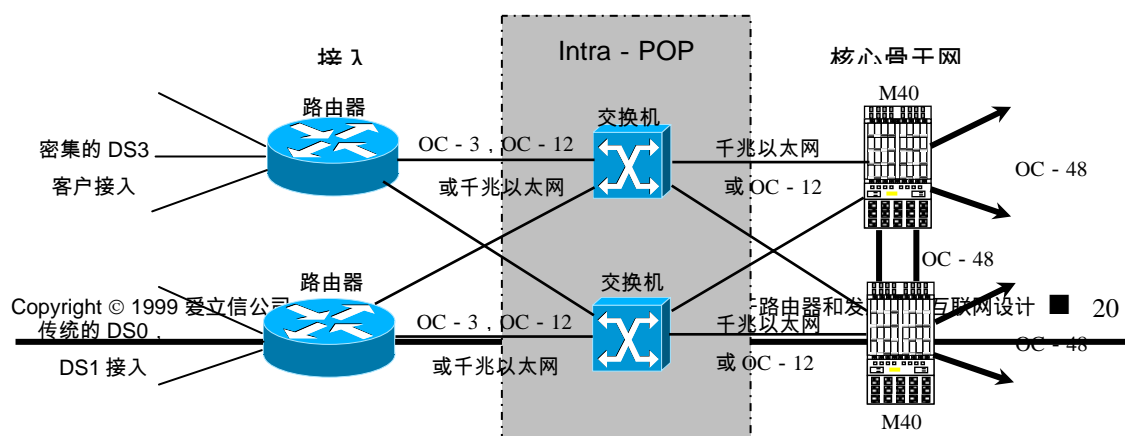
- 提供路由进程与包转发进程完全分离的结构，原子更新使得包转发引擎与路由引擎并行工作，不会影响转发性能。在外意情况下，原子更新使得 M40 系统避免了一些仍工作的链路的不稳定性，因此，消除了级联故障的主要原因。
- 由 JUNOS 互联网软件提供的流量工程特性允许 ISP 对网络故障进行管理。相对于缺少对故障的预留传输容量，M40 可提供一些工具，使 ISP 能够在有效资源上对现有的业务量进行最佳的分配，而不会在网络中引起阻塞及其它一些不稳定的情况。

M40 系统的配置

M40 系统是专为不断面临超级 POP 运行环境挑战的电信运营商及 ISP 而设计的，它可以使 ISP 将其现有的基于 OC - 3 和 OC - 12 速率的骨干网升级到基于 OC - 48 速率的骨干网。M40 Internet 骨干路由器在许多方面是行业领先的：

- 尺寸 - 每个互联网骨干路由器的机箱高度为 35 英寸，因此，一个高度为 7 英尺的机架可安装两个机箱。
- 功耗 - M40 系统的每机框英寸的消耗小于 1 安培
- 性能密度 - M40 系统在每机框英寸上提供高于 1Mpps 的转发能力，这是目前其它同类工作在 Internet 骨干网上的路由器性能的 10 倍。
- 插槽密度 - 一个完全安装的 M40 系统可提供每机框英寸一个插槽的插槽密度，是目前其它类似的用于超级 POP 环境中的系统插槽密度的 4 倍。

M40Internet 骨干路由器代表了一代全新的服务提供商系统，同时，它还具备一些为惯于工作在电信环境中的人们所熟悉的性能。例如，前面板允许技术人员监视系统状态，通过远



程的网络运行中心 (NOC) 帮助处理系统故障，它同时海提供其他一些相关的系统功能。

图 8：超级 POP 环境

结论

Internet 骨干路由器 - - M40，为电信运营商及 ISP 提供了一个全新的路由系统，它是为 ISP 实现从 OC - 3 和 OC - 12 骨干网到 OC - 48 骨干网的转换而专门设计的。Internet 网络的核心部分仍沿着两个方向被坚持不懈地发展着：软件的丰富性和带宽。M40 系统提供了所有下一代路由系统所必须具备的特性：

- 包转发引擎的包处理进程和富裕设计的交换结构，使 M40 系统可以轻松的同时支持八个满负荷工作的 OC - 48 速率接口。M40 系统能够在不影响网络控制单元的情况下，提供以前只能在交换机中才能发现的转发性能。
- 路由引擎运行着由经验丰富的、行业中的专家们设计和编写的行业级的全功能路由协议和流量工程软件。
- M40 的基本结构 - 路由功能与包转发功能完全分离 - 是以提供“性能保障”为目的的进行设计的，可为大型 ISP 网络在意外条件下提供增强的可靠性。M40 系统在提供网络可靠性及不影响网络其他部分而适应高波动环境方面的能力是独一无二的。
- 可编程的，与类似于计算机微处理器的 ASIC 允许 JUNOS Internet 软件可在将来使包转发引擎在不更换硬件的情况下支持新的功能。
- 多种高密度的接口模块，机械性能和适用性，使得 Internet 骨干路由器 M40 可在大型 ISP 网络的核心部分有杰出的表现。

- 通过提供强有力的四个基本单元 - 路由软件，包处理，交换结构和线路卡 - 的路由系统，M40 系统为 Internet 骨干网提供了一个完全的解决方案，它是您的 Internet 骨干网成功的保障！