

组织：中国互动出版网 (<http://www.china-pub.com/>)

RFC 文档中文翻译计划 (<http://www.china-pub.com/compters/emook/aboutemook.htm>)

E-mail: ouyang@china-pub.com

译者：cata_xu (cata_xu amethyst@theory.issp.ac.cn)

译文发布时间：2001-7-25

版权：本中文翻译文档版权归中国互动出版网所有。可以用于非商业用途自由转载，但必须保留本文档的翻译及版权信息。

Network Working Group

Request for Comments: 2105

Category: Informational

Y. Rekhter

B. Davie

D. Katz

E. Rosen

G. Swallow

Cisco Systems, Inc.

February 1997

Cisco 系统的标签交换体系结构纵览

(Cisco Systems' Tag Switching Architecture Overview)

本备忘录的状态

本备忘录提供了 Internet 社区的一些信息，但并没有详细讲述任何一种 Internet 标准。本备忘录的发布不受任何限制。

版权声明

Copyright (C) The Internet Society (1997). All Rights Reserved.

IESG 注意：

这个协议既不是 IETF 工作组的一个产品，也不是一个标准的追踪文档。此协议不必须从标准追踪文档中所收到的广泛的、并且深刻的社区评论中获益。

摘要

此文档提供了一种用来网络层分组转发的新型方法的纵览，此方法被称为标签交换。本文档将描述标签交换体系结构的两个主要组成部分：转发和控制组件。当现有的网络层路由协议加入绑定和发布用来控制的标签机制时，用简单的标志交换（label-swapping）技术可以完成转发。标签交换能保留 IP 的分级特性，并且也有助于改进 IP 网络的可升级性。当标签交换不依靠 ATM 时，它就可以直接应用于 ATM 交换。本文档将描述一定范围内的标签交换应用以及特定应用场景。

目录

1. 介绍.....	2
2. 标签交换部分.....	3
3. 转发部分.....	3
3.1 标签封装.....	4
4. 控制部分.....	4
4.1 基于目的地的路由.....	5
4.2 路由层技术.....	6
4.3 组播.....	7
4.4 灵活路由（清晰路由）.....	8
5. ATM 上的标签交换.....	8
6. 服务质量.....	9
7. 标签交换移植策略.....	9
8. 总结.....	10
9. 安全考虑.....	10
10. 知识产权考虑.....	10
11. 致谢.....	10
12. 作者地址.....	10

1. 介绍

持续的 Internet 发展要求 Internet 服务提供商们（ISP）能提供更多的带宽。然而，对更多带宽的推动因素不仅仅是 Internet 的发展，而且也来自于正在显现的多媒体应用的要求。对更多带宽的要求又反过来要求路由器对组播和单播通信来说都能有更高的转发性能（每秒的所发的分组数）。

Internet 的发展也要求改良 Internet 路由系统的分级特性。包含单个路由器所维护的路由信息量的能力，以及建立一个层次路由技术的能力都是支持一个高品质、可升级的路由系统所必须的。

我们看到了在改进转发性能的同时，也需要增加支持组播的路由功能，需要允许对如何路

由通信进行更灵活的控制，以及需要提供建立一个路由层技术的能力。此外，有一个能支持更完善地发展以满足新的和日见突出要求的路由系统正变得越来越重要。

标签交换是一种为这些挑战提供了一种有效解决方法的技术。标签交换结合了网络层路由提供的灵活性和丰富功能以及标志交换转发范例所提供的简单性。标签交换转发范例（标志交换）的简单性能在保持有竞争力的价格/性能的同时改良转发性能。通过将标签粘到一个宽泛的转发粒度上，同样的转发范例就可以用来支持许多种路由功能，诸如基于目的地的路由、路由层技术、组播、以及灵活路由控制。最终，在保留同样转发范例的同时，简单转发的连接，宽泛的转发粒度，以及进一步发展路由功能的能力能使一个路由系统更完善地发展以满足新的和日益显现的要求。

本文档的其余部分组织如下：第二节介绍了标签交换的主要组件：转发和控制。第三节描述了转发部分，而第四节讲述了控制部分。第五节形容了标签交换怎样和 ATM 一起使用。第六节描述了使用标签交换以帮助提供一些质量服务。第七节简要讲述了可能的应用场景。第八节对这些结果进行总结。

2. 标签交换部分

标签交换由两部分组成：转发和控制。转发部分使用分组所携带的标签信息（标签）和标签交换所维护的标签转发信息来执行分组转发。控制部分则负责在一组互联的标签交换体系中维护正确的标签转发信息。

3. 转发部分

标签交换使用的基本的转发部分范例是基于标志交换概念的。当一个标签交换收到一个带标签的分组时，此交换在它的标签信息基地（Tag Information Base TIB）中使用这个标签作为索引。TIB 中的每个输入端口都包括一个输入标签，以及一个或更多这种形式的子输入端口（输出标签、输出接口、和输出链路层信息）。如果此交换发现一个输入端口带有的输入标签与在这个分组里携带的标签一致的话，那么对于这个输入端口里每一个子输入端口（输出标签、输出接口、和输出连接层信息），交换都将用输出标签来替换分组里的标签，并且用输出链路层信息来替换分组里的链路层信息（如 MAC 地址），以及在输出接口上转发这个分组。

从上面转发部分的描述我们可以作出几个结论。首先，转发判决是基于一种精确匹配算法，此算法使用一固定长度且相当短的，用做索引的标签。相对于传统上用在网络层里的长匹配转发来说，这样做能使用一种简化的转发程序，而且反过来又能得到更高的转发性能（每秒转发的分组数更多）。这种转发程序是足够简单的以致于甚至允许以一种直接的硬件实现。

第二个结论是转发判决与标签的转发粒度无关。例如，应用在单播和组播上的转发算法是一样的，尽管一个单播输入端口仅有一个子输入端口（输出标签、输出接口、和输出连接级信息），而一个组播输入端口则可能有一个或多个子输入端口（输出标签、输出接口、和输出链路层信息）。（对于多路连接来说，输出链路层信息在这种情形下将分组包括一个组播 MAC 地址。）这就例证了为什么带标签转发的同样转发范例能被用于支持不同的路由功能（如单播，组播，等等）。

因而这个简化转发程序从本质上减弱了标签转发的控制部分。新的路由（控制）功能就能在没有发布转发范例的前提下很容易得得以应用。这也意味着在加入新路由功能时重新优化转发性能（通过修改硬件和软件）将不是必需的。

3.1 标签封装

一个分组里可以以几种方式来携带标签信息：

- 1) 作为一个小的“薄片”标签头嵌入第二层和网络层头之间；
- 2) 如果第二层头提供足够的语义的话（例如，如下所将要提到的 ATM），则作为第二层头的一部分。
- 3) 作为网络层头的一部分（例如，使用适当修改过语义的 IPv6 里的流程标志域（Flow Label field））。

因此实际上标签交换是可能在任何媒介类型上，包括点到点连接，多路连接，以及 ATM，得以实现。

我们也得出标签转发部分是和网络层无关的结论。对一个特定的网络层协议使用特殊的控制部分能使使用不同网络层协议的标签交换可用。

4. 控制部分

在一个标签和网络层路由（路由器）之间绑定的概念对标签交换来说是必需的。为了提供比较好的分级特性，同时也提供多种路由功能，标签交换需要支持宽泛的转发粒度。一种极端情况是一个标签（绑定）被连结到一组路由上去（更明确地说是连结到这个组中路由的网络层可抵达信息（Network Layer Reachability Information）上）。另一个极端就是一个标签被绑定到单个应用流上（如一个 RSVP 流）。一个标签也能被绑定到一个组播树上。

控制部分对创建标签绑定负责，然后在标签交换当中发布标签绑定信息。这个控制部分是由一些模块的集合组成的，每个模块都被设计来支持一种特定的路由功能。如果要支持新的路由功能就得添加新的模块。下面将讲述几种模块。

4.1 基于目的地的路由

这一小节我们将描述标签交换是如何支持基于目的地的路由，并追述一个使用基于目的地路由的路由器是如何作出一个转发判决的。这个转发判决是基于单个分组所携带的目的地地址以及由此路由器维护的转发信息基地（Forwarding Information Base FIB）所储有的信息之上的。一个路由器通过使用这个路由器从路由协议上（如 OSPF、BGP）收到的信息来建立它自己 FIB。

为了支持带标签交换的基于目的地路由，一个标签交换就象一个路由器一样参与到路由协议（如 OSPF、BGP）之中，并用它从这些协议收到的信息建立它的 FIB。

有三种允许的标签分配和标签信息基地（TIB）管理的方法：（a）下游标签分配，（b）依照需要的下游标签分配，以及（c）上游标签分配。在所有这些方法中，一个交换都要分配标签并将它们绑定到它们 FIB 里的地址前缀上。在下游分配中，这个在分组里携带的标签被产生并绑定到在连结下游末端处（相对于数据流的方向）转发的一个前缀上。在上游分配里，标签被分配并绑定在连接的上游末端处。“依照需要”分配意思是当标签被上游转换要求去分配时，标签将只能被下游交换分配和发布。方法（b）和（c）在 ATM 网络里是最有用的（见第 5 节）。请注意在下游分配中，一个交换要为创建能应用到输入数据分组上的标签绑定以及收到从它的邻近交换来的输出分组的标签绑定而负责。在上游分配中，一个交换要为创建输出标签的标签绑定（即应用到离开交换的数据分组上的标签）和收到从它的邻近交换来的输入标签的标签绑定而负责。

下游标签分配方案运作如下：对于这个交换 FIB 里的每一个路由，交换分配一个标签，并创建一个输入端口，这个端口在它的标签信息基地（TIB）里粘有已被设为已分配标签的输入标签，以及接着发布在（输入）标签和到邻近标签交换的路由之间进行绑定的消息。这种消息的发布既可以通过把绑定放在现有路由协议的顶部来实现，也可以通过使用一个孤立的标签发布协议 [TDP] 来实现。当标签交换为一个路由收到标签绑定信息，且这个信息是由这个路由的下一跳来组织时，此交换把这个标签（作为绑定信息的一部分被携带）放进和此路由相连的 TIB 输入端口的输出标签里。这样就创建了输出标签和路由之间的绑定。

依照需要的下游标签分配方案运作如下：对于交换的 FIB 中的每一个路由来说，交换将校验此路由的下一跳。然后交换给下一跳发布一个对此路由的一个标签绑定的请求（通过 TDP）。当下一跳收到这个请求时，下一跳分配一个标签，并创建一个在标签信息基地（TIB）里粘有已被设为已分配标签的输入标签的输入端口，以及接着返回在（输入）标签和到发送原始请求交换的路由之间进行绑定的消息。当交换收到这个绑定信息时，此交换在它的 TIB 里创建一个输入端口，并把输入端口里的输出标签设置为从下一跳收到的值。

上游标签分配方案运作如下：如果一个标签交换有一个或多个点到点接口，那么对于 FIB 里的每个路由来说，交换分配一个标签，并创建一个在标签信息基地（TIB）里粘有已被设为已分配标签的输入标签的输入端口，以及接着发布给下一跳（通过 TDP）在（输入）标签和

路由之间进行绑定的消息。这些路由是通过这些接口中的一个到达下一跳。当下一跳的标签转换收到标签绑定信息时，这个交换放置这个标签（作为此绑定信息的一部分被携带）到和粘在此路由上的 TIB 输入端口的输入标签位置上。

一旦一个 TIB 输入端口上既有输入和输出标签，则标签交换能转发分组给路由，这些路由是粘有使用标签交换转发算法（正如第 5 节里形容的一样）标签的。

当一个标签交换在输出标签和路由之间建立一个绑定时，这个交换在安置它自己的 TIB 之外，

也用此绑定信息对它的 FIB 进行升级。这样做能使交换在先前未标签的分组上加上标签。

为了理解与基于目的地路由联系的标签交换的分级特性，我们注意到一个标签交换必须维护的标签总数不能比交换 FIB 里路由的数目多。并且在某些情况下，单个标签要和一组路由相连，而不是只和一个路由相连。所以，如果标签被分配到单个流时将需要更少的组态。

通常来说，一个标签交换将试图把它带所有路由输入输出标签的 TIB 安置到它能被抵达的地方，以使所有的分组能被简单的标志交换所转发。所以标签配置是由拓扑（路由）推动的，而不是由通信来决定的，即一个 FIB 输入端口的存在促成标签配置，而不是由于数据分组的抵达而促成标签配置。

使用粘在路由上的标签而不是粘在流上的标签，也意味着没有必要对所有流执行流分类步骤去决定是否给一个流分配标签。这样反过来又简化了所有方案，并且当出现通信模式改变时可以使其更强壮和稳定。

注意到当标签交换被用于支持基于目的地的路由时，标签交换不会完全排除执行正常网络层转发的需要。首先，在一个先前没有标签的分组上加一个标签需要正常的网络层转发。这个功能可以由第一跳路由器来实现，或者由在能够参与标签交换的路径上第一个路由器来实现。另外，不管什么时候一个标签交换在单个标签里集成了一个路由集合（例如，通过使用分层路由技术的方法）且这些路由没有共享一个下一跳，交换都需要为携带标签的分组执行网络层协议。但是可以观察到的是路由集成的位置数要比必须做转发判决的位置数要少，而且更通常的是集成仅被应用于一个标签交换所维护的一子集路由，所以通常分组转发的大多数时间都是用标签交换算法。

4.2 路由层技术

IP 路由体系结构模型把一个网络作为一些路由域的集合。在一个域里，路由是通过内部路由（如 OSPF）来提供的，而跨域的路路由是由外部路由（如 BGP）来提供的。但是在携带运输通

信域（例如，由 Internet 服务提供商形成的域）里的所有路由器不是得维护内部路由提供的信息就是得维护外部路由提供的信息。这样就造成了一些问题。首先大量的这些信息并不是无关紧要的，所有这就增加了对路由器所需资源的额外要求。并且路由信息体积的增加也增加了路由集中时间。反过来这些问题也降低了系统的全部性能。

标签交换允许减弱内部和外部路由的功能，以使在一个域边缘的标签交换只需要去维护外部路由提供的路由信息，而同时域内部的所有交换则去维护域内部路由提供的信息（通常都比外部路由信息少很多）。反过来说，这样做就减少了路由在非边缘交换上的加载，也缩短了路由集中时间。

为了支持这种功能，标签交换允许一个分组不是携带一个标签而是携带组织为堆栈的一套标签。标签交换既能在堆栈顶部交换标签或者推出标签，也能交换标签并推一个或更多标签进入堆栈。

当不同域里的标签交换之间转发分组时，分组里的标签堆栈只包含一个标签。但是当在单个域里转发分组时，分组里的标签堆栈就不只有一个标签，而是两个了（第二个标签是由域入口边缘处的标签交换推入的。堆栈顶部的标签提供分组转发信息给一个合适的出口边缘处的标签交换，同时堆栈里的下一个标签在这个出口交换处提供正确的分组转发信息。然后这个堆栈被此出口交换和倒数第二个（相对于这个出口交换来说）交换推出。

这个场景中使用的控制器件和使用基于目的地路由的控制器件很相似。事实上，唯一关键性不同在于这个场景中的标签绑定信息是分布在实际邻近标签交换之间以及在单个域里边缘标签交换之间。另外也可以观察到后者（分布在边缘交换之间）一般都伴随一个非常小的 BGP 外延（通过一个分立的标签绑定 BGP 特性）。

4.3 组播

组播路由的实质是生成树的概念。组播路由过程（如 PIM）要建立这些树（带有作为叶子的接收器件）负责，与此同时，组播转发要为转发在这些树之间的组播分组而负责。

为了支持一个有标签交换的组播转发功能，每个标签交换按如下方式将一个标签和一个组播树粘到一起。当标签交换创建一个组播转发输入端口（既是为了一个共享树也是为了一个特殊资源树）以及这个端口的输出接口列表时，交换也创建了本地标签（每个输出端口一个）。此交换在它的 TIB 里创建一个输入端口并利用这个信息为每个输出接口进行配置（输出标签，输出接口，输出 MAC 头），并放置一个本地产生的标签在输出标签域里。这样就在一个组播树和标签之间创建了一个绑定。然后此交换在每个和输入端口相连的输出端口上发布标签（和此接口相连）和树之间进行绑定的消息。

当标签交换收到一个组播树和从另一个标签交换来的标签之间绑定时，并且如果另一个交换时上游邻居（相对于组播树来说），本地交换就把在绑定里携带的标签放进和这个树相连的 TIB 输入端口的输入标签器件里。

当一组标签交换通过一个多路子网互连时，组播的标签分配过程必须在这些交换里进行调整。在所有别的情况时，组播标签的分配过程应该和用基于目的地路由的标签分配过程一致。

4.4 灵活路由（清晰路由）

基于目的地路由的基本特性之一是从一个分组得来的用来转发这个分组的唯一信息是目的地地址。虽然这个性质使可高度升级的路由成为可能，但它也限制了对分组采用实际路径施加影响的能力。这也反过来限制了在多路连接中均分通信量的能力，即将负载从利用率高的连接移开，并移到利用率低的连接上。对于支持不同级别服务的 Internet 服务提供商（ISP）来说，基于目的地的路由也限制了他们分立与这些级别所使用的连接相关联的不同级别的能力。今天，一些 ISP 使用帧中继或 ATM 来消除由基于目的地路由所强加的限制。因为灵活的标签粒度，所以标签交换不使用帧中继或 ATM 就消除这些限制。

为了沿与基于目的地路由所决定的路径不同的路径提供转发，标签交换的控制器件在标签交换里安装标签绑定，这些标签交换不相应于基于目的地路由的路径。

5. ATM 上的标签交换

因为标签转发范例是基于标志交换的，并且 ATM 转发也是基于标志交换的，那么标签交换技术就能通过实现标签交换的控制器件而很容易地被应用于 ATM 交换机上。

标签交换所需的标签信息能被携带在 VCI 域里。尽管 VPI 域的大小限制了它所参与的网络的大小，但是如果需要两层标记的话，那么 VPI 域也能被使用。然而对大多数一层标记来说，VCI 就足够了。

为了获得必需的控制信息，交换应该能（最低程度上）平等参与到网络层路由协议（如 OSPF, BGP）中。但是如果此交换必需去执行路由信息集合，接着必需去支持基于目的地路由的话，则此交换应当也能对一些通信片段执行网络层转发。

在 ATM 交换上支持带标签交换的基于目的地路由功能可能需要交换不是去维护一个，而是维护几个粘在一个路由（或者是有相同下一跳的一组路由器）上的标签。必须要做的是避免从不同上游标签交换到达的分组发生交错，并且并发地发送到同一个下一跳。依照需求的下游标签分配和上游标签分配方案都能被用做标签分配以及在 ATM 交换机上的 TIB 维护过程。

所以，ATM 交换能支持标签交换，但是最低限度它需要实现网络层路由协议，以及在此交换上的标签交换控制器件，可能也需要支持一些网络层转发。

在 ATM 交换机上实现标签交换能够简化 ATM 交换和路由的集成，即一个能标签交换的 ATM 交换

可以对邻近路由显现成一个路由。这样就提供了一个可行的，更可升级性的选择给这个覆盖模型，并且也移去了对 ATM 地址、路由和发送信号方案的需要。因为在 4.1 节里形容的基于目的地转发方式是由拓扑学推动的，而不是由通信量推动的，所以这种对 ATM 交换方法的应

用不会增加很多的安装费用，也不会依靠数据流的寿命。

在 ATM 交换机上实现标签交换不会排除在同一个交换上支持一个传统 ATM 控制平面（如 PNNI）的能力。标签交换和 ATM 控制平面这两种组成器件都能在“夜间行船”（Ships In the Night）模式（具有 VPI/VCI 空间和别的参与资源导致器件之间没有相互作用）下运行。

6. 服务质量

需要两种机制提供一定的服务质量以使分组通过一个路由器或一个标签交换。首先我们需要把分组区分成不同的级别。其次我们需要保证分组的处理是能够提供给每个级别的分组合适的 QOS 特征（带宽，遗失，等等）。

在分组第一次被区分后标签交换提供了一种简单的方法给分组标上属于一个特定级别的记号。最初的区分是通过使用携带在网络层或更高层头里的信息来完成的。一个对应于分级结果的标签将被应用到这个分组上。然后标好标签的分组就能被标签交换路由器在路由器的路径里有效处理而不需要再做区分。实际分组的时间计划和队列主要是正交的，这里的关键在于标签交换能使用简单的逻辑去发现识别分组应该怎样被预定的状态。

用于 QOS 目的的标签交换地准确使用很大程度上依靠于怎样运用 QOS。如果使用 RSVP 去要求一个级别分组确定 QOS 的话，那么分配一个相应于每个 RSVP 交谈的标签给一个标签交换上所安装的状态就将是必要的。这可以通过 TDP 或 RSVP 的外延来完成。

7 标签交换移植策略

因为标签交换是执行在一对邻近的标签交换之间的，并且也因为标签绑定信息能分布在一成对的基础之上，所以能够以一种相当简单的增强模式引入标签交换。例如，一旦一对邻近的路由器被改变成标签交换的话，每一个交换都把预定分组标记到另一个交换上，这样就能使另一个交换使用标签交换。因为标签交换和路由器使用同一个路由协议，所以标签交换的引入不会影响路由器。事实上，在一个路由器看来，一个连接到这个路由器上的标签交换就起了一个路由器的作用。

随着越来越多的路由器被升级为可以使用标签交换，标签交换所提供的功能范围也在扩展。例如，一旦在一个域里的所有路由器都升级到可以支持标签交换的话，那么开始使用路由层技术的功能就将成为可能。

8. 总结

在这篇文档里我们形容了标签交换技术。标签交换没有被局限到一种特定网络层协议上，它是一种多协议解决方案。标签交换的转发器件应该是足够简单以促进高性能的转发，并可以在高性能的转发硬件如 ATM 交换机上得以实现。控制器件应该是足够灵活的以便支持广泛的路由功能，诸如基于目的地的路由、组播路由、层次路由技术、以及清晰定义路由。通过允许粘有标签的宽泛转发粒度，我们提供了可升级和功能丰富的路由。将宽泛的转发粒度和把控制器件发展为与转发器件基本无关的能力结合在一起就能产生一个解决方案以便能完美地引入新的路由功能以便符合迅速发展计算机网络环境的要求。

9. 安全考虑

本文档不讨论安全问题

10 知识产权考虑

Cisco 系统可以寻求本文档所发布之部分或全部技术的专利或别的知识产权保护。如果任何从本文档里派生的标准为授予 Cisco 系统一个或多个专利所保护的话，Cisco 打算公布这些专利并授权这些专利能被合理且无差别条件使用。

11. 致谢

感谢 Anthony Alles, Fred Baker, Paul Doolan, Dino Farinacci, Guy Fedorkow, Jeremy Lawrence, Arthur Lin, Morgan Littlewood, Keith McCloghrie, and Dan Tappan 为本职工作所做的重要贡献。

12. 作者地址

Yakov Rekhter
Cisco Systems, Inc.
170 Tasman Drive

San Jose, CA, 95134

EMail: yakov@cisco.com

Bruce Davie
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824

EMail: bsd@cisco.com

Dave Katz
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134

EMail: dkatz@cisco.com

Eric Rosen
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824

EMail: erosen@cisco.com

George Swallow
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824

EMail: swallow@cisco.com