[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 15/173 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200780007431.9

[43] 公开日 2009 年 3 月 25 日

[11] 公开号 CN 101395594A

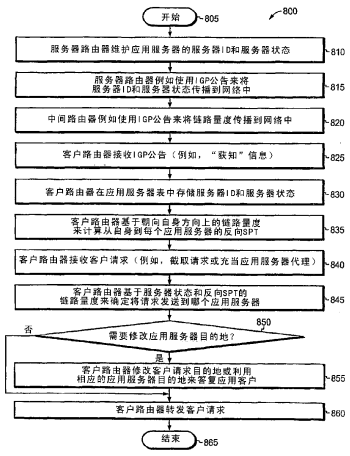
[22] 申请日 2007.2.22
[21] 申请号 200780007431.9
[30] 优先权
[32] 2006.3.1 [33] IT [31] T02006A000149
[86] 国际申请 PCT/US2007/004746 2007.2.22
[87] 国际公布 WO2007/106319 英 2007.9.20
[85] 进入国家阶段日期 2008.9.1
[71] 申请人 思科技术公司
地址 美国加利福尼亚州
[72] 发明人 斯特凡诺·B·普雷维蒂
戴维·D·沃德
[74] 专利代理机构 北京东方亿思知识产权代理有限责任公司
代理人 宋 鹤 南 霆
权利要求书 2 页 说明书 18 页 附图 8 页

[54] 发明名称

用于计算机网络中 IP 骨干上的数据流的最优路由的技术

[57] 摘要

一种技术优化计算机网络中因特网协议(IP)骨干上的应用数据流的路由。 根据该新技术，客户路由器获知多个应用服务器的服务器状态(例如，待决请求的数目等)，并且还确定应用服务器和客户路由器之间的中间链路的量度(中间链路量度)，例如，尤其是从应用服务器到客户路由器方向上的链路量度。 收到来自应用客户的应用请求(“客户请求”)之后，客户路由器基于服务器状态和中间链路量度来确定将该客户请求发送到多个应用服务器中的哪一个，并相应地发送该客户请求。



1. 一种方法，用于优化计算机网络中因特网协议（IP）骨干上的应用数据流的路由，所述方法包括：

获知多个应用服务器的服务器状态；

确定客户路由器和所述多个应用服务器之间的中间链路的量度（中间链路量度）；

在所述客户路由器处接收来自应用客户的应用请求；以及

基于所述服务器状态和所述中间链路量度来确定将所述客户请求发送到所述多个应用服务器中的哪一个。

2. 根据权利要求 1 所述的方法，还包括：

在所述客户路由器处执行所述获知、计算、接收和确定的步骤。

3. 一种装置，用于优化计算机网络中因特网协议（IP）骨干上的应用数据流的路由，所述装置包括：

用于获知多个应用服务器的服务器状态的装置；

用于计算从每个应用服务器到客户路由器的反向最短路径树（SPT）的装置；

用于在所述客户路由器处接收来自应用客户的应用请求的装置；以及

用于基于所述服务器状态和所述反向 SPT 来确定将所述客户请求发送到所述多个应用服务器中的哪一个的装置。

4. 一种计算机可读介质，包含用于优化计算机网络中因特网协议（IP）骨干上的应用数据流的路由的可执行程序指令，所述可执行程序指令包含程序指令用于：

获知多个应用服务器的服务器状态；

计算从每个应用服务器到客户路由器的反向最短路径树（SPT）；

在所述客户路由器处接收来自应用客户的应用请求；以及

基于所述服务器状态和所述反向 SPT 来确定将所述客户请求发送到所述多个应用服务器中的哪一个。

5. 一种节点，与用于计算机网络中因特网协议（IP）骨干上的应用数

据流的路由优化一起使用，所述节点包括：

一个或多个网络接口；

处理器，耦合到所述一个或多个网络接口并适合于执行软件处理；以及

存储器，适合于存储可由所述处理器执行的应用接口处理，所述应用接口处理被配置成：i) 获知多个应用服务器的服务器状态；ii) 计算从每个应用服务器到所述节点的反向最短路径树（SPT）；iii) 接收来自应用客户的应用请求；以及 iv) 基于所述服务器状态和所述反向 SPT 来确定将所述客户请求发送到所述多个应用服务器中的哪一个。

用于计算机网络中IP骨干上的数据流的最优路由的技术

技术领域

本发明涉及计算机网络，更具体地，涉及优化计算机网络中因特网协议（IP）骨干上的应用数据流的路由。

背景技术

计算机网络是多个节点的地理分布集合，所述节点通过用于在端节点之间传输数据的通信链路和网段来互连。有很多类型的网络可用，所述类型范围从局域网（LAN）到广域网（WAN）不等。LAN 通常通过位于同一普通物理位置（如大楼或校园）的专用私有通信链路来连接多个节点。另一方面，WAN 通常通过长距离通信链路来连接多个地理上分散的节点，所述链路如公共载体电话线、光路、同步光网络（SONET）或同步数字体系（SDH）链路。因特网是连接遍布世界的分散网络的 WAN 的示例，提供各种网络上的节点之间的全球通信。节点通常通过根据预定协议交换离散帧或数据分组来通过网络通信，所述协议如传输控制协议/因特网协议（TCP/IP）。在这样的情况下，协议由定义节点如何彼此交互的一组规则组成。计算机网络可以通过诸如路由器之类的中间网络节点来进一步互连，以扩大每个网络的有效“尺寸”。

由于能够证明互连的计算机网络的管理繁重，因此较小的一群计算机网络可以作为路由域或自治系统来维护。自治系统（AS）内的网络通常通过被配置成执行域内路由协议的常规“域内”路由器来耦合在一起，并且一般隶属于公共权限（common authority）。为了改善路由可伸缩性，服务供应商（例如，ISP）可将 AS 划分成多个“区域”或“级别”。但是，可能希望增加能够交换数据的节点的数量；这种情况下，运行域间路由协议的域间路由器用于互连各种 AS 的节点。另外，可能希望互连在不同管理域下操作的各种 AS。此处所使用的区域或级别，或者更具体地说，AS，

一般被称为“域”。

域内路由协议或者内部网关协议（IGP）的示例为开放最短路径优先（OSPF）路由协议和中间系统到中间系统（IS-IS）路由协议。OSPF 和 IS-IS 协议基于链路状态技术，因此通常被称为链路状态路由协议。链路状态协议定义路由信息和网络拓扑信息在域中被交换和处理的方式。该信息一般针对域内路由器的本地状态（例如，路由器的可用接口和可达邻居或邻接）。OSPF 协议被描述于日期为 1998 年 4 月、题为“*OSPF version 2*”的 RFC 2328，在 IP 情况下使用的 IS-IS 协议被描述于日期为 1990 年 12 月、题为“*Use of OSI IS-IS for routing in TCP/IP and Dual Environments*”的 RFC 1195 中，二者都通过引用方式结合于此。

中间网络节点通常将其路由信息存储在由路由信息库（RIB）来维护和管理的路由表中。路由表是一种可搜索数据结构，其中网络地址被映射到其关联路由信息。然而，本领域技术人员将理解路由表不需要被组织为表，而是也可以是另一类型的可搜索数据结构。尽管中间网络节点的路由表可以配备预定的一组路由信息，但是该节点也可以在其发送和接收数据分组时动态获取（“学习”）网络路由信息。当在中间网络节点处接收到分组时，分组的目的地地址可以用于标识包含与收到的分组相关联的路由信息的路由表条目。分组的路由信息指示分组的下一跳地址等等。

为了确保其路由表包含最新路由信息，中间网络节点可以与其他中间节点协作以散布代表当前网络拓扑的路由信息。例如，假定中间网络节点检测到其相邻节点（即，邻接网络节点）之一例如由于链路故障或相邻节点“离线”等等变得不可用。在这种情形下，中间网络节点可以更新存储在其路由表中的路由信息，以确保数据分组不被路由到不可用的网络节点。此外，中间节点还可以将该网络拓扑的变化传递给其他中间网络节点，以便它们也可以更新其本地路由表并绕过不可用节点。以这种方式，每个中间网络节点“获知”拓扑的变化。

一般来说，路由信息是根据预定的网络通信协议，例如链路状态协议（例如 IS-IS 或 OSPF）在中间网络节点之间散布的。通常的链路状态协议使用链路状态公告或链路状态分组（或“IGP 公告”）来在互连的中间网

络节点（IGP 节点）之间交换路由信息。这里所使用的 IGP 公告一般描述 IGP 路由协议用来在互连的 IGP 节点（即，路由器和交换机）之间传递路由信息的任何消息。在操作上，第一 IGP 节点可以生成 IGP 公告，并通过其每个耦合到其他 IGP 节点的网络接口“洪泛（flood）”（即，发送）分组。其后，第二 IGP 节点可以接收被洪泛的 IGP 公告，并基于包含在收到的 IGP 公告中的路由信息更新其路由表。接着，第二 IGP 节点可以通过其每个网络接口来洪泛收到的 IGP 公告，接收到 IGP 公告的接口除外。该洪泛处理可以重复，直到每个互连的 IGP 节点都已接收到 IGP 公告并更新其本地路由表为止。

在实际中，每个 IGP 节点一般生成并散布这样的 IGP 公告，该公告的路由信息包括中间节点的相邻网络节点以及与每个邻居相关联的一个或更多个“代价”值的列表。这里所使用的与相邻节点相关联的代价值是用于确定与该节点进行通信的相对难易程度的任意度量（“链路度量”）。例如，成本值可以按照到达相邻节点所需的跳数、分组到达相邻节点的平均时间、耦合到相邻节点的通信链路上的可用带宽或网络流量等等方面来量度。值得注意地，如本领域技术人员将理解的那样，IGP 公告的流量工程（TE）扩展可以用来传送各种链路度量，例如链路利用率等。用于 IGP 的 TE 扩展的示例在日期为 2004 年 6 月、题为“*Intermediate-System-to-Intermediate-System (IS-IS) Extensions for Traffic Engineering (TE)*”的 RFC 3784 和日期为 2003 年 9 月、题为“*Traffic Engineering (TE) Extensions to OSPF Version 2*”的 RFC 3630 中找到，二者都通过引用方式结合于此。

如上所述，通常洪泛 IGP 公告直至每个中间网络 IGP 节点都已从每个其他互连的中间节点收到 IGP 公告。然后，每个 IGP 节点可以（例如在链路状态协议中）通过聚集收到的邻近节点和代价值的列表来构造同一“视图”的网络拓扑。为此，每个 IGP 节点可将收到的该路由信息输入到“最短路径优先”（shortest path first，SPF）计算中，该计算确定将中间节点与每个其他网络节点耦合的最低代价的网络路径，即，因此而计算本领域技术人员所理解的“最短路径树”（SPT）。例如，Dijkstra 算法是用于执行这种 SPF 计算的常规技术，对该算法更详细的描述见 1999 年 9 月出版

的、Radia Perlman 所著课本“*Interconnections Second Edition*”的 12.2.4 小节，其通过引用方式被结合于此，如同在此处被完整提出一样。每个 IGP 节点基于其 SPF 计算结果来更新存储在其本地路由表中的路由信息。具体而言，RIB 更新路由表来将目的节点与 SPF 计算所确定的和到达那些节点的最低代价路径相关的下一跳接口联系起来。

在网络中转移的数据分组可包括固定尺寸的数据包和/或可变尺寸的数据帧。每个数据分组通常包括前置（“封装”）了根据网络通信协议格式化的至少一个网络首部的“有效载荷”数据。网络首部包括使得客户节点和中间节点能够通过计算机网络来有效地路由该分组的信息。经常，分组的网络首部至少包括数据链路（第 2 层）首部和互联网（第 3 层）首部，如开发系统互连（OSI）参考模型所定义的那样。OSI 参考模型一般被详细描述于日期为 1999 年 9 月出版的、Radia Perlman 所著题为“*Interconnections Second Edition*”的参考书的第 1.1 节，其通过引用方式被结合于此，如同在此处被完整提出一样。

操作时，客户节点将数据分组发送到中间网络节点的网络接口。之后，中间网络节点接收分组并将分组转发到其下一目的地。例如，中间网络节点可以执行第 2 层交换功能，仅基于分组的数据链路首部的内容来将分组从一个网络接口重新定向到另一个网络接口。或者，中间网络节点可以执行第 3 层路由功能或转发决定，基于分组的互联网首部来选择最合适的网络接口以转发分组。

数据分组用于通过网络和子网来传输多种形式的信息。例如，视频信息可以根据本领域技术人员公知的视频点播（VoD）标准来发送。VoD 指代用于通过数据网络从源节点（例如，VoD 应用服务器）向目的节点（例如，VoD 应用客户）发送视频信息的技术群。源节点和目的节点采用语音代理，将视频信息从其传统形式转换成适合分组发送的形式。换言之，源节点的视频将视频信息编码、压缩并封装成多个数据分组，目的节点的语音代理执行相反的功能以对 VoD 分组进行解封装、解压缩和解码。例如，VoD 内容服务器可以将视频数据流提供到用户的一个或更多个“机顶盒”。此外，音乐信息可以根据本领域技术人员公知的标准以类似于 VoD

的方式来发送。音乐代理的示例可包括音乐应用（例如，苹果®公司的iTunes®音乐程序）的个人计算机（PC），提供音乐服务的网络设备（例如，因特网自动点唱机）等。值得注意地，VoD 和音乐服务的使用是网络内的节点可以操作的（例如，应用层的）应用的示例。本领域技术人员将理解，其他应用也可以在网络节点处操作。

源节点（发送者）可以被配置成在数据网络中将单向数据分组流或“数据流”转移到目的节点（接收者）。该数据流例如可包括数据或视频/音乐信息。数据流是单向的因为数据从发送者到接收器单向传播。发送和接收从发送者到接收者的数据分组的中间网络节点的逻辑行列定义数据流的数据路径。数据流的数据路径中比流中的第二节点更靠近接收者的第一节点被称为第二节点的“下游”。同样地，数据流的路径中比流中的第二节点更靠近发送者的第一节点被称为第二节点的“上游”。

一般地，在现今的网络配置中，为了保证数据的冗余性以及提供来自应用客户的请求的负载平衡/共享，多个应用内容服务器（例如，VoD、音乐等）可以分散遍布于网络的不同部分。不过值得注意地，当在网络内使用标准的 IP 路由（例如，“IP 骨干”）时，路由层和应用层之间不存在通信来提供有效的负载平衡。这常常在应用层用来对多个应用服务器尝试客户请求的负载平衡的算法中造成不一致。

例如，VoD 客户可能向 VoD 服务器请求已被客户的 VoD 应用选择的视频数据流。然而，客户的 VoD 应用不知道来自所选服务器的数据流所使用的网络资源（例如，视频流从服务器到客户将遍历的链路/节点）。因此，没有网络资源及其当前状态的知识，应用层所做的任何对客户请求的负载平衡的尝试都是无效的，特别是当使用 IP 骨干时。另外，由于路由层不知道应用请求或应用服务器的状态，路由层所做的任何对客户请求的负载平衡的尝试也都是无效的。因此，仍需要有效的技术来基于应用层信息和路由层信息优化 IP 骨干上的应用数据流的路由。

发明内容

本发明针对用于优化计算机网络中因特网协议（IP）骨干上的应用数

据流的路由的技术。根据该新技术，客户路由器获知多个应用服务器的服务器状态（例如，待决请求的数目等），并且还确定应用服务器和客户路由器之间的中间链路的量度（中间链路量度），例如，尤其是从应用服务器到客户路由器方向上的链路量度。收到来自应用客户的应用请求（“客户请求”）之后，客户路由器基于服务器状态和中间链路量度来确定将该客户请求发送到多个应用服务器中的哪一个，并相应地发送该客户请求。

在此处描述的示意性实施例中，客户路由器通过例如开始于连接到应用服务器的一个或多个服务器路由器的网络（例如，域）内传播（“公告”）的内部网关协议（IGP）消息的方式获知应用服务器的服务器状态和中间链路量度。IGP 消息可以示意性地体现为内部系统到内部系统（IS-IS）链路状态分组（“IGP 公告 ”）。值得注意地，IGP 公告包括用于传送服务器状态和链路量度信息的可变长度字段或类型/长度/值（TLV）编码格式。

根据本发明的一个方面，服务器路由器知道每个连接的应用服务器的标识（ID）（服务器 ID）以及每个服务器的当前状态（例如，活动连接的数目，待决请求的数目，当前 CPU/存储器负载，当前使用带宽等）。服务器路由器例如在 IGP 公告的新应用服务器（APPL_SERVER）TLV 中将服务器 ID 和服务器状态公告给域中的路由器。每个 APPL_SERVER TLV 可以对应于单个服务器 ID，并且每个 TLV 可以具有一个或多个子 TLV 用于传送相应的应用服务器的服务器状态信息。

根据本发明的另一方面，客户路由器接收 IGP 公告，并示意性地计算从其自身到每个应用服务器的反向 SPT，即，使用从应用服务器发送到请求的应用客户的数据流的方向上的链路量度。值得注意地，反向 SPT 是基于例如使用 IGP 流量工程（TE）扩展从服务器路由器发送的 IGP 公告中获得的链路量度（例如，链路利用率、链路延时、差错率等）来计算的。客户路由器基于例如包括相关链路量度的改变在内的网络中的改变/更新来维护当前的反向 SPT。

根据本发明的又一个方面，客户路由器例如通过确定存在请求分组形式/类型或服务器 ID 是该请求的目的地来接收（“截取”）客户请求。或

者，客户路由器可以充当应用客户的应用服务器代理，在这种情况下，客户将请求提交给客户路由器，如本领域技术人员理解的那样。收到客户请求后，客户路由器基于例如服务器的当前负载之类的服务器状态并基于诸如来自反向 SPT 之类的链路量度来确定将该请求发送到哪个应用服务器。客户路由器随后可以修改客户请求以确保它被发送到相应的应用服务器（例如，改变目的地址），或者可以利用相应的服务器 ID 来答复应用客户（即，这样应用客户可以将请求重新发送到相应的应用服务器）。

有利地，该新技术优化了计算机网络中 IP 骨干上的应用数据流的路由。通过基于服务器状态和（例如，来自反向 SPT 的）中间链路量度确定将客户请求发送到何处，该新技术允许客户请求被发送到最优的应用服务器，因此对该请求进行负载平衡。特别地，请求可以被负载平衡，而无需应用层的路由和资源利用率知识，即网络层拥有足够信息来优化应用数据流。另外，新技术的动态性质减轻了麻烦的手动配置的需要。

附图说明

本发明的上述及其他特征通过结合附图来参考以下说明将得到最好的理解，图中相似的标号指代相同或功能相似的元素，其中：

图 1 是可以有利地和本发明一起使用的示例性计算机网络的示意性框图；

图 2 是可以有利地和本发明一起使用的示例性路由器的示意性框图；

图 3 是可以被路由器洪泛的示例性 IGP 公告的示意性框图；

图 4 是可以有利地和本发明一起使用的客户应用请求的部分的示意性框图；

图 5 是阐释可以有利地和本发明一起使用的可变长度字段的格式的示意性框图；

图 6 是可以有利地和本发明一起使用的示例性应用服务器表的示意性框图；

图 7 是示出根据本发明来计算的反向最短路径树（SPT）的图 1 中的计算机网络的示意性框图；并且

图 8 是阐释根据本发明用于优化 IP 骨干上的应用数据流的路由的过程的流程图。

具体实施方式

图 1 是可以有利地和本发明一起使用的示例性计算机网络 100 的示意性框图。网络 100 包括多个互连的网络节点，如应用客户和两个或更多个应用服务器（例如，应用服务器 1 和 2）。示意性地，应用客户可以互连到客户路由器，并且应用服务器可以互连到相应的服务器路由器（例如，分别为服务器路由器 1 和 2）。客户路由器和服务器路由器可以通过一个或多个中间节点（例如，中间路由器 A 和 B）的中间网络来互连，例如，通过广域网（WAN）链路（或局域网，“LAN”链路，点到点链路，无线 LAN 等），以形成网络 100。互连的网络节点可以根据预定的网络通信协议组来交换数据分组，所述协议例如是传输控制协议/因特网协议（TCP/IP）。本领域技术人员将理解，任何数目的节点、链路等可以在计算机网络 100 中使用，并通过各种方式来连接，并且此处示出的视图是为了简单起见。例如，客户路由器和服务器路由器可以分别被互连到多于一个应用客户或服务器。

图 2 是节点 200 的示意性框图，所述节点示意性地是可以有利地和本发明一起使用的路由器，例如作为客户或服务器路由器。该节点包括通过系统总线 250 互连的多个网络接口 210、处理器 220 和存储器 240。网络接口 210 包含机械的、电的和信令电路，用于通过耦合到网络 100 的物理链路传递数据。网络接口可以被配置成使用各种不同通信协议来发送和/或接收数据，所述协议包括 TCP/IP，UDP，ATM，同步光网络（SONET），无线协议，帧中继，以太网，光分布数据接口（FDDI）等。

存储器 240 包括可由处理器 220 和网络接口 210 寻址以存储和本发明相关的软件程序和数据结构的多个存储位置。处理器 220 可包括适合于运行软件程序并操控数据结构的必要元素或逻辑，所述数据结构如路由表 246、流量工程（TE）数据库 244、反向最短路径树（SPT）表 249 和应用服务器表 600。其多个部分通常驻留在存储器 240 中并由处理器运行的路

由器操作系统 242（例如思科系统公司的互联网操作系统，或 IOS™）通过调用支持路由器上运行的软件处理和/或服务的网络操作等方式来对路由器进行功能操作。这些软件处理和/或服务可包括路由服务 247、路由信息库（RIB）245、TE 服务 243 和应用接口服务 248。本领域技术人员将理解，其他处理器和存储装置，包括各种计算机可读介质，可用于存储和运行属于此处描述的发明技术的程序指令。

路由服务 247 包含由处理器运行以执行由一个或多个路由协议（如 IGP（例如，OSPF 和 IS-IS）、BGP 等）提供的功能的计算机可读指令。这些功能可以被配置成管理包含例如用于做出转发决定的数据的转发信息数据库（未示出）。路由服务 247 还可执行与虚拟路由协议有关的功能，如维护本领域技术人员所理解的 VRF 实例（未示出）。

网络拓扑的改变可以使用诸如通常的 IS-IS 和 OSPF 协议之类的链路状态协议来在路由器 200 中间传递。例如，假定 AS 内的通信链路故障或与网络节点相关的代价值改变。一旦网络状态的改变被路由器之一检测到，则该路由器就洪泛 IGP 公告，将改变传递到 AS 中的其他路由器。通过这种方式，每个路由器最终“收敛”到同一视图的网络拓扑。

图 3 阐释可以被路由器 200 洪泛的示例性 IGP 公告 300（例如，IS-IS 链路状态分组）。该分组包括域内路由协议鉴别字段 302 和长度指示符字段 304，前者存储标识消息的具体协议（例如 IS-IS）的值，后者存储指示用于公告的标准首部的长度的值。另外，版本/协议 ID 扩展（ext）字段 306 可用来进一步存储定义协议的特定版本的值。反向字段 308 和“R”字段留作将来与协议一起使用，ECO 和用户 ECO 字段 314 和 316 也是，它们都被接收路由器忽略直到被指示在将来的协议版本中解码。

类型字段 310（及相应的版本字段 312）存储指示被发送的 IGP 公告 300 的类型（和版本）的值，其可以定义公告内其他面向类型的字段 322 的存在。例如，公告的类型可以是本领域技术人员所理解的“Hello”（你好）分组或“LSP”分组。PDU 长度字段 318 存储指示包含首部、面向类型的字段和数据字段在内的整个 PDU（协议数据单元，或 IGP 公告 300）的长度的值。源 ID 字段 320 存储标识生成并且最初广播 IGP 公告 300 的

路由器的值。

其他面向类型的字段 322 可包括协议所定义的任何数目的字段，如本领域技术人员所理解的校验和字段、最大区域地址字段等。例如，序列号字段（未示出）可以存储指示 IGP 公告的相对版本的序列号。通常，该字段中存储的序列号对于每个 IGP 公告的新版本例如增加一。因此若 IGP 公告 300 的序列号小于在先前收到的 IGP 公告版本中存储的（即同一公告节点生成的）序列号，则认为该 IGP 公告“过期”（无效）。相应地，路由器 200 可被配置成只存储和转发最新版本的 IGP 公告，例如具有最大序列号的版本。还可以使用剩余寿命字段（未示出）来存储可用于确定 IGP 公告 300 是否有效的值。剩余寿命值通常被初始化为非零整数值，常以秒为单位。剩余寿命值例如可以每秒减少一，直到剩余寿命值达到零，从而指示该 IGP 公告已变得无效。即，存储或洪泛公告 300 的每个路由器 200 不断使分组变老直到剩余寿命值等于零。本领域技术人员将意识到，作为替代可以使用其他老化机制，如从例如等于零的初始值开始增加 IGP 公告剩余寿命值直到剩余寿命值达到已知上限。

数据段 330 包括一个或多个可变长度字段 500，每个字段具有此处将进一步描述的特定类型（或代码）、长度和值（TLV）。例如，为了公告网络拓扑，可以使用一对或多对相邻节点字段（未示出）和代价字段（未示出）。相邻节点字段可以存储诸如地址之类的值，该值指示从源 ID 字段 320 中存储的中间节点可以直接访问的网络节点。代价字段可以存储已经被例如公告节点关联到在相邻节点字段中标识的网络节点的值。注意在其他实施例中，单一相邻节点可以与多个代价值相关联。其他路由信息也可以被包含在 IGP 公告 300 的可变长度字段 500 中，如校验和值、填充字段、属性字段等，以及诸如新 APPL_SERVER 字段之类的应用服务器信息字段（下面进一步描述）。一般地，收到的 IGP 公告被存储在路由器 200 的链路状态数据库（LSDB）（未示出）中。

再次参考图 2，TE 服务 243 包含用于根据本发明来操作 TE 功能的计算机可执行指令。流量工程的示例描述于上面结合的 RFC 3784 和 RFC 3630。TE 数据库（TED）244 可用于根据本发明来（例如，在 TE 扩展的

可变长度字段 500 中）存储由诸如 IGP 之类的路由协议提供的 TE 信息，并且示意性地由 TE 服务 243 维护和管理。

应用接口服务 249 包含由处理器 220 运行以根据本发明来执行与一个或多个应用有关的功能的计算机可执行指令，所述功能例如是本领域技术人员所理解的视频点播（VoD）、音乐服务等。值得注意地，这些功能可以被配置成例如根据此处描述的本发明来与路由服务 247 和/或 TE 服务 243 协作。

操作时，应用客户一般通过“客户请求”来向应用服务器请求应用内容（例如，视频、音乐等）。图 4 是可以有利地和本发明一起使用的客户请求 400 的部分的示意性框图。请求 400 包含常用首部 410 以及应用请求字段和/或信息对象 420 等。常用首部 400 还可包含（例如发送请求的应用客户的）源地址 412 和请求被发往的目的地址 414（例如，应用服务器）。通常，应用客户上的应用服务例如可以通过本领域技术人员所理解的手动配置或动态学习而配置有适当的应用服务器的地点。因此，被配置的该应用服务器的地址可以被包含在客户请求 400 的目的地址 414 内。应用请求字段和/或信息对象 420 可以包含通常的面向应用的请求信息，如本领域技术人员所理解的请求类型、代码、特定字段、数据、信息等。

本发明针对用于优化计算机网络中 IP 骨干上的应用数据流的路由的技术。根据该新技术，客户路由器获知多个应用服务器的服务器状态（例如，待决请求的数目等），并且还确定应用服务器和客户路由器之间的中间链路的量度（中间链路量度），例如，尤其是从应用服务器到客户路由器方向上的链路量度。收到客户请求后，客户路由器基于服务器状态和中间链路量度来确定将该客户请求发送到多个应用服务器中的哪一个，并相应地发送该客户请求。

在此处描述的示意性实施例中，客户路由器通过在网络（例如，域）内传播（“公告”）的 IGP 消息的方式获知应用服务器的服务器状态和中间链路量度，所述 IGP 消息例如开始于连接到应用服务器的一个或多个服务器路由器，如下所述。IGP 消息可以示意性地体现为 IS-IS 链路状态分组（IGP 公告 300）。值得注意地，IGP 公告包括用于传送服务器状态和链路

量度信息的可变长度字段或 TLV 编码格式。

TLV 编码格式用于识别正在被传递（传送）的信息的类型（T）、待传送的信息的长度（L）以及被传送的实际信息的值（V）。长度字段中包含的长度（L）参数通常是面向实现方式的，并且可以表示从对象的类型字段的起点到终点的长度。然而，该长度一般表示值（V）字段而不是类型（T）或长度（L）字段的长度。

图 5 是阐释可以有利地和本发明一起使用的可变长度字段 500 的格式的示意性框图。可变长度字段 500 被示意性地体现为 IGP 公告 300 中包含的 TLV，并且被扩展以携带有关应用服务器的信息。为此，“应用服务器 TLV”500（“APPL_SERVER”）被组织以包括类型字段 505，该字段包含新 APPL_SERVER TLV 的预定类型的值。长度字段 510 是可变长度值。TLV 编码格式还可包含 TLV“有效载荷”（例如，值字段 515）内携带的一个或多个无序的子 TLV 550，每个子 TLV 具有类型字段 555、长度字段 560 和值字段 565。字段 TLV 500 和子 TLV 550（一个或多个）以各种方式被使用，包括此处根据本发明所描述的。值得注意地，还可使用本领域技术人员理解的通常的 TE 扩展来扩展可变长度字段 500 以携带有关网络链路的链路量度的信息（例如，链路利用率、链路延时、差错率等）。

根据本发明的一个方面，服务器路由器知道每个连接的应用服务器的标识（ID）（服务器 ID）（例如，IP 地址）以及每个服务器的当前状态。服务器的状态示例例如可以包括：活动应用连接的数目，待决应用请求的数目，当前 CPU/存储器负载，当前使用带宽等。值得注意地，服务器信息或属性还可包括可用应用数据流的指示，例如本领域技术人员所理解的服务器处可用的 VoD/音乐内容。服务器路由器可以基于与应用服务器的本地通信（例如，本领域技术人员所理解的路由器和服务器直接的特定消息交换）来知晓应用服务器信息，或者作为替代，服务器路由器可基于对进入的请求和外发的数据流的监视来维护服务器状态。示意性地，本地通信可以通过与特定本地通信协议内的可变长度字段 500 相似的方式来体现。

服务器路由器例如在 IGP 公告 300 的新 APPL_SERVER TLV 500 中将

服务器 ID 和服务器状态公告给域中的路由器。APPL_SERVER TLV 500 例如根据 IS-IS 网络范围的洪泛（或者，例如所理解的 OSPF 类型 10（区域范围）或类型 11（AS 范围）不透明 LSA）被洪泛到网络 100（即，域内）的每个路由器。示意性地，每个 APPL_SERVER TLV 500 可以对应于单个服务器 ID。（例如用于应用服务器 1 的）服务器 ID 可被包含在 TLV 500 的值字段 515 内的特定服务器 ID 字段 520 中，或者作为替代可被包含在单独的子 TLV 550 中。另外，每个 TLV 500 可以具有一个或多个子 TLV 550 用于传送相应的应用服务器的服务器状态信息。例如，特定服务器的多个活动连接可被包含在与用于该服务器的 TLV 500 相对应的第一子 TLV 中，第二子 TLV 550 可包含多个待决请求等。通过这种方式，服务器路由器将应用服务器信息（ID 和状态）的知识注入到被转发给网络的路由器（例如，客户路由器）的 IGP 公告中。

网络的每个路由器（例如，中间路由器和客户/服务器路由器）接收具有 APPL_SERVER TLV 500 的 IGP 公告 300，并相应地维护应用服务器信息表。图 6 是可以有利地和本发明一起使用的示例性应用服务器表 600 的示意性框图。应用服务器表 600 示意性地是路由器的存储器 240 中存储的数据结构并包括一个或多个条目 650，每个条目包含多个用于存储应用服务器 ID 605 的字段，以及一个或多个服务器状态信息字段，如活动连接 610、待决请求 615、CPU 负载 620、存储器负载 625、已用带宽 630 和其他状态 635。应用服务器表 600 示意性地由应用接口服务 248 维护并管理。收到 APPL_SERVER TLV 500 后，路由器在字段 605 中存储相应的应用服务器 ID（例如，服务器 1-N），还在各个服务器信息字段 610-635 中存储任何附加数据（例如，数据 1-N）。值得注意地，路由器通过每次在 IGP 公告 300 中收到新的/更新的 APPL_SERVER TLV 500 时更新该表来维护数据的当前值。另外，可对每个特定应用（例如，对于 VoD，音乐等）维护单独的表。

根据本发明的另一方面，客户路由器接收 IGP 公告，并示意性地计算从其自身到每个应用服务器的反向 SPT，即，使用从应用服务器发送到请求的应用客户的数据流的方向上的链路量度。值得注意地，反向 SPT 是基

于例如使用 IGP TE 扩展或其他传送链路量度的方法而从服务器路由器或网络中的任何中间路由器发送的 IGP 公告中获得的链路量度（例如，链路利用率、链路延时、差错率等）来计算的。另外，除了此处提到的链路状态，本领域技术人员将理解，根据本发明，还可以使用与链路相关的任何静态或动态的量度/属性（例如，可以反应链路的实时状态的量度/属性）。

反向 SPT 为客户路由器提供从每个应用服务器朝向客户路由器本身的考虑了朝向客户路由器的链路量度的一组最短路径。需要注意，因为由客户路由器实施的负载平衡算法应考虑从服务器到客户而不是从客户到服务器的路径上的链路量度，所以根据本发明，最适宜使用反向 SPT。图 7 是示出根据本发明来计算的反向 SPT 的图 1 中的计算机网络的示意性框图。示意性地，客户路由器以自身为根来计算反向 SPT，并确定从可能由被请求的应用数据流使用的应用服务器，例如应用服务器 1（或服务器路由器，例如，服务器路由器 1）开始的最短路径。反向 SPT（或“rSPT”）在本领域内公知，可通过与通常的 SPT 相似的方式来计算。然而，在反向 SPT 中，计算节点（例如，客户路由器）使用从网络中的其他路由器朝向其自身的方向上的链路量度。换言之，反向 SPT 定义从网络中的任何其他路由器（例如，应用服务器）一直到计算路由器（例如，客户路由器）的最短路径集合。值得注意地，未被选作到达客户节点的最短（最佳）路径的到其他中间路由器（未示出）的各种分支被示出用于说明。客户路由器基于例如包括相关链路量度的改变在内的网络中的改变/更新来将反向 SPT 保持为当前的。反向 SPT 被保持为当前的以便客户路由器能够根据预定准则来选择客户请求将被发往的最佳（例如，最优）的应用服务器，所述准则例如是被请求的数据流的已配置的链路利用率等，如下所述。

根据本发明的又一个方面，客户路由器例如通过确定存在请求分组形式/类型或服务器 ID 是该请求的目的地来接收（“截取”）客户请求。例如，客户路由器可以检查从应用客户接收的分组的内容，并在路由器/网络层处理该内容。具体而言，客户路由器搜索以寻找对应于客户请求 400 的分组格式、类型或特定内容，如本领域技术人员所理解的那样。例如，通

过对照（例如，应用服务器表 600 的字段 605 中的）应用服务器 ID 的列表参考分组的目的地址，客户路由器可以确定该分组是客户请求 400。应用层分组截取是本领域技术人员公知和理解的，并且根据本发明，可以使用任一这类截取技术来确定分组是客户请求 400。或者，客户路由器可以充当应用客户的应用服务器代理。换言之，客户路由器可以向应用客户公告其自身为这样的应用服务器，通常去往实际的应用服务器（例如，应用服务器 1 和 2）的请求可被发送到该应用服务器。在这种情况下，应用客户将请求提交给客户路由器，即假设客户路由器为应用服务器，如本领域技术人员理解的那样。

收到客户请求后，客户路由器基于例如服务器的当前负载（连接/请求的数目，CPU/存储器负载等）之类的服务器状态并基于诸如来自反向 SPT 之类的链路量度（例如，链路利用率）来确定将该请求发送到哪个应用服务器。可用服务器的服务器状态可以通过查阅应用服务器表 600 来确定。通过基于服务器状态将负载平衡算法应用到应用服务器以及另外在算法中包含例如来自反向 SPT 的中间链路量度，本发明允许客户请求的更理想的路由以及因此允许对得到的应用数据流的更理想的路由。

本领域技术人员所理解的各种负载平衡技术适合于与本发明一起使用以将服务器状态和链路量度考虑在内。例如，基于请求的预定义准则，例如，基于被请求的数据流的链路利用率的预计增长，客户路由器可以确定，与其他应用服务器（例如，应用服务器 1）相比，从某些应用服务器（例如，应用服务器 2）开始的反向 SPT 更拥挤，并可能无法容纳另外的流量。然而，客户路由器还必须考虑其他应用服务器（例如，服务器 1）在其他服务器状态（例如，CPU 和/或存储器负载）方面是否过载。其他比较和负载平衡技术将被本领域技术人员理解，这里提到的那些技术只是代表性示例。

一旦客户路由器确定（选择）了最佳应用服务器，客户路由器便尝试确保客户请求被转发到所选应用服务器。例如，客户路由器可以确定客户请求 400 是否已经（例如，就目的地址 414 而言）是去往所选应用服务器（例如，应用服务器 1）的，这种情况下，请求 400 被客户路由器转发。

另一方面，若客户请求 400 是去往不同应用服务器（例如，应用服务器 2）的，则客户路由器尝试修正该请求的目的地。

为了修正客户请求 400 的目的地，客户路由器可以修改客户请求以确保它被发送到相应的应用服务器（例如，改变目的地址 414）。例如，客户路由器可以将客户请求封装在到所选应用服务器的新 IP 首部 410 中，并代表客户来发送该请求（例如，源地址 412 仍为应用客户地址），即“身份替代”。或者，客户路由器例如可以通过例如 IGP 之类的本地通信协议（路由层）或者示意性地经由面向应用的协议（应用层）来利用相应的服务器 ID 来答复应用客户，如本领域技术人员所理解的那样。如果这样配置，则在这种情形下，应用客户可以将请求 400 重新发送到相应的应用服务器，即，目的地址字段 414 中含有所选应用服务器地址。

图 8 是阐释根据本发明用于优化 IP 骨干上的应用数据流的路由的过程的流程图。过程 800 开始于步骤 805，继续到步骤 810，其中服务器路由器（例如，服务器路由器 1 和 2）维护被连接的应用服务器（例如，分别是应用服务器 1 和 2）的服务器 ID 和服务器状态，如上所述。在步骤 815 中，服务器路由器例如使用 IGP 公告 300 中的 APPL_SERVER TLV 500 将服务器 ID 和服务器状态传播到网络中。中间路由器（例如，中间路由器 A 和 B）例如也使用 IGP 公告 300（如通过 TE 扩展）来在网络内传播链路的链路量度（例如，链路利用率），如上所述。客户路由器在步骤 825 中接收 IGP 公告 300，从而获知该信息，并在步骤 830 中将服务器 ID 和服务器状态存储在应用服务器表 600 中。在步骤 835 中，为了示意性地确定中间链路量度，客户路由器可以随后（如上所述，基于朝向其自身的链路量度）计算从其自身到每个应用服务器的反向 SPT，并将反向 SPT 存储在例如反向 SPT 表 249 中。值得注意地，客户路由器可以响应于收到的网络中的改变/更新来更新表 600 或计算出的反向 SPT。

在步骤 840 中，客户路由器例如通过截取客户请求 400 或通过充当应用服务器代理来接收请求，这两种方式如上所述。同样如上所述，作为响应，客户路由器在步骤 845 中基于服务器状态和例如反向 SPT 的中间链路量度来确定将请求发送到哪个应用服务器。在步骤 850 中若需要修改该请

求的应用服务器目的地（即，可以使用更理想的服务器），则客户路由器或者可以（例如，通过改变目的地址 414）修改客户请求目的地，或者代替地可以利用相应的应用服务器目的地（例如，更理想的服务器的地址和/或服务器 ID）来答复应用客户。客户路由器可以随后在步骤 860 中将客户请求转发到相应的应用服务器。值得注意地，响应于将相应的应用服务器目的地通知应用客户，客户路由器可以利用相应的应用服务器目的地来转发来自应用客户的返回请求（例如，“重接收”请求）。过程 800 终止于步骤 865。

有利地，新技术优化了计算机网络中 IP 骨干上的应用数据流的路由。通过基于服务器状态和（例如来自反向 SPT 的）中间链路量度来确定将客户请求发送到哪里，新技术允许客户请求被发送到最优的应用服务器，从而对该请求进行负载平衡。具体而言，请求可以被负载平衡，而无需应用层的路由和资源利用率知识，即，网络层拥有足够信息来优化应用数据流。另外，新技术的动态性质减轻了麻烦的手动配置的需要。

虽然已示出并描述了优化计算机网络中 IP 骨干上的应用数据流的路由的示例性实施例，但是应该理解，在本发明的精神和范围内可以做出各种其他变更和修改。例如，此处本发明是使用 IS-IS 链路状态分组作为 IGP 公告 300 来描述的。然而，本发明在其更广泛的意义上不限于此，实际上，它可以与其他 IGP 公告一起使用，如本领域技术人员所理解的 OSPF LSA（例如，类型 10 或 11 不透明 LSA）。另外，虽然上述说明描述了在连接到应用客户的客户路由器处执行本技术，但是给定本领域技术人员所理解的合适配置，该技术可以在网络内的其他节点处执行，如应用客户自身。另外，虽然示出并描述了本发明使用反向 SPT 来管理中间链路量度信息，但是本领域技术人员将理解，根据本发明，可以使用其他技术来收集/利用中间链路量度。

前述说明针对本发明的具体实施例。然而很明显，可以对所述实施例做出其他变更和修改，并取得它们的优点的一些或全部。例如，可以清楚地预料到，本发明的教导可以实现为软件（包括含有在计算机上运行的程序指令的计算机可读介质），硬件、固件或其组合。另外，可以生成电磁

信号来通过例如无线数据链路或诸如因特网之类的数据网络携带实现本发明的多个方面的计算机可执行指令。因此，本说明应仅通过示例的方式来看待，而不是要限制发明范围。因此，所附权利要求的目的是涵盖落入本发明的真实精神和范围内的所有这种变化和修改。

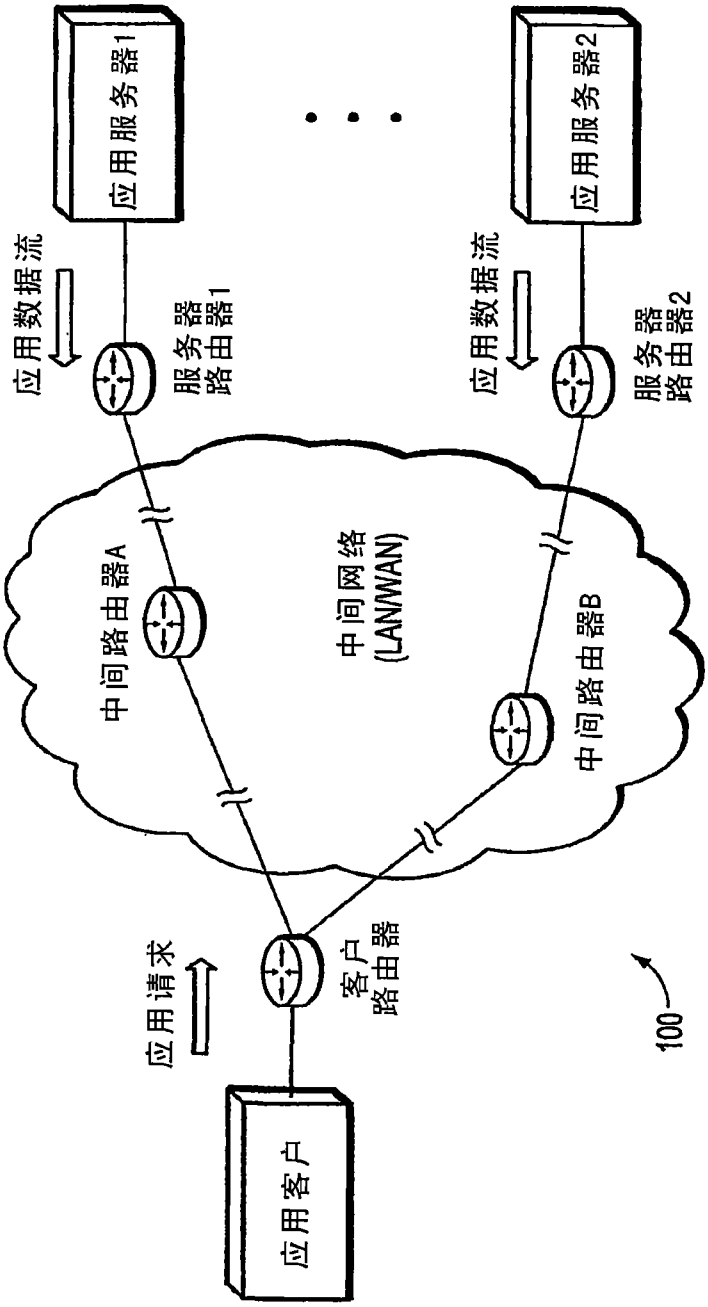

应用服务器1
应用服务器2
应用数据流
应用数据流
服务器路由器1
服务器路由器2
中间路由器A
中间网络(LAN/WAN)
中间路由器B
应用请求
客户路由器
应用客户
100

图1

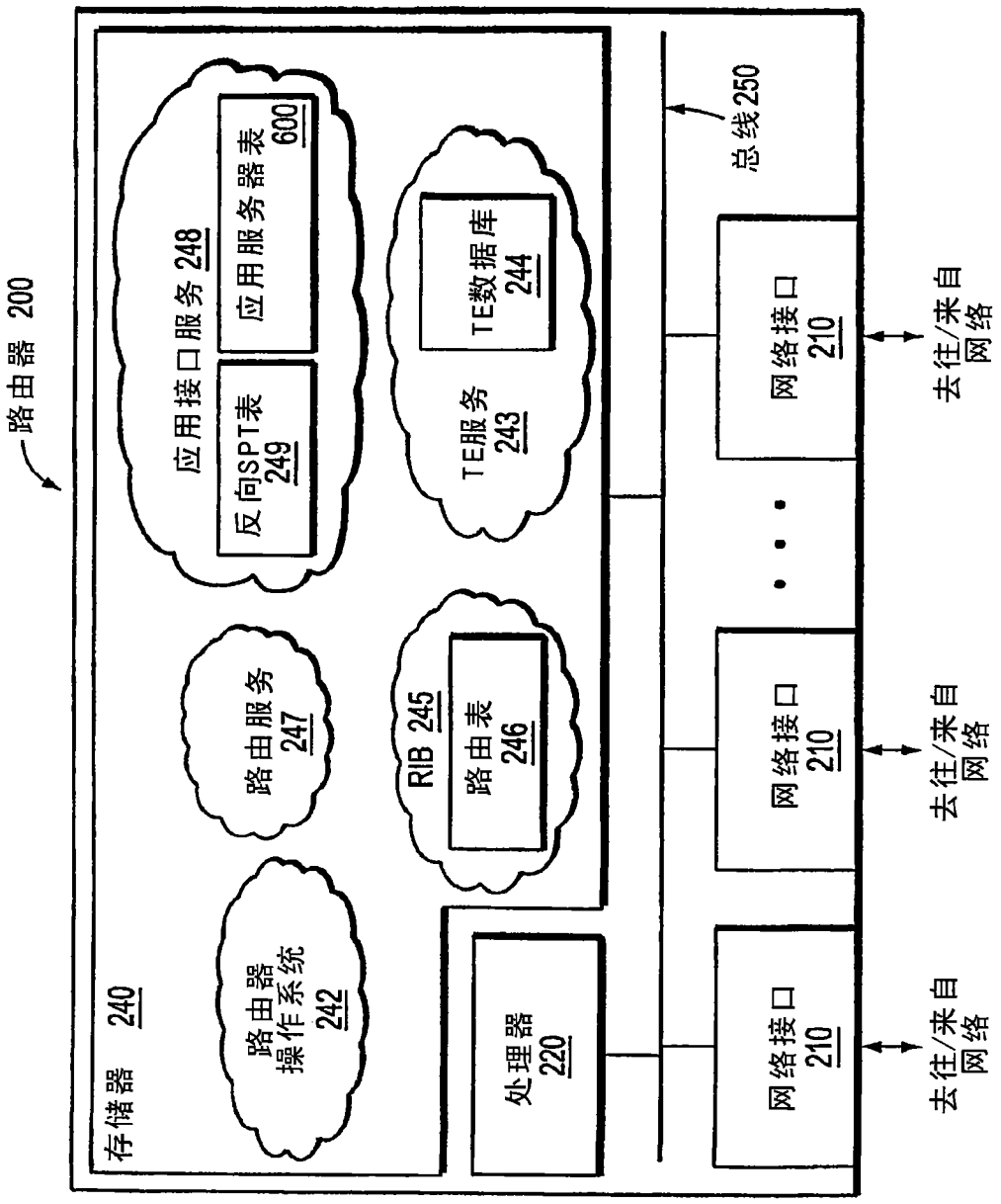

路由器 200
存储器 240
路由器操作系统 242
处理器 220
路由服务 247
RIB 245
路由表 246
应用接口服务 248
反向SPT表 249
应用服务器表 600
TE服务 243
TE数据库 244
总线 250
网络接口 210
网络接口 210
网络接口 210
去往/来自网络
去往/来自网络
去往/来自网络

图2

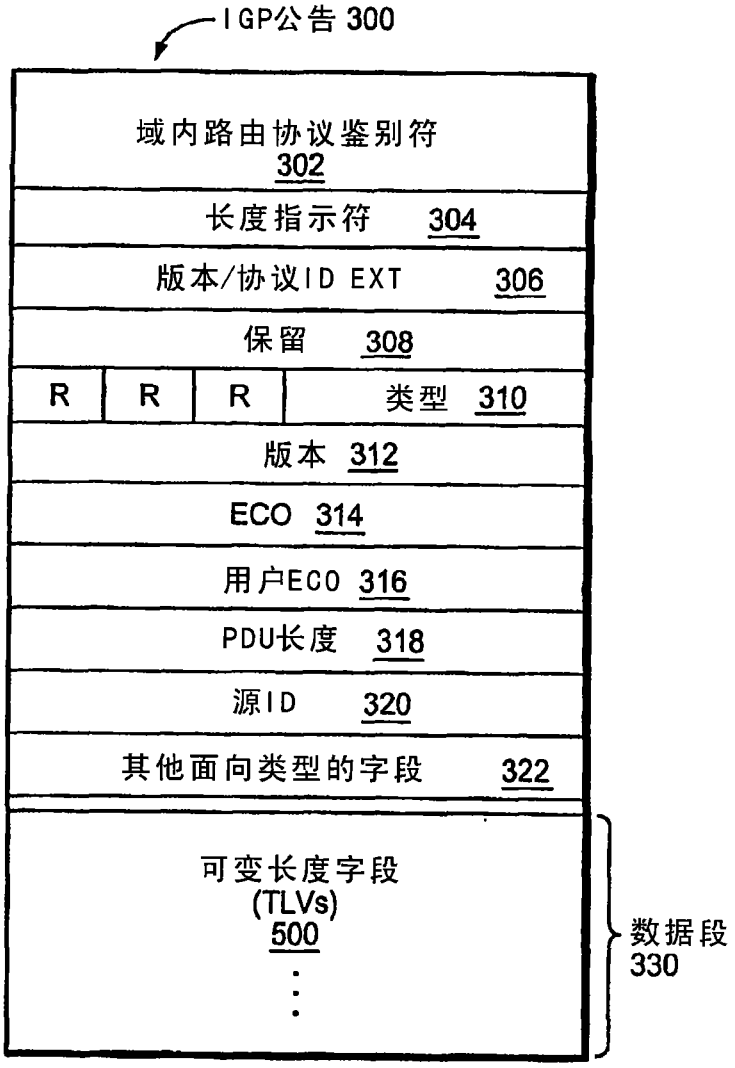

IGP公告 300
域内路由协议鉴别符
302
长度指示符 304
版本/协议ID EXT 306
保留 308
R R R 类型 310
版本 312
ECO 314
用户ECO 316
PDU长度 318
源ID 320
其他面向类型的字段 322
可变长度字段
(TLVs)
500
数据段
330

图3

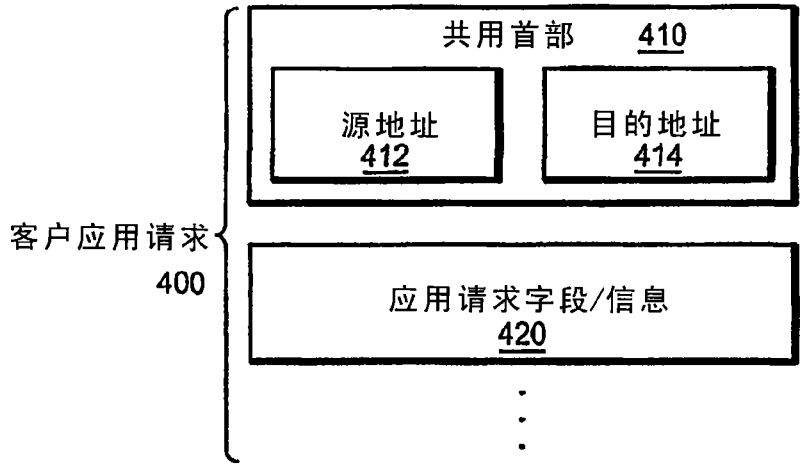


图4

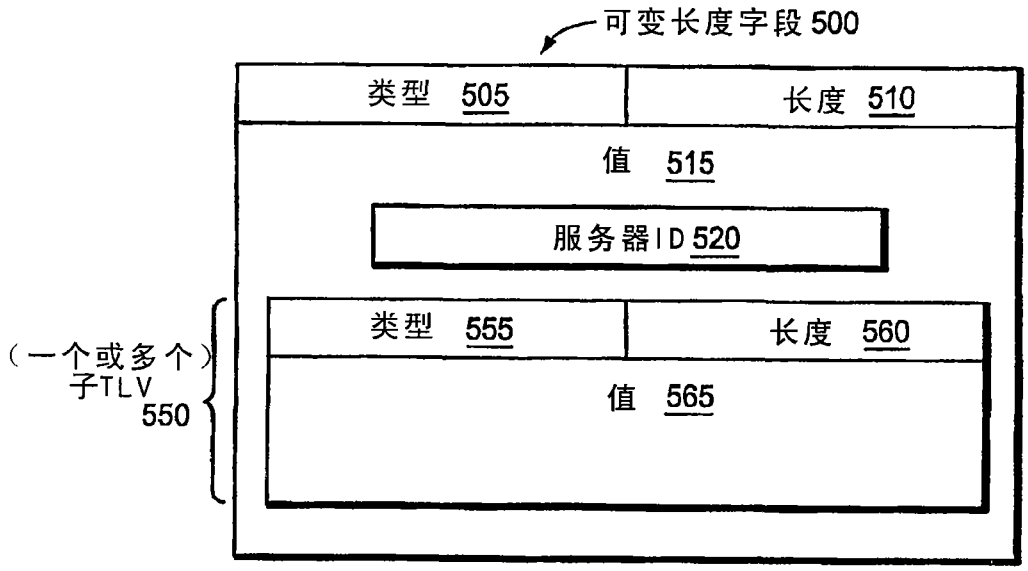

可变长度字段500
类型 505
长度 510
值 515
服务器ID 520
类型 555
长度 560
值 565
（一个或多个）
子TLV
550

图5

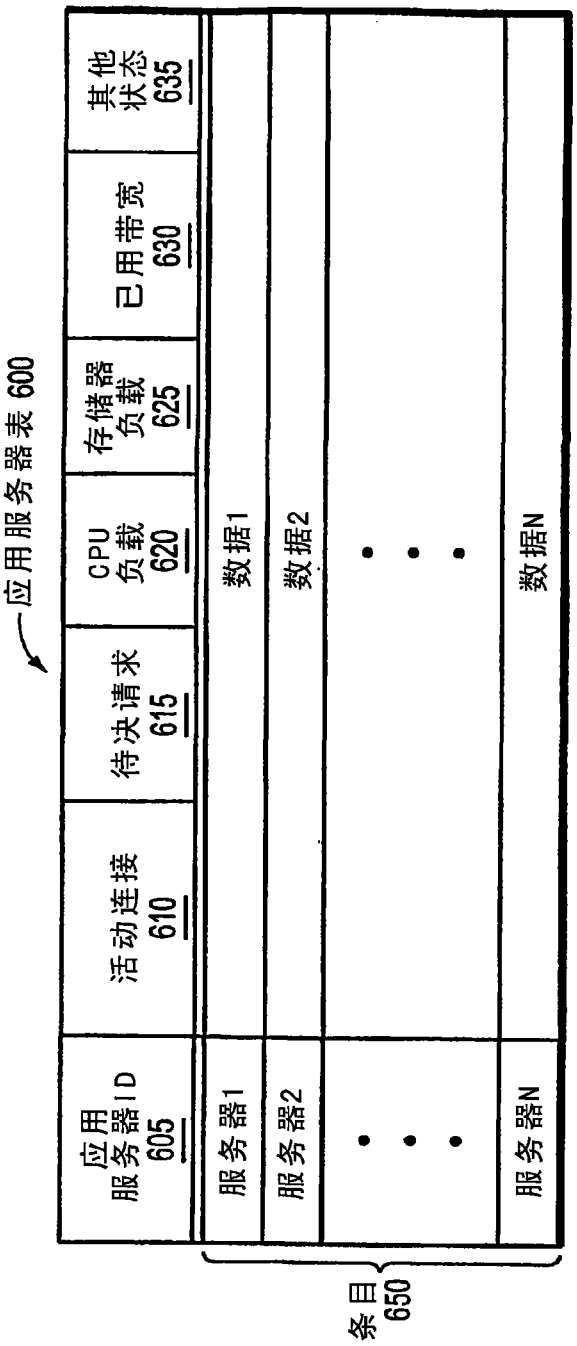

应用服务器表 600
应用服务器ID 605
活动连接 610
待决请求 615
CPU负载 620
存储器负载 625
已用带宽 630
其他状态 635
服务器1
服务器2
服务器N
数据1
数据2
数据N
条目 650

图6

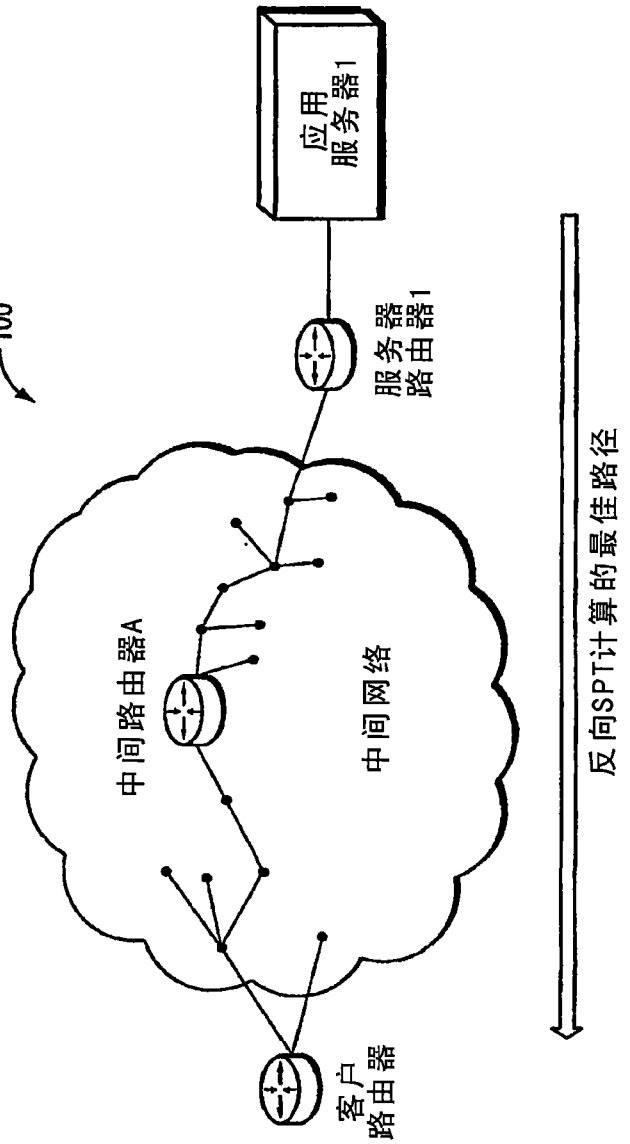


图7

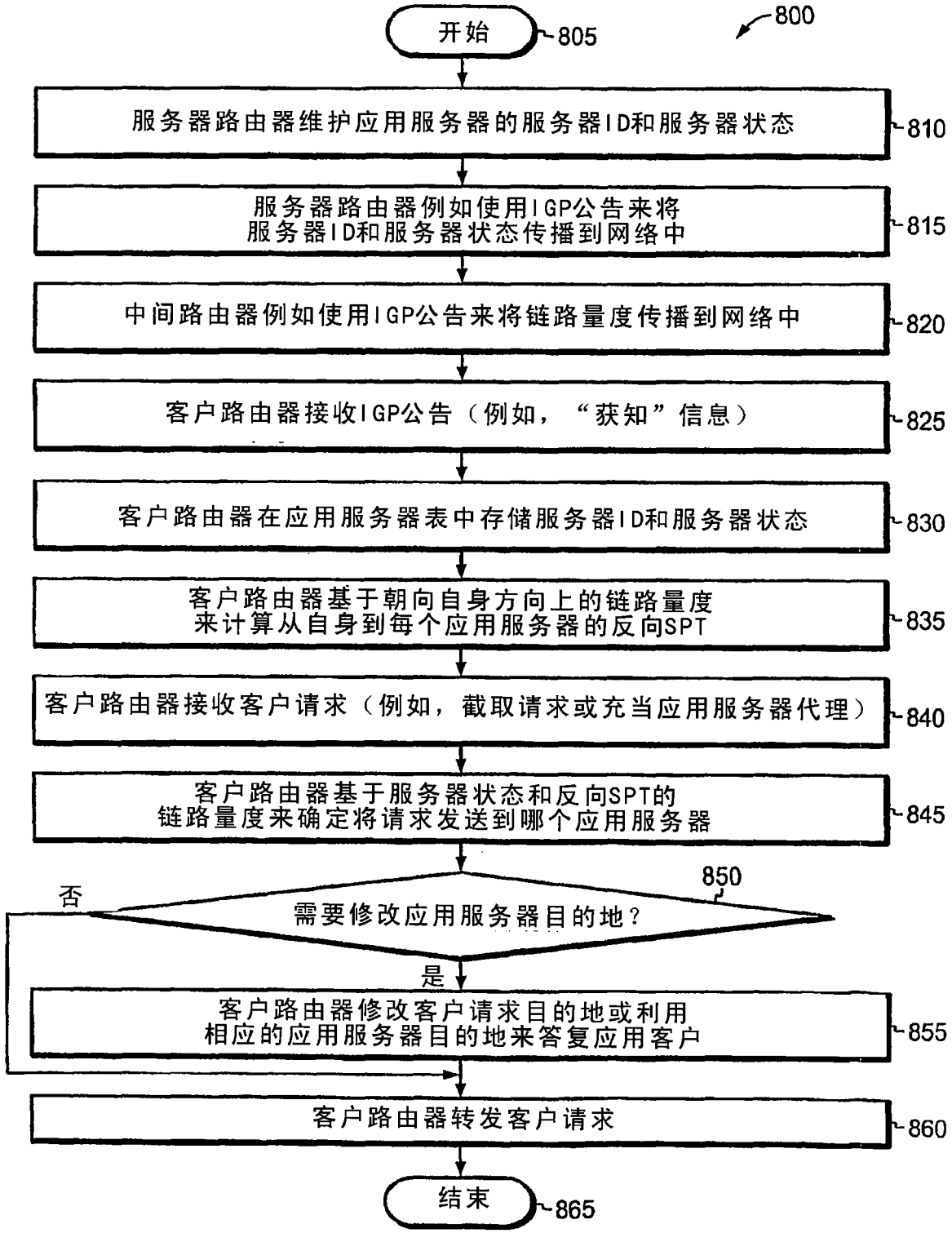

800
开始
805
服务器路由器维护应用服务器的服务器ID和服务器状态
810
服务器路由器例如使用IGP公告来将
服务器ID和服务器状态传播到网络中
815
中间路由器例如使用IGP公告来将链路量度传播到网络中
820
客户路由器接收IGP公告（例如，“获知”信息）
825
客户路由器在应用服务器表中存储服务器ID和服务器状态
830
客户路由器基于朝向自身方向上的链路量度
来计算从自身到每个应用服务器的反向SPT
835
客户路由器接收客户请求（例如，截取请求或充当应用服务器代理）
840
客户路由器基于服务器状态和反向SPT的
链路量度来确定将请求发送到哪个应用服务器
845
850
否
需要修改应用服务器目的地？
是
客户路由器修改客户请求目的地或利用
相应的应用服务器目的地来答复应用客户
855
客户路由器转发客户请求
860
结束
865

图8