(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104052669 A
(43)申请公布日 2014.09.17

(21)申请号 201410090589.9

(22)申请日 2014.03.12

(30)优先权数据
61/778,293 2013.03.12 US
14/194,567 2014.02.28 US

(71)申请人 西普联特公司
地址 美国加利福尼亚州

(72)发明人 王炜煌 M·巴兰 N·希瓦 Z·沙

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 王茂华 辛鸣

(51)Int.Cl.
H04L 12/745(2013.01)

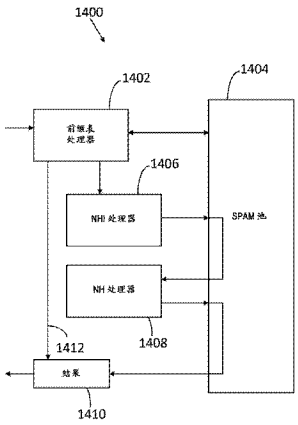
权利要求书1页 说明书4页 附图8页

(54)发明名称

用于处理交替配置的最长前缀匹配表的装置和方法

(57)摘要

本发明的各实施例涉及用于处理交替配置的最长前缀匹配表的装置和方法。一种网络交换机，包括存储器，存储器可配置用于存储字典树层级中的单个字典树的交替表表示。前缀表处理器使用输入网络地址并行访问单个字典树的交替表表示并且在每个交替表表示中搜索最长前缀匹配以获得本地前缀匹配。选择来自本地前缀匹配的最长前缀匹配。最长前缀匹配具有关联的下一跳索引基地址和偏移值。下一跳索引处理器利用下一跳索引基地址和偏移值访问存储器中的下一跳索引表以获得下一跳表指针。下一跳处理器使用下一跳表指针访问存储器中的下一跳表以获得目的地网络地址。



CN 104052669 A

1. 一种网络交换机，包括：

存储器，可配置用于存储字典树层级中的单个字典树的交替表表示；

前缀表处理器，用于：

使用输入网络地址并行访问所述单个字典树的所述交替表表示并且在每个交替表表示中搜索最长前缀匹配以获得本地前缀匹配，

从所述本地前缀匹配选择所述最长前缀匹配，其中所述最长前缀匹配具有关联的下一跳索引基地址和偏移值；

下一跳索引处理器，用于利用所述下一跳索引基地址和偏移值访问所述存储器中的下一跳索引表以获得下一跳表指针；以及

下一跳处理器，用于使用所述下一跳表指针访问所述存储器中的下一跳表以获得目的地网络地址。

2. 根据权利要求 1 所述的网络交换机，其中所述交替表表示包括标识选择的字典表节点的稀疏模式表示。

3. 根据权利要求 2 所述的网络交换机，其中所述稀疏模式表示包括分支标识和步幅值。

4. 根据权利要求 1 所述的网络交换机，其中所述交替表表示包括具有标识选择的字典树节点的位图的位图模式表示。

5. 根据权利要求 4 所述的网络交换机，其中所述位图模式表示包括分支标识和步幅值。

6. 根据权利要求 1 所述的网络交换机，其中所述交替表表示包括标识在字典树的底部的选择的字典树节点的叶推进表示。

7. 根据权利要求 6 所述的网络交换机，其中所述叶推进表示包括分支标识和步幅值。

8. 根据权利要求 1 所述的网络交换机，其中所述前缀表处理器是具有确定性查找延时的硬件资源。

9. 根据权利要求 1 所述的网络交换机，其中所述交替表表示包括在相同表中具有不同分组类型的表。

10. 根据权利要求 9 所述的网络交换机，其中所述不同分组类型包括 IPV4 分组和 IPV6 分组。

11. 根据权利要求 1 所述的网络交换机，其中所述前缀表处理器标识用于远程主机的最长前缀匹配和用于直接附接的主机的准确匹配。

12. 根据权利要求 1 所述的网络交换机，其中所述前缀表处理器在相同表中标识用于所述远程主机的所述最长前缀匹配和用于所述直接附接的主机的所述准确匹配。

13. 根据权利要求 1 所述的网络交换机，其中所述下一跳索引处理器处理下一跳表条目的块以有助于等价多路径寻路由。

14. 根据权利要求 13 所述的网络交换机，其中所述块指定上至 1024 个路径。

用于处理交替配置的最长前缀匹配表的装置和方法

[0001] 相关申请的交叉引用

[0002] 本申请要求对通过引用将其公开内容结合于此、于 2013 年 3 月 12 日提交的第 61/778,293 号美国临时专利申请的优先权。

技术领域

[0003] 本发明总体上涉及处理计算机网络中的流量。更特别地，本发明涉及具有交替配置的最长前缀匹配表的网络交换机。

背景技术

[0004] 图 1 图示了根据现有技术被利用的最长前缀匹配处理器 100。最长前缀匹配处理器 100 包括用于在转发表 104 中找到条目的搜索引擎 102。转发表中的每个条目具有关联的交换机标识符，交换机标识符表示用于网络分组的下一目的地(跳(hop))。

[0005] 图 2 图示了现有技术的转发表 200 的简化版本。转发表具有前缀列和下一跳列，前缀列具有条目的集合，下一跳列具有指明的交换机标识符。每个前缀条目是网络地址的分段。星号(*)指明“不在意”状态。如果没有前缀匹配，则下一跳被指明为交换机“A”。希望获得具有如下前缀的匹配，该前缀因为其表示到目标机器的最具体路径而最长。

[0006] 图 3 图示了由具有最长前缀匹配处理器 100 的交换机 302 接收的分组 300。最长前缀匹配处理器生成到另一机器的跳，另一机器比如为交换机 A、C 或者 E 或者机器 B、D、F 或者 G。

[0007] 图 4 图示了二叉树或者字典树(trie)400。该图的顶部图示了三角形 400，三角形 400 是在该图的底部示出的字典树 400 的简化的表示。字典树 400 具有由 0 和 1 表示的不同路径，其中 0 表示向左的路径并且 1 表示向右的路径。使用这一技术，可以用数字值指明字典树中的任何路径。如所示，字典树 400 的底部具有四个叶节点，其可以分支到八个目的地(0)至(7)。到目的地(0)的路径被表示为 000，因为三个左分支用来到达这一位置。到目的地(3)的路径被标识为 011，因为一个左分支(0)后接两个右分支(1)。以这一方式，可以使用二进制值指定每个路径。字典树 400 具有三层节点并且因此被称为具有为三的步幅(stride)。

[0008] 图 5 图示了字典树 400 可以与许多其他字典树具有任意复杂层级关系。例如，字典树 400 可以被视为层 0，在其下方具有在层 1 的八个字典树 400_1 至 400_8。这一字典树扩展可以推进用于任何数目的层，如图 5 中所示。每个字典树可以具有不同大小。因此，可以领会到，可以通过这样的结构定义任意长路径表达。

[0009] 图 6 图示了字典树 600 的每个节点可以具有指明的编号。在这一示例中，如所示，节点被编号为 1 至 7。

[0010] 图 7 图示了其中下一跳信息驻留在暗节点(节点 5 和 7)的场景。到节点 5 的路径可以用二进制被表达为 01，而到节点 7 的路径可以用二进制被表达为 11。参照图 5 可以领会到，节点路径可以遍历多个层并且将用长二进制值表达到特定节点的作为结果的路径。

[0011] 鉴于前述，存在对提供用于标识最长前缀匹配的改进技术的持续需求。

发明内容

[0012] 一种网络交换机，包括存储器，存储器可配置用于存储字典树层级中的单个字典树的交替表表示。前缀表处理器使用输入网络地址并行访问单个字典树的交替表表示并且在每个交替表表示中搜索最长前缀匹配以获得本地前缀匹配。选择来自本地前缀匹配的最长前缀匹配。最长前缀匹配具有关联的下一跳索引基地址和偏移值。下一跳索引处理器利用下一跳索引基地址和偏移值访问存储器中的下一跳索引表以获得下一跳表指针。下一跳处理器使用下一跳表指针访问存储器中的下一跳表以获得目的地网络地址。

附图说明

[0013] 连同附图结合以下具体描述来更完全地领会本发明，在附图中：

[0014] 图 1 图示了现有技术的最长前缀匹配处理器。

[0015] 图 2 图示了现有技术的转发表。

[0016] 图 3 图示了具有用于将分组路由到任何数目的交换机或者机器的最长前缀匹配处理器的交换机。

[0017] 图 4 图示了现有技术的二叉树。

[0018] 图 5 图示了二叉树的具有任意复杂路径的嵌套配置。

[0019] 图 6 图示了根据本发明的实施例被使用的节点编号。

[0020] 图 7 图示了在选择的节点具有下一跳信息的二叉树。

[0021] 图 8 图示了根据本发明的实施例被利用的稀疏模式表配置。

[0022] 图 9 图示了根据本发明的实施例被利用的位图模式表配置。

[0023] 图 10 图示了根据本发明的实施例被利用的叶推进（leaf-push）模式表配置。

[0024] 图 11 图示了根据本发明的实施例被利用的下一跳索引表。

[0025] 图 12 图示了根据本发明的实施例被利用的下一跳表。

[0026] 图 13 图示了根据本发明的实施例被执行的分组处理操作。

[0027] 图 14 图示了根据本发明的实施例被利用的处理组件。

[0028] 相似参考号贯穿附图的若干视图指示对应的部分。

具体实施方式

[0029] 本发明是在交换机中嵌入的最长前缀匹配处理器。最长前缀匹配处理器利用交替配置的最长前缀匹配表，可以针对不同最长前缀匹配搜索策略和存储器优化策略优化最长前缀匹配表。在图 8 中示出了一个最长前缀匹配表。该表的第一列指定分组类型，在这一实例中为 IPV4。有利地，IPV4 和 IPV6 分组类型可以被高效地存储在单个表中。下一列指定存储模式，在这一实例中为稀疏类型。稀疏类型有助于针对单个字典树的大步幅。即，稀疏模式有助于针对其中单个字典树大的配置的处理。下一列指定分支长度。分支长度是从层级字典树的根节点到单个字典树（比如图 5 中的任意字典树（也被称为子字典树））的任意根节点的路径的长度的指示符。步幅指定字典树的大小，比如 $2^{步幅}$。在这一示例中，步幅为 5 或者 2^5 或者 32。下一列是下一跳索引基地址。本发明将下一跳索引用于对下一跳表

的引用。这允许在配置存储器资源中的最大灵活性。该表的下一列是分支 ID，比如如以上讨论的通过树的路径的二进制表示。图 8 的最后列指定具有下一跳信息的字典树节点 ID。图 7 提供字典树节点和具有下一跳信息的对应的节点 ID 的示例。图 8 的前四个字段具有固定宽度。剩余的两个字段具有可变宽度。

[0030] 图 9 图示了用于前缀匹配表的交替配置。第一字段指定分组类型。下一字段指定位图模式。如以上讨论的，下一字段指定分支长度。下一字段指定步幅。在位图模式中，用于步幅 5 的字典树的所有节点被存储在 2^5-1 或者 31 位的数组中。下两个字段是 NHI 基地址和分支 ID。如以上讨论的，分支 ID 指定分支路径。NHI 基地址指定对应于字典树位图的最低有效位的基位置，其被标记为 1。其他节点使它们的指针从基位置递增进入 NHI 表。例如，如果基位置为零，则被标记为 1 的第五位将从基位置被递增 4 个条目。最后字段是字典树位图。在这一示例中，字典树位图为 31 位图。位图指定在指定的字典树中具有下一跳信息的不同位置。

[0031] 图 10 图示了用于前缀匹配表的交替配置。第一字段指定叶推进模式。在叶推进模式中，在数组中仅表示最底部节点。例如，关于图 6，在数组中将仅表示节点 4、5、6 和 7。如果图 6 的其他节点具有下一跳信息，则可以使用各种技术（比如受控前缀扩展）来将该信息定位在最底部节点中。NHI 表在这一实例中需要进行对应的复制。这一模式在其中用于直接连接的主机的网络地址被存储在相同表的实例中特别存储器高效，因为那些主机通常具有连续地址。这一模式也是有利的，因为所有搜索在相同层终止。

[0032] 图 10 的表也具有分支长度字段、步幅字段、NHI 基地址字段、分支 ID 字段和字典树底部位图字段。在 5 步幅的实例中，整个位图具有 2^5-1 或者 31 位，这导致 16 个节点的底部集合。因此，16 位图可以用来指定这些不同节点的位置。

[0033] 图 11 图示了下一跳索引表。表中的不同条目指定到下一跳表基位置加上块大小偏移的指针。基位置指定进入表的入口点而块大小指定从该基位置的附加字段。下一跳表条目的块大小被用于等价多路径（ECMP）寻路由，一种高级网络寻路由策略。连同 IP 地址，对最长前缀匹配引擎的每个请求伴随有用于路径选择的 ECMP 哈希值。ECMP 哈希值可以大于可用块大小，因此最后路径选择被计算为下一跳基地址加上下一跳块大小的输入 ECMP 哈希值取模。在一个实施例中，针对 ECMP 支持上至 1024 个路径。

[0034] NHI 表大小是字典树配置的函数。表大小可以被动态配置和优化。

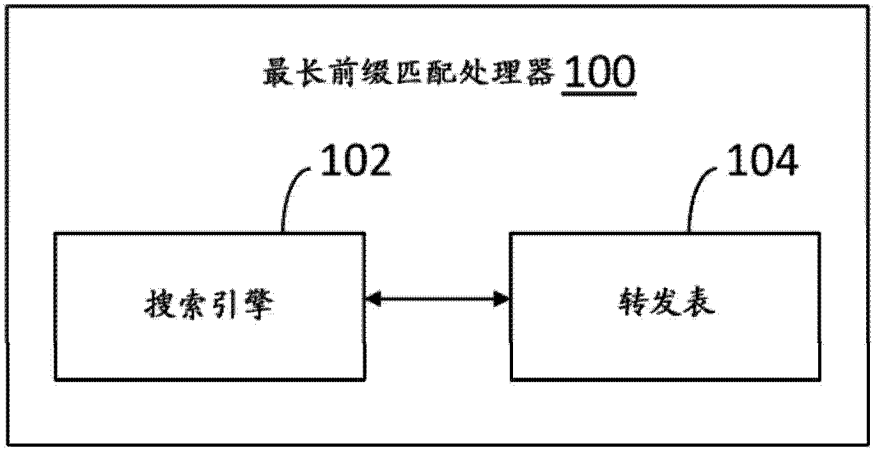
[0035] 图 12 图示了下一跳表。该表中的每个条目指定到目的地机器的跳。

[0036] 图 13 图示了根据本发明的实施例被执行的处理操作。入口网络地址（例如，网际协议地址）1300 应用于前缀表 0 至 N 的集合，全部或者部分表可以由软件配置。如结合图 8 至图 10 讨论的，每个前缀表包含前缀条目的集合，每个前缀条目指定到字典树的路径并且然后是该字典树的特性。使用前缀表 01302 作为示例，入口网络地址的位的软件配置的选择应用于哈希函数以映射到前缀表地址 1304。有利地，在硬件中利用固定延时完成所有匹配。然后读取前缀表存储器以到达子字典树 0（任意复杂层级字典树结构中的单个字典树，比如图 5 中所示的字典树）0136。然后在该子字典树执行最长前缀匹配 1308。跨所有前缀表 0 至 N 重复这一处理以创建本地前缀匹配。然后处理结果以选择最长前缀匹配 1310。然后读取 NHI 表 1312，这导致对下一跳表的访问 1314，这产生下一跳 1316。

[0037] 图 14 图示了根据本发明的实施例被配置的最长前缀匹配处理器 1400。处理器

1400 包括前缀表处理器 1402，其为被配置用于指定并且然后访问不同最长前缀匹配表(比如图 8- 图 10 中的最长前缀匹配表)的硬件资源。表驻留在存储器 1404 中，存储器 1404 可以是 SPAM 资源池。前缀表处理器 1402 协调图 13 的操作 1304 至 1310。下一跳索引处理器 1406 然后可以用来访问存储器 1404 中的下一跳索引表，下一跳索引表向下一跳处理器 1408 返回基地址和偏移(如果存在)。下一跳处理器 1408 访问在存储器 1404 中存储的下一跳表，其产生下一跳地址作为结果 1410。如果前缀表处理器 1402 具有字典树未命中或者其他错误，则可以如用箭头 1412 所示直接生成错误或者默认结果 1410。

[0038] 前述描述出于说明的目的使用具体命名提供对本发明的透彻理解。然而，对本领域技术人员将明显的是，并不需要具体细节以便实践本发明。因此，出于例示和描述的目的呈现本发明的具体实施例的前述描述。它们并不旨在于是详尽的或者将本发明限制到所公开的精确形式；明显地，鉴于以上教导许多修改和改变是可能的。选择并且描述实施例以便最佳地说明本发明的原理及其实践应用，它们因此使得本领域其他技术人员能够最佳地利用本发明以及具有在适合于预期的特定使用时的各种修改的各种实施例。旨在于以下权利要求及其等效物限定本发明的范围。



（现有技术）

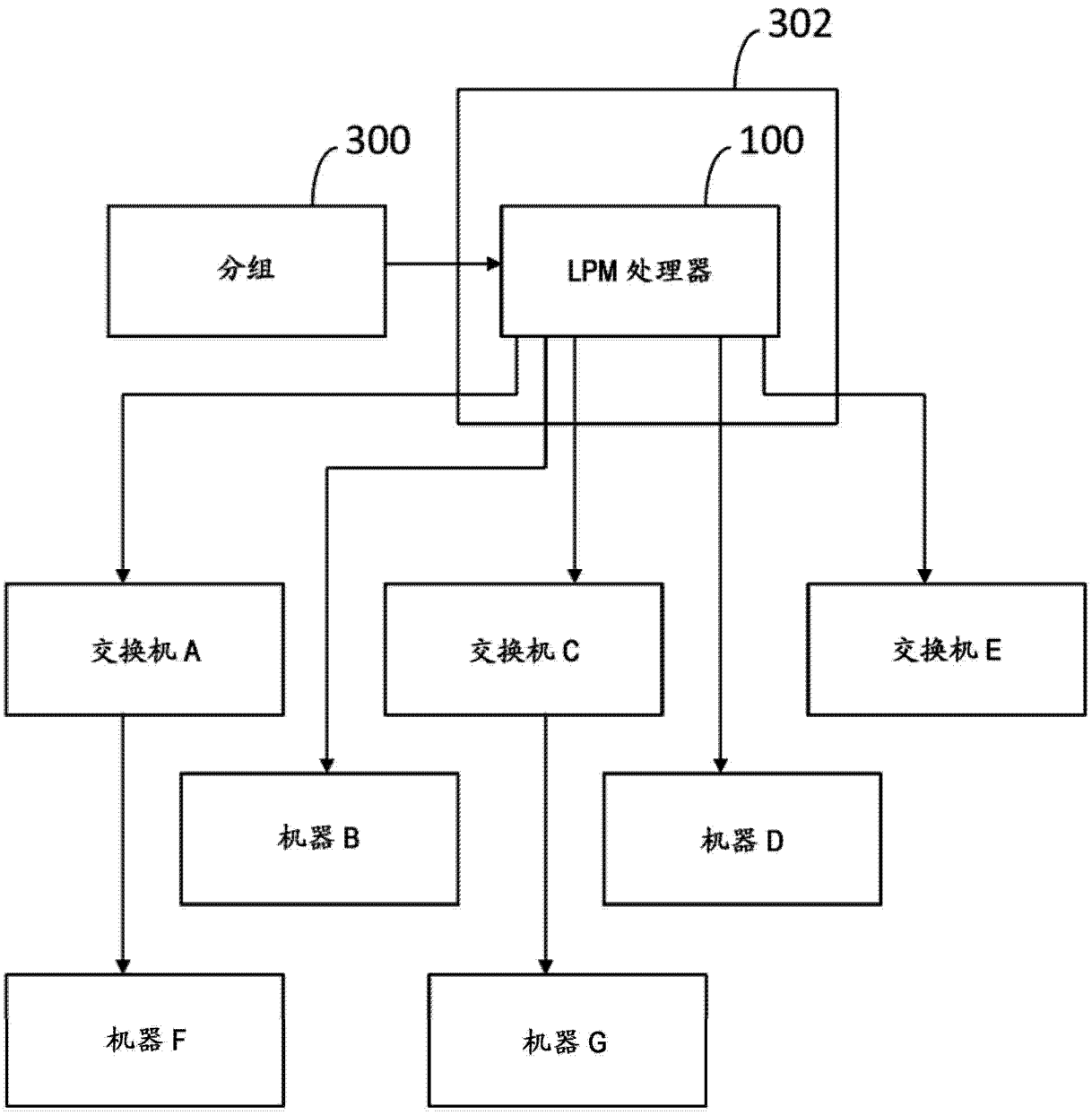
图 1

转发表 200

前缀	下一跳
*	A
0*	C
011*	D
1010*	F
11010*	G

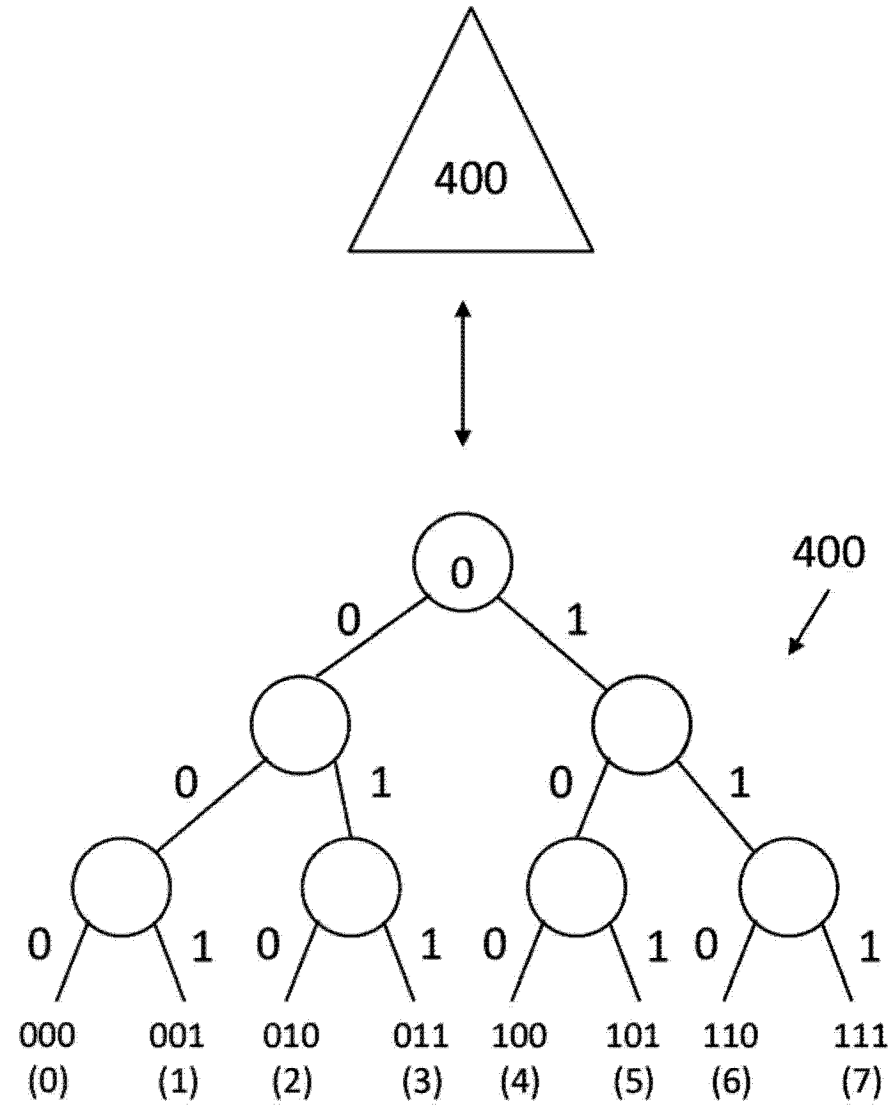
（现有技术）

图 2



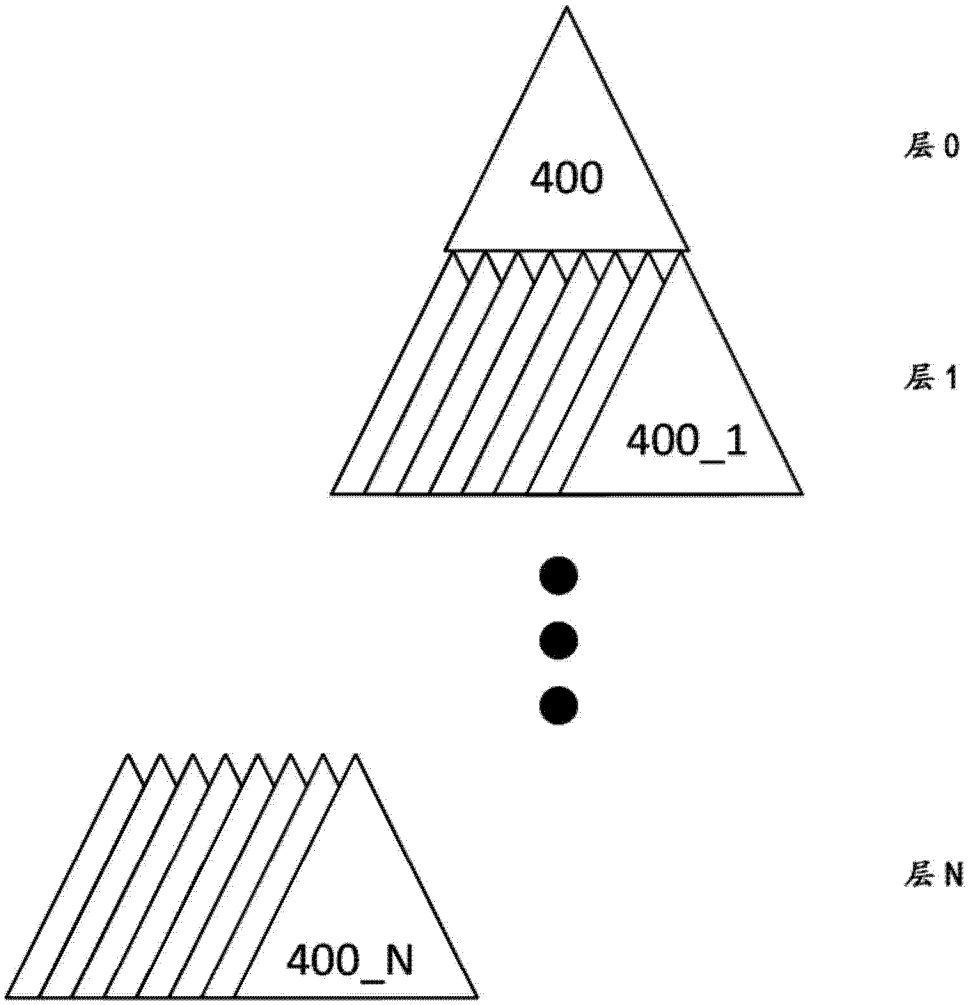
（现有技术）

图 3


400
400
0
0
1
0
1
0
1
0
1
0
1
0
1
0
1
000
(0)
001
(1)
010
(2)
011
(3)
100
(4)
101
(5)
110
(6)
111
(7)

（现有技术）

图 4



（现有技术）

图 5

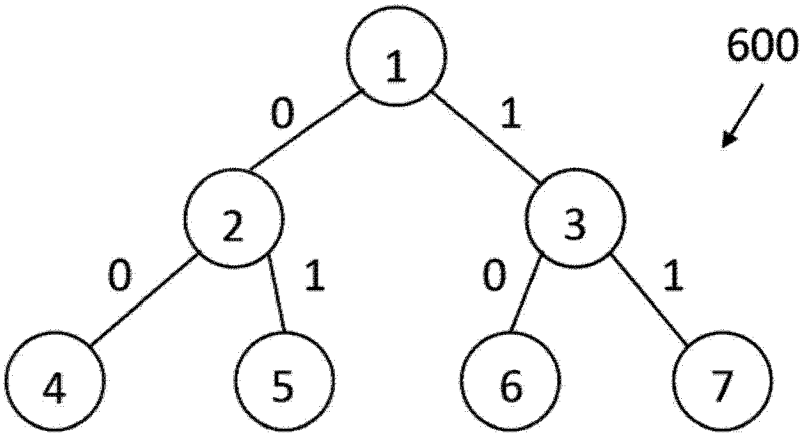


图 6

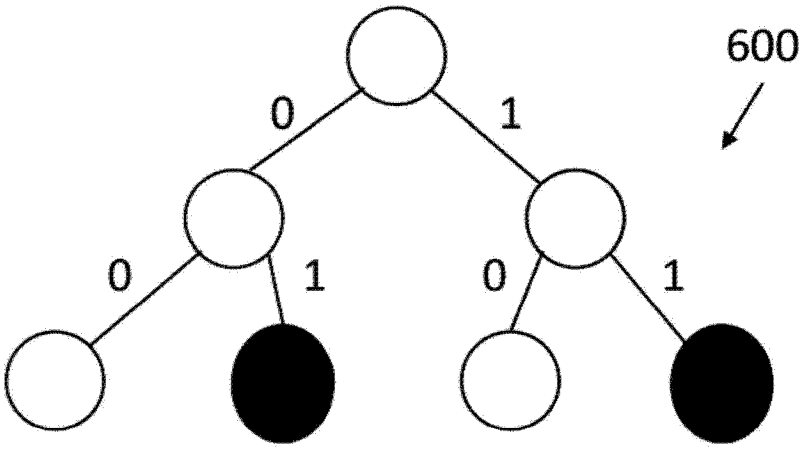


图 7

类型	模式	分支长度	步幅	NHI 基地址	分支 ID	字典树节点 ID
IPV4	稀疏	X	5	XY	ABC	A, B, C

图 8

类型	模式	分支长度	步幅	NHI 基地址	分支 ID	字典树位图
IPV6	位图	X	5	XY	ABC	31 位图

图 9

类型	模式	分支长度	步幅	NHI 基地址	分支 ID	字典树底部位图
IPV4	叶推进	X	5	XY	ABC	16 位图

图 10

NHI 表
到 NH 表加块大小的指针

图 11

NH 表
跳 1
跳 2
跳 3
跳 4
跳 5
跳 6
跳 7

图 12

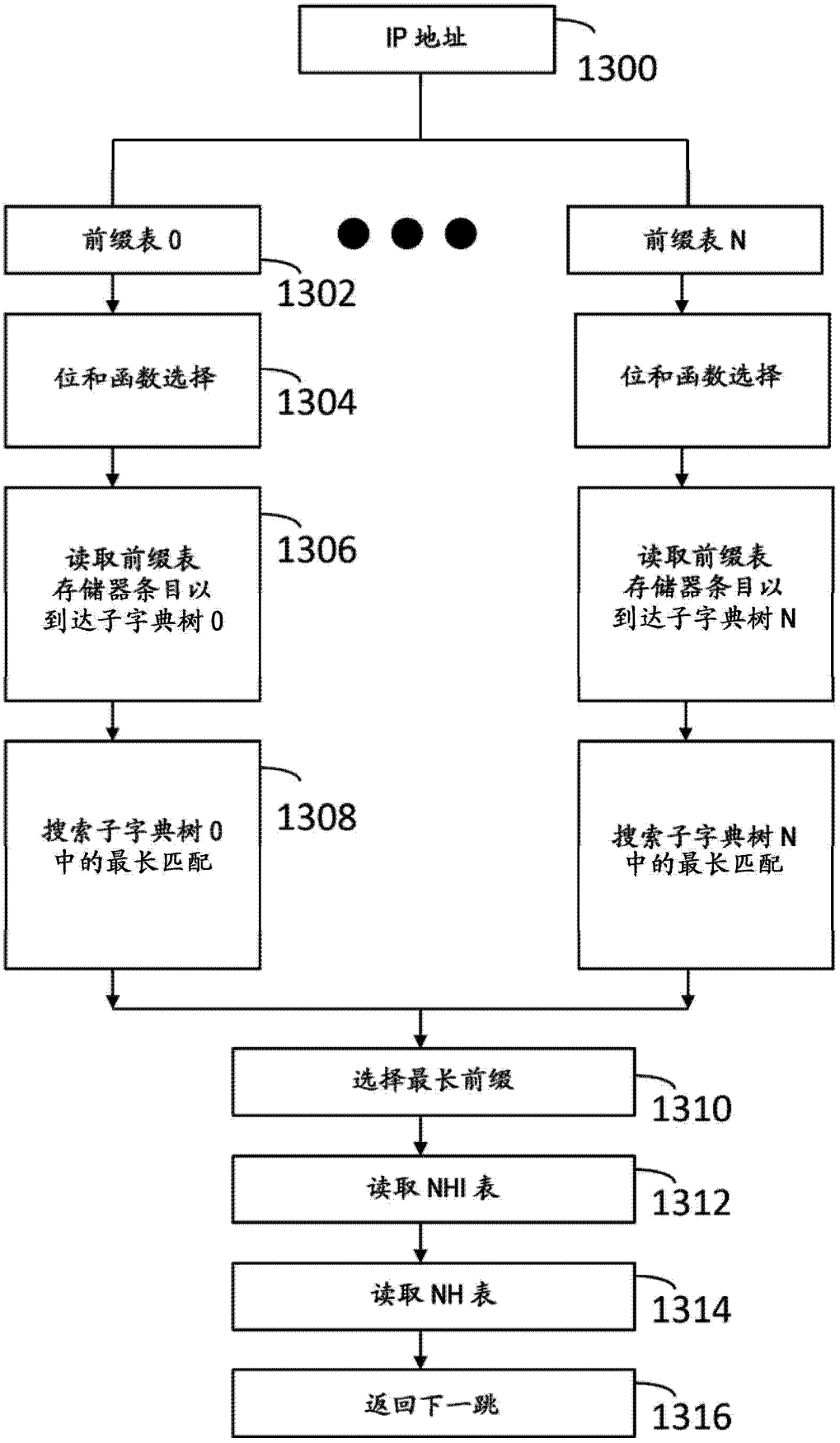

IP 地址
1300
前缀表 0
前缀表 N
1302
位和函数选择
位和函数选择
1304
读取前缀表
存储器条目以
到达子字典树 0
读取前缀表
存储器条目以
到达子字典树 N
1306
搜索子字典树 0
中的最长匹配
搜索子字典树 N
中的最长匹配
1308
选择最长前缀
1310
读取 NHI 表
1312
读取 NH 表
1314
返回下一跳
1316

图 13

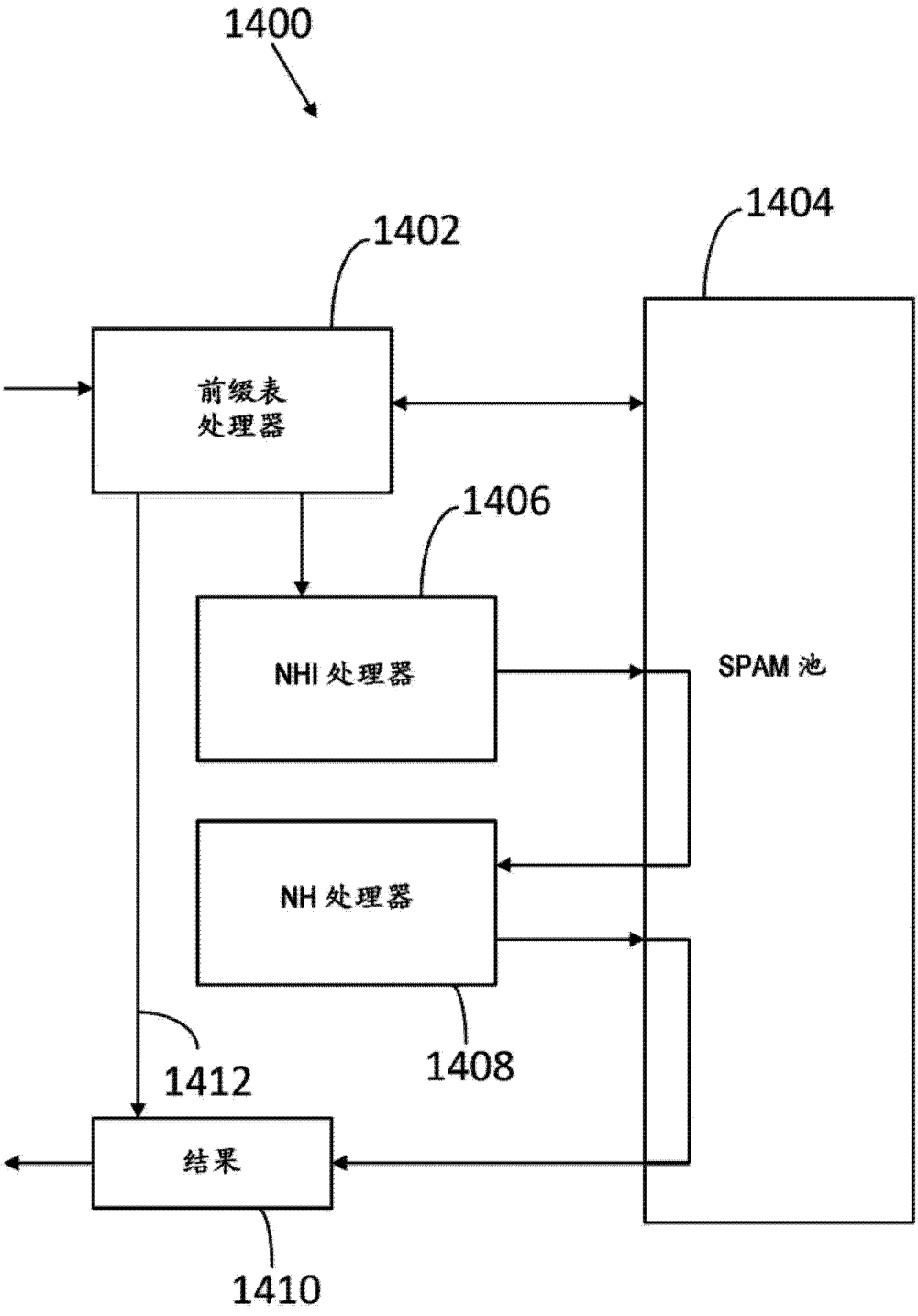

1400
1402
1404
前缀表
处理器
1406
NHI 处理器
SPAM 池
NH 处理器
1408
1412
结果
1410

图 14

Abstract

A network switch includes a memory configurable to store alternate table representations of an individual trie in a hierarchy of tries. A prefix table processor accesses in parallel, using an input network address, the alternate table representations of the individual trie and searches for a longest prefix match in each alternate table representation to obtain local prefix matches. The longest prefix match from the local prefix matches is selected. The longest prefix match has an associated next hop index base address and offset value. A next hop index processor accesses a next hop index table in the memory utilizing the next hop index base address and offset value to obtain a next hop table pointer. A next hop processor accesses a next hop table in the memory using the next hop table pointer to obtain a destination network address.