(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104641607 B
(45)授权公告日 2018.01.16

(21)申请号 201380048216.9
(22)申请日 2013.09.11
(65)同一申请的已公布的文献号
申请公布号 CN 104641607 A
(43)申请公布日 2015.05.20
(30)优先权数据
61/702,317 2012.09.18 US
13/708,200 2012.12.07 US
(85)PCT国际申请进入国家阶段日
2015.03.17
(86)PCT国际申请的申请数据
PCT/US2013/059171 2013.09.11
(87)PCT国际申请的公布数据
WO2014/046929 EN 2014.03.27
(73)专利权人 思科技术公司
地址 美国加利福尼亚州
(72)发明人 托马斯·J·埃兹尔
艾勒山多·弗里
普图·哈利·撒巴吉奥 李明哲
克里斯托弗·A·维尔德曼
林仪筹 黄之聪
(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 11258
代理人 李晓冬
(51)Int.Cl.
H04L 12/851(2006.01)
H04L 12/813(2006.01)
(56)对比文件
US 2012177046 A1,2012.07.12,
US 2012177046 A1,2012.07.12,
US 2009238184 A1,2009.09.24,
US 2010158009 A1,2010.06.24,
CN 1846409 A,2006.10.11,
审查员 王严冉

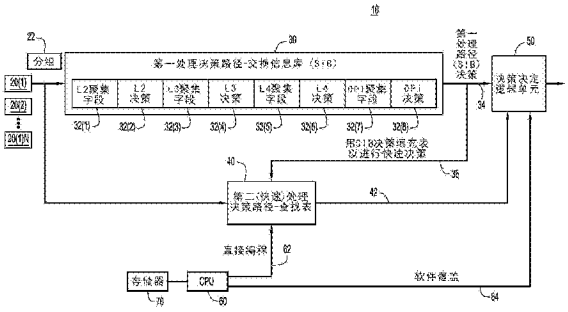
权利要求书2页 说明书5页 附图4页

(54)发明名称

一种用于做出超低延迟处理决策的方法及装置

(57)摘要

本文提出了针对网络设备中的分组实现处理决策的超低延迟确定的技术。分组在网络设备的端口处被接收。在第一处理决策路径中基于分组的内容和一个或多个网络策略来确定处理决策。在与第一处理路径并行的第二处理决策路径中，通过访问存储了处理决策的表来确定处理决策。针对与被包含在表中的一个或多个特定分组流参数相匹配的分组，第二处理决策路径能够比第一处理决策路径更快地输出处理决策。如果第二处理决策路径做出了处理决策，则使用由第二处理决策路径确定的处理决策，否则使用由第一处理决策路径确定的处理决策。



CN 104641607 B

1.一种用于做出超低延迟处理决策的方法,包括:

在网络设备的端口处接收分组;

在第一处理决策路径中基于所述分组的内容和一个或多个网络策略来确定处理决策;

在第二处理决策路径中通过访问存储了处理决策的表来确定处理决策,并且针对与被包含在所述表中的一个或多个特定分组流参数相匹配的分组,所述第二处理决策路径能够比所述第一处理决策路径更快地输出处理决策;以及

如果处理决策能够由所述第二处理决策路径做出,则使用由所述第二处理决策路径确定的处理决策,否则使用由所述第一处理决策路径确定的处理决策,

其中,所述第一处理决策路径中的处理决策和所述第二处理决策路径中的处理决策各自表示来自一组决策的选择,所述一组决策包括转发所述分组、交换所述分组、允许所述分组旁路、以及丢弃所述分组。

2.如权利要求1所述的方法,其中,在所述第二处理决策路径中的确定包括:使用掩蔽来执行逐位比较,以确定与被存储在所述表的字段中的数据的部分匹配或精确匹配。

3.如权利要求1所述的方法,其中,在所述第二处理决策路径中的确定包括:在所述表中执行单一查找,以在所述分组的一个或多个流参数和所述表中的条目中的一个或多个特定分组流参数之间进行匹配。

4.如权利要求3所述的方法,其中,在所述第二处理决策路径中的确定包括:使用所述分组的第2层字段、第3层字段、第4层字段、深度分组检测字段以及用户定义的参数中的一个或多个的任何组合来在所述表中执行所述单一查找,以确定与被存储在所述表中的一个或多个特定分组流参数的匹配。

5.如权利要求4所述的方法,其中,在所述第二处理决策路径中的确定包括:使用包括所述分组的第2层字段、第3层字段、第4层字段、深度分组检测字段以及用户定义的参数中的一个或多个的特定组合的静态手动配置关键词,来确定与被存储在所述表中的一个或多个特定分组流参数的匹配。

6.如权利要求1所述的方法,其中,在所述第一处理决策路径中的确定包括:检查所述分组的第2层、第3层、第4层和深度分组检测报头字段,以对所述分组做出处理决策。

7.如权利要求1所述的方法,还包括:用与由所述第一处理决策路径做出的处理决策相关联的信息填充所述第二处理决策路径中的所述表,以供用于处理随后接收的分组。

8.如权利要求1所述的方法,还包括:用针对分组做出的满足一个或多个特定分组流参数的一个或多个处理决策的信息对所述第二处理决策路径的所述表进行编程。

9.如权利要求1所述的方法,还包括:如果所述第二处理决策路径做出了处理决策,则覆盖由所述第二处理决策路径确定的处理决策,并且使用由所述第一处理决策路径确定的处理决策。

10.如权利要求1所述的方法,还包括:覆盖由所述第一处理决策路径确定的处理决策,并且使用由所述第二处理决策路径确定的处理决策。

11.如权利要求1所述的方法,其中,所述第一处理决策路径中的确定和所述第二处理决策路径中的确定包括访问共享的访问控制列表表。

12.一种用于做出超低延迟处理决策的装置,包括:

多个端口,所述多个端口被配置为从网络接收分组和将分组输出到所述网络;

第一处理决策路径，所述第一处理决策路径被配置为基于所述分组的内容和一个或多个网络策略来确定针对接收的分组的处理决策；

与所述第一处理决策路径并行的第二处理决策路径，所述第二处理决策路径包括存储了处理决策的表，并且针对与被包含在所述表中的一个或多个特定分组流参数相匹配的分组，所述第二处理决策路径能够比所述第一处理决策路径更快地输出处理决策；以及

决策决定逻辑单元，所述决策决定逻辑单元被配置在所述第二处理决策路径做出了处理决策的情况下，使用由所述第二处理决策路径确定的处理决策，否则使用由所述第一处理决策路径确定的处理决策，

其中，所述第一处理决策路径中的处理决策和所述第二处理决策路径中的处理决策各自表示来自一组决策的选择，所述一组决策包括转发所述分组、交换所述分组、允许所述分组旁路、以及丢弃所述分组。

13.如权利要求12所述的装置，其中，所述第二处理决策路径被配置为在所述表中执行单一查找，以在所述分组的一个或多个流参数和所述表中的条目中的一个或多个特定分组流参数之间进行匹配。

14.如权利要求13所述的装置，其中，所述第二处理决策路径被配置为使用所述分组第2层字段、第3层字段、第4层字段、深度分组检测字段以及用户定义的参数中的一个或多个的任何组合来在所述表中执行所述单一查找，以确定与被存储在所述表中的一个或多个特定分组流参数的匹配。

15.如权利要求12所述的装置，其中，所述第一处理决策路径包括逻辑，该逻辑被配置为检查所述分组的第2层、第3层、第4层和深度分组检测报头字段以针对所述分组做出处理决策。

16.如权利要求12所述的装置，其中，所述第一处理决策路径的输出被耦合到所述第二处理决策路径，以用与由所述第一处理决策路径做出的处理决策相关联的信息填充所述第二处理决策路径中的所述表，以供用于处理随后接收的分组。

17.如权利要求12所述的装置，还包括：处理器单元，所述处理器单元被配置为用针对分组做出的满足一个或多个特定分组流量参数的一个或多个处理决策的信息对所述第二处理决策路径的所述表进行编程。

18.如权利要求12所述的装置，其中，所述决策决定逻辑单元被配置为：如果所述第二处理决策路径做出了处理决策，则覆盖由所述第二处理决策路径确定的处理决策，并且使用由所述第一处理决策路径确定的处理决策。

一种用于做出超低延迟处理决策的方法及装置

[0001] 相关申请的交叉引用

[0002] 本申请要求于2012年9月18日提交的、名称为“Ultra Low Latency Multi-Protocol Networking Device(超低延迟多协议网络设备)”的美国临时申请No.61/702,317的优先权,该临时申请的整体通过引用结合于此。

技术领域

[0003] 本公开一般地涉及减少网络设备中的延迟。

背景技术

[0004] 超低延迟(ULL)网络对诸如高频交易(HFT)用户(其中按每纳秒计数)之类的某些用户来说是关键的。更具体地,比对手更快使得HFT客户能够在波动性周期期间捕捉机会,增加订单流、流动性以及加速价格发现。

[0005] 诸如交换机之类的传统的网络设备建立于顺序地做出决策的传统方法之上。虽然这样简化了设计考虑,但是由于决策被延期并且重要的资源 (即,复制表)是需要的,因此顺序的方法也引入了内在的延迟。

附图说明

[0006] 图1是具有多个处理决策路径的网络设备的框图,多个处理决策路径包括被配置为对分组做出超低延迟处理决策的快速处理决策路径。

[0007] 图2是示出了用于图1的网络设备的快速处理决策路径中的表的示例的示图。

[0008] 图3是示出了共享访问控制列表表的多个处理决策路径的框图。

[0009] 图4是示出了在网络设备中使用多个处理决策路径执行的操作的流程图。

具体实施方式

[0010] 概览

[0011] 本文提出了针对网络设备中的分组实现处理决策的超低延迟确定的技术。分组在网络设备的端口处被接收。在第一处理决策路径中基于分组的内容和一个或多个网络策略来确定处理决策。在与第一处理路径并行的第二处理决策路径中,通过访问存储了处理决策的表来确定处理决策。针对与被包含在表中的一个或多个特定分组流参数相匹配的分组,第二处理决策路径能够比第一处理决策路径更快地输出处理决策。如果第二处理决策路径做出了处理决策,则使用由第二处理决策路径确定的处理决策,否则使用由第一处理决策路径确定的处理决策。

[0012] 示例实施例

[0013] 在计算机网络中,数据以分组的形式从源被传送到目的地,分组通常穿过一个或多个网络设备(例如,交换机、路由器、防火墙等)。在传送期间,网络设备可能执行了一个或多个操作,这将延迟引入到分组传送过程中。

[0014] 参照图1。图1示出了网络设备10的框图并且主要地示出了与本文提出的超低延迟处理决策技术相关的网络设备10的那些组件。网络设备10 包括多个端口20(1)-20(N),在这些端口处分组从网络到达网络设备,并且分组从这些端口离开网络设备到网络。端口20(1)-20(N)中的任何端口都可以作为进入端口或外出端口。单个分组在图1中以标号22示出。

[0015] 网络设备10还包括第一处理决策路径30、第二处理决策路径40和决策决定逻辑单元50。第一处理决策路径30例如是交换信息库(SIB),包括多个处理单元32(1)-32(8),如下文中所进一步描述的,这些处理单元基于分组的内容和一个或多个网络策略来顺序地执行决策操作。如果分组具有与被存储在第二处理决策路径40的表中的一个或多个特定流参数相匹配的流参数,则第二处理决策路径40可以比第一处理决策路径30更快地产生针对分组的处理决策。也就是说,第二处理决策路径40主要地包含表(如下文中关于图2所进一步描述的)。第二处理决策路径40可能不总是能够针对分组产生处理决策,并且事实上仅当分组具有与被存储在第二处理决策路径40的表中的一个或多个特定流参数相匹配的流参数时,第二处理决策路径40才对分组产生处理决策。

[0016] 在传统的网络设备中,仅SIB或等价的功能组件可用于基于到达分组的内容和网络策略来做出分组处理决策。SIB可以处理诸如开放最短路径优先(OSPF)之类的协议控制分组和边界网关协议(BGP)分组。一旦这些协议集中于一定决策,则对未来到达的匹配分组采取交换动作。通过这些交换信息库组件,间接实现了扩展(scale)。例如,对交换决策的匹配被置于交换信息库中,并且随后的动作能在结果数据库中找到。多个匹配可以指向相同的结果数据库以采取相同的动作。该方法引起了延迟,但是实现了可扩展性。本文提出的是通过将可选择的处理决策编程到第二处理决策路径40中来实现超低延迟的技术,第二处理决策路径40与第一处理决策路径30并行操作。第二处理决策路径40可以被称为可配置的交换机单元。

[0017] 现在描述第一处理决策路径的处理单元32(1)-32(8)。如本领域中所已知的,跨网络被传输的分组包括报头部分和有效载荷。报头通常包括关于分组的源和目的地的信息、以及在第2层(L2)、第3层(L3)和第4层 (L4)处以及在深度分组检测(DPI)字段中的其它信息。因此,在任何给定的分组中,在第2层字段、第3层字段、第4层字段和深度分组检测字段中具有分组流参数信息,该信息可用于确定针对分组做出怎样的处理决策。因此,第一处理决策路径30包括顺序地检查分组的报头中的所有这些字段以便针对分组做出处理决策的逻辑。具有L2聚集字段单元 32(1),其聚集了用于做出L2处理决策的所有的L2字段。L2决策单元 32(2)基于L2字段来做出L2处理决策。具有L3聚集字段单元32(3),其聚集了所有的L3字段,并且L3决策单元32(4)基于L3字段来做出L3处理决策。类似地,具有聚集L4字段的L4聚集字段单元32(5)和基于L4字段来做出L4处理决策的L4决策单元32(6)。最后,具有聚集DPI字段的 DPI聚集字段单元32(7)和基于DPI字段来做出DPI处理决策的DPI决定单元32(8)。

[0018] 被供应到第一处理决策路径30的关于分组的分组流信息(例如,第2 层字段、第3层字段等)也被并行供应到第二处理决策路径40。然而,如图1所示,由于所有的相关字段被聚集和处理,因此使用第一处理路径30 对分组做出处理决策所需的时间量是相当多的。第一处理决策路径30通常能够针对被期待通过网络设备处理的任何分组做出处理决策。相对

地，第二处理决策路径40使用了存储适用于网络设备接收的某些分组(例如，那些被预料为相对更有可能重复发生的分组)的处理决策的表。针对与被存储在第二处理决策路径40的表中的一个或多个特定分组流参数相匹配的分组，第二处理决策路径40能够比第一处理决策路径30更快地输出处理决策，决策决定逻辑单元50被配置为在第二处理决策路径40能够做出处理决策的情况下选择由第二处理决策路径40做出(输出)的处理决策以供使用，否则使用由第一处理决策路径30做出的处理决策。因此，在一些情况中，第二处理决策路径40在它的表中可能不具有允许它针对分组做出处理决策的条目。在这样的情况中，决策决定逻辑单元50 简单地使用由第一处理决策路径30输出的处理决策。然而，在其它情况中，第二处理决策路径40可以具有与接收的分组的一个或多个流参数相匹配的条目并且可以在第一处理决策路径30之前非常迅速地输出针对分组的处理决策。在这种情况中，决策决定逻辑单元50将使用来自第二处理决策路径40的处理决策。由第一处理决策路径做出的处理决策以及第二处理决策路径做出的处理决策可以包括下列各项中的至少一个：转发、交换(桥接)、旁路、丢弃等。

[0019] 第一处理决策路径30的处理决策输出34被耦合到决策决定逻辑单元 50并且第二处理决策路径40的处理决策输出42也被耦合到决策决定逻辑单元50。另外，如标号36处所示，第一处理决策路径30的处理决策输出 34被反馈到第二处理决策路径40以用处理决策输出34(即，SIB决策) 填充第二处理决策路径40的表，从而使得第二处理决策路径40能够做出快速的处理决策以供处理随后接收的具有得出该特定处理决策的流参数的分组。因此，由第一处理决策路径40(即，SIB)实现的学习被用于填充第二处理决策路径40的表。

[0020] 仍然参照图1，在一个示例中，第一处理决策路径30、第二处理决策路径40和决策决定逻辑单元50都被实施在具有数字逻辑门的硬件中。网络设备10还包括中央处理单元(CPU)60和存储器70。CPU 60可以通过执行被存储或被编码在存储器70中的软件指令来控制网络设备10的操作。例如，如62处所示，CPU 60可以直接将第二处理决策路径40编程以用一个或多个条目填充第二处理决策路径40的表，从而使得第二处理决策路径40可以对其他类型的分组迅速地做出处理决策。

[0021] 可能存在当第二处理决策路径40做出了处理决策，但是希望覆盖 (override)由第二处理决策路径40做出的处理决策的情况，而是替代使用由第一处理决策路径做出的处理决策。相反，可能存在希望覆盖由第一处理路径30确定的处理决策并且使用由第二处理决策路径40确定的处理决策的情况。为此，在62处CPU 60被耦合到决策决定逻辑单元50以使决策决定逻辑50覆盖由第二处理决策路径40做出的决策并且使用由第一处理决策路径30做出的处理决策，反之亦然。

[0022] 存储器70可以包括只读存储器(ROM)、随机存取存储器 (RAM)、磁盘存储介质设备、光学存储介质设备、闪速存储器设备、电的、光学的、或其它物理的/有形的存储器存储设备。CPU 60例如是微处理器或微控制器。因此，通常，存储器70可以包括用包括计算机可执行指令的软件编码的一个或多个有形的(非暂态)计算机可读存储介质(例如，存储器设备)，并且当软件被执行(通过CPU 60)时，软件可操作执行本文描述的操作。

[0023] 现在参照图2。图2示出了第二处理决策路径40的一个示例实施方式，其包括可用于做出分组处理决策的表匹配逻辑44和表41。表41包括多个条目43(1)-43(P)。具有与每个条目相关联的多个字段，例如源地址 (SA)字段46(A)、目的地地址(DA)字段46(B)、尺寸字

段46(C)和用于各种其它分组流参数的字段。在一个示例中，这些字段可以包含分组的第2层流参数，从而使得SA字段46(A)是第2层SA，即介质访问控制 (MAC)SA，DA字段46(B)是MAC DA等。字段47(D)包含针对具有与给定条目相匹配的流参数的分组的处理决策。

[0024] 表匹配逻辑44包括数字比较逻辑，数字比较逻辑将待处理的分组的参数与表41中的相应的字段进行比较以确定是否存在匹配。例如，如果具有SA为“11”、DA为“83”和尺寸(例如，小于)128比特的分组到达，则匹配被宣布并且处理决策“桥接”立即被输出。图2中示出的其它示例遵从相似的逻辑，并且应该理解的是表41的字段中使用的数字仅仅用于例示的目的并且不反映这样的参数的真实值。可以是：SA和DA的匹配可以产生处理决策的输出，而与分组的尺寸无关。

[0025] 表匹配逻辑44也用(从第一处理决策路径30或从CPU 60接收的) 条目填充表41，并且从表41移除在预定时间周期(例如，最后一小时、最后一天等)没有导致任何匹配的过时的条目。表匹配逻辑44可以被实施在软件中，或可以被实施为软件和硬件的组合。

[0026] 第二处理决策路径40可以涉及表41中的单一查找，并且如上所述，涉及使用由分组第2层字段、第3层字段、第4层字段、深度分组检测字段以及用户定义的参数中的一个或多个的任意组合定义的关键词进行的单一表查找，以确定与被存储在表中的一个或多个特定分组流参数的匹配。关键词可以是针对第2层字段、第3层字段、第4层字段、深度分组检测字段以及用户定义的参数中一个或多个的特定组合的静态手动配置的关键词，以确定与被存储在表中的一个或多个特定分组流参数的匹配。在第二处理决策路径40做出了处理决策的情况中，单一表查找的使用大大地缩短了在第二处理决策路径40中获得处理决策所需的时间量。然而，表41 存储了针对具有被预期通常由网络设备接收的分组的流参数的分组或具有应该以超低延迟方式处理的流参数的分组的处理决策，并且针对这些分组，处理决策应该由第二处理决策路径做出。

[0027] 第二处理决策路径40的表41中的字段可以由用户(经由CPU)通过协议手动地编程或可以从来自优化到第一处理决策路径30的单一条目的 SIB决策导出。另外，有可选的掩蔽47(例如，被实施为三元内容可寻址存储器)，以忽略通过逐位比较而不相关的字段，以允许对被存储在表41 中的字段的数据的部分的或准确的匹配。掩蔽47优化第二处理决策路径关键词，以匹配不止一种形式的分组。为了避免分组复制，除非明确地被标记为失败，否则第二处理决策路径的表中的匹配通常成功。在此结构中单播和多播分组的处理是相同的。

[0028] 在一些实施方式中，如果第二处理决策路径是满负载的，较慢的第一处理决策路径30可以被用于第2层硬件学习、L2未知单播洪泛、延迟不敏感流量和溢出。在学习模式的一个实施方式中，针对通过第一处理决策路径30的每个分组类型，如何处理分组类型的信息被获得，并且该信息可以被提供给第二处理决策路径40，用于在第二处理决策路径40中创建新的表条目。与指定分组类型所需的最小分组信息量相对应的“关键词”则被用于针对待使用如图1中示出的并行路径结构处理的未来分组来通过第二处理决策路径进行表查找。因此，在学习模式中，可以使用第一处理决策路径30基于分组类型来获得一组唯一的标识符，并且该信息可以提供将被用于在第二处理决策路径40中创建表条目的一组“关键词”。

[0029] L2硬件学习使得CPU减少了大量CPU存取和处理负荷。由于每个分组需要被存储在

存储器中并且被软件处理，因此软件学习通常比介质访问控制(MAC)表学习请求效率低。当存储器是满的(例如，当第二处理决策路径40中的表达到最大可接受尺寸)时，进一步的学习请求可以被忽略。

[0030] 现在转向图3，该示图示出了图1中所示的配置的变型。在该配置中，第一处理决策路径30和第二处理决策路径40共享访问控制列表 (ACL)表80用于做出处理决策。因此，被第二处理决策路径40使用的表41是ACL表80，ACL表80也被第一处理决策路径30使用。

[0031] 现在参照图4，图4是针对概述了被配置为本文提出的用于做出超低延迟处理决策的网络设备的操作的流程图100的描述。在110处，分组在网络设备的端口处被接收。通常，分组被存储在缓冲器中并且被排队用于处理。在120处，使用从分组获得的信息，在第一处理决策路径中基于分组的内容和一个或多个网络策略来针对分组确定处理决策。在130处，与由第一处理决策路径执行的操作并行，在第二处理决策中通过访问存储处理决策的表来确定处理决策，并且其中，针对与被包含在表中的一个或多个特定分组流参数相匹配的分组，第二处理决策路径能够比第一处理决策路径更快地输出处理决策。在140处，如果第二处理决策路径能够做出处理决策，则使用由第二处理决策路径确定的处理决策，否则使用由第一处理决策路径确定的处理决策。如上面关于图1所解释的，当一个处理决策路径覆盖另一处理决策路径时，可以执行覆盖，第一处理决策路径覆盖第二处理决策路径，或第二处理决策路径覆盖第一处理决策路径。

[0032] 总之，当存在表匹配时，第二处理决策路径的单一表查找功能用于以超低延迟处理(例如，交换)分组。虽然由于所需的SIB条目的每个组合必须被枚举(状态激增)，表是不可扩展的，但是使用表进行快速处理决策具有极低延迟，需要最少的一次表访问。

[0033] 描述了主题的特定的实施方式。其它实施方式在下述权利要求的范围内。在一些情况中，权利要求中列举的动作可以按不同的顺序被执行并且仍然实现期望的结果。另外，附图中示出的过程不一定要求所示出的特定顺序、或相继顺序来实现期望的结果。在某些实施方式中，多任务和并行处理可能是有优势的。

[0034] 上述描述仅作为示例。

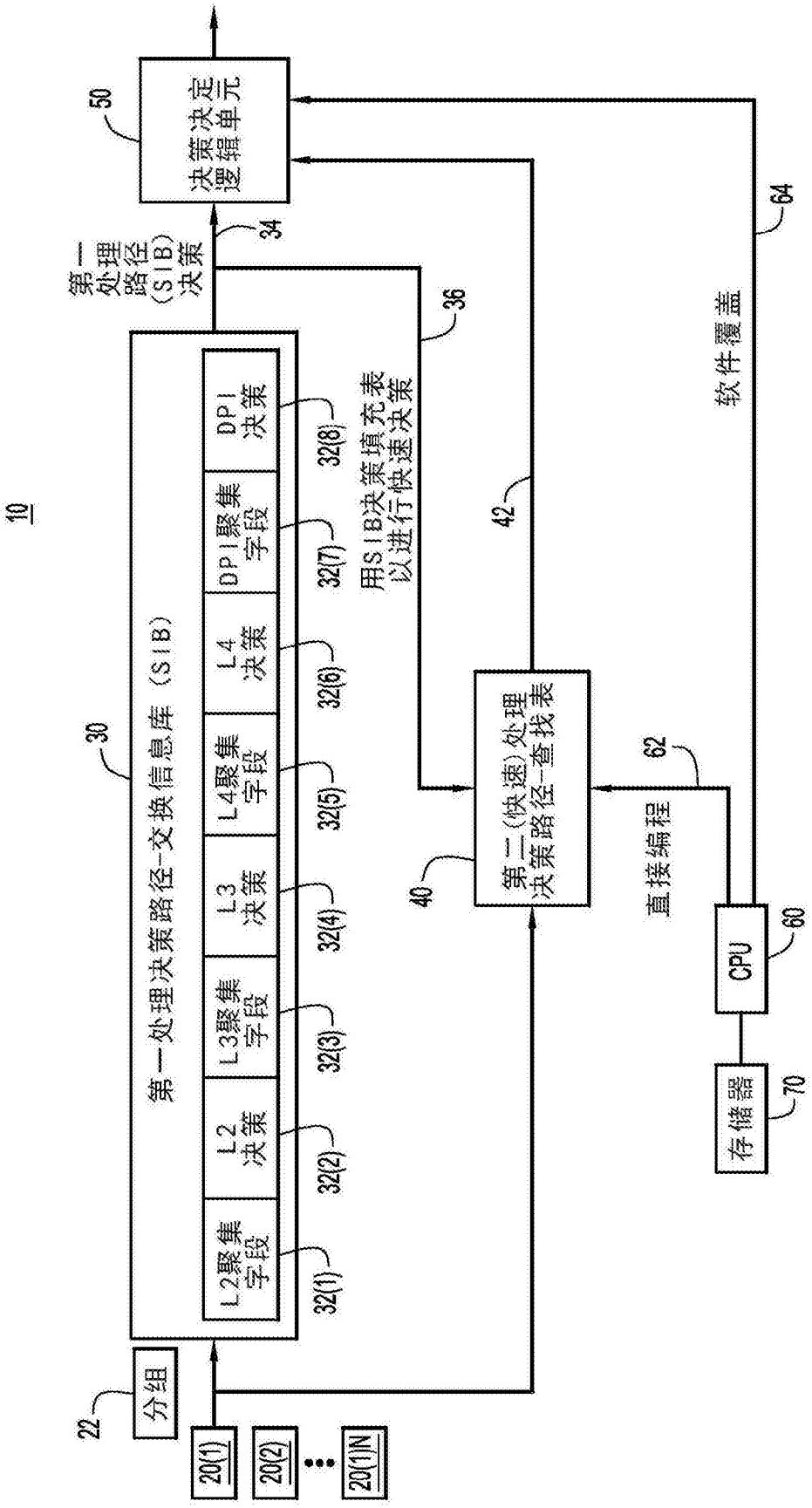

10
50
决策决定
逻辑单元
34
第一
处理
路径
(SIB)
决策
36
用SIB决策填充表
以进行快速决策
64
软件覆盖
42
30
第一处理决策路径-交换信息库(SIB)
DPI
决策
32(8)
DPI聚集
字段
32(7)
L4
决策
32(6)
L4聚集
字段
32(5)
L3
决策
32(4)
L3聚集
字段
32(3)
L2
决策
32(2)
L2聚集
字段
32(1)
40
第二(快速)处理
决策路径-查找表
62
直接编程
CPU
60
存储器
70
22
分组
20(1)
20(2)
20(1)N

图1

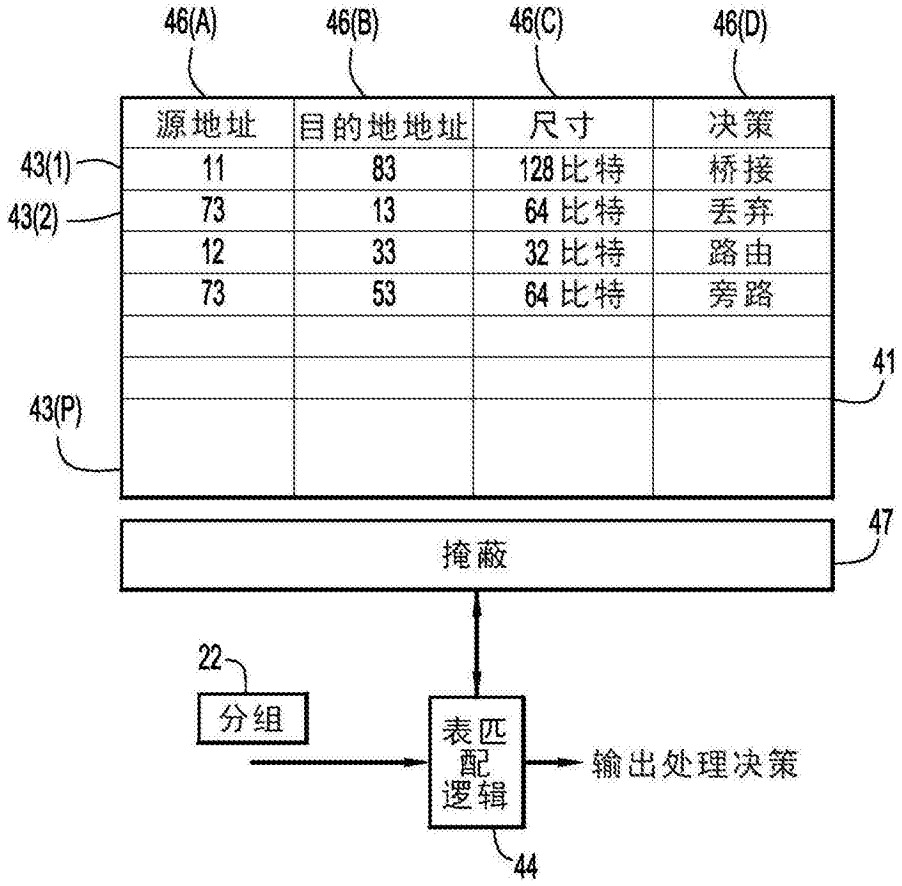

46(A)
46(B)
46(C)
46(D)
源地址
目的地地址
尺寸
决策
43(1)
11
83
128 比特
桥接
43(2)
73
13
64 比特
丢弃
12
33
32 比特
路由
73
53
64 比特
旁路
41
43(P)
掩蔽
47
22
分组
表匹配逻辑
输出处理决策
44

图2

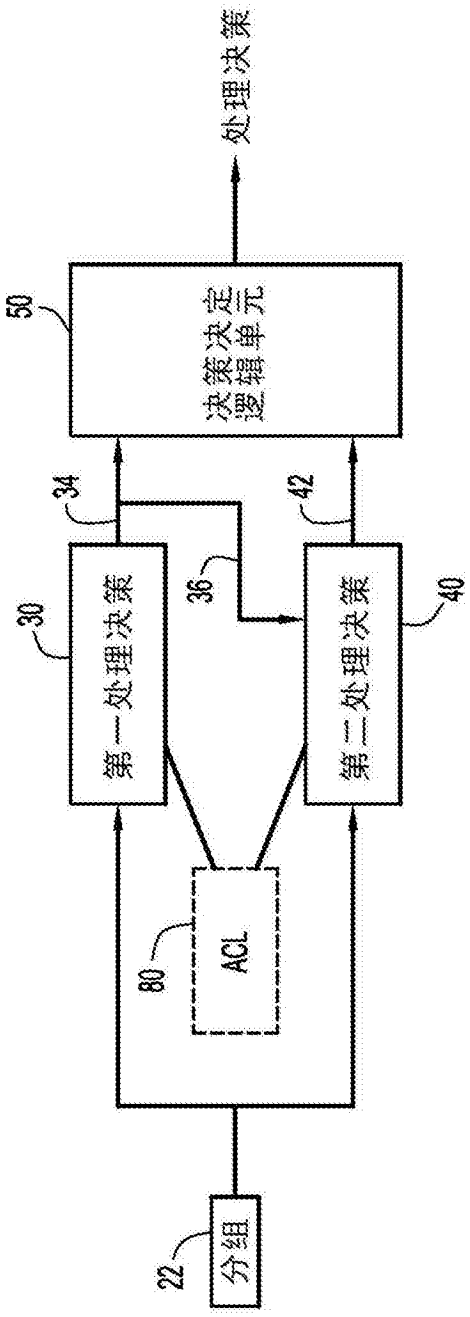


图3

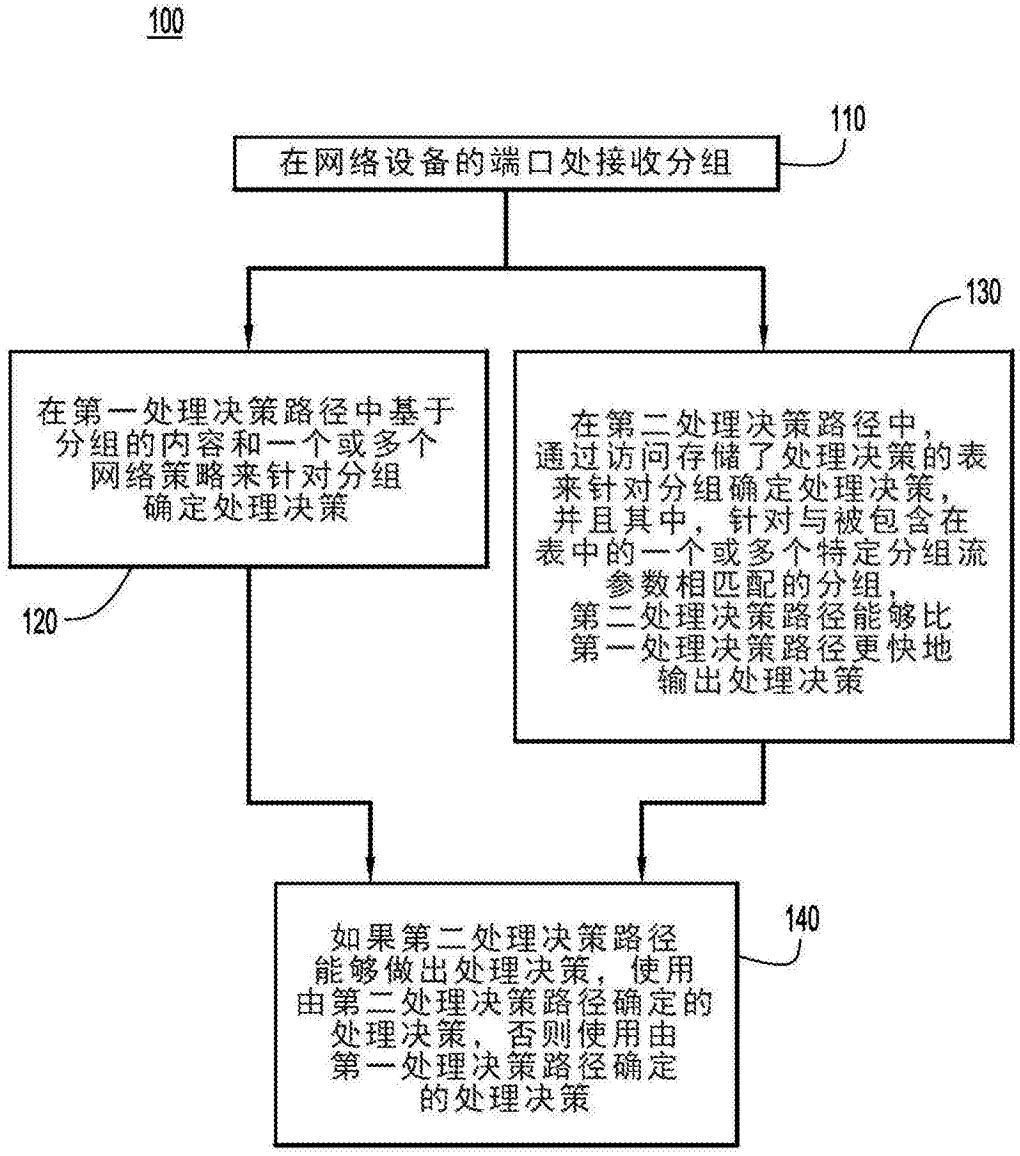

100
110
在网络设备的端口处接收分组
130
在第一处理决策路径中基于分组的内容和一个或多个网络策略来针对分组确定处理决策
120
在第二处理决策路径中，通过访问存储了处理决策的表来针对分组确定处理决策，并且其中，针对与被包含在表中的一个或多个特定分组流参数相匹配的分组，第二处理决策路径能够比第一处理决策路径更快地输出处理决策
140
如果第二处理决策路径能够做出处理决策，使用由第二处理决策路径确定的处理决策，否则使用由第一处理决策路径确定的处理决策

图4