

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2017年5月4日 (04.05.2017)



(10) 国际公布号
WO 2017/071667 A1

- (51) 国际专利分类号:
H04L 12/931 (2013.01)
- (21) 国际申请号: PCT/CN2016/103943
- (22) 国际申请日: 2016年10月31日 (31.10.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510721081.9 2015年10月30日 (30.10.2015) CN
- (71) 申请人: 新华三技术有限公司 (NEW H3C TECHNOLOGIES CO., LTD.) [CN/CN]; 中国浙江省杭州市滨江区长河路466号, Zhejiang 310052 (CN)。
- (72) 发明人: 赵志宇 (ZHAO, Zhiyu); 中国北京市海淀区上地信息产业基地创业路2号东方电子大厦730室, Beijing 100085 (CN)。 慕长林 (MU, Changlin); 中国北京市海淀区上地信息产业基地创业路2号东方电子大厦730室, Beijing 100085 (CN)。 左彦峰 (ZUO, Yanfeng); 中国北京市海淀区上地信息产业基地创业路2号东方电子大厦730室, Beijing 100085 (CN)。

- (74) 代理人: 北京德琦知识产权代理有限公司 (DEQI INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区知春路1号学院国际大厦7层, Beijing 100083 (CN)。
- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。
- (84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

[见续页]

(54) Title: PACKET FORWARDING

(54) 发明名称: 报文转发

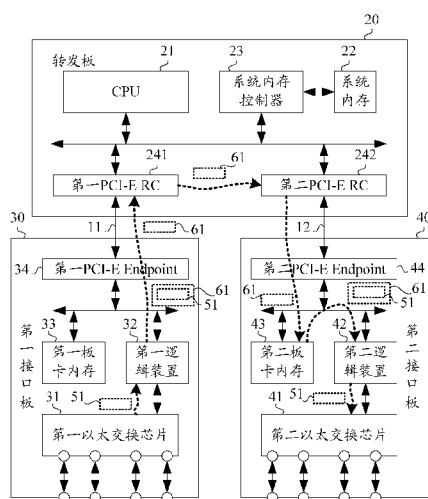


图 3a

- | | |
|----------------------------------|-----------------------------------|
| 20 Forwarding board | 33 First board card memory |
| 22 System memory | 34 First PCI-E endpoint |
| 23 System memory controller | 40 Second interface board |
| 241 First PCI-E RC | 41 Second Ethernet switching chip |
| 242 Second PCI-E RC | 42 Second logic device |
| 30 First interface board | 43 Second board card memory |
| 31 First Ethernet switching chip | 44 Second PCI-E endpoint |
| 32 First logic device | |

(57) Abstract: When a first interface board receives from an Ethernet switching chip a first Ethernet data packet, once the first Ethernet data packet is determined as taking a second interface board as a target board card, a logic device of the first interface board encapsulates the first Ethernet data packet as a first PCI-E packet that takes the PCI-E memory space address of the board card memory in the second interface board as a target address, so as to enable a PCI-E endpoint to forward the first PCI-E packet to a network device; when the first interface board acquires from board card memory a second PCI-E packet coming from a third interface board, the logic device of the first interface board parses out a second Ethernet data packet from the second PCI-E packet and delivers it to the Ethernet switching chip, the third interface board and the second interface board being same or not same.

(57) 摘要: 当第一接口板从以太交换芯片接收到第一以太网数据报文时, 当确定该第一以太网数据报文以第二接口板为目的的板卡, 则该第一接口板的逻辑装置将该第一以太网数据报文封装为以所述第二接口板中的板卡内存的 PCI-E 内存空间地址为目的地址的第一 PCI-E 报文, 以使得 PCI-E 端点将该第一 PCI-E 报文向网络设备转发板转发; 当第一接口板从板卡内存中获取来自第三接口板的第二 PCI-E 报文时, 该第一接口板的逻辑装置从该第二 PCI-E 报文中解析出第二以太网数据报文送至以太交换芯片, 其中所述第三接口板与第二接口板相同, 或者不同。



本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

报文转发

背景技术

网络设备中可以包括转发板和至少两块接口板，转发板是负责网络报文转发处理的单板。接口板是一个出物理网络接口，用于负责网络报文接收和发送。并且，
5 转发板和每块接口板之间可以利用 PCI-E（Peripheral Component Interconnect Express，高速外设部件互连标准）总线实现互连。

当选用 PCI-E 总线实现转发板和接口板之间的互连时，要考虑提高不同接口板之间的转发性能、并降低转发板的 CPU 占用率。

附图简要说明

- 10 图 1 为一个例子中的网络设备的架构图；
图 2 为如图 1 所示的网络设备中的 PCI-E 内存空间的映射图；
图 3a 和图 3b 为基于如图 2 所示的 PCI-E 内存空间映射的报文转发原理的示意图；
图 4 为如图 1 所示的网络设备中的 PCI-E 内存空间的扩展映射图；
15 图 5a 和图 5b 为基于如图 4 所示的 PCI-E 内存空间映射的报文转发原理的示意图；
图 6a 和图 6b 为如图 1 所示的网络设备中的表项配置的示意图；
图 7 为如图 1 所示的网络设备的扩展架构示意图；
图 8 为如图 7 所示的网络设备中的 PCI-E 内存空间的映射图；
20 图 9 为基于如图 8 所示的 PCI-E 内存空间映射的报文转发原理的示意图；
图 10 为如图 1 或图 7 所示的网络设备中的第一逻辑装置的逻辑结构图；
图 11 为如图 10 所示的第一逻辑装置的工作原理的示意图；
图 12 为如图 1 所示的网络设备中的 PCI-E 内存空间基于分片机制的映射图；
图 13a 和图 13b 为基于如图 12 所示的 PCI-E 内存空间映射的报文转发原理的示意图；
25 图 14 为如图 13a 和图 13b 所示的网络设备中的第一逻辑装置的逻辑结构图；

图 15a 和图 15b 为如图 14 所示的第一逻辑装置的工作原理的示意图；

图 16 为一个例子中的报文转发方法的流程图。

实施方式

请参见图 1，在一个例子中，网络设备 10 包括转发板 20、第一接口板 30 以及
5 第二接口板 40。

转发板 20 包括 CPU 21、系统内存 22、系统内存控制器 23、以及第一 PCI-E RC
(Root Complex, 根复合体) 241 和第二 PCI-E RC 242。其中，CPU 21、系统内存
控制器 23、以及第一 PCI-E RC 241 和第二 PCI-E RC 242 彼此互连，并且，系统内
存 22 与系统内存控制器 23 相互连接。

10 在该例子中，转发板的 CPU 21、系统内存控制器 23、以及第一 PCI-E RC 241
和第二 PCI-E RC 242 可以集成为 SOC (System-on-a-Chip, 片上系统) CPU；或者，
转发板的系统内存控制器 23 可以集成于 CPU 21 中，并且，转发板的第一 PCI-E
RC 241 和第二 PCI-E RC 242 可以集成为独立于 CPU 21 的 PCH (Platform Controller
Hub, 平台控制器)。

15 并且，图 1 中示出的第一 PCI-E RC 241 和第二 PCI-E RC 242 相互独立，但这并
不意味着第一 PCI-E RC 241 和第二 PCI-E RC 242 必然为两个相互独立的实体，即，
图 1 中示出的第一 PCI-E RC 241 和第二 PCI-E RC 242 可以是由一个具有 PCI-E RC
功能的实体通过外接 PCI-E Switch (交换器) 在逻辑上形成的两个等效 PCI-E RC。
并且，当第一 PCI-E RC 241 和第二 PCI-E RC 242 为两个相互独立的实体时，其中的
20 每一个也可以外接 PCI-E Switch。

第一接口板 30 包括第一以太交换芯片 31、第一逻辑装置 32、第一板卡内存 33、
以及通过第一 PCI-E 总线 11 连接第一 PCI-E RC 241 的第一 PCI-E Endpoint 34。其
中，第一以太交换芯片 31 和第一逻辑装置 32 相互连接 (例如通过以太总线)，并且，
第一逻辑装置 32、第一板卡内存 33、以及第一 PCI-E Endpoint 34 可以彼此互连。

25 第二接口板 40 包括第二以太交换芯片 41、第二逻辑装置 42、第二板卡内存 43、
以及通过第二 PCI-E 总线 12 连接第二 PCI-E RC 242 的第二 PCI-E Endpoint 44。其
中，第二以太交换芯片 41 和第二逻辑装置 42 可以相互连接 (例如通过以太总线)，
并且，第二逻辑装置 42、第二板卡内存 43、以及第二 PCI-E Endpoint 44 可以彼此
互连。

请参见图 2，在该网络设备 10 中，第一板卡内存 33 和第二板卡内存 43 均映射在网络设备 10 的 PCI-E 内存空间 60 中，并且，第一板卡内存 33 在 PCI-E 内存空间 60 中映射的地址区间 60b 不同于第二板卡内存 43 在 PCI-E 内存空间 60 中映射的地址区间 60a。从而，第一板卡内存 33 和第二板卡内存 43 对第一 PCI-E RC 241 和第二 PCI-E RC 242、以及第一 PCI-E Endpoint 34 和第二 PCI-E Endpoint 44 均可见。

请参见图 3a 并结合图 2，对于从第一接口板 30 向第二接口板 40 转发的情况：

当第一接口板 30 中的第一逻辑装置 32 从第一以太交换芯片 31 接收以太数据报文 51 时，第一逻辑装置 32 可以确定该以太数据报文 51 的目的板卡。其中，该以太数据报文 51 可以是由第一以太交换芯片 31 从网络设备 10 的外部接收，第一逻辑装置 32 可以根据维护有以太 MAC 地址与板卡标识的映射表确定以太数据报文 51 的目的板卡。

在一个例子中，无论以太数据报文 51 的目的板卡是第二接口板 40 还是转发板 20，该以太数据报文 51 都将以 PCI-E 格式转发至转发板 20 中的系统内存 22，并由 CPU 21 从系统内存 22 中读取后处理。

在本例中，当确定从第一以太交换芯片 31 接收到的以太数据报文 51 是以第二接口板 40 为目的板卡时，第一逻辑装置 32 可以将该以太数据报文 51 封装为以第二板卡内存 43 的 PCI-E 内存空间地址（即，图 2 中示出的地址区间 60a 内的地址）为目的地址的 PCI-E 报文（本文中的“PCI-E 报文”可以是指 PCI-E 写报文的简称）61，以供第一 PCI-E Endpoint 34 将该 PCI-E 报文 61 向转发板 20 转发、并使转发板 20 根据该 PCI-E 报文 61 的目的地址将该 PCI-E 报文 61 向作为目的板卡的第二接口板 40 的第二板卡内存 43 转发。例如，第一 PCI-E Endpoint 34 将该 PCI-E 报文 61 发送至第一 PCI-E RC 241，并由第一 PCI-E RC 241 根据该 PCI-E 报文 61 的目的地址将该 PCI-E 报文 61 转发给第二 PCI-E RC 242；封装有以太数据报文 51 的 PCI-E 报文 61 被第二 PCI-E RC 242 转发至第二 PCI-E Endpoint 44，并被第二 PCI-E Endpoint 44 写入第二板卡内存 43。

相应地，第二逻辑装置 42 可以从第二板卡内存 43 中获取该 PCI-E 报文 61，并从该 PCI-E 报文 61 中解析出以太数据报文 51 发送至第二以太交换芯片 41，从而可以由第二以太交换芯片 41 发出至网络设备 10 的外部。

请参见图 3b 并结合图 2，与从第一接口板 30 向第二接口板 40 转发同理，对于从第二接口板 40 向第一接口板 30 转发的情况：

当第二接口板 40 中的第二逻辑装置 42 从第二以太交换芯片 41 接收以太数据报文 52 时，第二逻辑装置 42 可以确定该以太数据报文 52 的目的板卡。其中，该以太数据报文 52 可以是由第二以太交换芯片 41 从网络设备 10 的外部接收。

当确定从第二以太交换芯片 41 接收到的以太数据报文 52 是以第一接口板 30 为目的板卡时，第二逻辑装置 42 可以将该以太数据报文 52 封装为以第一板卡内存 33 的 PCI-E 内存空间地址（即，图 2 中示出的地址区间 60b 内的地址）为目的地址的 PCI-E 报文 62，以供第二 PCI-E Endpoint 44 将该 PCI-E 报文 62 向转发板 20 转发，并使转发板 20 根据该 PCI-E 报文 62 的目的地址将该 PCI-E 报文 62 向作为目的板卡的第一接口板 30 的第一板卡内存 33 转发。例如，第二 PCI-E Endpoint 44 将该 PCI-E 报文 62 发送至第二 PCI-E RC 242，并由第二 PCI-E RC 242 根据该 PCI-E 报文 62 的目的地址将该 PCI-E 报文 62 转发给第一 PCI-E RC 241；封装有以太数据报文 52 的 PCI-E 报文 62 被第一 PCI-E RC 241 转发至第一 PCI-E Endpoint 34，并被第一 PCI-E Endpoint 34 写入第一板卡内存 33。

相应地，第一逻辑装置 32 可以从第一板卡内存 33 中获取该 PCI-E 报文 62，并从该 PCI-E 报文 62 中解析出以太数据报文 52 转发至第一以太交换芯片 31，从而可以由第一以太交换芯片 31 发出至网络设备 10 的外部。

如上可见，以太数据报文 51 或 52 在第一接口板 30 与第二接口板 40 之间的转发可以无需转发板 20 的 CPU 21 参与，并且，以太数据报文 51 或 52 在第一接口板 30 与第二接口板 40 之间的转发性能可以得到一定程度的提升。

另外，请参见图 4，转发板 20 中的系统内存 22 映射在网络设备 10 的 PCI-E 内存空间 60 中，即，内存空间 60 对第一 PCI-E RC 241 和第二 PCI-E RC 242、以及第一 PCI-E Endpoint 34 和第二 PCI-E Endpoint 44 均可见。并且，系统内存 22 在 PCI-E 内存空间中映射的地址区间 60c 不同于第一板卡内存 33 在 PCI-E 内存空间 60 中映射的地址区间 60b、以及第二板卡内存 43 在 PCI-E 内存空间 60 中映射的地址区间 60a。其中，系统内存 22 的映射，可以是系统内存 22 的全部内存空间或部分内存空间映射在 PCI-E 内存空间 60 中。

请参见图 5a 并结合图 4，对于从第一接口板 30 向转发板 20 转发的情况：

当第一接口板 30 中的第一逻辑装置 32 从第一以太交换芯片 31 接收以太数据报文 53 时，第一逻辑装置 32 可以确定该以太数据报文 53 的目的板卡，例如，第一逻辑装置 32 可以根据以太 MAC 地址与板卡标识的映射表中由转发板 20 的 CPU 21 配

置的表项确定以太数据报文 53 的目的板卡。

当确定从第一以太交换芯片 31 接收到的以太数据报文 53 以转发板 20 为目的板卡时，第一逻辑装置 32 可以将该以太数据报文 53 封装为以系统内存 22 的 PCI-E 内存空间地址（即，图 4 中示出的地址区间 60c 内的地址）为目的地址的 PCI-E 报文 63，以供第一 PCI-E Endpoint 34 将该 PCI-E 报文 63 发送至第一 PCI-E RC 241、并可以由第一 PCI-E RC 241 根据该 PCI-E 报文 63 的目的地址将该 PCI-E 报文 63 写入至系统内存 22 中，从而，当 CPU 21 通过系统内存控制器 23 访问系统内存 22 时即可读取到该 PCI-E 报文 63、并从该 PCI-E 报文 63 中解析得到以太数据报文 53。

请参见图 5b 并结合图 4，与从第一接口板 30 向转发板 20 转发同理，对于从第二接口板 40 向转发板 20 转发的情况：

当第二接口板 40 中的第二逻辑装置 42 从第二以太交换芯片 41 接收以太数据报文 54 时，第二逻辑装置 42 可以确定该以太数据报文 54 的目的板卡。

当确定从第二以太交换芯片 41 接收到的以太数据报文 54 以转发板 20 为目的板卡时，第二逻辑装置 42 可以将该以太数据报文 54 封装为以系统内存 22 的 PCI-E 内存空间地址（即，图 4 中示出的地址区间 60c 内的地址）为目的地址的 PCI-E 报文 64，以供第二 PCI-E Endpoint 44 将该 PCI-E 报文 64 发送至第二 PCI-E RC 242、并可以由第二 PCI-E RC 242 根据该 PCI-E 报文 64 的目的地址将该 PCI-E 报文 64 写入至系统内存 22 中，从而，当 CPU 21 通过系统内存控制器 23 访问系统内存 22 时即可读取到该 PCI-E 报文 64、并从该 PCI-E 报文 64 中解析得到以太数据报文 54。

如上可见，这个例子中采用的第一接口板 30 与第二接口板 40 之间的转发，可以不影响第一接口板 30 和第二接口板 40 中的每一个与转发板 20 之间的转发。

请参见图 6a 和图 6b，第一逻辑装置 32 和第二逻辑装置 42 中可以分别维护以太 MAC 地址与板卡标识的映射表 71a 和 71b，用于根据以太数据报文 51 或 52 或 53 或 54 的目的 MAC 地址确定该以太数据报文 51 或 52 或 53 或 54 的目的板卡。

并且，仍参见图 6a 和图 6b，第一逻辑装置 32 和第二逻辑装置 42 中还可以维护板卡标识与 PCI-E 内存空间地址的映射表 72，用于根据以太数据报文 51 或 52 或 53 或 54 的目的板卡确定封装有该以太数据报文 51 或 52 或 53 或 54 的 PCI-E 报文 61 或 62 或 63 或 64 的目的地址。

其中，图 6a 中示出了以太 MAC 地址与板卡标识的映射表 71a 和 71b 以及板卡标识与 PCI-E 内存空间地址的映射表 72 在网络设备 10 初始化后的状态，如图 6a 所

示, 以太 MAC 地址与板卡标识的映射表 71a 和 71b 中对应转发板 20 的表项、以及板卡标识与 PCI-E 内存空间地址的映射表 72 中的所有表项均可以由 CPU 21 在网络设备 10 初始化时利用 PCI-E 报文 60 下发。图 6a 和图 6b 中示出的“MAC0”表示转发板 20 的 MAC 地址。

5 在该例子中, 用于下发表项的 PCI-E 报文 60 是由 CPU 21 利用第一 PCI-E 总线 11 和第二 PCI-E 总线 12 转发的, 即, CPU 21 将第一 PCI-E 总线 11 和第二 PCI-E 总线 12 复用为用于下发表项的管理总线。可以理解的是, 网络设备 10 中也可以设置独立于第一 PCI-E 总线 11 和第二 PCI-E 总线 12、并专用于下发表项的管理总线, 此时, 下发的表项可以携带在独立的管理总线的协议报文中。

10 另外, 图 6b 中示出的以太 MAC 地址与板卡标识的映射表 71a 中对应第二接口板 40 的表项则可以由第一逻辑装置 32 通过来自第二接口板 40 的 PCI-E 报文 62 中封装的以太数据报文 52 学习得到, 图 6b 中示出的“MAC2”表示以太数据报文 52 的源 MAC 地址。

同理, 图 6b 中示出的以太 MAC 地址与板卡标识的映射表 71b 中对应第一接口板 30 的表项可以由第二逻辑装置 42 通过来自第一接口板 30 的 PCI-E 报文 61 中封装的以太数据报文 51 学习得到, 图 6b 中示出的“MAC1”表示以太数据报文 51 的源 MAC 地址。

当第一板卡内存 33 中被写入由第二逻辑装置 42 封装的 PCI-E 报文 62 时, 第一逻辑装置 32 可以从第一板卡内存 33 中获取该 PCI-E 报文 62, 该 PCI-E 报文 62 中封装有第二逻辑装置 42 从第二以太交换芯片 41 接收到的以太数据报文 52、并携带有第二接口板 40 的板卡标识。并且, 若以太 MAC 地址与板卡标识的映射表 71a 中此时不存在对应第二接口板 40 的表项, 则第一逻辑装置 32 可以根据从该 PCI-E 报文 62 中解析出的以太数据报文 52 的源 MAC 地址和该 PCI-E 报文 62 中携带的板卡标识, 在以太 MAC 地址与板卡标识的映射表 71a 中创建对应第二接口板 40 的表项。

25 同理, 当第二板卡内存 43 中被写入由第一逻辑装置 32 封装的 PCI-E 报文 61 时, 第二逻辑装置 42 可以从第二板卡内存 43 中获取该 PCI-E 报文 61, 该 PCI-E 报文 61 中封装有第一逻辑装置 32 从第一以太交换芯片 31 接收到的以太数据报文 51、并携带有第一接口板 30 的板卡标识。并且, 若以太 MAC 地址与板卡标识的映射表 71b 中此时不存在对应第一接口板 30 的表项, 则第二逻辑装置 42 可以根据从该 PCI-E 报文 61 中解析出的以太数据报文 51 的源 MAC 地址和该 PCI-E 报文 61 中携带的板

卡标识,在以太 MAC 地址与板卡标识的映射表 71b 中创建对应第一接口板 30 的表项。

其中,第一逻辑装置 32 或者第二逻辑装置 42 有可能不能根据以太 MAC 地址与板卡标识的映射表查找到一个以太数据报文的目的板卡。

5 例如,以太数据报文 51 是广播报文,由于其目的 MAC 地址为广播地址,所以其不能够通过以太 MAC 地址与板卡标识的映射表唯一地确定以太数据报文 51 的目的板卡,导致目的板卡查询失败。此时,可以采用向除本板之外的其它所有接口板逐一转发 PCI-E 报文的方式。在如下示例中,仍继续使用标识“51”表示第一以太交换芯片 31 接收到的以太数据报文,但可以理解的是,如下示例中利用标识“51”
10 表示的以太数据报文可以不同于前文描述的以太数据报文 51。

 请参见图 7,网络设备 10'中可以进一步包括第三接口板 80,该第三接口板 80 包括第三 PCI-E Endpoint 84 和第三板卡内存 83。相应地,转发板 20'可以进一步包括第三 PCI-E RC 243,第三 PCI-E RC 243 通过第三 PCI-E 总线 13 连接第三 PCI-E Endpoint 84。其中,与第一 PCI-E RC 241 和第二 PCI-E RC 242 同理,第三 PCI-E RC
15 243 可以是具有 PCI-E RC 的独立实体,也可以是逻辑上形成的等效 PCI-E RC。

 如图 8 所示,第三板卡内存 83 可以映射在 PCI-E 内存空间 60 中,并且,第三板卡内存 83 在 PCI-E 内存空间 60 中映射的地址区间 60d 不同于系统内存 22、第一板卡内存 33、以及第二板卡内存 43 在 PCI-E 内存空间中分别映射的地址区间 60c、60b 以及 60a。其中,CPU 21 在初始化时下发的板卡标识与 PCI-E 内存空间地址的
20 映射表 72 中的所有表项进一步包括了对应第三接口板 80 的表项。并且,第三 PCI-E RC 243 和第三 PCI-E Endpoint 84 也可以映射在 PCI-E 内存空间 60 中。

 请参见图 9 并结合图 8:

 当从第一以太交换芯片 31 接收到以太数据报文 51 时,第一逻辑装置 32 通过匹配以太 MAC 地址与板卡标识的映射表 71a 中的表项来确定该以太数据报文 51 的目的
25 板卡;

 若在以太 MAC 地址与板卡标识的映射表 71a 中匹配表项失败,即,第一逻辑装置 32 确定该以太数据报文 51 不是以转发板 20 为目的板卡(例如,映射表 71a 中只有初始的 MAC0 转发板 20 的表项),此时以太数据报文 51 的目的 MAC 可以是广播地址,则第一逻辑装置 32 可以根据板卡标识与 PCI-E 内存空间地址的映射表 72,
30 将该以太数据报文 51 封装为分别以第二板卡内存 43 的 PCI-E 内存空间地址(即,

图 8 中示出的地址区间 60a 内的地址)和第三板卡内存 83 的 PCI-E 内存空间地址(即, 图 8 中示出的地址区间 60d 内的地址)为目的地址的多于一个 PCI-E 报文 61 和 61', 以供第一 PCI-E Endpoint 34 将 PCI-E 报文 61 和 61' 发送至第一 PCI-E RC 241、并使第一 PCI-E RC 241 根据 PCI-E 报文 61 和 61' 的目的地址将 PCI-E 报文 61 和 61' 分别
5 转发给第二 PCI-E RC 242 和第三 PCI-E RC 243。

从而, 封装有以太网数据报文 51 的 PCI-E 报文 61 可以到达第二接口板 40, 并使第二接口板 40 中的第二逻辑装置 42 在以太 MAC 地址与板卡标识的映射表 71b 中创建对应第一接口板 30 的表项。同理, PCI-E 报文 61' 可以到达第三接口板 80, 并且, 第三接口板 80 中可以具有与第一逻辑装置 32 或第二逻辑装置 42 类似的装置(图
10 7 和图 9 中均未示出), 该装置可以解析出以太网数据报文 51, 在以太 MAC 地址与板卡标识的映射表中创建对应第一接口板 30 的表项。

基于上述的原理, 请参见图 10, 第一逻辑装置 32 可以包括: 以太总线控制器 320、以太网报文接收处理模块 321、第一映射表维护模块 322、PCI-E 报文发送处理模块 323、第二映射表维护模块 324、PCI-E 报文接收处理模块 325、以太网报文发送
15 处理模块 326、以及 CPU 交换寄存器 327。

请参见图 11, 基于如图 10 所示的结构, 第一逻辑装置 32 的工作原理如下(由于以太网数据报文 51 和 53 在第一逻辑装置 32 中的处理方式基本相同、并且 PCI-E 报文 61 和 63 在第一逻辑装置 32 中的处理方式基本相同, 因此, 为了简化视图, 图 11 中省略了以太网数据报文 53 和 PCI-E 报文 63):

20 以太总线控制器 320 连接本板的第一以太交换芯片 31, 从而第一逻辑装置 32 与第一以太交换芯片 31 之间的以太网数据报文 51 和 52 可以交互。

以太网报文接收处理模块 321 从以太总线控制器 320 接收来自第一以太交换芯片 31 的以太网数据报文 51。

第一映射表维护模块 322 中维护有以太 MAC 地址与板卡标识的映射表 71a, 供
25 以太网报文接收处理模块 321 通过匹配以太 MAC 地址与板卡标识的映射表 71a 中的表项确定该以太网数据报文 51 的目的板卡。

以太网报文接收处理模块 321 在完成匹配后, 会将以太网数据报文 51 与匹配结果一起发送至 PCI-E 报文发送处理模块 323。其中, 此处所述的匹配结果可以为成功匹配到的目的板卡的板卡标识、或匹配失败, 图 11 中以匹配到第二接口板 40 的板卡
30 标识为例(如图 11 中的 S1111 所示)。

PCI-E 报文发送处理模块 323 对接收到的以太数据报文 51 进行封装。

第二映射表维护模块 324 中维护有板卡标识与 PCI-E 内存空间地址的映射表 72，供 PCI-E 报文发送处理模块 323 通过匹配板卡标识与 PCI-E 内存空间地址的映射表 72 中的表项确定封装形成的 PCI-E 报文 61 的目的地址。

5 PCI-E 报文发送处理模块 323 完成匹配后，会将以太数据报文 51 封装在 PCI-E 报文 61 中，并根据匹配结果确定 PCI-E 报文 61 的目的地址，然后，将 PCI-E 报文 61 提供给第一 PCI-E Endpoint 34。

其中，此处所述的匹配结果可以为已确定的一块目的板卡的板卡标识、或者为所有其它接口板的板卡标识（在确定目的板卡失败时），图 11 中以匹配到第二板卡
10 内存 43 的地址区间 60a 为例（如图 11 中的 S1112 和 S1113 所示）。

对于确定目的板卡失败的情况：第一映射表维护模块 322 中可以维护有对应网络设备 10 中除本板 30 之外的所有其它接口板（即第二接口板 40 和第三接口板 80）的表项，以太报文接收处理模块 321 可以向 PCI-E 报文发送处理模块 323 上报匹配
15 结果为网络设备 10 中除本板 30 之外的所有其它接口板，使 PCI-E 报文发送处理模块 323 可以逐一匹配到除本板 30 之外的所有其它接口板对应的表项；或者，以太报文接收处理模块 321 可以向 PCI-E 报文发送处理模块 323 上报匹配失败的通知，使 PCI-E 报文发送处理模块 323 可以轮询除本板 30 之外的所有其它接口板对应的表项。

另外，CPU 交换寄存器 327 中可以存放由 CPU 21 下发的本板的板卡标识，PCI-E
20 报文发送处理模块 323 可以将本板的板卡标识封装在 PCI-E 报文 61 中（如图 11 中的 S1114 所示），以供接收到该 PCI-E 报文 61 的其它接口板（包括但不限于第二接口板 40）能够利用本板的板卡标识和以太数据报文 51 的源 MAC 地址学习对应于本板的以太 MAC 地址与板卡标识的映射表项。

PCI-E 报文接收处理模块 325 可以从第一板卡内存 33 中读取以本板为目的板卡
25 的 PCI-E 报文 60 和 62。

PCI-E 报文接收处理模块 325 可以从 PCI-E 报文 60 中解析得到 CPU 21 在网络设备 10 初始化时下发的表项、并写入至第一映射表维护模块 322 和第二映射表维护模块 324 中（如图 11 中的 S1121 所示）；并且，PCI-E 报文接收处理模块 325 可以从 PCI-E 报文 60 中解析得到 CPU 21 在网络设备 10 初始化时下发的本板的板卡标
30 识、并写入至 CPU 交换寄存器 327 中（如图 11 中的 S1122 所示）。

PCI-E 报文接收处理模块 325 可以从 PCI-E 报文 62 中解析得到来自第二接口板 40 的以太网数据报文 52 以及第二接口板 40 的板卡标识, 并将解析得到的以太网数据报文 52 和板卡标识转发给以太网报文发送处理模块 326。

以太网报文发送处理模块 326 从接收到的以太网数据报文 52 中提取源 MAC 地址, 并根据提取的源 MAC 地址与接收到的第二接口板 40 的板卡标识在第一映射表维护模块 322 中创建对应第二接口板 40 的表项。

以太网报文发送处理模块 326 还将接收到的以太网数据报文 52 发送至以太网总线控制器 320。

第二逻辑装置 42 可以具有与第一逻辑装置 32 基本相同的结构。

请参见图 12, 在另一个例子中, 第二板卡内存 43 在 PCI-E 内存空间 60 中映射的地址区间 60a 中可以包括多个数据报文缓存片 60a_1~60a_m (m 为大于 1 的正整数) 以及至少一个控制报文缓存片 60a_ctr, 同理, 第一板卡内存 33 在 PCI-E 内存空间 60 中映射的地址区间 60b 中可以包括多个数据报文缓存片 60b_1~60b_n (n 为大于 1 的正整数) 以及至少一个控制报文缓存片 60b_ctr。

相应地, 第一逻辑装置 32 维护的板卡标识与 PCI-E 内存空间地址的映射表 72 中包括的对应第二接口板 40 的表项 91, 被划分为分别对应多个数据报文缓存片 60a_1~60a_m 的子表项, 并且每个子表项具有一个标志位 Flag, 用于表示该子表项对应的数据报文缓存片 60a_i (i 为大于等于 1 且小于等于 m 的正整数) 的状态是占用还是空闲。同理, 第二逻辑装置 42 维护的板卡标识与 PCI-E 内存空间地址的映射表 72 中包括的对应第一接口板 30 的表项 92 中, 也被划分为分别对应多个数据报文缓存片 60b_1~60b_n 的子表项, 并且每个子表项具有一个标志位 Flag, 用于表示该子表项对应的数据报文缓存片 60b_j (j 为大于等于 1 且小于等于 n 的正整数) 的状态是占用还是空闲。

另外, 控制报文缓存片 60a_ctr 和 60b_ctr 可以作为板卡标识与 PCI-E 内存空间地址的映射表 72 中的不具有标志位 Flag 的一个子表项, 或者, 可以独立于板卡标识与 PCI-E 内存空间地址的映射表 72 单独存放。

请参见图 13a 并结合图 12, 对于从第一接口板 30 向第二接口板 40 转发的情况:

当第一接口板 30 中的第一逻辑装置 32 从第一以太网交换芯片 31 接收以太网数据报文 51 时, 第一逻辑装置 32 可以确定该以太网数据报文 51 的目的板卡。

当确定从第一以太网交换芯片 31 接收到的以太网数据报文 51 是以第二接口板 40 为

目的板卡时,第一逻辑装置 32 从第二板卡内存 43 在 PCI-E 内存空间 60 中映射的多个数据报文缓存片 60a_1~60a_m 中选择一片空闲的数据报文缓存片 60a_i、并将其对应的子表项的标志位 Flag 置为占用状态,然后将该以太网数据报文 51 封装为以数据报文缓存片 60a_i 的 PCI-E 内存空间地址为目的地址的 PCI-E 报文 61,以供第一

5 PCI-E Endpoint 34 将该 PCI-E 报文 61 发送至第一 PCI-E RC 241、并可以由第一 PCI-E RC 241 根据该 PCI-E 报文 61 的目的地址将该 PCI-E 报文 61 转发给第二 PCI-E RC 242。

即,第一逻辑装置 32 为 PCI-E 报文 61 设定的目的地址为第二板卡内存 43 映射的一个空闲的数据报文缓存片 60a_i 的 PCI-E 内存空间地址。

10 相应地,当封装有以太网数据报文 51 的 PCI-E 报文 61 被第二 PCI-E RC 242 转发至第二 PCI-E Endpoint 44、并被第二 PCI-E Endpoint 44 写入第二板卡内存 43 中对应数据报文缓存片 60a_i 的地址区间时,第二逻辑装置 42 可以从第二板卡内存 43 中获取该 PCI-E 报文 61、并从该 PCI-E 报文 61 中解析出以太网数据报文 51 发送至第二以太网交换芯片 41,从而可以由第二以太网交换芯片 41 发出至网络设备 10 的外部。

15 并且,第二逻辑装置 42 还构造以控制报文缓存片 60b_ctr 为目的地址的 PCI-E 报文 65,以供第二 PCI-E Endpoint 44 将该 PCI-E 报文 65 发送至第二 PCI-E RC 242、并可以由第二 PCI-E RC 242 根据该 PCI-E 报文 65 的目的地址将该 PCI-E 报文 65 转发给第一 PCI-E RC 241。其中,PCI-E 报文 65 中还携带有表示释放数据报文缓存片 60a_i 的数据报文缓存片释放信息。

20 相应地,当封装有数据报文缓存片释放信息的 PCI-E 报文 65 被第一 PCI-E RC 241 转发至第一 PCI-E Endpoint 34、并被第一 PCI-E Endpoint 34 写入第一板卡内存 33 中对应控制报文缓存片 60b_ctr 的地址区间时,第一逻辑装置 32 可以从第一板卡内存 33 中获取该 PCI-E 报文 65、并从该 PCI-E 报文 65 中解析出数据报文缓存片释放信息,从而可以根据数据报文缓存片释放信息将数据报文缓存片 60a_i 对应的子

25 表项的标志位 Flag 置为空闲状态。

请参见图 13b 并结合图 12,与从第一接口板 30 向第二接口板 40 转发同理,第二逻辑装置 42 为 PCI-E 报文 62 设定的目的地址可以为第一板卡内存 33 映射的一个空闲的数据报文缓存片 60b_j 的 PCI-E 内存空间地址,并将数据报文缓存片 60b_j 对应的子表项的标志位 Flag 置为占用状态。相应地,第一逻辑装置 32 也可以构造

30 以控制报文缓存片 60a_ctr 为目的地址的 PCI-E 报文 66,以供第二逻辑装置 42 将数

据报文缓存片 60b_j 对应的子表项的标志位 Flag 置为空闲状态。

如上可见，通过对 PCI-E 内存地址空间的分片划分，可以对以太网数据报文在第一接口板 30 与第二接口板 40 之间的转发实施反压流控。

请参见图 14，当对 PCI-E 内存地址空间采用分片划分时，第一逻辑装置 32 中可以进一步包括释放执行模块 328 和释放通告模块 329。

请参见图 15a，PCI-E 报文发送处理模块 323 在将以太网数据报文 51 封装为以太网数据报文缓存片 60a_i 的 PCI-E 内存空间地址为目的地址 PCI-E 报文 61 时，可以将数据报文缓存片 60a_i 对应的子表项的标志位 Flag 置为占用状态(如图 15a 中的 S1411 所示)。并且，PCI-E 报文接收处理模块 325 在从第一板卡内存 33 中对应本板的控制报文缓存片 60b_ctr 的地址区间读取到 PCI-E 报文 65 时，可以从 PCI-E 报文中解析出数据报文缓存片释放信息提供给释放执行模块 328。从而，释放执行模块 328 可以根据 PCI-E 报文接收处理模块 325 提供的该数据报文缓存片释放信息将数据报文缓存片 60a_i 对应的子表项的标志位 Flag 置为空闲状态(如图 15a 中的 S1412 所示)。

请参见图 15b，PCI-E 报文接收处理模块 325 从第一板卡内存 33 中对应数据报文缓存片 60b_j 的地址区间读取到 PCI-E 报文 62 时，从该 PCI-E 报文 62 中解析出以太网数据报文 52，并提取该 PCI-E 报文 62 的目的地址和携带的第二接口板 40 的板卡标识提供给以太网报文发送处理模块 326；以太网报文发送处理模块 326 在将以太网数据报文 52 发出时，可以将 PCI-E 报文 62 的目的地址和第二接口板 40 的板卡标识提供给释放通告模块 329。从而，释放通告模块 329 根据 PCI-E 报文 62 的目的地址和第二接口板 40 的板卡标识，通知 PCI-E 报文发送处理模块 323 构造以对应第二接口板 40 的控制报文缓存片 60a_ctr 的 PCI-E 报文 66，并将 PCI-E 报文 62 的目的地址(即数据报文缓存片 60b_j 的 PCI-E 内存空间地址)作为数据报文缓存片释放信息携带在该 PCI-E 报文 66 中。

以上是对上述例子中的网络设备的详细说明。下述的例子中，还提供了用于网络设备中的报文转发方法。

请参见图 16，当该报文转发方法应用于如图 1 所示的网络设备 10 中的第一接口板 30 或第二接口板 40 时，该报文转发方法可以包括在第一逻辑装置 32 或第二逻辑装置 42 执行的如下步骤：

S1601，当从以太网交换芯片接收到来自网络设备的外部的以太网数据报文时，确定

该以太网数据报文的目的板卡；

S1602, 当成功确定从以太交换芯片接收到的以太网数据报文的目的板卡为网络设备的其它接口板时, 将该以太网数据报文封装为以该其它接口板的板卡内存的 PCI-E 内存空间地址为目的地址的 PCI-E 报文, 以供 PCI-E Endpoint 将该 PCI-E 报文向转发板转发、并使转发板根据该 PCI-E 报文的目的地址将该 PCI-E 报文向目的板卡的板卡内存转发;

S1603, 当从板卡内存获取来自其它接口板的 PCI-E 报文时, 从获取到的该 PCI-E 报文中解析出以太网数据报文转发至以太交换芯片。

另外, 当 S1601 确定从第一以太交换芯片接收到的以太网数据报文以转发板为目的的板卡时, 报文转发方法可以进一步包括: 将该以太网数据报文封装为以系统内存的 PCI-E 内存空间地址为目的地址的 PCI-E 报文, 以供 PCI-E Endpoint 将该 PCI-E 报文向转发板转发、并被写入转发板的系统内存中。

在该例子中, 该报文转发方法可以进一步维护板卡标识与 PCI-E 内存空间地址的映射表, 用于根据从以太交换芯片接收到的以太网数据报文的目的板卡, 确定封装有该以太网数据报文的 PCI-E 报文的目的地址; 并且, 根据转发板的 CPU 的配置, 在板卡标识与 PCI-E 内存空间地址的映射表中创建表项。

在该例子中, 报文转发方法可以进一步维护以太 MAC 地址与板卡标识的映射表, 并且, 报文转发方法可以从来自其它接口板的 PCI-E 报文中解析出其它接口板从网络设备的外部接收到的以太网数据报文、以及该其它接口板的板卡标识, 并且根据解析出的以太网数据报文的源 MAC 地址和该 PCI-E 报文中携带的板卡标识, 在以太 MAC 地址与板卡标识的映射表创建对应该其它接口板的表项; 以及, 报文转发方法可以根据 CPU 在网络设备初始化时的配置, 在以太 MAC 地址与板卡标识的映射表中对应转发板的表项。此时, 上述的 S1601 可以根据以太网数据报文的目的 MAC 地址确定该以太网数据报文的目的板卡。

在该例子中, 该报文转发方法还可以进一步在确定从以太交换芯片接收到的以太网数据报文的目的板卡失败时, 将该以太网数据报文封装为分别以网络设备的每一块其它接口板的卡内存的 PCI-E 内存空间地址为目的地址的多于一个 PCI-E 报文, 以供 PCI-E Endpoint 将多于一个 PCI-E 报文向转发板转发、并使转发板根据多于一个 PCI-E 报文的目的地址将多于一个 PCI-E 报文分别转发给对应的其它接口板。

另外, 该例子中的报文转发方法可以按照与如图 12 原理相同的方式采用反压控

制的流控机制，相应地：

各接口板的板卡内存在 PCI-E 内存空间中映射的地址区间中包括多个数据报文缓存片、以及至少一个控制报文缓存片；并且，该报文转发方法进一步包括：

- 当确定从以太交换芯片接收到的以太数据报文以其它接口板为目的板卡时，选择该其它接口板的板卡内存映射的一个空闲的数据报文缓存片的 PCI-E 内存空间地址作为 PCI-E 报文设定的目的地址；并且，根据该 PCI-E 报文的地址将该其它接口板的对应的数据报文缓存片的状态记录由空闲变更为占用；

- 当从来自其它接口板的 PCI-E 报文中解析出的以太数据报文转发至以太交换芯片时，构造携带有数据报文缓存片释放信息的 PCI-E 报文，并且构造的 PCI-E 报文以该其它接口板的板卡内存映射的控制报文缓存片为目的地址，以供 PCI-E Endpoint 将构造的该 PCI-E 报文向转发板转发，并使转发板根据构造的该 PCI-E 报文的地址将该 PCI-E 报文向该其它接口板转发；

- 当从来自其它接口板的 PCI-E 报文中解析出数据报文缓存片释放信息时，根据解析出的数据报文缓存片释放信息将该其它接口板的对应的数据报文缓存片的状态记录由占用重置为空闲。

- 如上可见，基于上述的例子，当接口板从以太交换芯片接收到以太数据报文时，若确定该以太数据报文以其它接口板为目的板卡，则该接口板的逻辑装置可以将该以太数据报文封装为以其它接口板中的板卡内存的 PCI-E 内存空间地址为目的地址的 PCI-E 报文，以使得该 PCI-E 报文可以通过转发板被转发至其它接口板的板卡内存；并且，当接口板从板卡内存中获取到来自其它接口板的 PCI-E 报文时，该接口板的逻辑装置可以从该 PCI-E 报文中解析出以太数据报文并转发。从而，以太数据报文在接口板之间的转发无需转发板的 CPU 参与、并且可以提升转发性能。

权利要求书

1、一种网络设备的第一接口板，其特征在于，包括 PCI-E 端点、以太交换芯片、板卡内存、以及逻辑装置；

5 PCI-E 端点通过 PCI-E 总线连接网络设备的转发板中对应该第一接口板的 PCI-E 根复合体；

以太交换芯片从网络设备的外部接收第一以太网数据报文；

板卡内存映射在网络设备的 PCI-E 内存空间中，并且，该板卡内存在 PCI-E 内存空间中映射的地址区间不同于其它接口板的板卡内存、以及转发板的系统内存在 PCI-E 内存空间中映射的地址区间；

10 逻辑装置从以太交换芯片接收第一以太网数据报文，并确定该第一以太网数据报文的板卡；其中，当确定从以太交换芯片接收到的第一以太网数据报文以网络设备的第二接口板为目的板卡时，逻辑装置将该第一以太网数据报文封装为以该第二接口板的板卡内存的 PCI-E 内存空间地址为目的地址的第一 PCI-E 报文，以供 PCI-E 端点将该第一 PCI-E 报文向网络设备转发板转发；

15 逻辑装置还从板卡内存获取来自第三接口板的第二 PCI-E 报文，并从获取到的该第二 PCI-E 报文中解析出第二以太网数据报文发送至以太交换芯片，其中所述第三接口板与第二接口板相同，或不同。

2、根据权利要求 1 所述的第一接口板，其特征在于，逻辑装置中维护有以太 MAC 地址与板卡标识的第一映射表，用于根据第一以太网数据报文的 MAC 地址
20 确定该第一以太网数据报文的板卡；

并且，来自第三接口板的第二 PCI-E 报文中封装有第三接口板从网络设备的外部接收到的第二以太网数据报文、并携带有该第三接口板的板卡标识；

以及，逻辑装置根据从该第二 PCI-E 报文中解析出的第二以太网数据报文的源 MAC 地址和该第二 PCI-E 报文中携带的板卡标识，在以太 MAC 地址与板卡标识的
25 第一映射表创建对应该第三接口板的表项。

3、根据权利要求 1 所述的第一接口板，其特征在于，逻辑装置维护有板卡标识与 PCI-E 内存空间地址的第二映射表，用于根据从以太交换芯片接收到的第一以太网数据报文的板卡，确定封装有该第一以太网数据报文的第一 PCI-E 报文的地址；并且，根据转发板 CPU 的配置，在板卡标识与 PCI-E 内存空间地址的第二映射

表中创建表项。

4、根据权利要求 1 所述的第一接口板，其特征在于，当确定从以太交换芯片接收到的第一以太网数据报文目的板卡失败时，逻辑装置将该第一以太网数据报文封装为分别以网络设备每一块其它接口板的板卡内存的 PCI-E 内存空间地址为目的地址的
5 PCI-E 报文，以供 PCI-E 端点向转发板转发。

5、根据权利要求 1 所述的第一接口板，其特征在于，该第一接口板与其它接口板的板卡内存存在 PCI-E 内存空间中映射的地址区间中包括多个数据报文缓存片、以及至少一个控制报文缓存片；

当确定从以太交换芯片接收到的第一以太网数据报文以第二接口板为目的板卡
10 时，逻辑装置为所述第一 PCI-E 报文设定的目的地址为该第二接口板的板卡内存映射的一个空闲的数据报文缓存片的 PCI-E 内存空间地址，并且，逻辑装置还根据该第一 PCI-E 报文的地址将该第二接口板的对应的数据报文缓存片的状态记录由空闲变更为占用；

当从来自第三接口板的第二 PCI-E 报文中解析出的第二以太网数据报文发送至以
15 太交换芯片时，逻辑装置构造携带有数据报文缓存片释放信息的第三 PCI-E 报文，并且构造的第三 PCI-E 报文以该第三接口板的板卡内存映射的控制报文缓存片为目的地址，以供 PCI-E 端点将构造的该第三 PCI-E 报文向转发板转发，其中所述释放信息对应的数据报文缓存片为所述第二 PCI-E 报文设定的目的地址；

当从来自第三接口板的第四 PCI-E 报文中解析出数据报文缓存片释放信息时，
20 逻辑装置根据解析出的数据报文缓存片释放信息将该第三接口板的对应的数据报文缓存片的状态记录由占用重置为空闲。

6、根据权利要求 1 所述的第一接口板，其特征在于，当确定从以太交换芯片接收到的第一以太网数据报文的板卡为网络设备的转发板时，逻辑装置将该第一以太网数据报文封装为以系统内存的 PCI-E 内存空间地址为目的地址的 PCI-E 报文，以
25 供 PCI-E 端点将该 PCI-E 报文向转发板转发、以由转发板写入转发板的系统内存中。

7、根据权利要求 6 所述的第一接口板，其特征在于，逻辑装置中维护有以太 MAC 地址与板卡标识的第一映射表，用于根据所述第一以太网数据报文的地址确定该第一以太网数据报文的板卡；

其中，逻辑装置根据 CPU 在网络设备初始化时的配置，在以太 MAC 地址与板
30 卡标识的第一映射表中创建对应转发板的表项；

并且, 逻辑装置从板卡内存获取的来自第三接口板的第二 PCI-E 报文中, 封装有所述第三接口板从网络设备的外部接收到的第二以太网数据报文、并携带有该第三接口板的板卡标识;

以及, 逻辑装置根据从该第二 PCI-E 报文中解析出的第二以太网数据报文的源
5 MAC 地址和该第二 PCI-E 报文中携带的板卡标识, 在以太 MAC 地址与板卡标识的第一映射表创建对应该第三接口板的表项。

8、一种网络设备, 其特征在于, 包括转发板和至少两块接口板, 转发板包括 CPU、系统内存、系统内存控制器、以及分别对应各接口板的 PCI-E 根复合体, 每块接口板包括以太交换芯片、逻辑装置、板卡内存、以及通过 PCI-E 总线连接转发
10 板中对应的 PCI-E 根复合体的 PCI-E 端点;

各接口板的以太交换芯片从网络设备的外部接收第一以太网数据报文;

各接口板的板卡内存以及转发板的系统内存映射在网络设备的 PCI-E 内存空间中, 并且, 各接口板的板卡内存以及转发板的系统内存存在 PCI-E 内存空间中映射的地址区间互不相同;

15 每块接口板的逻辑装置从该接口板的以太交换芯片接收所述第一以太网数据报文、并确定该第一以太网数据报文的目的板卡; 其中, 当确定从以太交换芯片接收到的所述第一以太网数据报文以网络设备的第二接口板为目的板卡时, 逻辑装置将该第一以太网数据报文封装为以该第二接口板的板卡内存的 PCI-E 内存空间地址为目的地址的第一 PCI-E 报文, 以供该接口板的 PCI-E 端点将该第一 PCI-E 报文向转发板的
20 对应 PCI-E 根复合体转发、并使转发板的对应 PCI-E 根复合体根据该第一 PCI-E 报文的地址的目的地址将该第一 PCI-E 报文向目的板卡的板卡内存转发;

每块接口板的逻辑装置还从该接口板的板卡内存获取来自第三接口板的第二 PCI-E 报文, 并从获取到的该第二 PCI-E 报文中解析出第二以太网数据报文发送至该接口板的以太交换芯片, 其中所述第三接口板与第二接口板相同, 或不同。

25 9、根据权利要求 8 所述的网络设备, 其特征在于, 每块接口板的逻辑装置在确定从以太交换芯片接收到的第一以太网数据报文的地址的目的板卡为转发板时, 将该第一以太网数据报文封装为以系统内存的 PCI-E 内存空间地址为目的地址的 PCI-E 报文, 以供 PCI-E 端点将该 PCI-E 报文向转发板的对应 PCI-E 根复合体转发、并由转发板的对应 PCI-E 根复合体写入转发板的系统内存中。

30 10、根据权利要求 9 所述的网络设备, 其特征在于, 每块接口板的逻辑装置中

维护有以太 MAC 地址与板卡标识的第一映射表, 用于根据所述第一以太数据报文的目的 MAC 地址确定该第一以太数据报文的目的板卡;

其中, 每块接口板的逻辑装置根据 CPU 在网络设备初始化时的配置, 在以太 MAC 地址与板卡标识的第一映射表中创建对应转发板的表项;

5 并且, 每块接口板的逻辑装置从该接口板的板卡内存获取的来自所述第三接口板的第二 PCI-E 报文中, 封装有所述第三接口板从网络设备的外部接收到的第二以太数据报文、并携带有该第三接口板的板卡标识;

以及, 每块接口板的逻辑装置根据从该第二 PCI-E 报文中解析出的第二以太数据报文的源 MAC 地址和该第二 PCI-E 报文中携带的板卡标识, 在以太 MAC 地址与板卡标识的第一映射表创建对应该第三接口板的表项。

11、根据权利要求 8 所述的网络设备, 其特征在于, 每块接口板的逻辑装置维护有板卡标识与 PCI-E 内存空间地址的第二映射表, 用于根据从以太交换芯片接收到的第一以太数据报文的目的板卡, 确定封装有该第一以太数据报文的第一 PCI-E 报文的目的地址; 并且, 每块接口板的逻辑装置根据 CPU 在网络设备初始化时的配置, 在板卡标识与 PCI-E 内存空间地址的第二映射表中创建对应转发板和其它接口板的表项。

12、一种用于网络设备中的报文转发方法, 其特征在于, 网络设备包括转发板和至少两块接口板; 每块接口板包括 PCI-E 端点、以太交换芯片、以及板卡内存; 各接口板的板卡内存、以及转发板的系统内存共同映射在网络设备的 PCI-E 内存空间中, 并且, 各接口板的板卡内存、以及转发板的系统内存存在 PCI-E 内存空间中映射的地址区间互不相同;

该报文转发方法应用在任一块接口板、并包括:

当从以太交换芯片接收到来自网络设备的外部的第一以太数据报文时, 确定该第一以太数据报文的目的板卡;

25 当确定从以太交换芯片接收到的第一以太数据报文的目的板卡为网络设备的第二接口板时, 将该第一以太数据报文封装为以该第二接口板的板卡内存的 PCI-E 内存空间地址为目的地址的第一 PCI-E 报文, 以供 PCI-E 端点将该第一 PCI-E 报文向转发板转发、并使转发板根据该第一 PCI-E 报文的目的地址将该第一 PCI-E 报文向目的板卡的板卡内存转发;

30 当从板卡内存获取来自第三接口板的第二 PCI-E 报文时, 从获取到的该第二

PCI-E 报文中解析出第二以太网数据报文转发至以太网交换芯片。

13、根据权利要求 12 所述的报文转发方法，其特征在于，该报文转发方法进一步包括：

维护以太 MAC 地址与板卡标识的第一映射表，用于根据第一以太网数据报文的
5 目的 MAC 地址确定该第一以太网数据报文的板卡；

从来自第三接口板的第二 PCI-E 报文中解析出第三接口板从网络设备的外部接收到的第二以太网数据报文、以及该第三接口板的板卡标识；

根据解析出的第二以太网数据报文的源 MAC 地址和该第二 PCI-E 报文中携带的板卡标识，在以太 MAC 地址与板卡标识的第一映射表创建对应该第三接口板的表
10 项。

14、根据权利要求 12 所述的报文转发方法，其特征在于，该报文转发方法进一步包括：维护板卡标识与 PCI-E 内存空间地址的第二映射表，用于根据从以太网交换芯片接收到的第一以太网数据报文的板卡，确定封装有该第一以太网数据报文的第
15 一 PCI-E 报文的地址；并且，根据转发板的 CPU 的配置，在板卡标识与 PCI-E 内存空间地址的第二映射表中创建表项。

15、根据权利要求 12 所述的报文转发方法，其特征在于，该报文转发方法进一步包括：当确定从以太网交换芯片接收到的第一以太网数据报文的板卡失败时，将该第一以太网数据报文封装为分别以网络设备的每一块其它接口板的卡内存的 PCI-E 内存空间地址为目的地址的多于一个 PCI-E 报文，以供 PCI-E 端点将多于一
20 个 PCI-E 报文向转发板转发、并使转发板根据多于一个 PCI-E 报文的地址将多于一个 PCI-E 报文分别转发给对应的其它接口板。

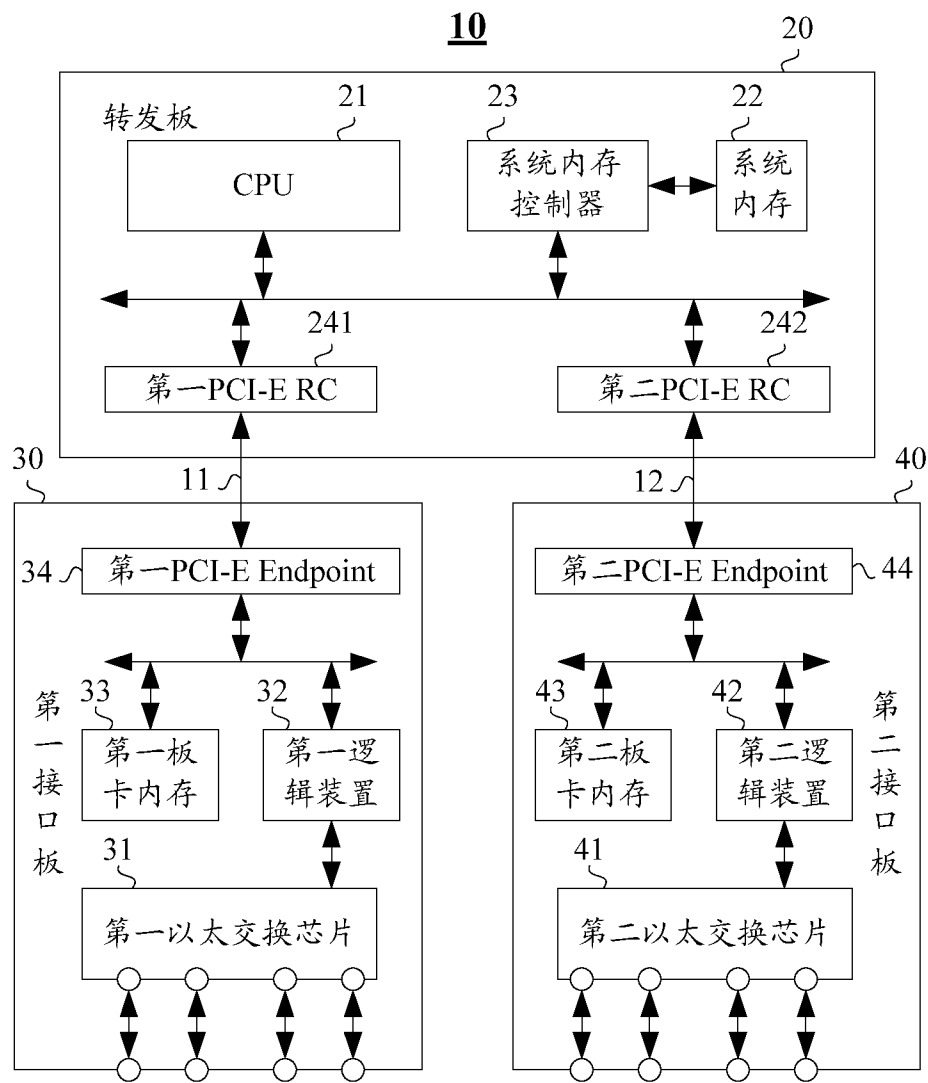
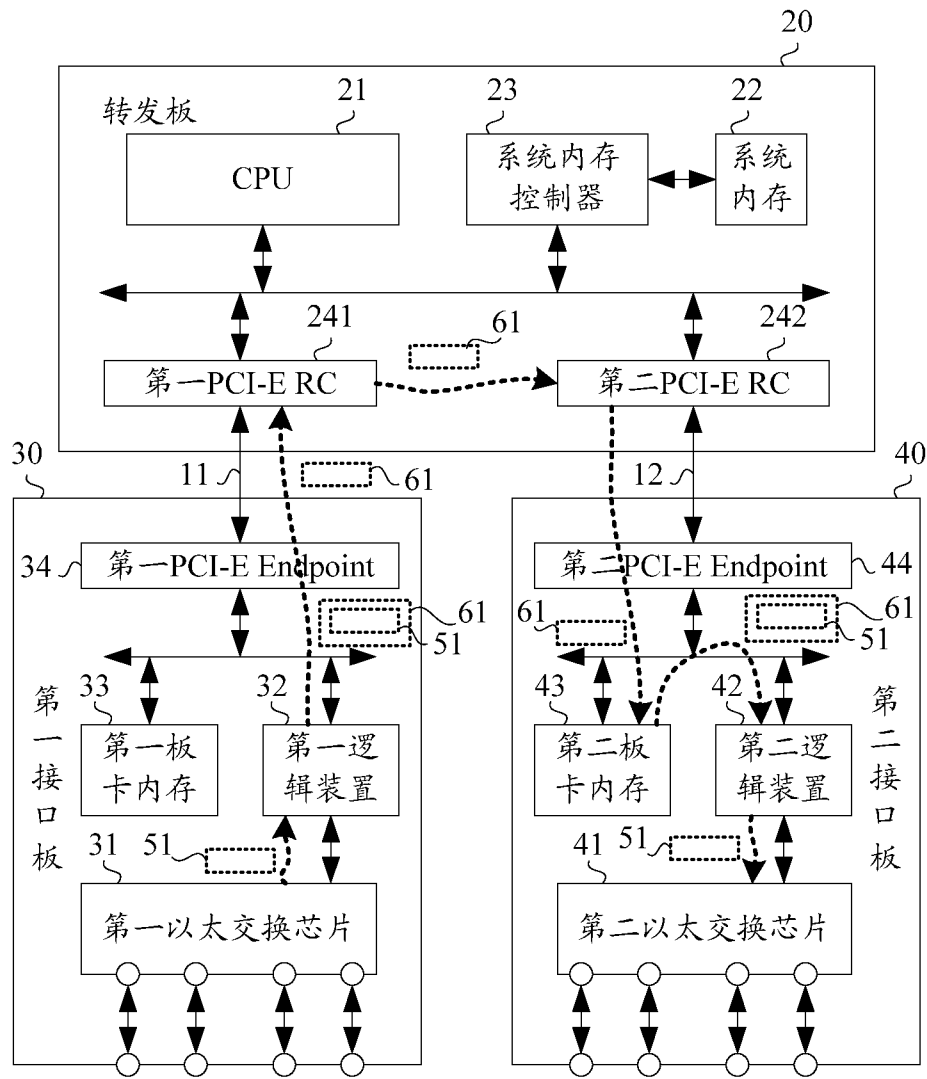
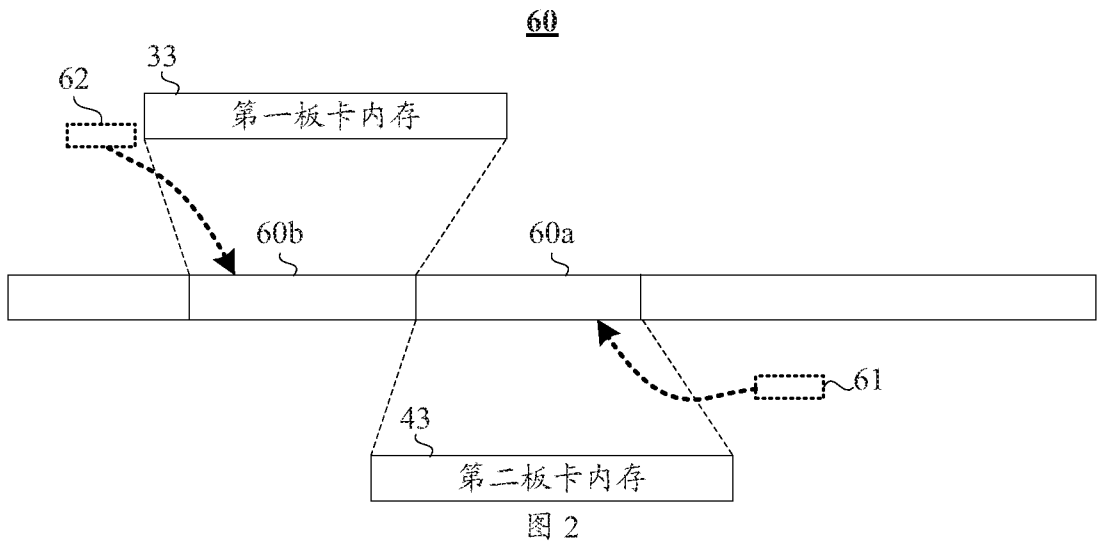


图 1



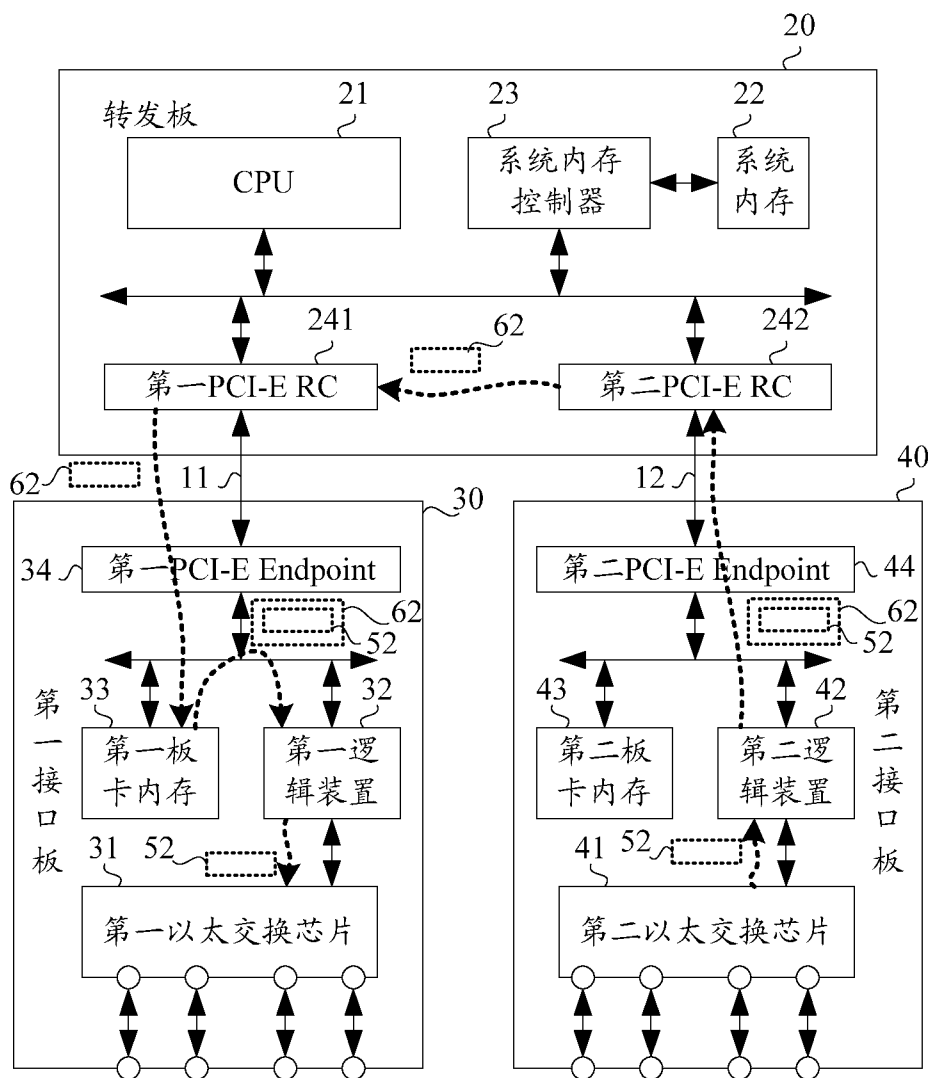


图 3b

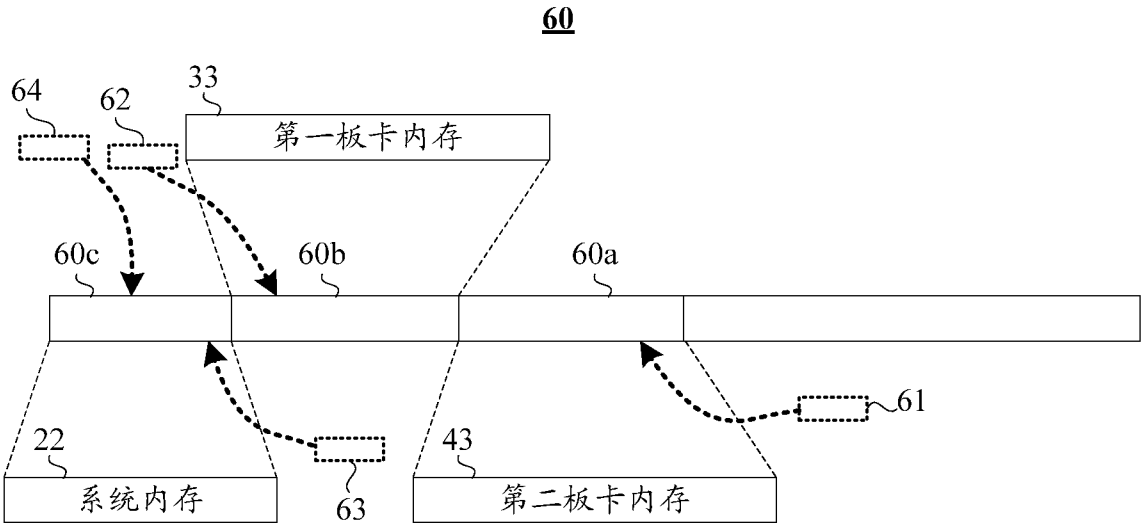


图 4

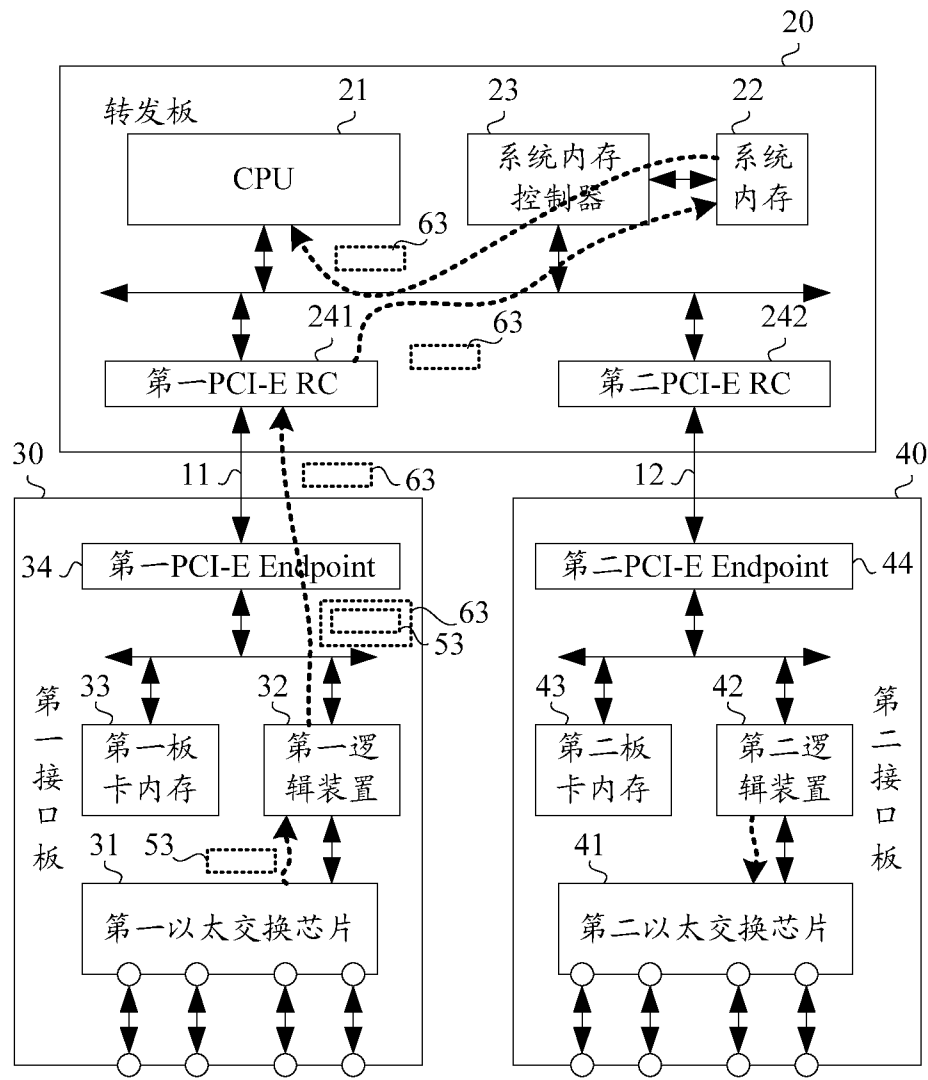


图 5a

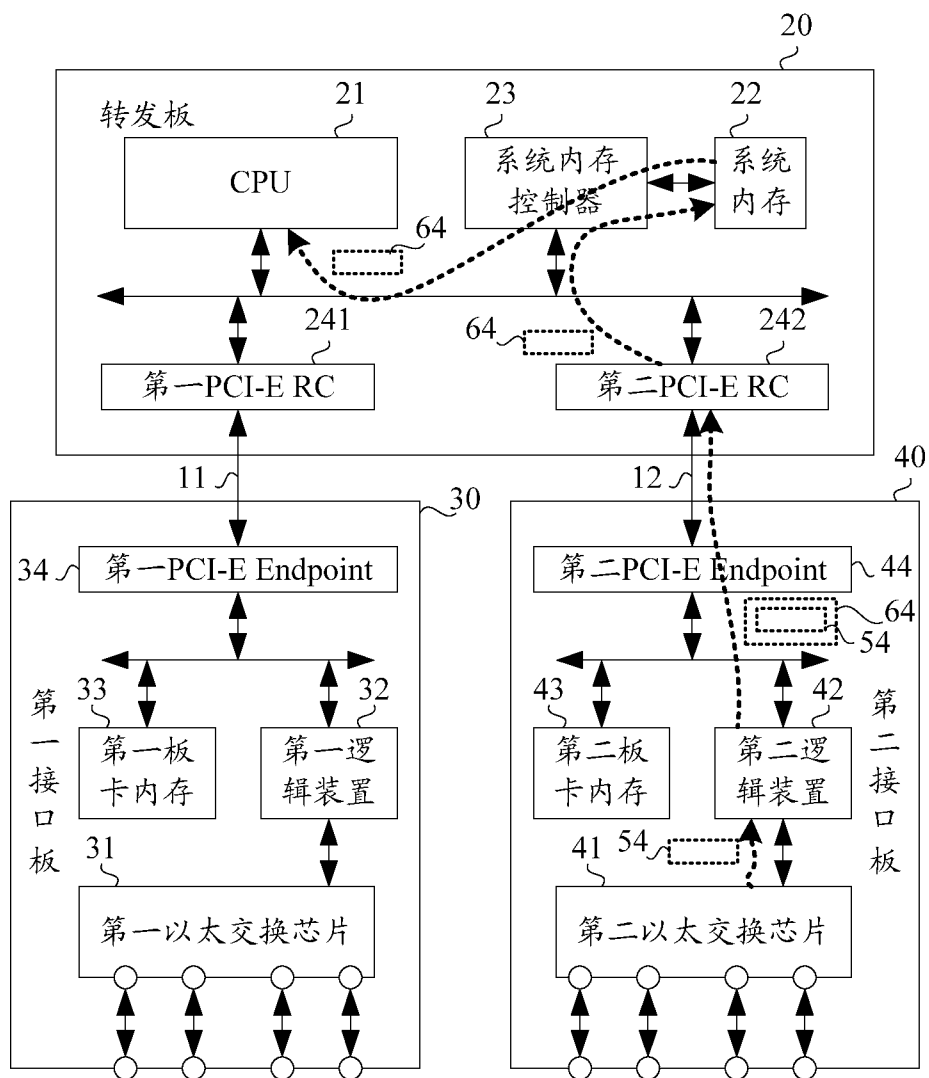


图 5b

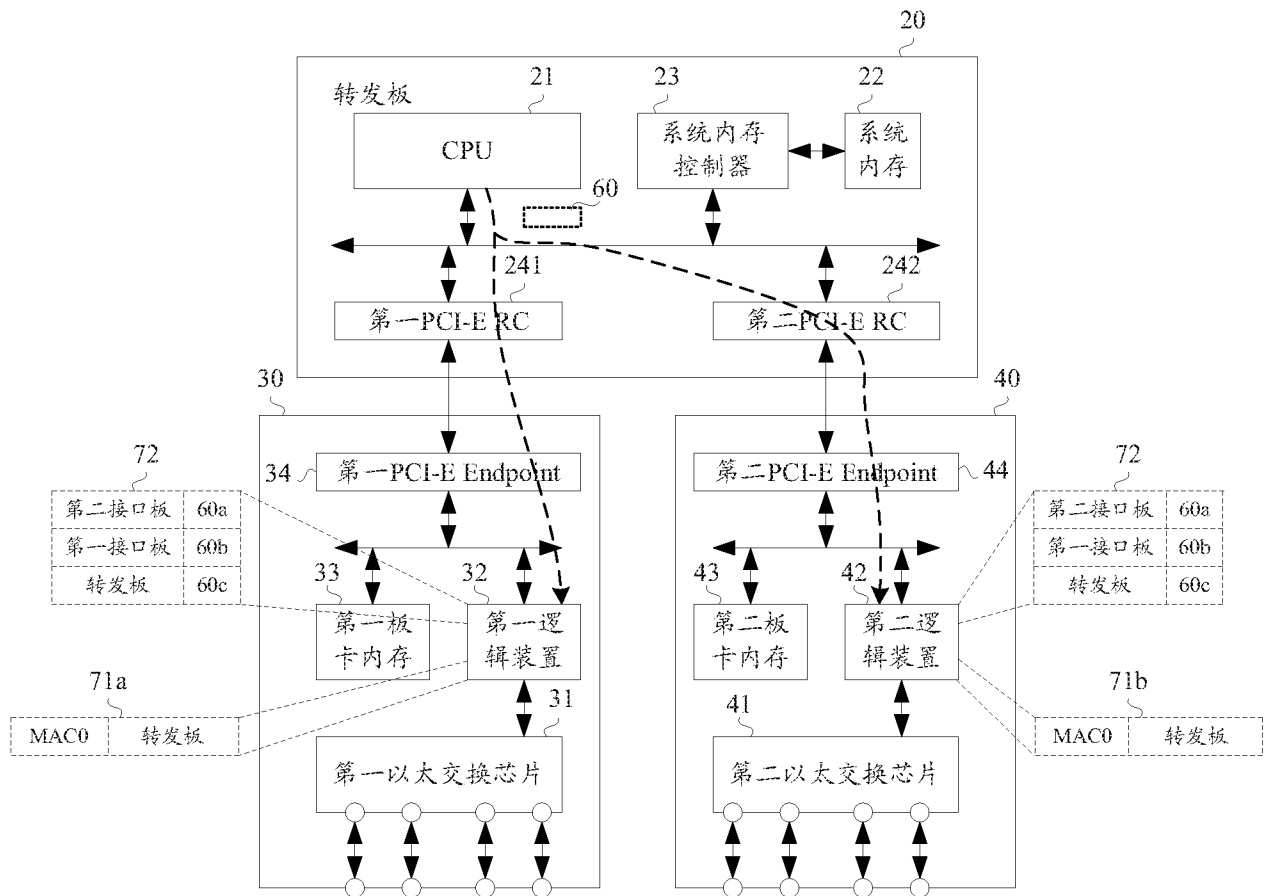


图 6a

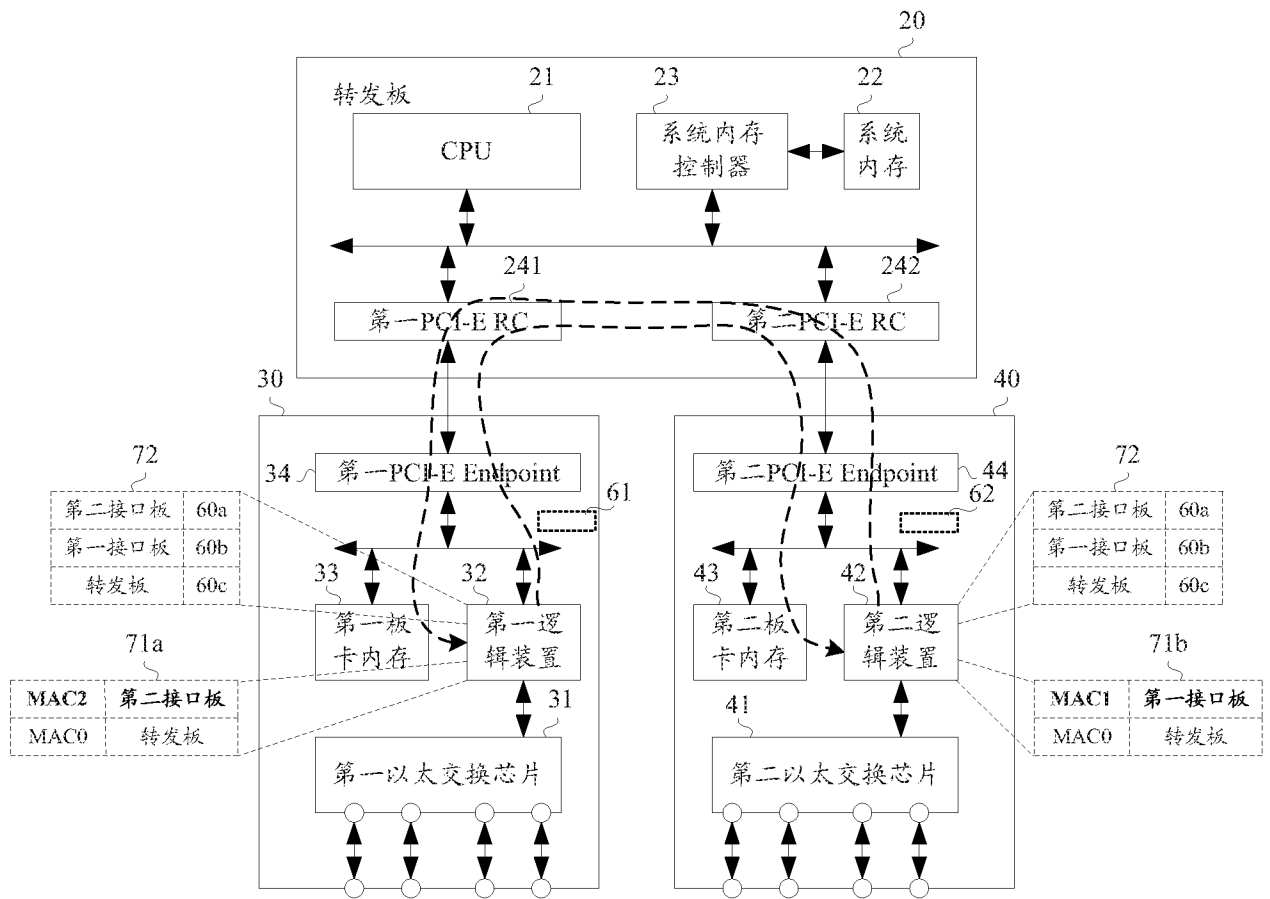


图 6b

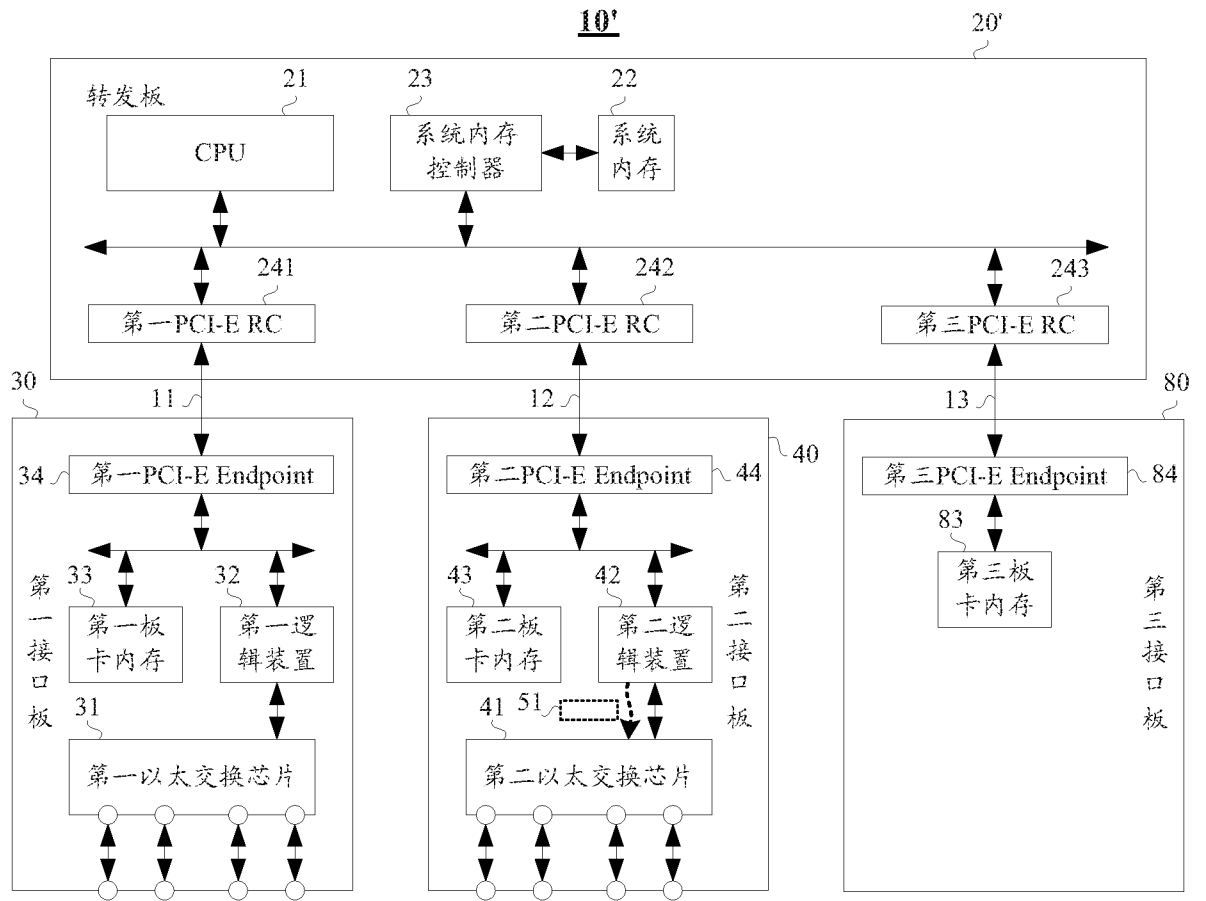


图 7

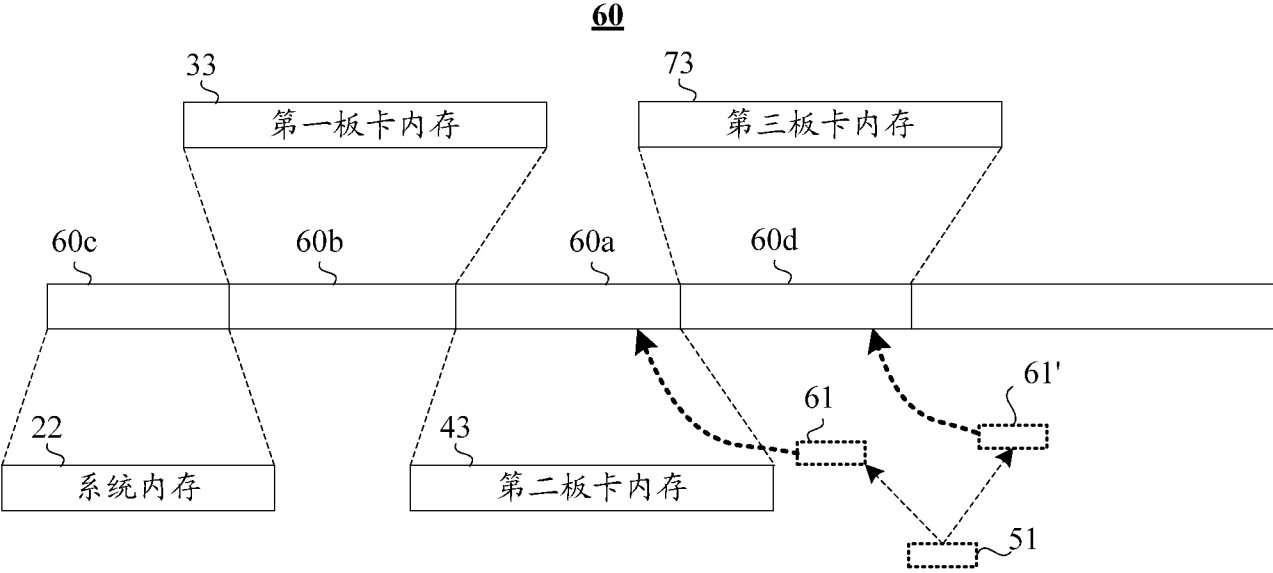


图 8

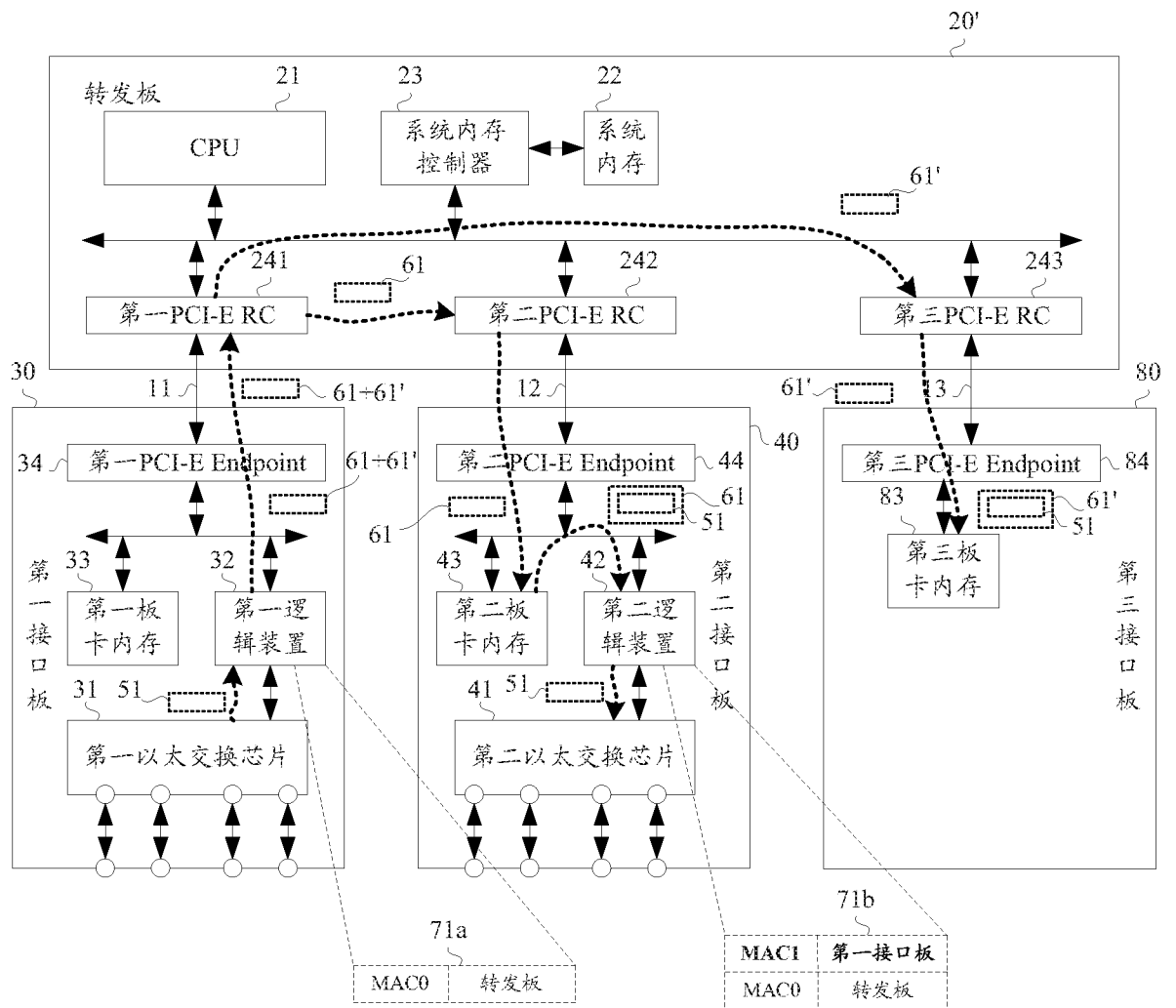


图 9

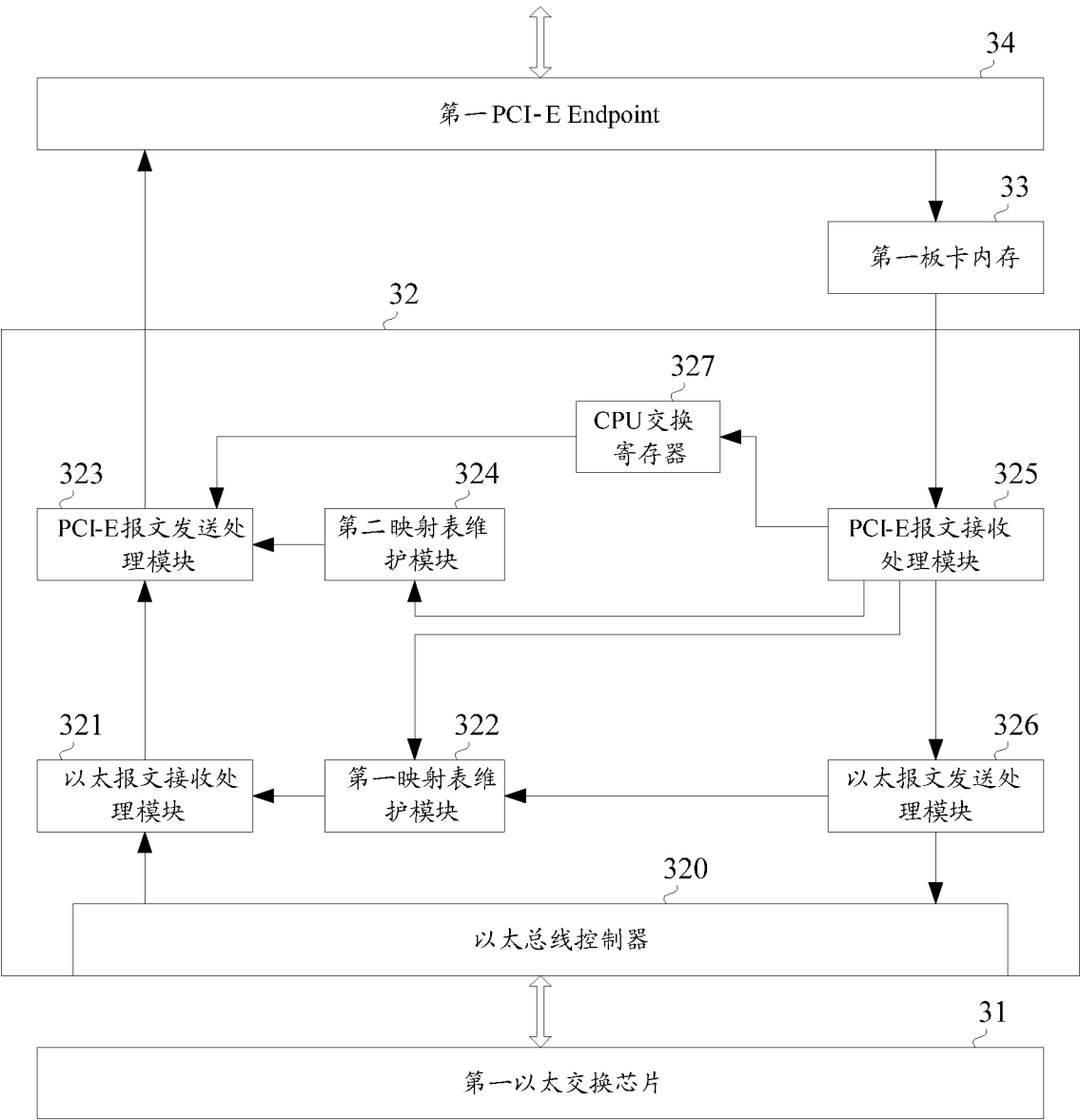


图 10

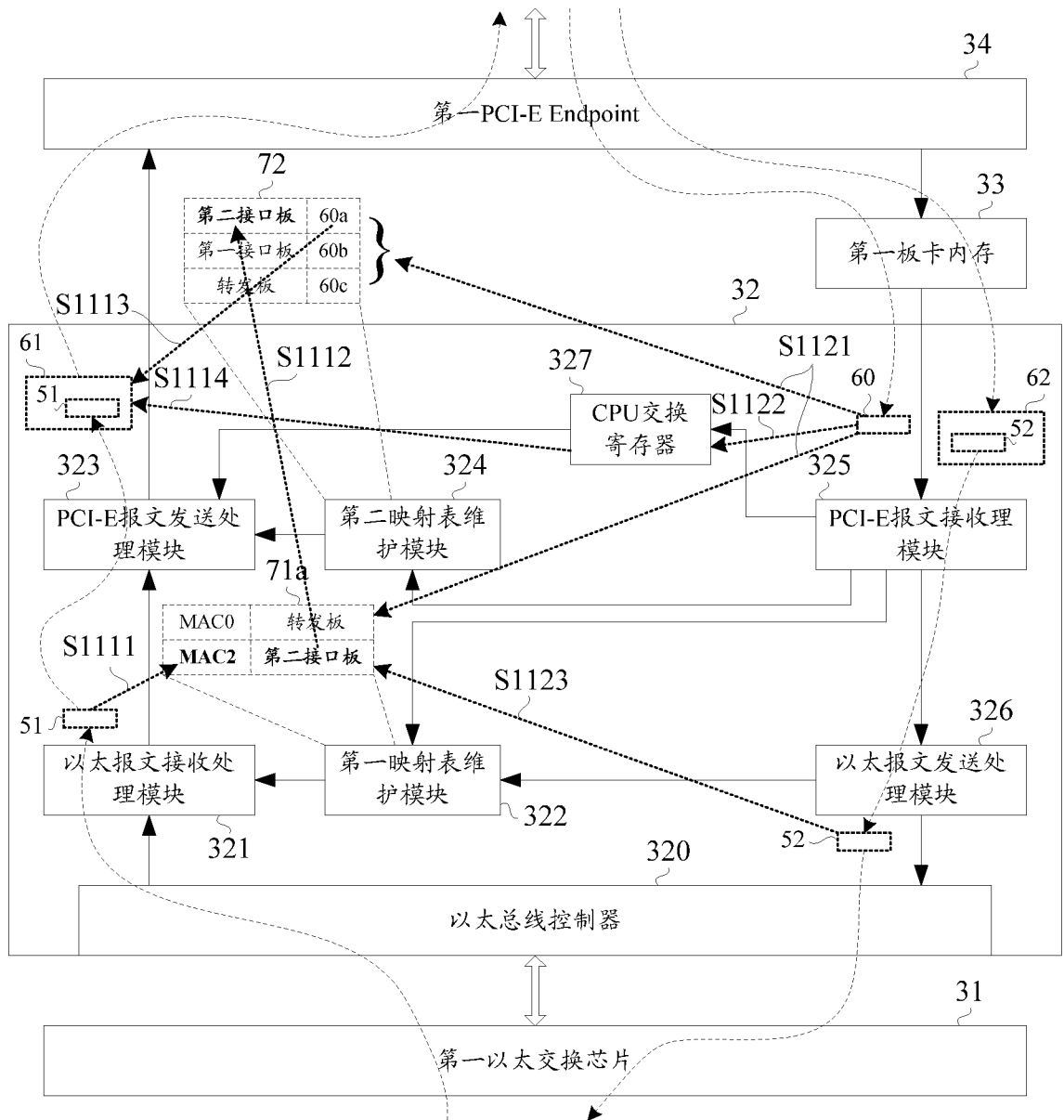


图 11

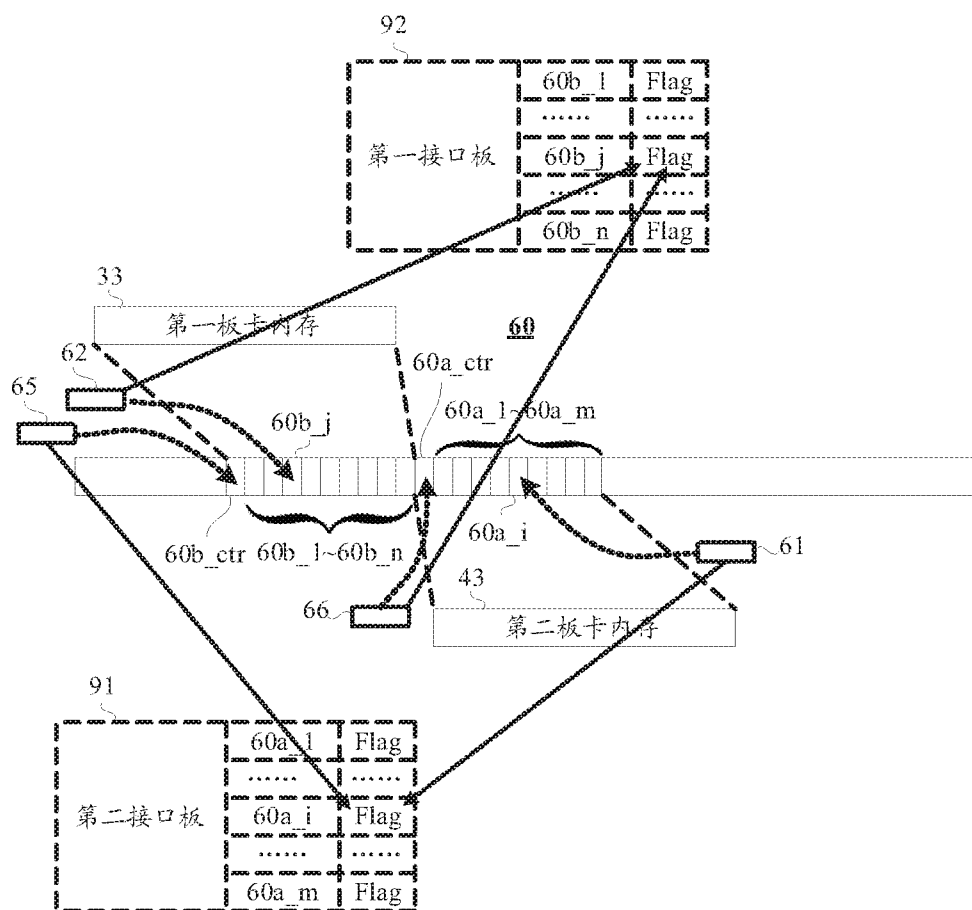


图 12

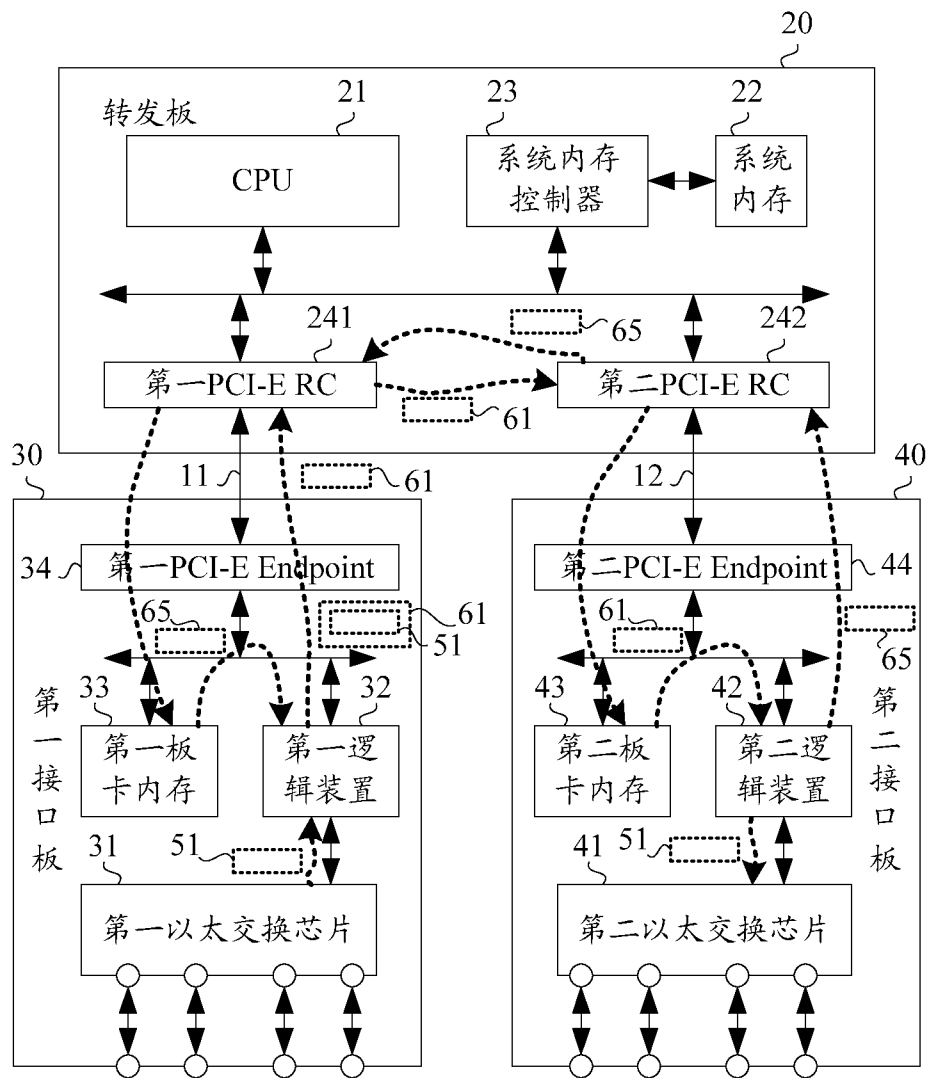


图 13a

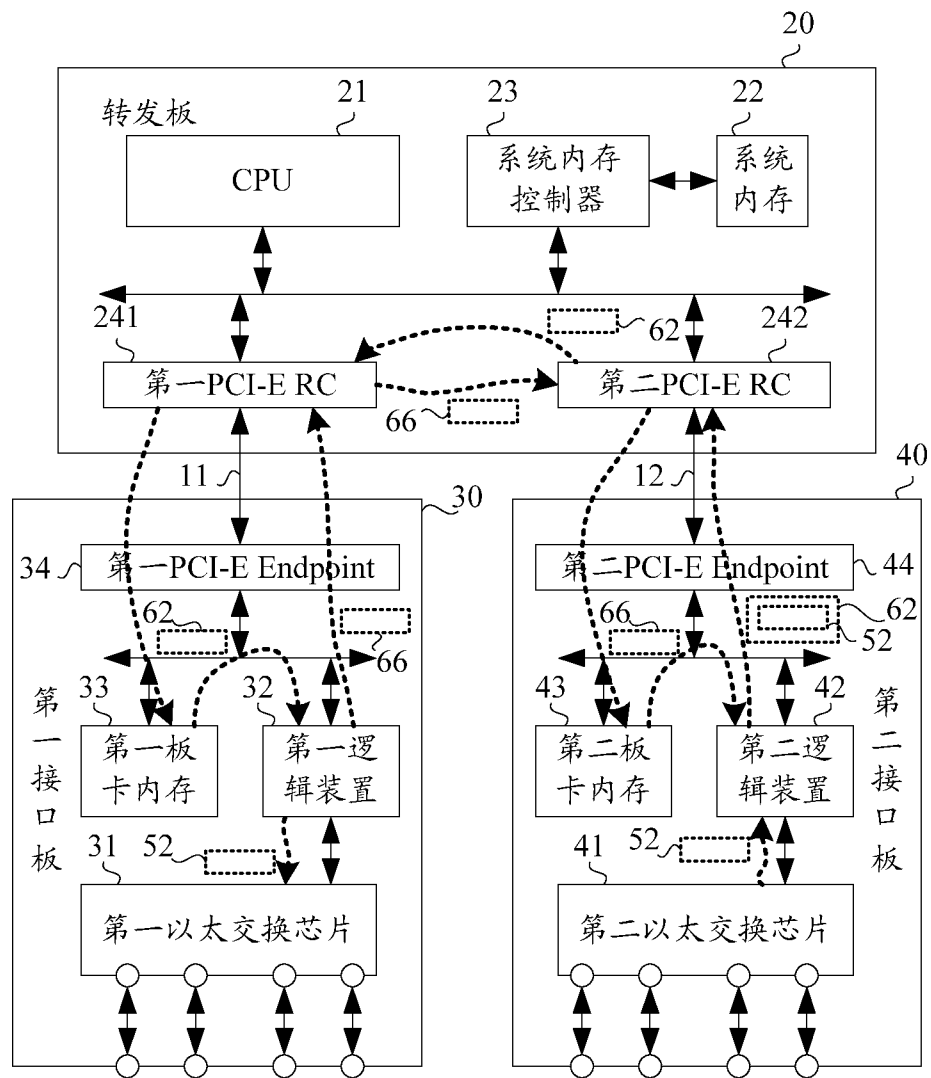


图 13b

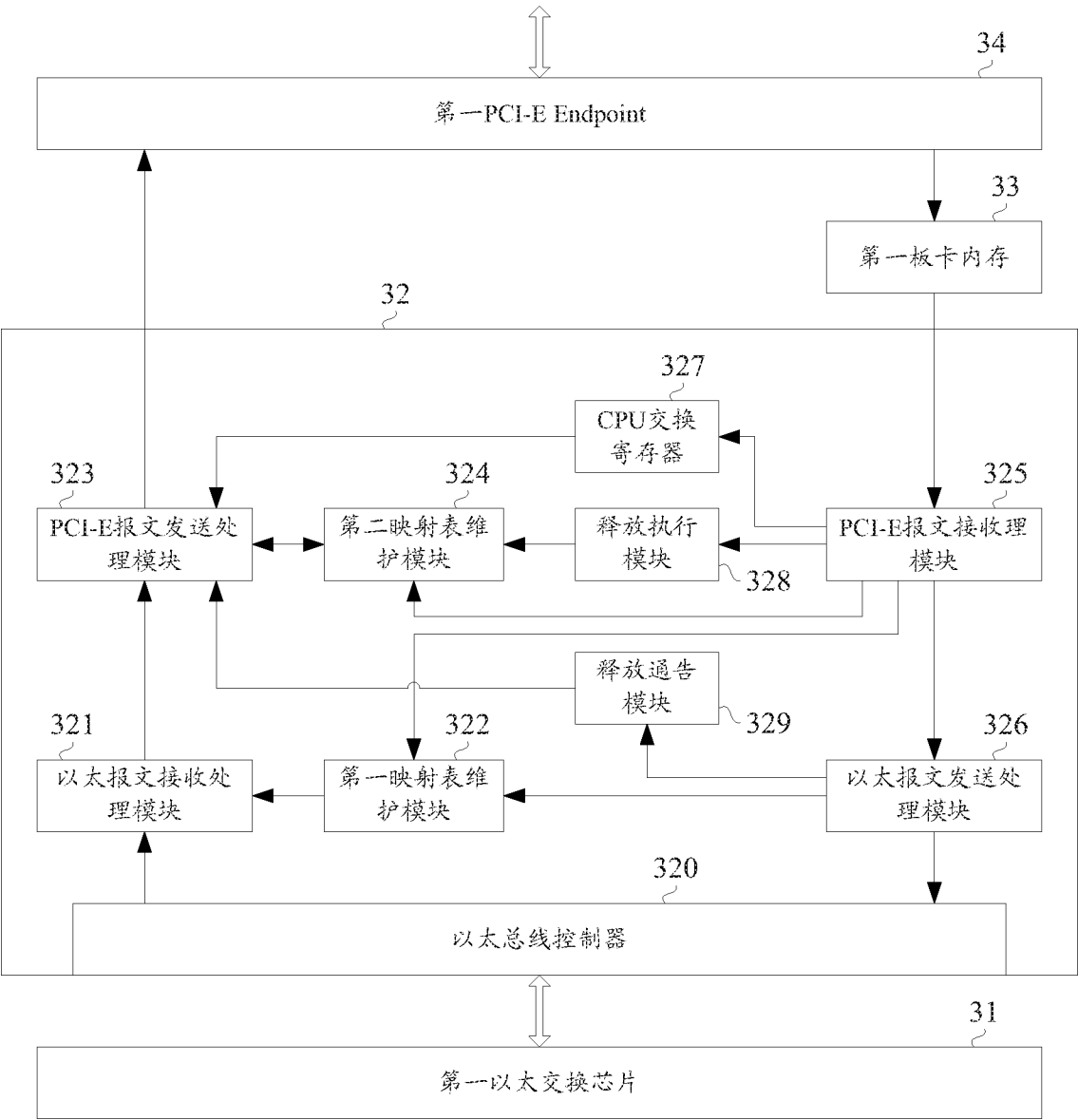


图 14

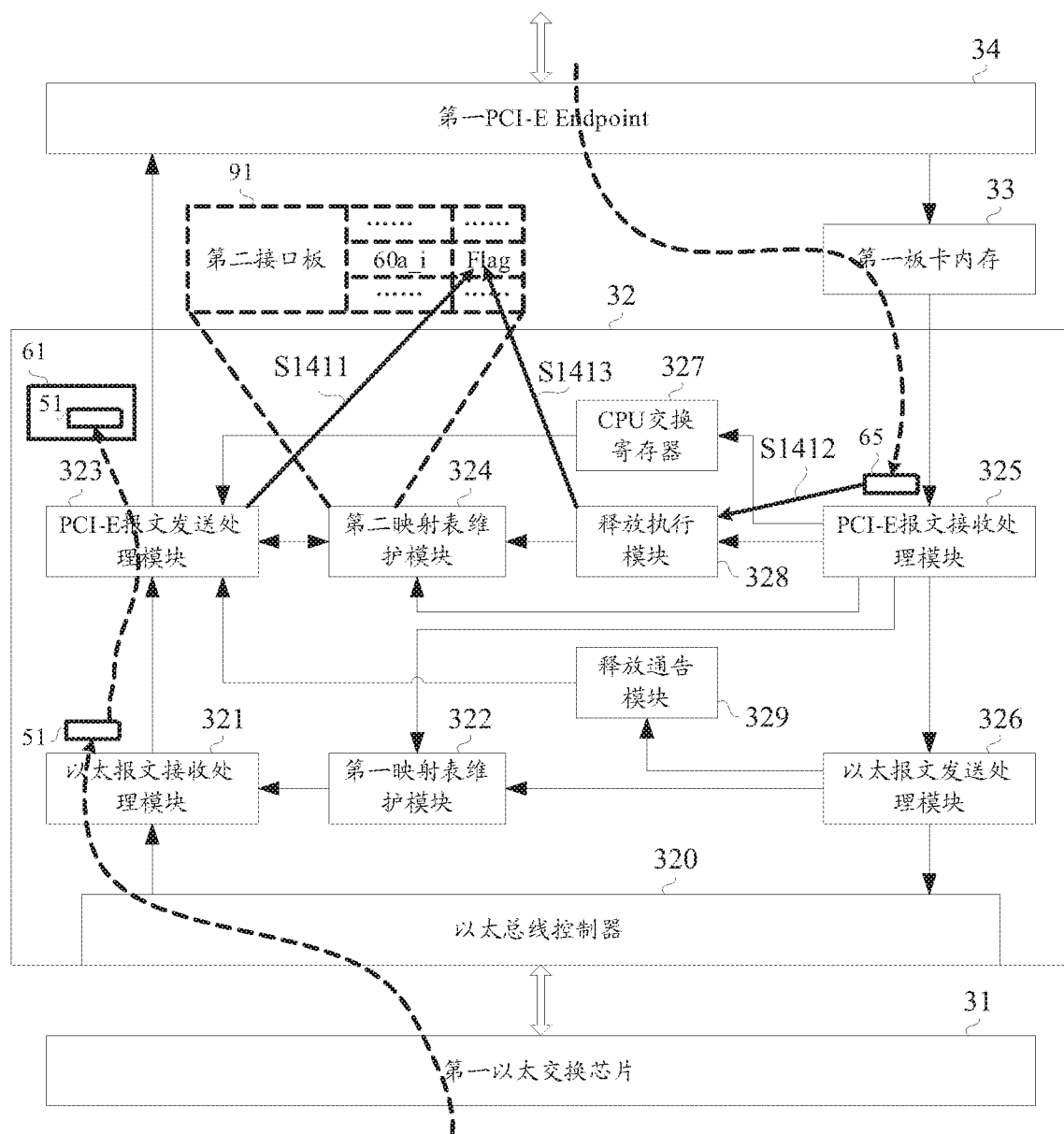
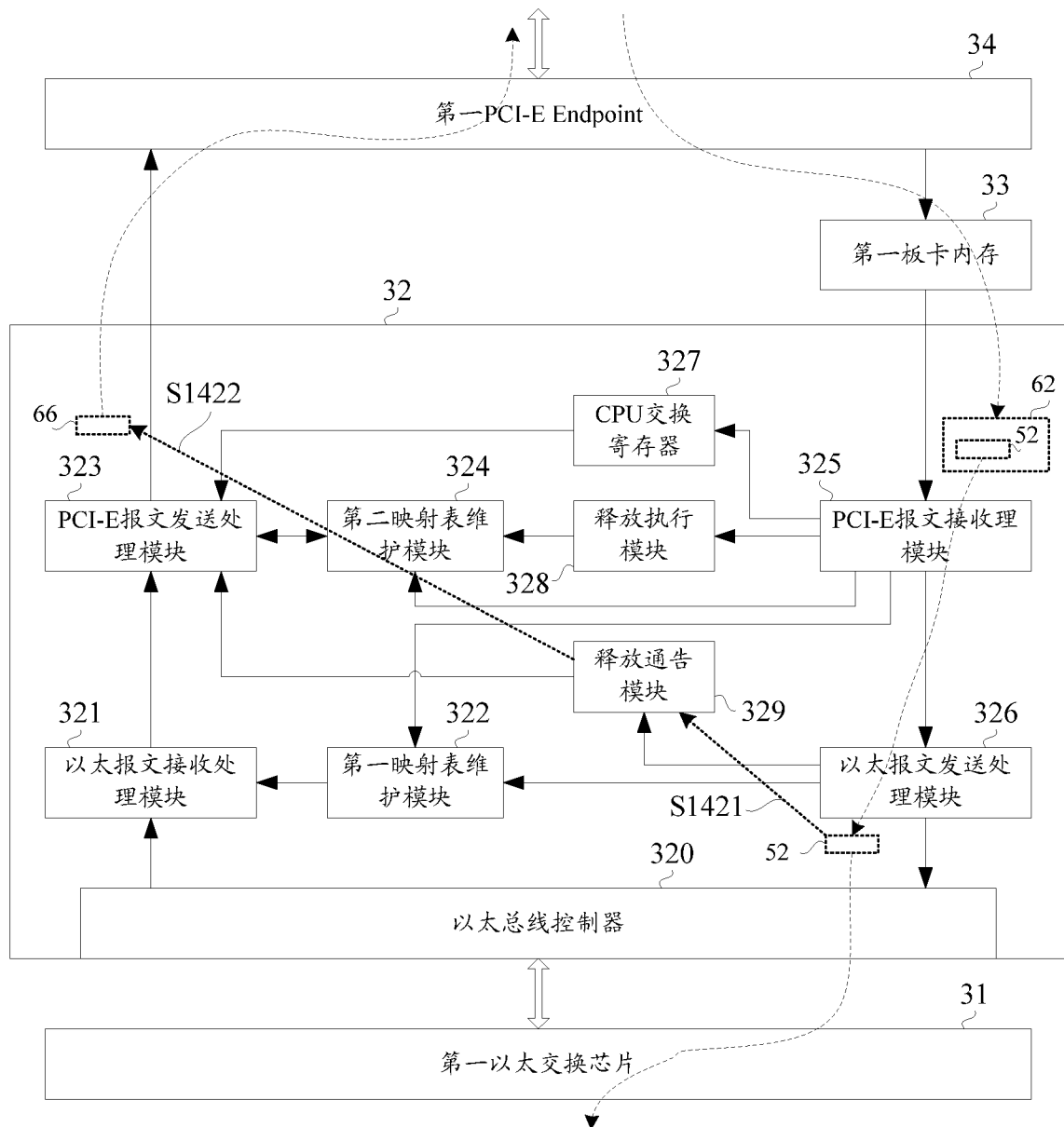


图 15a



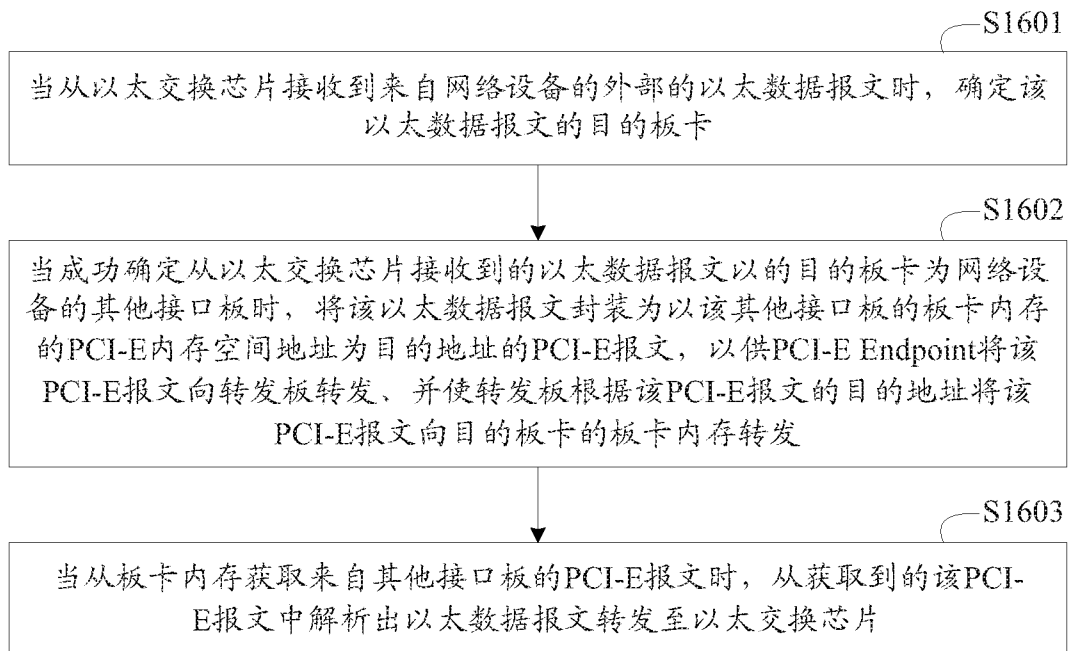


图 16

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/103943**A. CLASSIFICATION OF SUBJECT MATTER**

H04L 12/931 (2013.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L; G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CNTXT, CNKI, VEN, USTXT: high speed peripheral component interconnect standard, interface board, board card identification, PCI 1w E, PCIE, board, ethernet, memory, logic+, map+, forward???, packag???, identifier, id

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 101277195 A (H3C TECHNOLOGIES CO., LIMITED), 01 October 2008 (01.10.2008), see description, page 4, 6 th line from the bottom to page 10, line 15, and figures 2-9	1-15
A	CN 102393838 A (HUAWEI TECHNOLOGIES CO., LTD.), 28 March 2012 (28.03.2012), see the whole document	1-15
A	CN 103490961 A (H3C TECHNOLOGIES CO., LIMITED), 01 January 2014 (01.01.2014), see the whole document	1-15
A	US 2013031288 A1 (AGILENT TECHNOLOGIES, INC.), 31 January 2013 (31.01.2013), see the whole document	1-15

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 06 January 2017 (06.01.2017)	Date of mailing of the international search report 13 January 2017 (13.01.2017)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer ZOU, Feifei Telephone No.: (86-10) 62411263

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/103943

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 101277195 A	01 October 2008	CN 101277195 B	07 December 2011
		WO 2008119221 A1	09 October 2008
		CN 101277196 A	01 October 2008
		CN 101277196 B	28 September 2011
CN 102393838 A	28 March 2012	CN 102393838 B	11 March 2015
		US 2013238871 A1	12 September 2013
		WO 2012103768 A1	09 August 2012
		EP 2620876 A1	31 July 2013
CN 103490961 A	01 January 2014	None	
US 2013031288 A1	31 January 2013	US 8661178 B2	25 February 2014

国际检索报告

国际申请号

PCT/CN2016/103943

A. 主题的分类

H04L 12/931(2013.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

H04L; G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNABS, CNTXT, CNKI, VEN, USTXT: 高速外设部件互联标准, 接口板, 以太, 内存, 逻辑, 映射, 转发, 封装, 板卡标识, PCI 1w E, PCIE, board, ethernet, memory, logic+, map+, forward???, packag???, identifier, id

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 101277195 A (杭州华三通信技术有限公司) 2008年 10月 1日 (2008 - 10 - 01) 参见说明书第4页倒数第6行-第10页第15行、附图2-9	1-15
A	CN 102393838 A (华为技术有限公司) 2012年 3月 28日 (2012 - 03 - 28) 参见全文	1-15
A	CN 103490961 A (杭州华三通信技术有限公司) 2014年 1月 1日 (2014 - 01 - 01) 参见全文	1-15
A	US 2013031288 A1 (安捷伦科技有限公司) 2013年 1月 31日 (2013 - 01 - 31) 参见全文	1-15

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2017年 1月 6日

国际检索报告邮寄日期

2017年 1月 13日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

邹菲菲

电话号码 (86-10)62411263

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/103943

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	101277195	A	2008年 10月 1日	CN	101277195	B	2011年 12月 7日
				WO	2008119221	A1	2008年 10月 9日
				CN	101277196	A	2008年 10月 1日
				CN	101277196	B	2011年 9月 28日
CN	102393838	A	2012年 3月 28日	CN	102393838	B	2015年 3月 11日
				US	2013238871	A1	2013年 9月 12日
				WO	2012103768	A1	2012年 8月 9日
				EP	2620876	A1	2013年 7月 31日
CN	103490961	A	2014年 1月 1日	无			
US	2013031288	A1	2013年 1月 31日	US	8661178	B2	2014年 2月 25日

表 PCT/ISA/210 (同族专利附件) (2009年7月)