

[19] 中华人民共和国国家知识产权局



## [12] 发明专利申请公布说明书

[21] 申请号 200810114082.7

[43] 公开日 2008 年 10 月 15 日

[51] Int. Cl.  
H04L 12/18 (2006.01)  
H04L 12/56 (2006.01)

[11] 公开号 CN 101286866A

[22] 申请日 2008.5.30

[21] 申请号 200810114082.7

[71] 申请人 杭州华三通信技术有限公司

地址 310053 浙江省杭州市高新技术产业开发区之江科技工业园六和路 310 号华为杭州生产基地

[72] 发明人 田 浩

[74] 专利代理机构 北京德琦知识产权代理有限公司

代理人 宋志强 麻海明

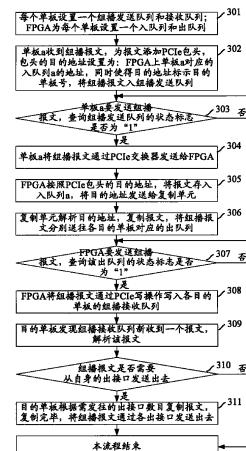
权利要求书 4 页 说明书 11 页 附图 4 页

### [54] 发明名称

基于高速周边元件扩展接口交换网的组播实现方法及系统

### [57] 摘要

本发明公开了基于 PCIe 交换网的组播实现方法、系统及 FPGA。方法包括：在交换网板上设置 FPGA，FPGA 与交换网板上的 PCIe 交换器之间的链路为 PCIe 链路，当源单板要向目的单板发送组播报文时，将目的单板标识信息放入组播报文，将该组播报文发往 FPGA，FPGA 根据组播报文中的目的单板标识信息复制报文，复制完毕，将各组播报文发往各目的单板。本发明将复制组播报文的工作转移到 FPGA 上完成，大大降低了单板的处理负担。



1、一种基于高速周边元件扩展接口 PCIe 交换网的组播实现方法，其特征在于，在交换网板上设置现场可编程门阵列 FPGA，FPGA 与交换网板上的 PCIe 交换器之间的链路为 PCIe 链路，方法包括：

源单板要向目的单板发送组播报文，将目的单板标识信息放入组播报文，将该组播报文发往 FPGA，FPGA 根据组播报文中的目的单板标识信息复制报文，复制完毕，将各组播报文发往各目的单板。

2、如权利要求 1 所述的方法，其特征在于，所述 FPGA 将各组播报文发往各目的单板之后进一步包括：目的单板收到组播报文，解析该报文，判断该报文是否要发往自身的出接口，若是，根据要发往的出接口数复制报文，复制完毕，将各组播报文从各出接口发送出去。

3、如权利要求 1 或 2 所述的方法，其特征在于，预先在 FPGA 上为每个单板设置一个入队列，

所述源单板将目的单板标识信息放入组播报文之后、将组播报文发往 FPGA 之前进一步包括：源单板判断 FPGA 上该源单板对应的入队列的可用空间是否大于预设第一阈值，若大于，则执行所述将组播报文发往 FPGA 的动作；否则，等待至源单板对应的入队列的可用空间大于预设第一阈值时，再执行所述将组播报文发往 FPGA 的动作。

4、如权利要求 3 所述的方法，其特征在于，在每个单板上设置一个状态标志，用于标示 FPGA 上的该单板对应的入队列的可用空间是否大于预设第一阈值，该状态标志初始化为“空”，且，当 FPGA 发现一个单板对应的入队列的可用空间从大于预设第一阈值变为小于预设第一阈值时，将该单板上的状态标志置为“满”，当 FPGA 发现一个单板对应的入队列的可用空间从小于预设第一阈值变为大于预设第一阈值时，将该单板上的状态标志置为“空”，

所述源单板判断 FPGA 上该源单板对应的入队列的可用空间是否大于预设第一阈值为：源单板判断自身的状态标志是否为“空”，若是，确定源单板对应

---

的入队列的可用空间大于预设第一阈值。

5、如权利要求3或4所述的方法，其特征在于，所述预先在FPGA上为每个单板设置一个入队列为：预先在FPGA上为每个单板分配一块内存空间作为该单板的入队列，

所述源单板将目的单板标识信息放入组播报文包括：源单板在组播报文中添加一个PCIe包头，包头中的目的地址为FPGA上该源单板对应的入队列的一个内存地址，并在目的地址中设置目的单板标志，

所述源单板将组播报文发往FPGA之后、FPGA根据组播报文中的目的单板标识信息复制报文之前进一步包括：源单板解析该组播报文的PCIe包头的目的地址中的目的单板标志，得到各目的单板标识，并得到目的单板数。

6、如权利要求1或2所述的方法，其特征在于，预先在每个单板上设置一个组播接收队列，

所述FPGA复制报文之后、将各组播报文发往各目的单板之前进一步包括：FPGA分别判断各目的单板的组播接收队列的可用空间是否大于预设第二阈值，若大于，则执行所述将组播报文发送给目的单板的动作；否则，等待至目的单板的组播接收队列的可用空间大于预设第二阈值时，再执行所述将组播报文发送给目的单板的动作。

7、如权利要求6所述的方法，其特征在于，预先在FPGA上为每个单板设置一个状态标志，分别用于标示每个单板上的组播接收队列的可用空间是否大于预设第二阈值，该状态标志初始化为“空”，且，当单板发现自身的组播接收队列的可用空间从大于预设第二阈值变为小于预设第二阈值时，将FPGA上该单板对应的状态标志置为“满”，当单板发现自身的组播接收队列的可用空间从小于预设第二阈值变为大于预设第二阈值时，将FPGA上该单板对应的状态标志置为“空”，

所述FPGA分别判断各目的单板的组播接收队列的可用空间是否大于预设第二阈值为：FPGA分别判断为各目的单板设置的状态标志是否为“空”，若是，确定目的单板的组播接收队列的可用空间大于预设第二阈值。

8、一种基于 PCIe 交换网的组播实现系统，其特征在于，包括：单板、PCIe 交换器、FPGA，其中：

单板，当要向目的单板发送组播报文时，将目的单板标识信息放入组播报文，将该组播报文通过 PCIe 交换器发往 FPGA；

FPGA，根据单板发来的组播报文中的目的单板标识信息复制报文，复制完毕，将各组播报文发往各目的单板。

9、如权利要求 8 所述的系统，其特征在于，所述单板包括：

组播发送处理单元，为待发送的组播报文添加 PCIe 包头，该包头中包含目的单板标识信息，将组播报文放入组播发送队列，从组播发送队列取出组播报文，通过 PCIe 交换器发送给 FPGA；

组播接收处理单元，发现组播接收队列中有新写入的组播报文，解析该报文，判断该报文是否需要发往本单板的出接口，若是，按照需发往的出接口数复制报文，复制完毕，将各报文从各出接口发送出去。

10、如权利要求 9 所述的系统，其特征在于，所述单板进一步包括：入队列状态设置模块，用于为 FPGA 上本单板对应的入队列设置一个状态标志，并初始化状态标志为“空”，并根据 FPGA 的对该状态标志的写操作更改该状态标志，

且所述组播发送处理单元进一步用于，当要从组播发送队列取出组播报文时，向入队列状态设置模块查询状态标志，若状态标志为“空”，则从组播发送队列取出组播报文；否则，继续向入队列状态设置模块查询状态标志，直至状态标志为“空”时，从组播发送队列取出组播报文。

11、如权利要求 9 所述的系统，其特征在于，所述单板进一步包括：组播接收队列状态维护模块，用于在发现组播接收队列的可用空间从大于预设第二阈值变为小于预设第二阈值时，将 FPGA 上本单板对应的状态标志置为“满”；在发现组播接收队列的可用空间从小于预设第二阈值变为大于预设第二阈值时，将 FPGA 上本单板对应的状态标志置为“空”。

12、一种 PCIe 交换网中的 FPGA，其特征在于，包括：

---

组播接收处理单元，接收单板通过 PCIe 交换器发来的组播报文，解析该报文的 PCIe 包头中的目的地址，得到源单板信息，将组播报文放入源单板对应的入队列，将 PCIe 包头中的目的地址发送给复制单元；

复制单元，根据 PCIe 包头中的目的地址，解析出组播报文在入队列中的地址和各目的单板标识，根据目的单板数复制报文，复制完毕，将组播报文在入队列中的地址放入各目的单板对应的出队列；

组播发送处理单元，从目的单板的出队列中取出组播报文，将报文发往各目的单板。

13、如权利要求 12 所述的 FPGA，其特征在于，所述 FPGA 进一步包括：流控管理单元，分别在本 FPGA 上为每个单板上的组播接收队列设置一个状态标志，初始化每个状态标志为“空”，并根据单板的更改状态标志的写操作更改单板的状态标志，

所述组播发送单元进一步用于，当要从目的单板对应的出队列中取报文时，向流控管理单元查询该目的单板的状态标志，若状态标志为“空”，则从出队列取出报文；否则，继续向流控管理单元查询该目的单板的状态标志，直至状态标志为“空”，再从出队列中取出报文。

14、如权利要求 12 所述的 FPGA，其特征在于，所述流控管理单元进一步用于，检测到一个单板对应的入队列的状态标志更改，通过写操作更改该单板上的状态标志。

## 基于高速周边元件扩展接口交换网的组播实现方法及系统

### 技术领域

本发明涉及组播技术领域，具体涉及一种基于高速周边元件扩展接口（PCIe, Peripheral Component Interconnect Express）交换网的组播实现方法、系统及现场可编程门阵列（FPGA, Field Programmable Gate Array）。

### 背景技术

组播是指在因特网协议（IP, Internet Protocol）网络中将数据包以尽力传送的形式发送到某个确定的节点集合即：组播组，其基本思想是：源主机即：组播源只发送一份数据，其目的地址为组播组地址；组播组中的所有主机都可收到同样的数据拷贝，并且只有组播组内的主机可以接收该数据，而其它主机则不能收到。

组播技术有效地解决了单点发送、多点接收的问题，实现了 IP 网络中点到多点的高效数据传送，能够大量节约网络带宽、降低网络负载。更重要的是，可以利用网络的组播特性方便地提供一些新的增值业务，包括在线直播、网络电视、远程教育、远程医疗、网络电台、实时视频会议等互联网的信息服务领域。

目前商用的 PCIe 交换芯片均没有提供组播传送的功能，基于 PCIe 总线技术的系统均通过软件实现数据的组播传送，其基本过程如图 1 所示，单板 0 在收到数据包后解析该报文，判断该报文是否需要组播发送，在确认需要组播发送且目的单板为 1、2、3 后，在单板 0 内会通过软件将报文复制为 3 份，随后将复制完毕的报文入输出队列，通过单板 0 与 PCIe 交换器间的 PCIe 链路将 3 份不同目的地址的报文发出，此时发送过程类似单播发送。如果每个单板又分成多个业务出接口，比如单板 1 有 20 个出接口，而单板 0 收到

的报文不仅要到达单板 1，还要从单板 1 的 15 个出接口发出，这样如果选择在单板 0 中复制报文，就意味着需要通过软件复制更多数量的报文。

可以看出通过软件实现组播的方式主要存在以下两个问题：

1、需要单板处理器参与报文复制工作，增加了单板的处理负担，降低整板的性能。

2、由于复制的多份报文均需要通过单板与 PCIe 交换器之间的 PCIe 链路发送，这会极大地降低单板与 PCIe 交换器之间的带宽利用率。

3、由于组播报文需要通过单板与 PCIe 交换器之间的 PCIe 链路多次发送，即增大该组播报文通过该端口的延迟，这使得组播报文之后的单播报文也得不到及时发送，造成该端口的报文延迟增大甚至拥塞。

## 发明内容

本发明提供基于 PCIe 交换网实现组播的方法、系统及 FPGA，以减轻单板的处理负担。

本发明的技术方案是这样实现的：

一种基于 PCIe 交换网的组播实现方法，其特征在于，在交换网板上设置 FPGA，FPGA 与交换网板上的 PCIe 交换器之间的链路为 PCIe 链路，方法包括：

源单板要向目的单板发送组播报文，将目的单板标识信息放入组播报文，将该组播报文发往 FPGA，FPGA 根据组播报文中的目的单板标识信息复制报文，复制完毕，将各组播报文发往各目的单板。

所述 FPGA 将各组播报文发往各目的单板之后进一步包括：目的单板收到组播报文，解析该报文，判断该报文是否要发往自身的出接口，若是，根据要发往的出接口数复制报文，复制完毕，将各组播报文从各出接口发送出去。

预先在 FPGA 上为每个单板设置一个入队列，

所述源单板将目的单板标识信息放入组播报文之后、将组播报文发往 FPGA 之前进一步包括：源单板判断 FPGA 上该源单板对应的入队列的可用空间是否大于预设第一阈值，若大于，则执行所述将组播报文发往 FPGA 的动作；

否则，等待至源单板对应的入队列的可用空间大于预设第一阈值时，再执行所述将组播报文发往 FPGA 的动作。

在每个单板上设置一个状态标志，用于标示 FPGA 上的该单板对应的入队列的可用空间是否大于预设第一阈值，该状态标志初始化为“空”，且，当 FPGA 发现一个单板对应的入队列的可用空间从大于预设第一阈值变为小于预设第一阈值时，将该单板上的状态标志置为“满”，当 FPGA 发现一个单板对应的入队列的可用空间从小于预设第一阈值变为大于预设第一阈值时，将该单板上的状态标志置为“空”，

所述源单板判断 FPGA 上该源单板对应的入队列的可用空间是否大于预设第一阈值为：源单板判断自身的状态标志是否为“空”，若是，确定源单板对应的入队列的可用空间大于预设第一阈值。

所述预先在 FPGA 上为每个单板设置一个入队列为：预先在 FPGA 上为每个单板分配一块内存空间作为该单板的入队列，

所述源单板将目的单板标识信息放入组播报文包括：源单板在组播报文中添加一个 PCIe 包头，包头中的目的地址为 FPGA 上该源单板对应的入队列的一个内存地址，并在目的地址中设置目的单板标志，

所述源单板将组播报文发往 FPGA 之后、FPGA 根据组播报文中的目的单板标识信息复制报文之前进一步包括：源单板解析该组播报文的 PCIe 包头的目的地址中的目的单板标志，得到各目的单板标识，并得到目的单板数。

预先在每个单板上设置一个组播接收队列，

所述 FPGA 复制报文之后、将各组播报文发往各目的单板之前进一步包括：FPGA 分别判断各目的单板的组播接收队列的可用空间是否大于预设第二阈值，若大于，则执行所述将组播报文发送给目的单板的动作；否则，等待至目的单板的组播接收队列的可用空间大于预设第二阈值时，再执行所述将组播报文发送给目的单板的动作。

预先在 FPGA 上为每个单板设置一个状态标志，分别用于标示每个单板上的组播接收队列的可用空间是否大于预设第二阈值，该状态标志初始化为“空”，

且，当单板发现自身的组播接收队列的可用空间从大于预设第二阈值变为小于预设第二阈值时，将 FPGA 上该单板对应的状态标志置为“满”，当单板发现自身的组播接收队列的可用空间从小于预设第二阈值变为大于预设第二阈值时，将 FPGA 上该单板对应的状态标志置为“空”，

所述 FPGA 分别判断各目的单板的组播接收队列的可用空间是否大于预设第二阈值为：FPGA 分别判断为各目的单板设置的状态标志是否为“空”，若是，确定目的单板的组播接收队列的可用空间大于预设第二阈值。

一种基于 PCIe 交换网的组播实现系统，包括：单板、PCIe 交换器、FPGA，其中：

单板，当要向目的单板发送组播报文时，将目的单板标识信息放入组播报文，将该组播报文通过 PCIe 交换器发往 FPGA；

FPGA，根据单板发来的组播报文中的目的单板标识信息复制报文，复制完毕，将各组播报文发往各目的单板。

所述单板包括：

组播发送处理单元，为待发送的组播报文添加 PCIe 包头，该包头中包含目的单板标识信息，将组播报文放入组播发送队列，从组播发送队列取出组播报文，通过 PCIe 交换器发送给 FPGA；

组播接收处理单元，发现组播接收队列中有新写入的组播报文，解析该报文，判断该报文是否需要发往本单板的出接口，若是，按照需发往的出接口数复制报文，复制完毕，将各报文从各出接口发送出去。

所述单板进一步包括：入队列状态设置模块，用于为 FPGA 上本单板对应的入队列设置一个状态标志，并初始化状态标志为“空”，并根据 FPGA 的对该状态标志的写操作更改该状态标志，

且所述组播发送处理单元进一步用于，当要从组播发送队列取出组播报文时，向入队列状态设置模块查询状态标志，若状态标志为“空”，则从组播发送队列取出组播报文；否则，继续向入队列状态设置模块查询状态标志，直至状态标志为“空”时，从组播发送队列取出组播报文。

所述单板进一步包括：组播接收队列状态维护模块，用于在发现组播接收队列的可用空间从大于预设第二阈值变为小于预设第二阈值时，将 FPGA 上本单板对应的状态标志置为“满”；在发现组播接收队列的可用空间从小于预设第二阈值变为大于预设第二阈值时，将 FPGA 上本单板对应的状态标志置为“空”。

一种 PCIe 交换网中的 FPGA，包括：

组播接收处理单元，接收单板通过 PCIe 交换器发来的组播报文，解析该报文的 PCIe 包头中的目的地址，得到源单板信息，将组播报文放入源单板对应的入队列，将 PCIe 包头中的目的地址发送给复制单元；

复制单元，根据 PCIe 包头中的目的地址，解析出组播报文在入队列中的地址和各目的单板标识，根据目的单板数复制报文，复制完毕，将组播报文在入队列中的地址放入各目的单板对应的出队列；

组播发送处理单元，从目的单板的出队列中取出组播报文，将报文发往各目的单板。

所述 FPGA 进一步包括：流控管理单元，分别在本 FPGA 上为每个单板上的组播接收队列设置一个状态标志，初始化每个状态标志为“空”，并根据单板的更改状态标志的写操作更改单板的状态标志，

所述组播发送单元进一步用于，当要从目的单板对应的出队列中取报文时，向流控管理单元查询该目的单板的状态标志，若状态标志为“空”，则从出队列取出报文；否则，继续向流控管理单元查询该目的单板的状态标志，直至状态标志为“空”，再从出队列中取出报文。

所述流控管理单元进一步用于，检测到一个单板对应的入队列的状态标志更改，通过写操作更改该单板上的状态标志。

与现有技术相比，本发明通过在交换网板上设置 FPGA，FPGA 与交换网板上的 PCIe 交换器之间的链路为 PCIe 链路，当源单板要向目的单板发送组播报文时，将目的单板标识信息放入组播报文，将该组播报文发往 FPGA，FPGA 根据组播报文中的目的单板标识信息复制报文，复制完毕，将各组播报文发往各目的单板。本发明将复制组播报文的工作转移到 FPGA 上完成，

大大降低了单板的处理负担，提高了整板性能，并提高了单板与 PCIe 交换器之间的带宽利用率；同时，由于不需要在单板中复制组播报文，即源单板只需向交换网板发送一次组播报文，这使得源单板上在组播报文之后的报文，比如单播报文能够及时得到发送，降低了后续报文的链路延迟，从而降低了拥塞。

### 附图说明

图 1 为现有的通过软件实现基于 PCIe 交换网的组播传送的示意图；  
图 2 为本发明实施例提供的基于 PCIe 交换网实现组播的示意图；  
图 3 为本发明实施例提供的基于 PCIe 交换网实现组播的流程图；  
图 4 为本发明实施例提供的基于 PCIe 交换网实现组播的系统组成图；  
图 5 为本发明实施例提供的单板的结构示意图；  
图 6 为本发明实施例提供的 FPGA 的结构示意图。

### 具体实施方式

下面结合附图及具体实施例对本发明再作进一步详细的说明。

图 2 为本发明实施例提供的基于 PCIe 交换网实现组播的示意图，如图 2 所示，在本发明实施例中，在交换网板上设置一个 FPGA，FPGA 与 PCIe 交换器之间的链路为 PCIe 链路。当一个单板如：单板 0 要向目的单板如：单板 1、2、3 发送组播报文 M 时，单板 0 先将组播报文 M 通过 PCIe 交换器发给 FPGA，FPGA 再根据目的单板数，复制报文 M，将复制得到的报文 M1、M2、M3 通过 PCIe 交换器分发给单板 1、2、3，图 2 中的 N 为单板数。

图 3 为本发明实施例提供的基于 PCIe 交换网实现组播的流程图，如图 3 所示，其具体步骤如下：

步骤 301：每个单板在初始化时设置一个组播发送队列和一个组播接收队列；FPGA 在初始化时为每个单板分别设置一个入队列和一个出队列，并为每个入队列分配一个内存基址和偏移地址，FPGA 将每个单板的入队列的

内存基址和偏移地址分别发送给各单板。

为每个单板上的组播发送队列设置一个状态标志，用于标示 FPGA 上与该单板对应的入队列是否有足够的可用空间，该状态标志初始化为“1”，当 FPGA 上一个单板 a 的入队列 a 没有足够的可用空间时，FPGA 通过 PCIe 写操作将单板 a 上的组播发送队列的状态标志置“0”，此后，当 FPGA 上单板 a 对应的入队列 a 又有足够的可用空间时，则 FPGA 通过 PCIe 写操作将单板 a 上的组播发送队列的状态标志置“1”。

为 FPGA 上每个单板对应的出队列设置一个状态标志，用于标示该单板上的组播接收队列是否有足够的可用空间，该状态标志初始化为“1”，当一个单板 a 上的组播接收队列没有足够的可用空间时，单板 a 通过 PCIe 写操作将 FPGA 上该单板 a 对应的出队列 a 的状态标志置“0”，当单板 a 上的组播接收队列重新又有足够的可用空间时，单板 a 通过 PCIe 写操作将 FPGA 上单板 a 对应的出队列 a 的状态标志置“1”。

具体地，可为 FPGA 上每个单板对应的入队列设置一个第一阈值，若入队列的可用空间大于该第一阈值，则认为该入队列有足够的可用空间；同样，可为每个单板上的组播接收队列设置一个第二阈值，若该组播接收队列的可用空间大于该第二阈值，则认为组播接收队列有足够的可用空间。

由于 PCIe 是基于地址路由的，因此，只有为 FPGA 上每个单板对应的入队列分配地址，才能使得来自单板的组播报文正确入队。入队列的地址即为内存空间的地址。例如：FPGA 中，单板 1 对应的入队列 1 的内存基址为 0xA1000000，偏移地址为 0xFF，则，单板 1 要发送组播报文时，需要将组播报文的 PCIe 包头中的目的地址设置在 0xA1000000~0xA10000FF 之间，这样，FPGA 收到该组播报文时，就能根据该目的地址将报文送往入队列 1。

步骤 302：一个单板 a 接收到一个报文，解析该报文，发现该报文为组播报文，并得到该报文发往的目的单板号，为报文添加 PCIe 包头，PCIe 包头中的目的地址设置为：FPGA 上单板 a 对应的入队列 a 的一个地址，同时，要使得该目的地址能够标示目的单板号，将组播报文入组播发送队列。

可预先设定 PCIe 包头中的目的地址从最后 1 位起，从后往前，每一位代表一个目的单板，若该位为 1 表示该单板为报文的目的单板，若该位为 0 表示该单板不为报文的目的单板。例如：FPGA 中，单板 1 对应的入队列 1 的内存基址为 0xA1000000，偏移地址为 0xFF，单板 1 收到一个组播报文后，发现报文的目的单板为单板 1、2、3，则单板 1 需要将组播报文的 PCIe 包头中的目的地址设置为满足条件：1、在 0xA1000000~0xA10000FF 之间；2、最后 2~4 位为 1，则该目的地址可以为：0xA100000E，可以看出，该地址满足条件 1，同时，由于  $0Xe=0b1110$ ，即倒数第 2~4 位为 1，因此，可知：组播报文的目的单板为单板 1、2、3。

步骤 303：单板 a 要发送一个组播报文，查询本单板的组播发送队列的状态标志是否为“1”，若是，执行步骤 304；否则，返回本步骤执行查询本单板的组播发送队列的状态标志是否为“1”的动作。

步骤 304：单板 a 将组播报文通过 PCIe 交换器发送给 FPGA。

步骤 305：FPGA 接收到单板 a 发来的组播报文，解析该报文的 PCIe 包头，按照 PCIe 包头中的目的地址，将报文存入该目的地址对应的入队列 a，并将 PCIe 包头中的目的地址发送给复制单元。

步骤 306：FPGA 的复制单元接收 PCIe 包头中的目的地址，解析该目的地址，得到组播报文在入队列 a 中的存储地址，并得到组播报文的目的单板号，根据目的单板数目，复制报文，复制完毕，将组播报文分别送往各目的单板对应的出队列。

复制单元将组播报文分别送往各目的单板对应的出队列实际上是将报文在入队列 a 中的存储地址分别放入各目的单板对应的出队列中，

步骤 307：FPGA 确定要发送一个出队列中的组播报文，查询该出队列的状态标志是否为“1”，若是，执行步骤 308；否则，返回本步骤执行查询该出队列的状态标志是否为“1”的动作。

步骤 308：FPGA 将组播报文通过 PCIe 写操作写入各目的单板的组播接收队列。

步骤 309：目的单板发现组播接收队列新收到一个报文，解析该报文。

步骤 310：目的单板判断组播报文是否需要从自身的出接口发送出去，若是，执行步骤 311；否则，目的单板对报文作后续处理，本流程结束。

步骤 311：目的单板根据需发往的出接口数目复制报文，复制完毕，将组播报文通过各出接口发送出去，本流程结束。

在实际应用中，经常遇到由多打一即：多个单板同时向一个单板发送报文而引起拥塞的情形，在基于 PCIe 交换网的系统中，多打一通常分两种情况，一、仅由组播报文引起的多打一；二、单播和组播报文均存在时引起的多打一。本发明实施例提供的组播实现技术方案可有效降低多打一引起的拥塞，具体如下：

对于第一种情况，比如单板 0 发送的组播报文的目的单板为 1 和 2，单板 2 发送的组播报文的目的单板为 1 和 3，单板 3 发送的组播报文的目的单板为 0 和 1，此时单板 0、2、3 均向单板 1 打流量，若每个单板的带宽均等且不采取任何防止拥塞措施，则单板 1 必然会拥塞。而根据本发明实施例，目的单板 1 在多打一的情况下，单板 1 的组播接收队列的可用空间很快会小于预设第二阈值，此时，单板 1 会通知 FPGA，FPGA 会暂停向单板 1 发送组播报文，但 FPGA 依然会继续接收来自源单板的组播报文直到源单板的入队列的可用空间小于预设第一阈值，此时 FPGA 会通知源单板暂停发送组播报文，从而完成从目的单板到 FPGA 再到源单板的逐级流控，实现仅由组播报文引起的多打一时的交换网内不丢包。

对于第二种情况，比如单板 0 发送的组播报文的目的单板为 1 和 2，单板 2 发送的单播报文的目的单板为 1，单板 3 发送的单播报文的目的单板为 1，此时单板 0、2、3 均向单板 1 发流量，若每个单板的带宽均等且不采取任何防止拥塞措施，单板 1 必然会拥塞。而目的单板中，通常会为每个单板设置一个单播接收队列，即共有 N-1 个单播接收队列（N 为单板数），为所有单板共同设置 1 个组播接收队列，根据本发明实施例提供的技术方案，组播和单播分别流控，即当目的单板的组播接收队列的可用空间小于预设第二

阈值时，FPGA 会暂停向目的单板发送组播报文，并在入队列的可用空间小于预设第一阈值时，通知源单板暂停发送组播报文；当目的单板上某个源单板的单播接收队列的可用空间小于预设阈值时，可直接通知源单板暂停发送单播报文。可见，通过从目的单板到 FPGA 再到源单板的逐级组播流控和从目的单板到源单板的单播流控，可实现单播和组播报文均存在所引起的多打一时的交换网内不丢包。

图 4 为本发明实施例提供的基于 PCIe 交换网的组播实现系统，如图 4 所示，其主要包括：单板 0~N-1（N 为单板总数）、PCIe 交换器和 FPGA，其中，PCIe 交换器和 FPGA 位于交换网板上，PCIe 交换器和 FPGA 之间的链路为 PCIe 链路。

如图 5 所示，在实际应用中，单板可由入队列状态设置模块 411、组播发送处理单元 412、组播接收队列状态维护模块 413 和组播接收处理单元 414 组成，其中：

入队列状态设置模块 411：用于为 FPGA 上本单板对应的入队列设置一个状态标志，并初始化状态标志为“空”，并根据 FPGA 的对该状态标志的写操作更改该状态标志。

组播发送单元 412：当要发送一个组播报文时，根据 FPGA 为本单板设置的入队列的内存基址和偏移地址，为待发送的组播报文添加 PCIe 包头，该包头的目的地址为所述入队列的一个地址，且该目的地址包含目的单板号信息，将组播报文放入组播发送队列。当要发送组播报文时，向入队列状态设置模块 411 查询状态标志，若状态标志为“空”，则从组播发送队列取出组播报文并通过 PCIe 交换器发送给 FPGA；否则，继续向入队列状态设置模块 411 查询状态标志，直至状态标志为“空”时，从组播发送队列取出组播报文并通过 PCIe 交换器发送给 FPGA。

组播接收队列状态维护模块 413：用于在发现本单板的组播接收队列的可用空间从大于预设第二阈值变为小于预设第二阈值时，将 FPGA 上本单板对应的状态标志置为“满”；在发现本单板的组播接收队列的可用空间从小于预设第

二阈值变为大于预设第二阈值时，将 FPGA 上本单板对应的状态标志置为“空”。

组播接收处理单元 414：发现本单板的组播接收队列中有新写入的组播报文，解析该报文，判断该报文是否需要发往本单板的出接口，若是，按照需发往的出接口数复制报文，复制完毕，将各报文从各出接口发送出去。

如图 6 所示，在实际应用中，FPGA 可由组播接收处理单元 421、复制单元 422、组播发送处理单元 423 和流控管理单元 424 组成，其中：

组播接收处理单元 421：接收单板通过 PCIe 交换器发来的组播报文，解析该报文的 PCIe 包头中的目的地址，得到源单板号，将组播报文放入源单板对应的入队列，将 PCIe 包头中的目的地址发送给复制单元 422。

复制单元 422：根据 PCIe 包头中的目的地址，解析出组播报文在入队列中的存储地址和各目的单板号，根据目的单板数复制报文，复制完毕，将组播报文在入队列中的存储地址放入各目的单板对应的出队列。

组播发送处理单元 423：当要向一个目的单板发送组播报文时，向流控管理单元 424 查询该目的单板的状态标志，若状态标志为“空”，则从该目的单板对应的出队列取出报文并发送给目的单板；否则，继续向流控管理单元 424 查询该目的单板的状态标志，直至状态标志为“空”，再从该目的单板对应的出队列中取出报文并发送给目的单板。

流控管理单元 424：分别在本 FPGA 上为每个单板上的组播接收队列设置一个状态标志，初始化每个状态标志为“空”，并根据单板的用于更改状态标志的 PCIe 写操作更改单板的状态标志；当检测到一个单板对应的入队列的状态标志更改时，通过 PCIe 写操作更改该单板上的状态标志。

以上所述仅为本发明的过程及方法实施例，并不用以限制本发明，凡在本发明的精神和原则之内所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

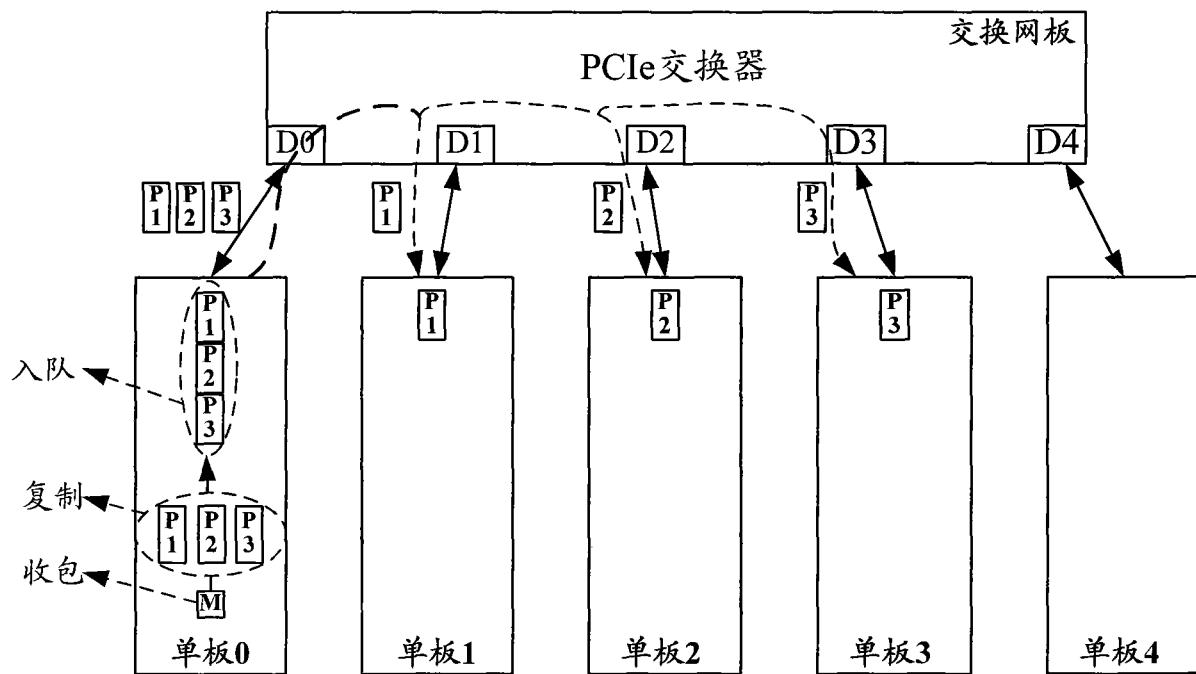


图 1

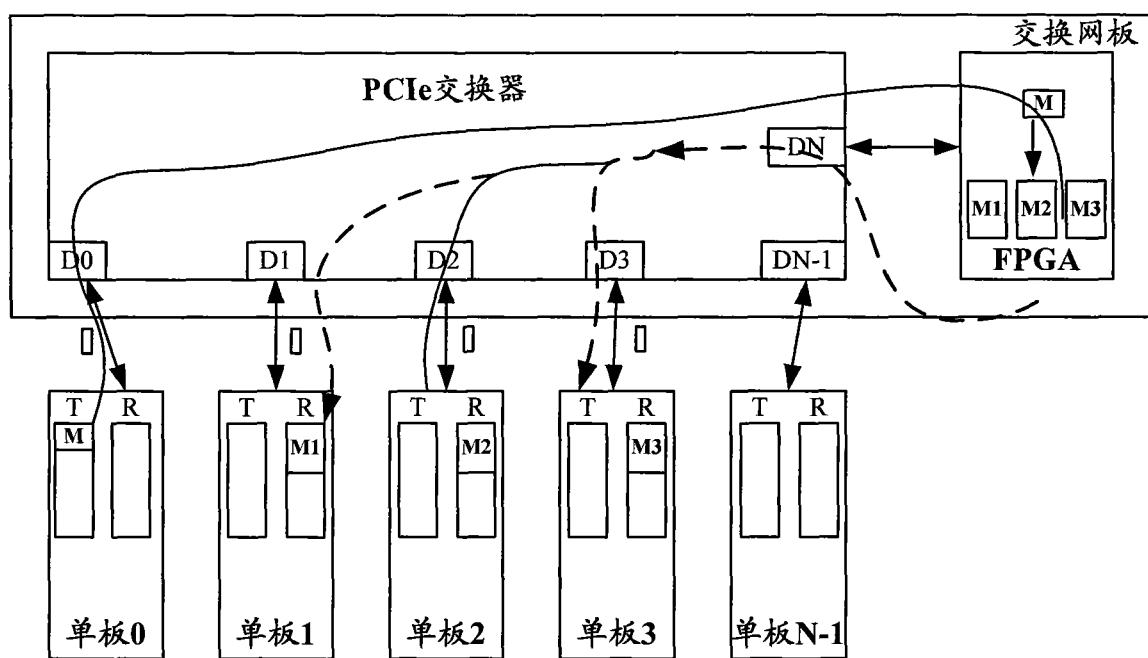


图 2

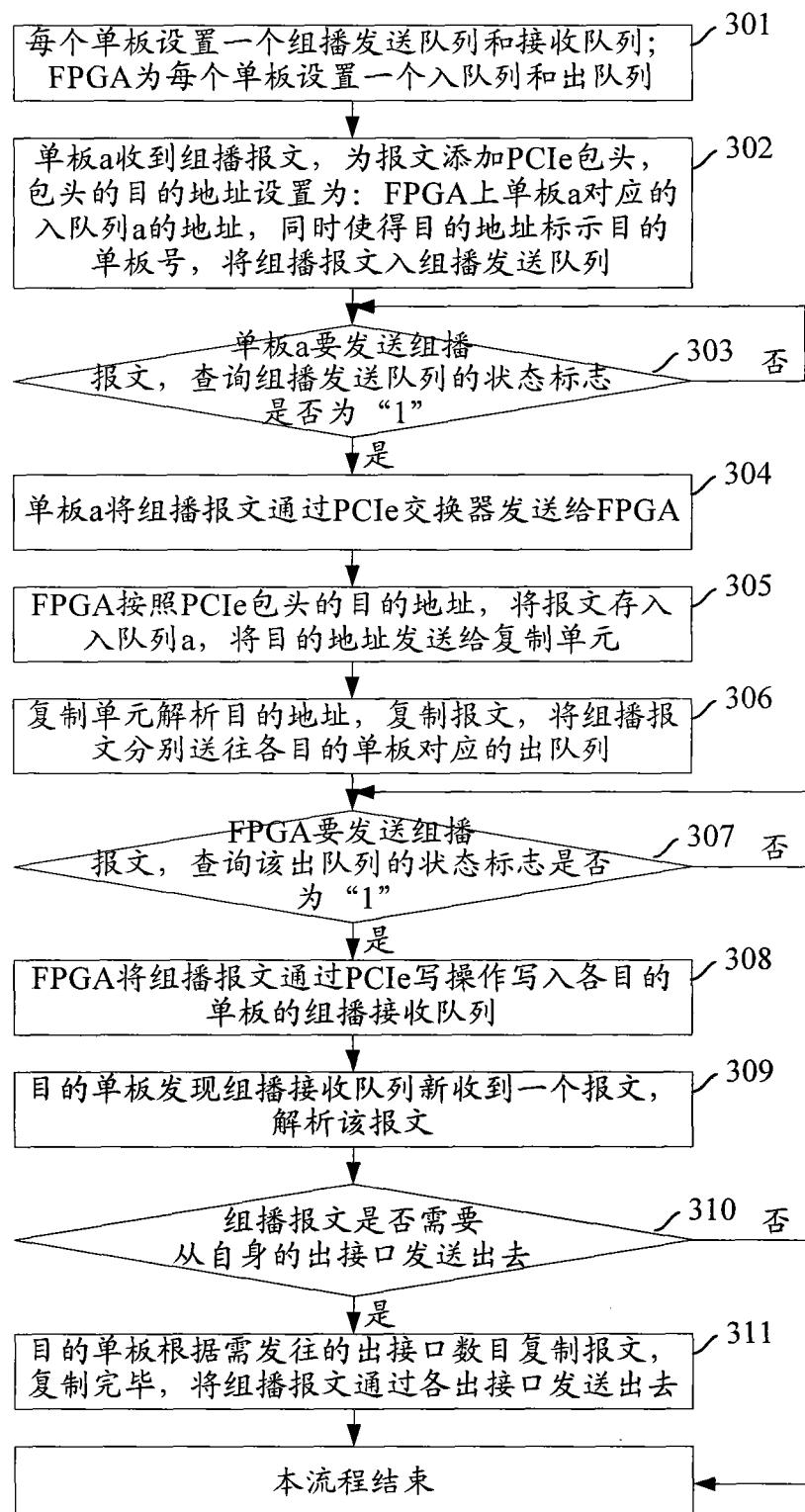


图 3

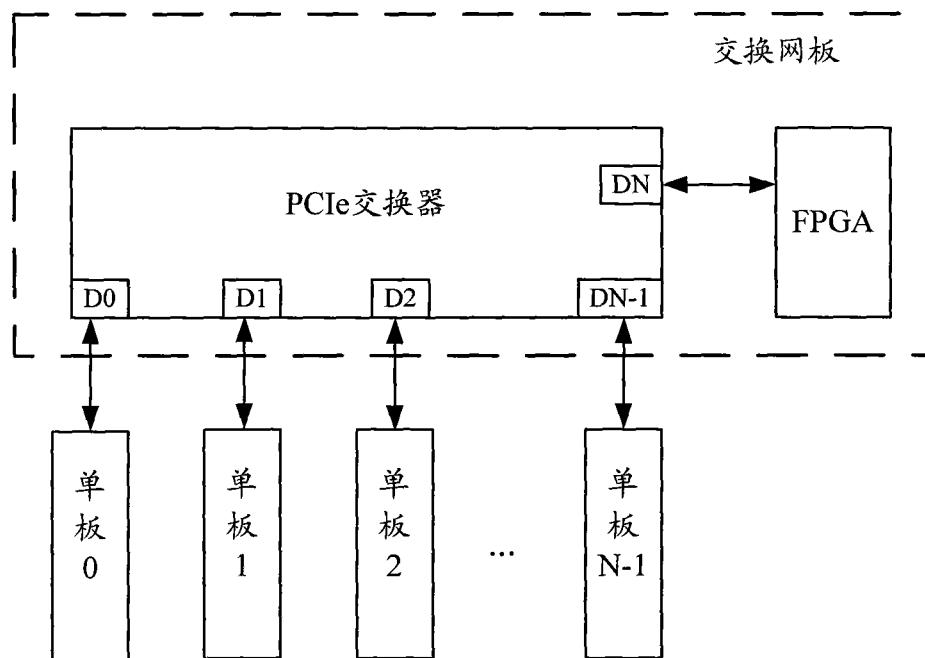


图 4

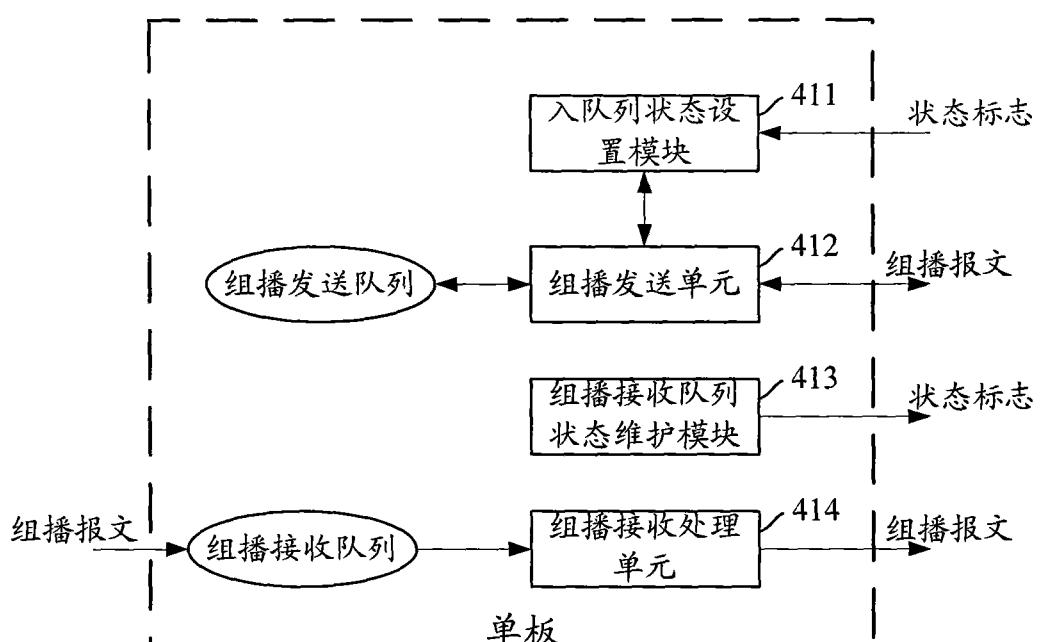


图 5

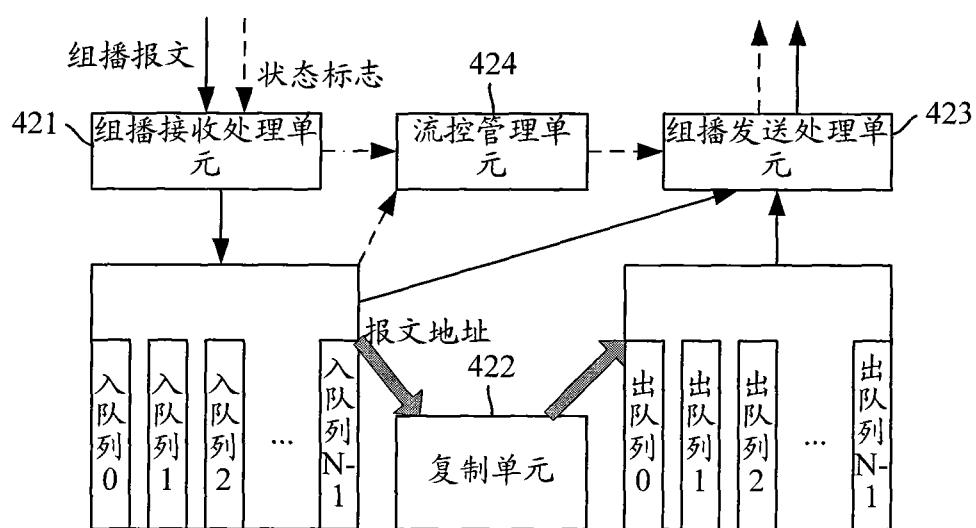


图 6