[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
H04L 12/54

[12] 发明专利说明书

[21] ZL 专利号 00118313.3

[45] 授权公告日 2004 年 1 月 21 日

[11] 授权公告号 CN 1135798C

[22] 申请日 2000.6.12 [21] 申请号 00118313.3

[30] 优先权

[32] 1999.6.12 [33] KR [31] 21940/1999

[32] 1999.12.22 [33] KR [31] 60235/1999

[71] 专利权人 三星电子株式会社

地址 韩国京畿道

[72] 发明人 郑镇宇 禹庆逸 都基钟

审查员 朱 琦

[74] 专利代理机构 中国专利代理(香港)有限公司

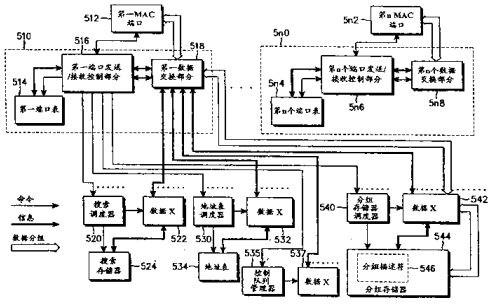
代理人 程天正 傅 康

权利要求书 1 页 说明书 18 页 附图 13 页

[54] 发明名称 数据网络中的分组交换设备和方法

[57] 摘要

数据网络中的分组交换设备包括多个负责输入/输出分组发送/接收命令和数据分组的端口；多个响应分组发送/接收命令访问分成组的信息资源和在分组存储器中存储相应数据分组或向相应端口发送这些数据分组的发送/接收控制部分；多个成组地存储分组交换所需信息并向发送/接收控制部分提供所存储信息的信息资源；和多个与各信息资源相连、用于调度对发送/接收控制部分的访问的信息资源调度器。



ISSN1008-4274

1. 数据网络中的一种分组交换设备包括:

负责输入/输出分组发送/接收命令和数据分组的多个端口;

用于响应分组发送/接收命令来访问分成组的信息资源、和把相应的数据分组存储在分组存储器中或向相应的端口发送存储在分组存储器中的相应的数据分组的多个发送/接收控制部分，所述多个发送/接收控制部分被映射到所述多个端口中的每一个;

多个用于成组地存储分组交换所需的信息并向发送/接收控制部分提供所存储信息的信息资源; 和

多个与各信息资源相连的、用于调度对发送/接收控制部分的访问的信息资源调度器。

2. 权利要求 1 所要求的分组交换设备，其特征在于，其中所述多个发送/接收控制部分被配置成用于映射到所述多个端口中的每一个端口。

3. 权利要求 1 所要求的分组交换设备，其特征在于，其中多个信息资源是分组描述符、链路存储器、搜索存储器和端口表。

4. 权利要求 3 所要求的分组交换设备，其特征在于，其中所述端口表被配置成用于映射到所述多个端口中的每一个。

5. 数据网络中的分组交换方法，包括:

第一步骤: 多个发送/接收控制部分向分成组的信息资源调度器输出相应的访问信号以便访问各信息资源;

第二步骤: 各信息资源调度器执行对访问信号的调度，以便多个发送/接收控制部分能够一次访问其中一个信息资源; 以及

第三步骤: 多个发送/接收控制部分存储接收到的数据分组，或如果访问路径被连接，就发送相对于相应的信息资源所存储的数据分组。

6. 权利要求 5 所要求的分组交换方法，其特征在于，将多个发送/接收控制部分提供给各个端口。

7. 权利要求 5 所要求的分组交换方法，其特征在于，其中多个信息资源是分组描述符、链路存储器、搜索存储器和端口表。

8. 权利要求 7 所要求的分组交换方法，其特征在于，其中将端口表提供给各个端口。

数据网络中的分组交换设备和方法

本发明涉及数据网络中的分组交换系统，更具体而言涉及以并行方式交换分组的设备和方法。

在各种网络（不包括点对点网络）中都存在数据采集和分配设备。其中最好的例子是交换机和路由器。通常这类设备至少有两个端口。设备通过这些端口接收数据，执行必要的数据处理，然后通过一个或多个的端口输出处理后的数据。在这些过程中，肯定会发生拥塞，并且这将引起数据传输等待。在产生拥塞的各种原因中最重要的一个原因是数据处理需要时间。

在分组交换系统中传统的分组处理方法如下所述：首先在步骤1：某个端口接收数据分组，接着在步骤2：先进先出（FIFO）部分暂存输入数据分组，步骤3：输入数据分组等待对先前输入的数据分组的处理，步骤4：数据分组处理部分对存储在FIFO中的输入数据分组执行必要的处理。这时，数据处理需要复杂的确定过程，以及这种确定需要在一个确定执行者（即控制器）和一个信息资源之间传递信息。步骤5：分组处理结束后，数据分组处理部分检查在相应的输出端口是否有其它先前处理过的分组。步骤6：如果有先前处理过的分组，数据分组处理部分就在缓冲区中存储处理过的分组。步骤7：如果先前处理过的分组全部输出了，则数据分组处理部分向输出端口发送存储在缓冲区中的处理过的分组。

根据传统的数据分组处理方法，由于一个数据分组处理部分控制多个端口，并且在一个时刻只能处理一个分组，因此可以用简单的结构实现它。

但是，如果与数据处理时间相比输入的分组数较大时（实际上，大部分分组交换机和路由器有这个特点），数据线实际处于空闲状态，即经过数据线并没有传送数据，而分组处理延时发生在数据分组处理部分。尤其在严重时可能产生数据丢失。

同时，在分组处理中可以考虑两个单元。即：一个用于控制和判断整个处理过程的控制部分，和一个用于存储和提供控制部分进行判断所需的信息的信息资源。在多数情况下，信息资源为寄存器和存储

器的形式。这时，在分组交换系统中为何传统分组数据处理方法在一个时刻只处理一个分组的原因是采用单个存储器实现该资源。

因此，为了解决在相关技术中所涉及的问题和提供快速分组处理，应该把在资源中存储的信息分成组， 以便各信息组存储在不同的资源中，并且应该分配多个发送/接收控制部分（大于各个相应组的资源数）以减少输入数据分组的控制开销。也可以把发送/接收控制部分分配给各端口。发送/接收控制部分通过同时访问各个组的信息资源就能够减少控制开销和加速分组处理。

同时，发送/接收控制部分应当能够共享信息资源。因此判优器和调度器应当使各发送/接收控制部分在一个时刻访问一个资源。如果发送/接收控制部分过多地访问一个指定信息资源，应通过重新调整各个组来保持访问负荷平衡。

图 1 表示了一个传统分组交换设备实施方案的结构。参看图 1，主机 100 控制分组交换设备的全部操作。主机 100 负责最高层，并执行向分组交换设备输入的命令。第一 MAC 端口 110 至第 n MAC 端口 1n0 可被连接到另一个分组交换设备、路由器或 PC，并执行标准 MAC 控制以便把数据分组发送/接收命令输出到发送/接收控制部分 120.数据交换部分 130 在发送/接收控制部分 120 的控制下确定在主机 100、第一 MAC 端口 110 至第 n 端口 1n0、和分组存储器 150 之间的数据和控制信号的路径。数据交换部分 130 可以用复用器/解复用器来实现。

搜索存储器 140 存储用于判断与接收到的分组的目的地址相对应的输出 MAC 端口的信息，从而使注册过的 MAC 地址能够被找出。分组存储器 150 具有多个信息资源，例如地址表 152，端口表 154 和分组描述符 156。分组存储器 150 存储输入的数据分组。地址表 152 存储有关 MAC 地址的信息，端口表 154 存储状态信息、使能信息和关于完成接收操作的信息。分组描述符 156 存储有关存储在分组存储器 150 中各个分组的信息（如分组连接信息）。

发送/接收控制部分 120 根据分组发送/接收命令控制经第一 MAC 端口至第 n MAC 端口 1n0 输入/输出的分组的发送/接收。特别地，发送/接收控制部分 120 暂存接收的数据分组，通过访问搜索存储器 140 检查接收到的分组头的目的地址是否是已注册的地址，并且找出已注

册的 MAC 地址信息存储在地址表 152 的位置。然后发送/接收控制部分 120 确定接收到的分组输出的 MAC 端口。

发送/接收控制部分 120 还通过访问地址表 152、端口表 154 和分组描述符 156 将接收的数据分组存储在分组存储器 150 中。在分组发送期间，通过访问地址表 152、端口表 154 和分组描述符 156，发送/接收控制部分 120 经相应输出端口发送存储在分组存储器 150 中的分组数据。

图 2 表示了另一个传统分组交换设备实施方案的结构。参看图 2，总线接口 212 从主机总线 210 接收一个数据分组，并把该数据分组输出到第一 MAC 端口 211 到第 n 个 MAC 端口 21n。总线接口 212 还向主机总线 210 输出从 MAC 端口发送来的数据分组。第一 MAC 端口 211 到第 n 个 MAC 端口 21n 执行标准 MAC 控制并且向发送/接收控制部分 120 输出数据分组发送/接收命令。MAC 端口接口 238 是各 MAC 端口和发送/接收控制部分 228 之间的接口。MAC 端口接口 238 为每个 MAC 端口提供发送/接收 FIFO，并且暂存子分组。

复用器 224 在从 MAC 端口接口 238 输出的各端口的数据分组中选择相应的数据分组，并向发送/接收控制部分 228 输出相应的数据分组。解复用器 226 解复用从发送/接收控制部分 228 输出的数据分组，并向相应端口输出解复用的数据分组。

搜索存储器 236 存储用于判断与接收到的分组的目的地址相对应的输出 MAC 端口的信息。分组存储器 234 具有多个信息资源如地址表、端口表和分组描述符。分组存储器 234 存储输入的数据分组。

发送/接收控制部分 228 根据分组发送/接收命令控制经第一 MAC 端口到第 n 个 MAC 端口 21n 输入/输出的分组的发送/接收。特别地，发送/接收控制部分 228 暂存接收的数据分组，通过访问搜索存储器 236 检查接收到的分组头的目的地址是否是已注册的地址，并且找出已注册的 MAC 地址信息存储在分组存储器 234 中地址表（未说明）的位置。然后发送/接收控制部分 238 确定接收分组输出的 MAC 端口。

发送/接收控制部分 228 还通过访问分组存储器 234 中的地址表，端口表和分组描述符（未说明）把接收的数据分组存储在分组存储器 234 中。在分组发送期间，通过访问地址表、端口表和分组描述符，发送/接收控制部分 228 经相应的输出端口发送存储在分组存储

器 234 中的分组数据.

如图 1 和 2 所示，根据传统的分组交换设备，由于单个发送/接收控制部分从多个端口接收数据分组发送/接收命令，并且各种信息资源如地址表、端口表等都存储在一个分组存储器中，可知在一个时刻只处理一个分组。因此，尽管数据线实际处于空闲状态，但是在数据分组处理期间会产生分组时延。例如，如果发送/接收控制部分正在执行从某一端口来的命令，从另一端口来的分组就应该等待直到执行命令完成。

图 3 为说明在传统分组交换系统中的接收控制状态流程图。图 1 和 3 中，'Rx 控制'含义是基于在发送/接收控制部分 120 的搜索操作之后得到的信息所执行的一系列操作。特别地，图 3 说明发送/接收控制部分 120 从第一 MAC 端口 110 到第 n 个 MAC 端口 1n0 接收数据分组并将接收到的数据分组存储在分组存储器 150 中的一系列控制操作. 除了处理各种错误、地址不匹配和滤波等以外，图 3 是一个最简单的状态图。假设发送/接收控制部分 120 以频率 50MHz 工作，在图 3 中的每个状态标明了处理 64 字节分组所需的时间。如图 3 所示，可知从空闲状态 300 到分组存储器 150 发送实际数据分组发送的分组发送（Xfer_pkt）状态 332 之间总共有多个控制状态。

下面的表 1 表示由图 3 的传统分组交换方法控制的各状态在接收期间执行的操作。表 1 还表示在相应状态下发送/接收控制部分 120 将会经由数据交换部分 130 访问分组存储器 150、地址表 152 和端口表 154 中的哪一个，以及特别是当发送/接收控制部分以频率 50MHz 工作并接收 64 字节分组时，各状态的数据处理时间。

表 1

状态	操作	资源	时间
取得 Rx 信息	读接收端口表信息	端口表	420ns
Stc 查找	读地址表（源地址）	地址表	300ns
dst 查找	读地址表（源地址）	地址表	320ns
取得 pkt 计数	读 ATM 端口表分组计数 可跳入以太网操作	端口表	40ns

DeQ EB	使空缓冲器散列	端口表	220ns
初始化 desc	初始化分组描述符	分组存储器	200ns
取得当前地址	确定写数据的地址	端口表	20ns
取得 pkt 长度	从分组描述符读各种信息	分组存储器	60ns
更新 scrAT	在源地址表中更新统计值	地址表	80ns
更新 dst AT	在目的地址表中更新统计值	地址表	120ns
Xfer pkt	传递分组（子分组）	分组存储器	460ns
DeQ Rx	使 Rx 队列散列	端口表	40ns
EnQ Tx	使 Rx 队列排队	端口表	80ns

如表 1 所示，可知根据传统的分组交换方法，发送/接收控制部分 120 需要大量的时间以便通过访问端口表 154 和地址表 152 来发送/接收信息，除此之外它还需要有用于在分组存储器 150 中实际存储数据分组的时间。

在一个接收周期期间，各状态访问端口表 154、地址表 152 和分组存储器 150 所需的各时间总结如下：如果实际采用图 1 的传统分组交换设备以 50MHz 的频率进行工作，并接收 64 字节的数据分组，则访问端口表 154 所需的时间总共是 820ns，访问分组存储器所需的时间是 720ns，及访问地址表 152 所需的时间总共是 820ns。

因此，例如如果第一 MAC 端口 110 至第 n 个 MAC 端口 1n0 的发送/接收控制部分 120 是各自独立的，及各 MAC 端口的端口表分布在各自的发送/接收控制部分 120 中，则端口表访问时间将大大减少。实际上，整个接收控制周期所需的时间将减小到 820ns 左右（基于地址表 152 的访问时间）。

如果地址表 152 与分组存储器 150 分开，且各发送/接收控制部分可同时访问地址 152 和/或分组存储器 150，则不同端口的发送/接收控制部分就能够同时访问地址表 152 和分组存储器 150。因此，可以减小分组传输的时延，并使数据有效地传输。特别是，如果保证地址表 152 和各端口的发送/接收控制部分设置在同一芯片内，并且对

地址表 152 的访问是 32 比特或更多，则访问地址表 152 所需的时间将会小于 820ns。因此，在整个接收控制周期内，瓶颈将是访问分组存储器所需的时间（720ns）。换言之，接收控制周期所需的时间将减小到 720ns 以下。

下面的表 2 表示：假设采用图 1 的传统分组交换设备执行发送控制时在各状态执行的操作。

表 2

空闲	读端口表	读pkt desc	Pkt Xfer	更新pkt desc	空闲
	端口表	分组描述符	分组存储器	Pkt desc	
	220ns	300ns	540ns	160ns	

在表 2 中，在读端口状态可以执行以下操作。发送/接收控制部分 120 通过访问端口表 154 来读当前发送地址指针。如果待发送的分组是分组开始（SOP），发送/接收控制部分 120 就初始化端口表 154 的发送字节，并通过访问分组描述符 156 来读分组数据指针。如果待发送的分组是多址传送的，就读多址传送数据指针。

表 2 中在分组发送（Xfer_pkt）状态还可执行以下操作。发送/接收控制部分 120 通过访问分组存储器 150 读待发送的子分组。如果待发送的分组是分组开始（SOP），发送/接收控制部分 120 就使分组存储器 150 中的发送缓冲区散列，并使一个空缓冲区排队。然后发送/接收控制部分 120 减小当前分组计数。如果当前分组计数是‘0’，发送/接收控制部分 120 就禁止相应的端口队列。

同时假设图 1 的分组交换设备执行发送控制操作，与实际数据分组的发送操作相比，控制开销并没有那么大。但是如果以和接收控制操作同样的方式来将控制操作和发送操作相互分开，则能够减小处理数据分组的时间。例如如果在各端口发送/接收控制部分的发送块中提供分组描述符 156，则能够减少整个发送周期所需时间。

图 4 为一定时图，说明图 2 的传统分组交换设备中在 MAC 接口和发送/接收控制部分之间发送和接收分组的情形。在图 2 中，发送和接收的各分组的长度为 64 字节，这样一个分组就成为 SOP 或 EOP。工作频率也是 50MHz，时钟频率为 1/20ns。

发送/接收部分 228 处理来自先前在数据接收状态 424 搜索到的指定 MAC 端口的分组。在搜索和发送状态 426，对下一个将处理的分组执行搜索操作（该分组例如是从另一个 MAC 端口输出、而不是从上面的 MAC 端口输出的分组），并且执行向相应 MAC 端口发送当前待发送的分组的操作。如果搜索和发送状态 426 结束，发送/接收控制部分 228 就进入发送状态 428，并执行分组发送。然后在发送状态 428 结束之后，一个分组处理周期就结束。这时，数据接收状态 424 的时间段是 1480ns，增加的搜索和发送状态及发送状态的时间段是 1520ns。

图 4 中，接收（Rx）控制开销由下式给出：

[等式 1]

（1-320/2480）=87%

这里‘2480’表示数据接收状态 424 的时间段，‘320’表示发送/接收控制部分 228 把从相应的 MAC 端口实际接收的数据分组存储在分组存储器 150 中所需的时间。

图 4 中发送（Tx）控制开销由下式给出：

[等式 2]

（1-320/1520）=79%

这里‘1520’表示增加的搜索和发送状态 426 及发送状态 428 附加的时间段，‘320’表示发送/接收控制部分 228 向相应的 MAC 端口发送来自分组存储器 234 的实际发送数据分组所需的时间。

图 4 中整个控制开销由下式给出：

[等式 3]

（1-640/4000）=84%

这里‘4000’表示一个分组处理周期的时间。‘640’表示通过发送/接收控制部分 228 访问分组存储器 234 发送实际数据分组所需的时间。

从等式 3 可知如果图 2 的传统分组交换设备的分组长度是 64 字节，则控制开销是 84%。具体地，输出、处理和然后输出一个数据分组所需时间的 84%被用于控制操作，其余的 16%用于实际数据发送。

因此，本发明致力于解决相关技术中的问题，本发明的目的是提供一种在数据网络中通过减小控制开销来执行高速分组交换的设

备。

本发明的另一个目的是提供一种在数据网络中快速处理分组的设备，其中控制开销的减小是通过把信息资源分成组、存储多个不同的资源组、并分别通过多个发送/接收控制部分独立地访问这些信息资源而实现的。

本发明的还有一个目的是提供一种在数据网络中快速处理分组的设备，其中控制开销的减小是通过把信息资源分成组、存储多个不同的资源组、并分别通过多个端口发送/接收控制部分独立地访问这些信息资源而实现的。

本发明还有另一个目的是提供一种在数据网络中执行高速分组交换的设备和方法，这是通过把分组交换所需的信息资源例如分组描述符、端口表、链路存储器、地址表等分成组；并通过调度多个发送/接收控制部分的操作来并行地访问这些信息资源而实现的。

为了达到以上目的，根据本发明，数据网络中的一种分组交换设备包括多个负责输入/输出分组发送/接收命令和数据分组的端口；多个用于根据分组发送/接收命令来访问分成组的信息资源、和把相应的数据分组存储在分组存储器中或向相应的端口发送存储在分组存储器中的相应的数据分组的发送/接收控制部分；多个用于成组地存储分组交换所需的信息并向发送/接收控制部分提供所存储信息的信息资源；和多个与各信息资源相连的、用于调度对发送/接收控制部分的访问的信息资源调度器。

根据本发明的另一方面，提供了一种在数据网络中的分组交换方法，它包括：第一步骤：多个发送/接收控制部分向分成组的信息资源调度器输出相应的访问信号以便访问各信息资源，第二步骤：各信息资源调度器执行对访问信号的调度，以便多个发送/接收控制部分能够一次访问其中一个信息资源，第三步骤：多个发送/接收控制部分存储接收到的数据分组，或如果访问路径被连接，就发送相对于相应的信息资源所存储的数据分组。

参考以下附图详细描述本发明的优选实施方案，本发明的上述目的和优势将更加明显：

图 1 为一个传统分组交换设备实施方案的方框图；

图 2 为另一个传统分组交换设备实施方案的方框图；

图 3 说明传统分组交换设备中接收控制状态的流程图;

图 4 为一定时图，说明图 2 的传统分组交换设备中在 MAC 接口和发送/接收控制部分之间发送和接收分组的情形;

图 5 为根据本发明的第一个实施方案的数据网络中的分组交换设备的方框图;

图 6 为根据本发明的第二个实施方案的数据网络中的分组交换设备的方框图;

图 7A 到图 7C 为说明根据本发明的一个实施方案的分组交换设备的整个接收控制操作流程图。

图 8 为说明根据本发明的一个实施方案的分组交换设备的整个发送控制操作流程图.

图 9 为根据本发明的第三个实施方案的数据网络中的分组交换设备的方框图;

图 10 为根据本发明的第四个实施方案的数据网络中的分组交换设备的方框图; 及

图 11 为根据本发明的第五个实施方案的数据网络中的分组交换设备的方框图。

现在将更详细地讨论本发明的优选实施方案. 在以下本发明的描述中，将忽略这里合并的已知功能和结构的详细描述，因为它可能使本发明的主题模糊。以下我们参考附图解释本发明.

图 5 表示根据本发明的第一个实施方案的数据网络中分组交换设备。

第一 MAC 端口 512 至第 n 个 MAC 端口 5n2 可以分别连接到不同的分组交换设备、路由器或 PC 上。第一 MAC 端口 512 至第 n 个 MAC 端口 5n2 执行标准 MAC 控制，并向分别与之相连接的第一端口发送/接收控制部分 516 至第 n 个端口发送/接收控制部分 5n6 输出相应的分组发送/接收命令。MAC 端口还向分别与之相连接的发送/接收控制部分发送所接收到的数据分组，并从相应的发送/接收控制部分向相应的协议控制部分输出数据分组。协议控制部分可由其它分组交换设备、路由器或 PC 提供。

第一端口发送/接收控制部分 516 至第 n 个端口发送/接收控制部分 5n6 根据从相应的 MAC 端口输出的分组发送/接收命令来执行分组

发送/接收控制。

第一数据交换部分 518 至第 n 个数据交换部分 5n8 在相应的端口发送/接收控制部分的控制之下提供数据分组和控制信号路径。第一端口表 514 至第 n 个端口表 5n4 存储关于相应 MAC 端口的信息，并在各端口上进行分配。有关各 MAC 端口的信息相互独立，不需要其它端口分享这些信息。图 5 中第一端口表 514 至第 n 个端口表 5n4 与相应端口发送/接收控制部分相连。

应当认为当第一端口发送/接收控制部分 516 至第 n 个端口发送/接收控制部分 5n6 对相应的端口执行分组的发送/接收控制时，信息资源被分成组。根据图 5 的实施方案，这些组是第一端口表 514 至第 n 个端口表 5n4 和搜索存储器 524、地址表 534、以及控制队列管理器 535 和分组存储器 544。第一端口发送/接收控制部分 516 至第 n 个端口发送/接收控制部分 5n6 通过独立地访问搜索存储器 524、地址表 534、控制队列管理员 535、和分组存储器 544，从而将接收到的分组存储在分组存储器 150 中或者把存储在分组存储器 150 中的数据分组通过相应的输出端口进行发送。也即，为了执行分组接收/发送控制，各端口发送/接收控制部分在分开的信息资源中访问除了相应的端口表之外的四个调度器。各调度器可以采用循环（Round-Robin）系统。

搜索调度器 520 使得第一端口发送/接收控制部分 516 至第 n 个端口发送/接收控制部分 5n6 能够共享搜索存储器 524。也即，搜索调度器 520 在某个时刻只使能一个端口发送/接收控制部分访问搜索存储器 524。采用与搜索调度 520 同样的方式，地址表调度 530 和分组存储器调度 540 也使得第一端口发送/接收控制部分 516 至第 n 个端口发送/接收控制部分 5n6 能够共享地址表 534 和分组存储器 544。控制队列管理器 535 使得第一端口发送/接收控制部分 516 至第 n 个端口发送/接收控制部分 5n6 共享其本身。控制队列管理器 535 还存储对于分组存储器 544 的排队操作的各队列的指针信息，并且根据排队操作更新指针信息。控制队列管理器 535 还向所选的相应端口发送/接收控制部分输出指针信息，以便端口发送/接收控制部分执行排队操作。

同时，图 5 的结构除了分组存储器 544 之外可以被制成一个芯片。分组描述符 546 可以从分组存储器 544 中分离出来以便各端口发

送/接收控制部分能共享它，并且各端口发送/接收控制部分能够通过创建多个任务来执行发送/接收控制。

现在，我们将解释上面建造的根据本发明实施方案的分组交换设备的操作。各端口发送/接收控制部分暂存所接收到的数据分组。各端口发送/接收控制部分还检查所接收到的分组头的目的地址是否是已注册的地址，并经搜索调度器 520 访问搜索存储器 524 从而找出注册的 MAC 地址信息被存储在地址表 534 中的位置。然后，各端口发送/接收控制部分确定接收到的分组将输出到哪个 MAC 端口。

各端口发送/接收控制部分通过经地址表调度器 530 访问地址表 534 来检查接收到的分组的源地址和目的地址。之后，各端口发送/接收控制部分通过访问与其本身直接相连的相应端口表来检查端口信息，并且通过经地址表调度器 530 和分组存储器调度器 540 访问地址表 534 和分组描述符 540 来检查 MAC 地址信息和分组信息。然后各端口发送/接收控制部分把暂存的分组存储在分组存储器 546 中。在分组发送时，各端口发送/接收控制部分根据与其本身相连的相应地址表，通过经由地址表调度器 530 和分组存储器调度器 530 访问地址表 534 和分组描述符 546，从而经相应的输出端口发送存储在分组存储器 544 中的数据分组。

之后各端口发送/接收控制部分执行错误检测操作。特别地，如果相对于分组出现了 MAC 错误、未知的源地址、地址变动和目的地址，则各端口发送/接收控制部分作出舍弃、广播或向主机发送的判定。

同时我们将解释各端口发送/接收控制部分把数据分组存储在分组存储器 544 中、并向相应的 MAC 端口输出存储在分组存储器 544 中的数据分组这一过程的例子。

如果接收到分组，当各端口发送/接收控制部分访问相应的端口表时就存储接收到的分组，并根据控制队列管理器 535 的指针信息使空缓冲区散列并使接收队列排队。然后各端口发送/接收控制部分通过访问相应的端口表并根据控制队列管理器 535 的指针信息执行排队操作，从而利用该指针连接存储在分组存储器中的分组。分组的信息存储在分组描述符 546 中。当从相应端口接收到分组时，如果当前处理的分组是 EOP，各端口发送/接收控制部分根据控制队列管理器 535 的指针信息使接收（Rx）队列排队。如果当前处理的分组是 EOP，端

口发送/接收控制部分根据控制队列管理器 535 的指针信息使接收队列散列，并使在分组存储器 544 中提供的发送（Tx）队列排队。

在分组发送时，各发送/接收控制部分通过访问分组描述符 546 查阅分组信息，并且向输出端口的 MAC 发送存储在分组存储器中的相应分组。这时各端口发送/接收控制部分访问相应的端口表，根据控制队列管理器 535 的指针信息使发送（Tx）队列散列并使空缓冲器排队。

图 6 表示根据本发明第二个实施方案的数据网络中的分组交换设备。

图 6 的分组交换设备可以经总线接口 600 连接到一个主机（未示出）和多个分组交换设备（未示出）。图 6 的分组交换设备还可以经总线接口 600 连接到路由器或 PC。

第一 MAC 端口 604 至第 n 个 MAC 端口 606 执行标准的 MAC 控制，并输出分组发送/接收命令。各 MAC 端口负责数据分组的输入/输出。尤其是，各 MAC 端口向与之连接的发送/接收控制部分发送接收到的数据分组，并从相应的发送/接收控制部分向相应的协议控制部分输出数据分组。各 MAC 端口能够执行全双工操作或半双工操作。各 MAC 端口安置在分组交换设备之外。

第一 MAC 接口部分 608 至第 n 个接口部分 614 分别作为 MAC 端口和端口发送/接收控制部分之间的接口，并负责子分组发送。各 MAC 接口部分具有发送和接收 FIFO，并能暂存子分组。当能够发送或接收时，各 MAC 接口部分向相应的端口发送/接收控制部分输出分组发送/接收命令。

第一端口发送/接收控制部分 620 至第 n 个端口发送/接收控制部分 624 可以提供给每个 MAC 端口。各端口发送/接收控制部分有第一端口表 622 至第 n 个端口表 626。如果输入分组发送/接收命令，各端口发送/接收控制部分通过访问所提供的端口表执行地址搜索操作。各端口发送/接收控制部分还向分组存储器调度器 628、地址表调度器 630 或搜索调度器 632 输出连接请求信号，以便访问地址存储器 642、地址表 644 或搜索存储器 646。如果完成与所需信息资源的连接，则各端口发送/接收控制部分执行子分组发送、各子分组的 SOP 处理和 EOP 处理、分组排队和分组散列。各端口发送/接收控制部分还更新有

关源/目的地址的统计信息.

分组存储器调度器 628 被连接到各端口发送/接收控制部分. 地址表调度器 630 通过从各端口发送/接收控制部分调度连接请求信号而将所选的相应端口发送/接收控制部分连接到地址表 644.在本发明的实施方案中，分组存储器调度 628 可以控制控制队列管理器 634 提供的空队列、‘0’和‘1’的主机队列和多址传送队列。分组存储器调度器 628 还可以控制对控制队列管理器 634 的队列(例如接收(Rx)队列、发送(Tx)队列等)的排队和散列操作。

控制队列管理器 634 与各端口发送/接收控制部分相连，并通过调度从各端口发送/接收控制部分来的连接请求信号而执行对例如 Rx 队列、Tx 队列等队列的排队和散列操作。队列管理器 634 具有空队列、多址传送队列、0’和‘1’的主机队列和扩充队列。如果主机(未示出)与总线接口 600 相连，则扩充队列就包括主机队列.

地址表调度器 630 与各端口发送/接收控制部分相连。地址表调度 630 通过调度来自各端口发送/接收控制部分的连接请求信号而将所选的相应端口发送/接收控制部分与地址表 644 相连接。

搜索调度器 632 与各端口发送/接收控制部分相连。搜索调度器 632 通过调度来自各端口发送/接收控制部分的连接请求信号而将所选的相应端口发送/接收控制部分与搜索存储器 646 相连。

分组存储器接口 636 是分组存储器调度器 628 和分组存储器 642 之间的接口。地址表接口 638 为地址表调度器 630 和地址表 644 之间的接口。搜索存储器接口 640 为搜索调度器 632 和搜索存储器 646 之间的接口。

第一端口表 622 至第 n 个端口表 626 中存储状态信息、使能信息和有关各 MAC 端口完成接收操作的信息。分组存储器 642 中存储子分组，及分组描述符 648 中存储有关各子分组的信息。地址表 644 中存储已注册分组的目的 MAC 地址的源 MAC 地址信息。并且，搜索存储器 646 存储用于判断与接收到的分组的目的地址对应的输出 MAC 端口。

下面我们将解释图 6 所示的根据本发明实施方案的分组交换设备的操作。接收控制指的是把在各端口的 MAC 接口部分中存储的子分组存储在分组存储器 642 中的过程。如果接收到的子分组被输入到 FIFO 中去，则相应的 MAC 接口部分输出相应的分组接收命令. 如果输入该

命令，则端口发送/接收控制部分就检查接收到的子分组的头信息以获得所需的信息。

如果接收到的子分组与 SOP 一致，端口发送/接收控制部分就通过经由搜索调度器 632 访问搜索存储器而执行搜索操作。这时，在本发明的实施方案中，当执行搜索操作时，如果输入分组发送命令与搜索操作不一致，则端口发送/接收控制部分可以部分地执行分组发送命令。

端口发送/接收控制部分基于从搜索操作、地址表 644 和相应端口表获得的信息来运行状态机并因此传递所需的判断。接收到的子分组存储在分组存储器 646 中。

同时，各状态向分组存储器 628、地址表调度器 630 和队列管理器 634 输出各自所需的命令，并且如果相应的命令被调度器所选择，就通过访问分组描述符 648 和地址表 644 而获得所需的信息。端口发送/接收控制部分通过向分组存储器调度器 628 请求分组发送（Xfer_pkt）命令而把相应的 MAC 接口中存储的子分组存储在分组存储器 642 中的。如果接收到的子分组与 EOP 一致，即如果整个帧（例如以太网帧）的分组存储完成，端口发送/接收控制部分就对目的 MAC 端口使接收（Rx）队列散列，并使发送（Tx）队列排队。

同时，如果完成对于整帧子分组的接收，相应的发送/接收控制部分根据相应目的 MAC 端口的命令而执行分组发送控制。这时，在子分组单元中执行分组发送，并且发送期间所需的信息从分组描述符 648 和相应端口表中获得。

图 7A 至 7C 为说明根据本发明实施方案的分组交换设备的全部接收控制操作的流程图。这里所示的各个过程与分组交换系统中的一般数据分组过程一致。

图 8 为说明根据本发明实施方案的分组交换设备所执行的全部发送控制过程流程图。这里所示的各个过程与分组交换系统中的一般数据分组过程一致。

图 9 说明根据本发明的第三个实施方案的数据网络中的分组交换设备的结构。图 9 的设备结构与图 5 的设备结构相似，但是根据图 9 的设备，在图 5 的分组存储器 544 中的信息资源即分组连接信息是分开的。明确地说，链路存储器 934 与分组存储器 944 分开，并且其中

存储分组连接信息。分组连接信息可以由下一个描述符和发送队列指针组成。各下一描述符与分组存储器944的各空间对应，并且可以具有下一个链接空间的地址信息。发送队列指针可以含有一个头、尾信息和有关相应队列的当前空间数的信息。

控制队列管理器930使得第一端口发送/接收控制部分916至第n个端口发送/接收控制部分能共享其本身。控制队列管理器930通过访问链路存储器934而咨询和更新分组连接信息，存储用于对分组存储器944的进行排队操作的各队列的指针信息，并根据排队操作更新指针信息。控制队列管理器930还向所选的相应端口发送/接收控制部分输出指针信息，并且这使得该端口发送/接收控制部分执行排队操作。

再看图9，搜索存储器924包括了图5的地址表534。根据图9的结构，各端口发送/接收控制部分通过访问搜索调度器920来获得地址信息。

图10说明根据本发明的第四个实施方案的数据网络中的分组交换设备的结构。图10的设备结构与图6的类似，但是根据图10的设备，在图6的分组存储器642中提供的分组信息资源即分组连接信息是分开的。明确地说，链路存储器1044与分组存储器1042分开，并且其中存储分组连接信息。分组连接信息可以由下一个描述符和发送队列指针组成。各下一描述符与分组存储器1042的各空间对应并且具有下一个链接空间的地址信息。发送队列指针可以含有一个头、尾信息和有关相应队列的当前空间号码的信息。

控制队列管理器1030使得第一端口发送/接收控制部分916至第n个端口发送/接收控制部分1024共享其本身。控制队列管理器1030与各端口发送/接收控制部分相连。控制队列管理器1030从各端口发送/接收控制部分调度连接请求信号，并执行对各队列（如接收（Rx）队列、发送（Tx）队列等）的排队和散列操作。

控制队列管理器1030还通过访问链路存储器1044来咨询和更新分组连接信息，存储用于对分组存储器1044进行的排队操作的各队列的指针信息，并根据排队操作更新指针信息。

再看图10，搜索存储器1046包括了图6的地址表644。根据图10的结构，各端口发送/接收控制部分通过访问搜索调度器1032获得

地址信息。

同时，分组交换设备的性能可以由各种因素评估，其中最重要一个因素是吞吐量。这里吞吐量指的是每单位时间所能处理的数据量。由于可变分组（例如以太网分组）的长度，64 字节长度分组的处理能力（这是最低的吞吐量）可以一般地表达分组交换设备的性能。特别地，当 64 字节单点传送分组输入到所有端口并且然后从不同于输入端口的端口输出时，如果输入速率与输出速率相等时，则分组交换机支持线路速率。

下面的表 3 说明具有图 10 结构的分组交换设备对单个标准单点传送分组的发送/接收处理的例子。

表 3

接收控制操作

状态	描述	所需时钟	处理块
DeQ EB, EnQ Rx	在空缓冲区中提取一个空间以便在其中存储接收到的分组	14	COM
Inin Desc	存储描述符，它是各分组的信息	Lclk[1clk(调度 + 3clk(在 PCU 和 PMI 之间交换信号) +4clk(数据 Xfer) +3clk(预充电 SGRAM)]	PMI
Xfer 分组	在分组存储器中存储实际数据	23(1+3+16+3)	PMI
DeQ Rx EnQ Tx	在 Tx 队列中存储完成接收的空间	17	CQM

接收控制操作

状态	描述	所需时钟	处理块
读 Des	待发送的分组信息	12clk[1clk(调度)+ 4clk(信号交换)+ 4clk(数据 Xfer)+ 3clk(预充电 SQGAM)]	PMI
Xfer 分组	读待发送数据	24(1+4+16+4)	PMI
DeQ Rx EnQ Tx	使空缓冲区中的完成发送空间排队以便将其再次用作为空的空间	20	CQM

表 3 中，'CQM' 是控制队列管理器的缩写，及 'PMI' 是分组存储器接口的缩写。另外，'PCU' 是端口控制单元的缩写，它指的是端口发送/接收控制部分。

从表 3 中可知，分组存储器接口处理单个单点传送分组所需的时间总共是 70 个时钟，控制队列管理器所需的时间总共是 51 个时钟。因此整个处理中的瓶颈是 PMI，结果发送/接收 64 字节分组所需的时间总共是 70 个时钟。同时，假设是 64 字节的以太网分组，各单点传送分组包括一个 12 字节的帧间间歇和 8 字节的分组头，这样就包括 672 比特（=84×8 比特）。

因此，如果图 10 的设备的 MAC 端口数是 8，则 66MHz 频率的 672 比特分组的吞吐量可以表达为

[等式 4]

（672 比特×66Mclk 每秒/70clk）×2（包括 Rx 和 Tx）=1.267Gbps

在等式 4 中，假设 MAC 端口数是 8，图 10 的分组交换设备的处理速率应当为 1.6Gbps 以便支持线路速率。

图 11 说明根据本发明的第五个实施方案的数据网络中的分组交换设备。图 11 设备的结构与图 10 的类似，但是根据图 11 的设备，在图 10 的分组存储器 1042 中的信息资源即分组描述符 1048 是分开的。也即，增加了一个单独存储分组描述符的存储器。这个增加的存储器可在芯片的内部或外部，这使得分组存储器 1144 的负荷降低，

在图 10 的分组存储器 1042 中的信息资源即分组描述符 1048 是分开的。也即，增加了一个单独存储分组描述符的存储器。这个增加的存储器可在芯片的内部或外部，这使得分组存储器 1144 的负荷降低，因此减小了发送/接收控制时间。存储在分组描述符存储器 1146 的各分组信息能够以一对一的方式映射到实际存储在分组存储器 1144 中的各分组上。这样如果知道在存储器中的地址，也总是能知道另一个地址。

看图 11，分组描述符存储器 1146 与分组存储器 1144 分开，其中存储各分组的信息。分组描述符存储器调度器 1130 与各端口发送/接收控制部分相连。分组描述符存储器调度器 1130 从各端口发送/接收控制部分调度连接请求信号，并访问存储在分组描述符存储器 1146 中各分组的信息。

图 11 的分组交换设备对单点传送分组的发送/接收处理所需的时间如下：在分组发送（Tx）和接收（Rx）期间，分组描述符存储器接口 1138 单纯传输分组需要 47 个时钟，及分组描述符存储器接口 1138 初始化和访问分组描述符存储器 1146 需要 23 个时钟。如上所述，控制队列管理器 1132 也需要 51 个时钟。图 11 的拥塞瓶颈是控制队列管理器 1132，并且这时吞吐量可以表达为

[等式 5]

（672 比特×66Mclk 每秒/51clk）×2（包括 Rx 和 Tx）=1.74Gbps

在等式 5 中假设 8 个 MAC 端口，图 11 的对单点传送分组的分组交换设备处理速率等于或大于 1.6Gbps，这样，它可支持线路速率。

如上所述，根据本发明的分组交换设备和方法，将分组交换所需的信息资源（例如分组描述符、端口表、链路存储器、地址表等分成组，通过调度多个发送/接收控制部分而并行地访问信息资源，从而可以减小控制开销。因此利用结构上修改的传统分组交换设备，本发明能够进行高速分组交换。

虽然已经结合当前被认为是最实际和优选的实施方案描述本发明，但是应该理解可以在不偏离本发明范围的情况下对其作其它的修改。因此本发明不应当局限于所发表的实施方案，而应当由附加的权利要求和其等同物所定义。

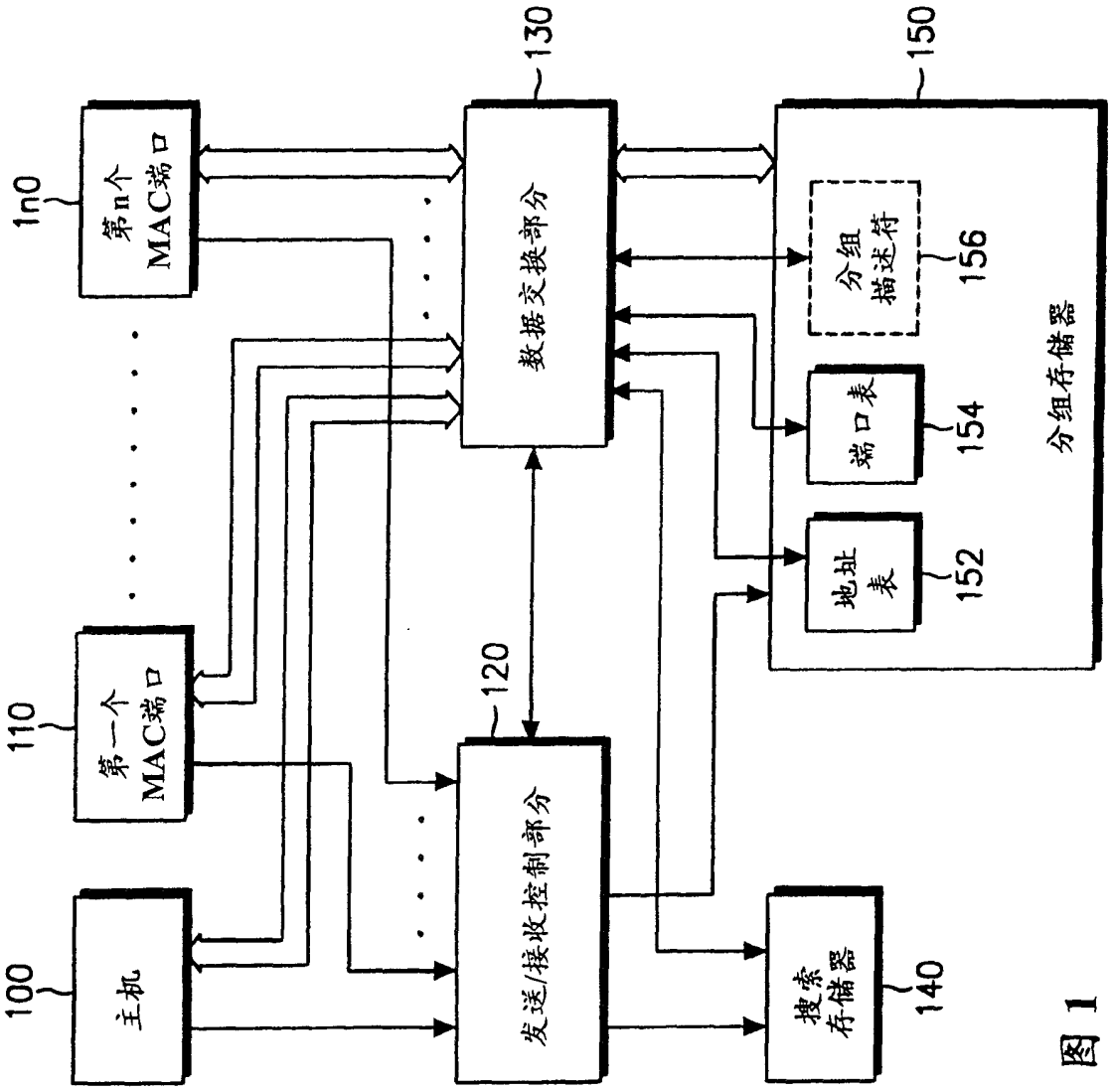

主机
100
第一个MAC端口
110
第n个MAC端口
1n0
发送/接收控制部分
120
数据交换部分
130
搜索存储器
140
分组存储器
150
地址表
152
端口表
154
分组描述符
156

图 1

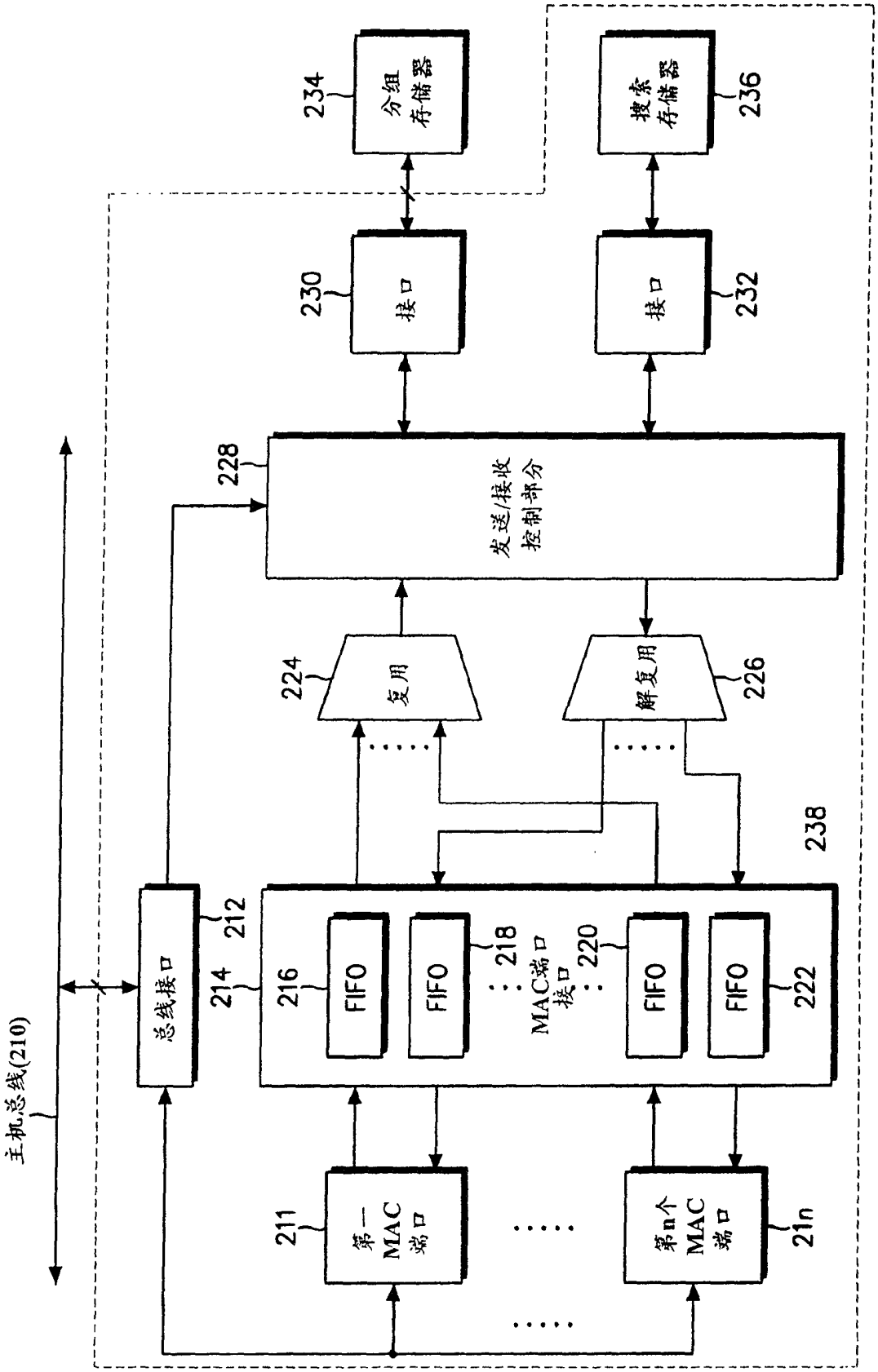

主机总线(210)
总线接口
212
214
216
FIFO
FIFO
218
MAC端口接口
220
FIFO
FIFO
222
238
211
第一MAC端口
第n个MAC端口
21n
224
复用
解复用
226
228
发送/接收控制部分
230
接口
232
接口
234
分组存储器
236
搜索存储器

图 2

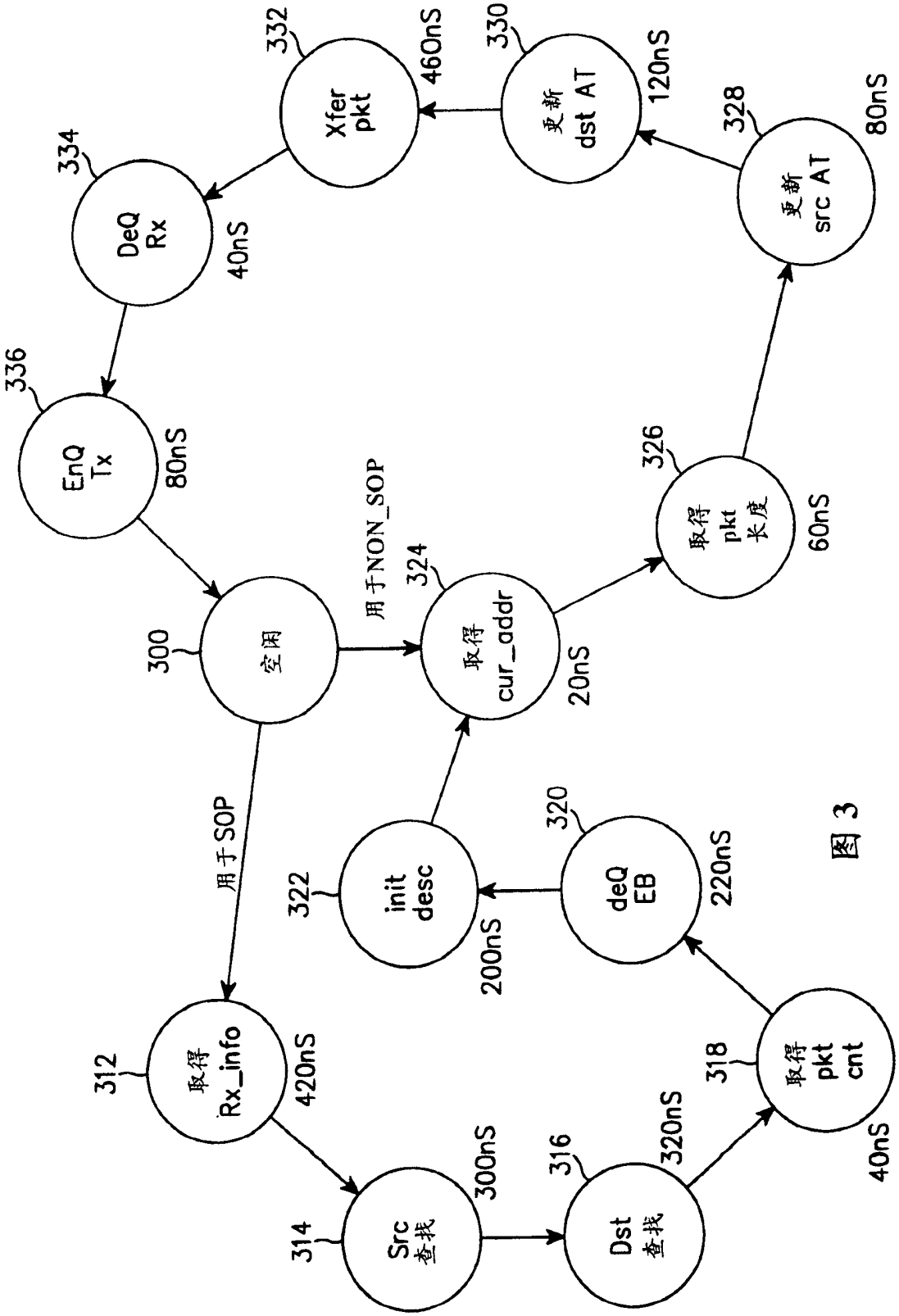

300
空闲
312
取得 Rx_info
420nS
314
Src 查找
300nS
316
Dst 查找
320nS
318
取得 pkt cnt
40nS
320
deQ EB
220nS
322
init desc
200nS
324
取得 cur_addr
20nS
326
取得 pkt 长度
60nS
328
更新 src AT
80nS
330
更新 dst AT
120nS
332
Xfer pkt
460nS
334
DeQ Rx
40nS
336
EnQ Tx
80nS
用于SOP
用于NON_SOP

图 3

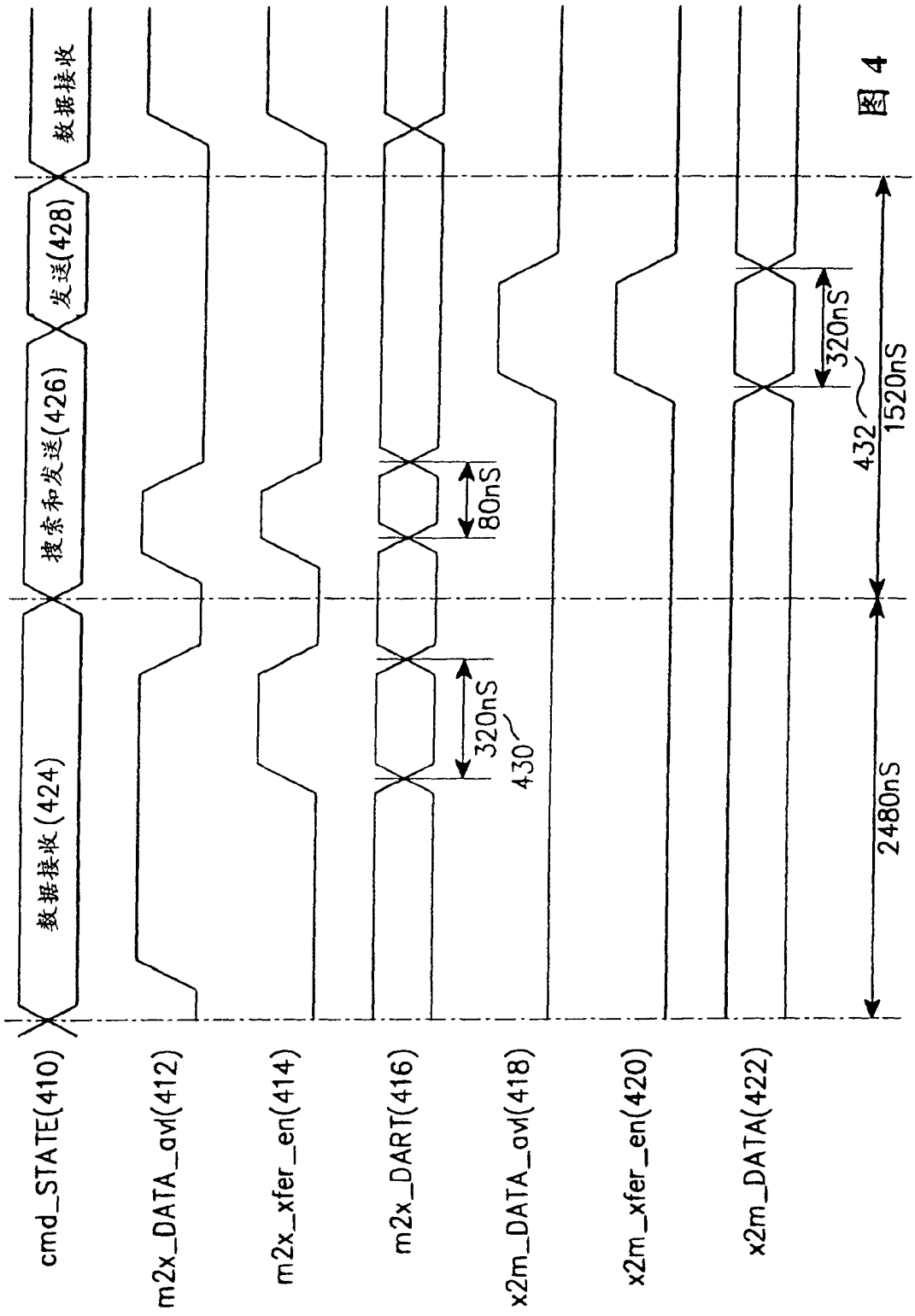

cmd_STATE(410)
m2x_DATA_avl(412)
m2x_xfer_en(414)
m2x_DART(416)
x2m_DATA_avl(418)
x2m_xfer_en(420)
x2m_DATA(422)
数据接收(424)
搜索和发送(426)
发送(428)
数据接收
320nS
430
80nS
320nS
432
1520nS
2480nS

图 4

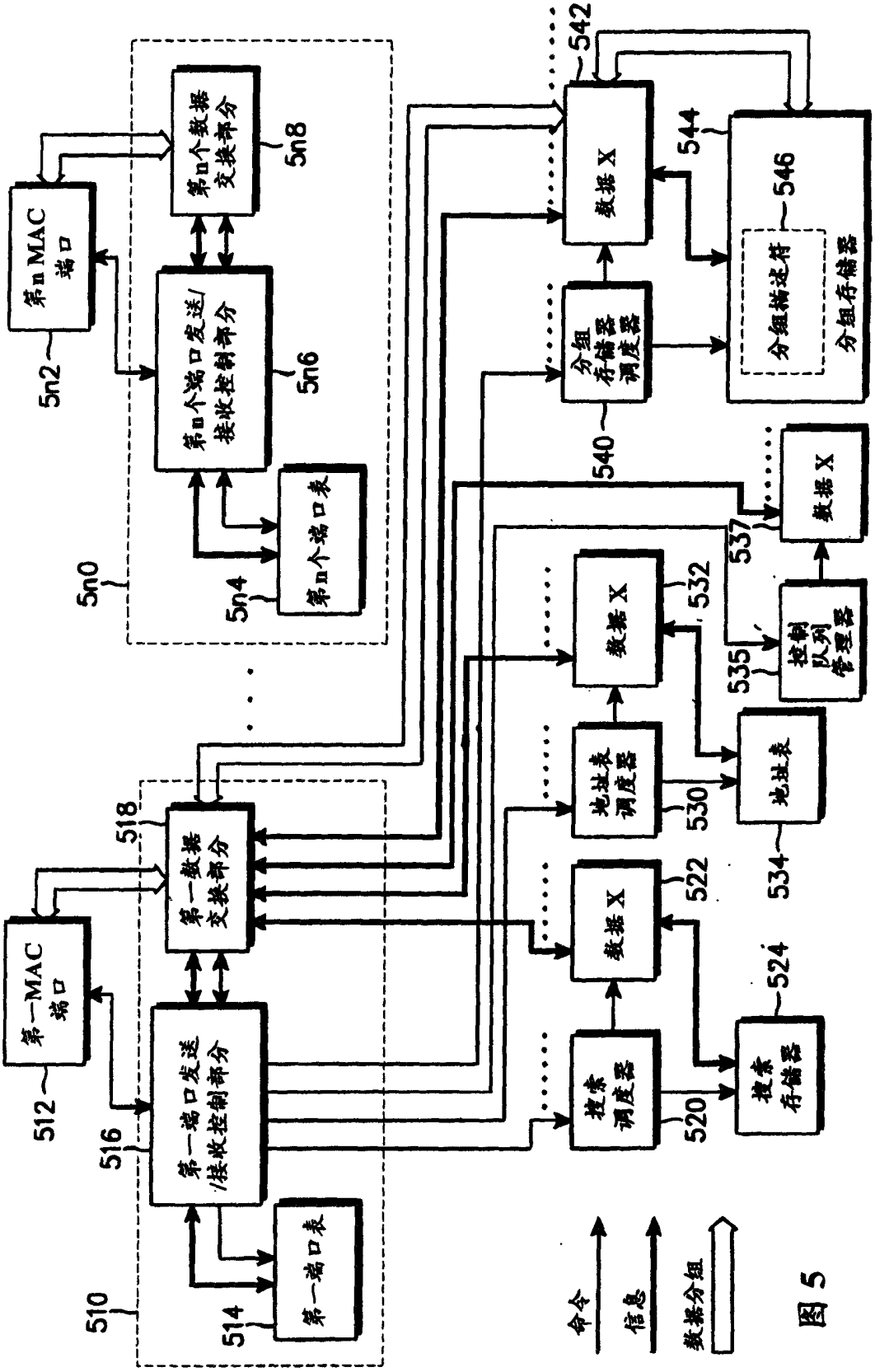

第一MAC端口
第一端口发送/接收控制部分
第一端口表
第一数据交换部分
第n MAC端口
第n个端口发送/接收控制部分
第n个端口表
第n个数据交换部分
5n0
5n2
5n4
5n6
5n8
510
512
514
516
518
搜索调度器
搜索存储器
地址表调度器
地址表
分组存储器调度器
分组存储器
分组描述符
控制队列管理器
数据X
520
522
524
530
532
534
535
537
540
542
544
546
命令
信息
数据分组

图 5

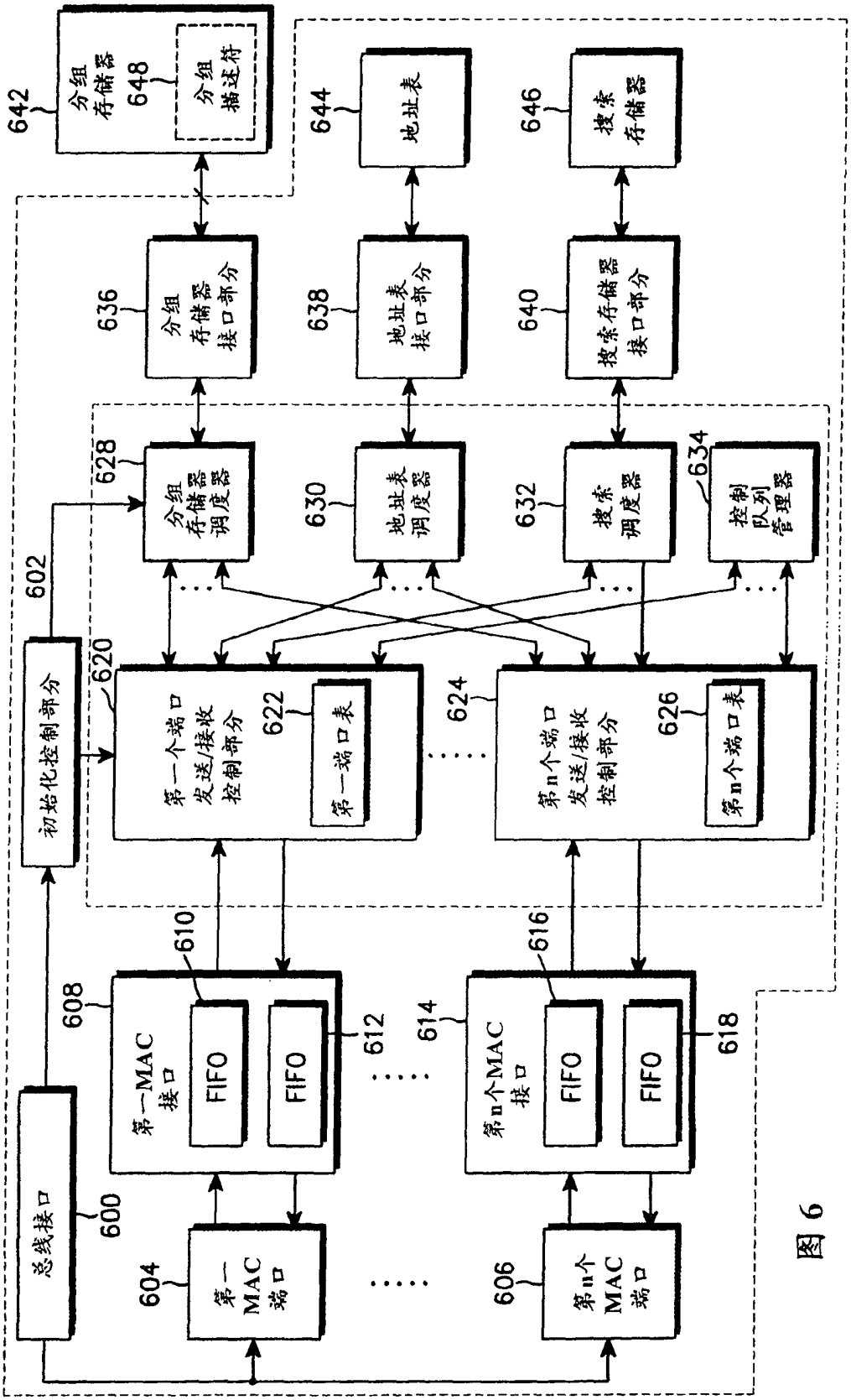

642
分组存储器 648
分组描述符
644
地址表
646
搜索存储器
636
分组存储器接口部分
638
地址表接口部分
640
搜索存储器接口部分
628
分组存储器调度器
630
地址表调度器
632
搜索调度器
634
控制队列管理器
602
初始化控制部分
620
第一个端口发送/接收控制部分
622
第一端口表
624
第n个端口发送/接收控制部分
626
第n个端口表
610
608
第一MAC接口
FIFO
FIFO
612
614
616
第n个MAC接口
FIFO
FIFO
618
总线接口
600
604
第一MAC端口
606
第n个MAC端口

图 6

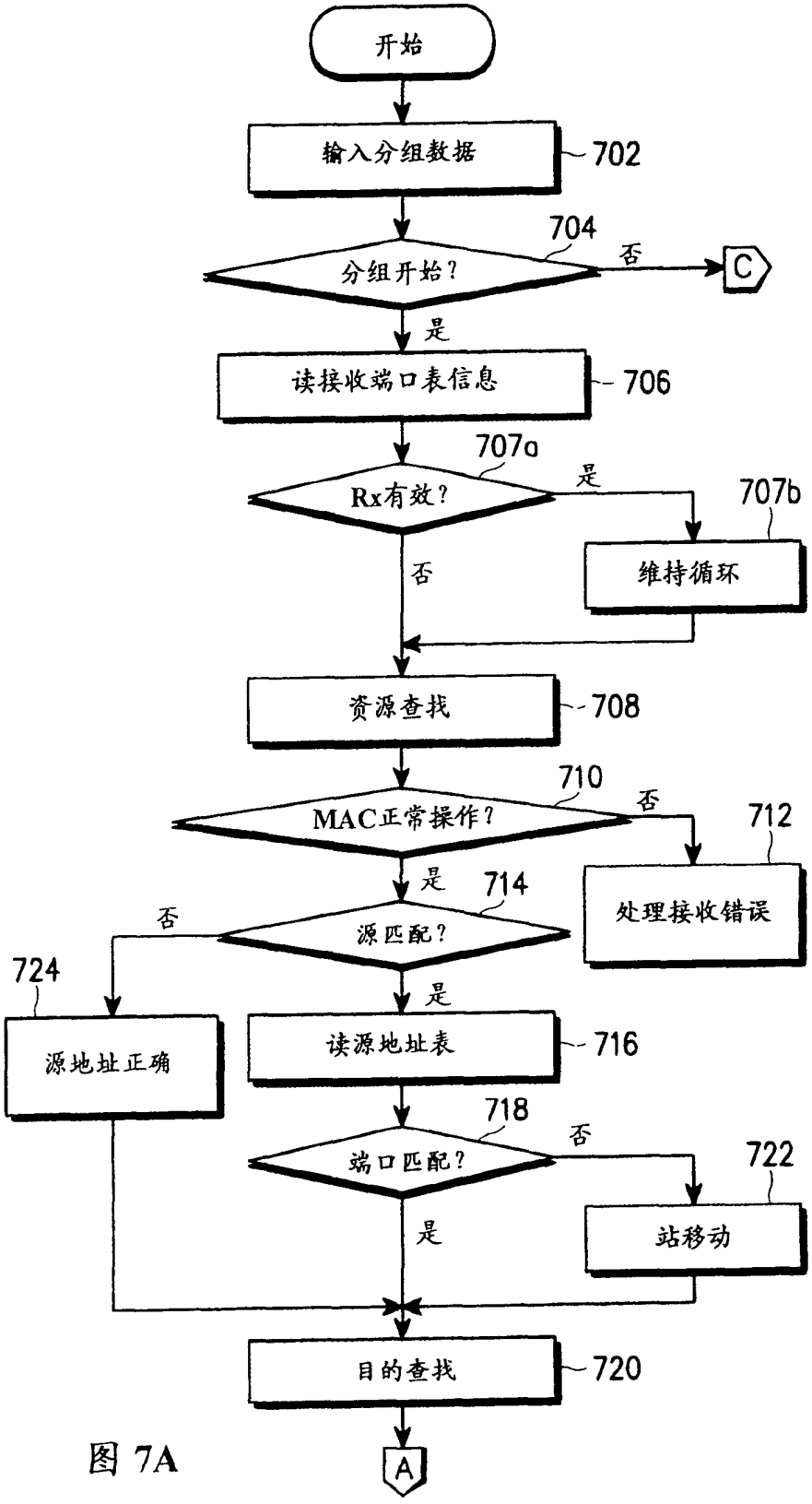

开始
输入分组数据
702
704
分组开始?
否
C
是
读接收端口表信息
706
707a
Rx有效?
是
707b
维持循环
否
资源查找
708
710
MAC正常操作?
否
712
处理接收错误
是
714
源匹配?
否
724
源地址正确
是
读源地址表
716
718
端口匹配?
否
722
站移动
是
目的查找
720
A

图 7A

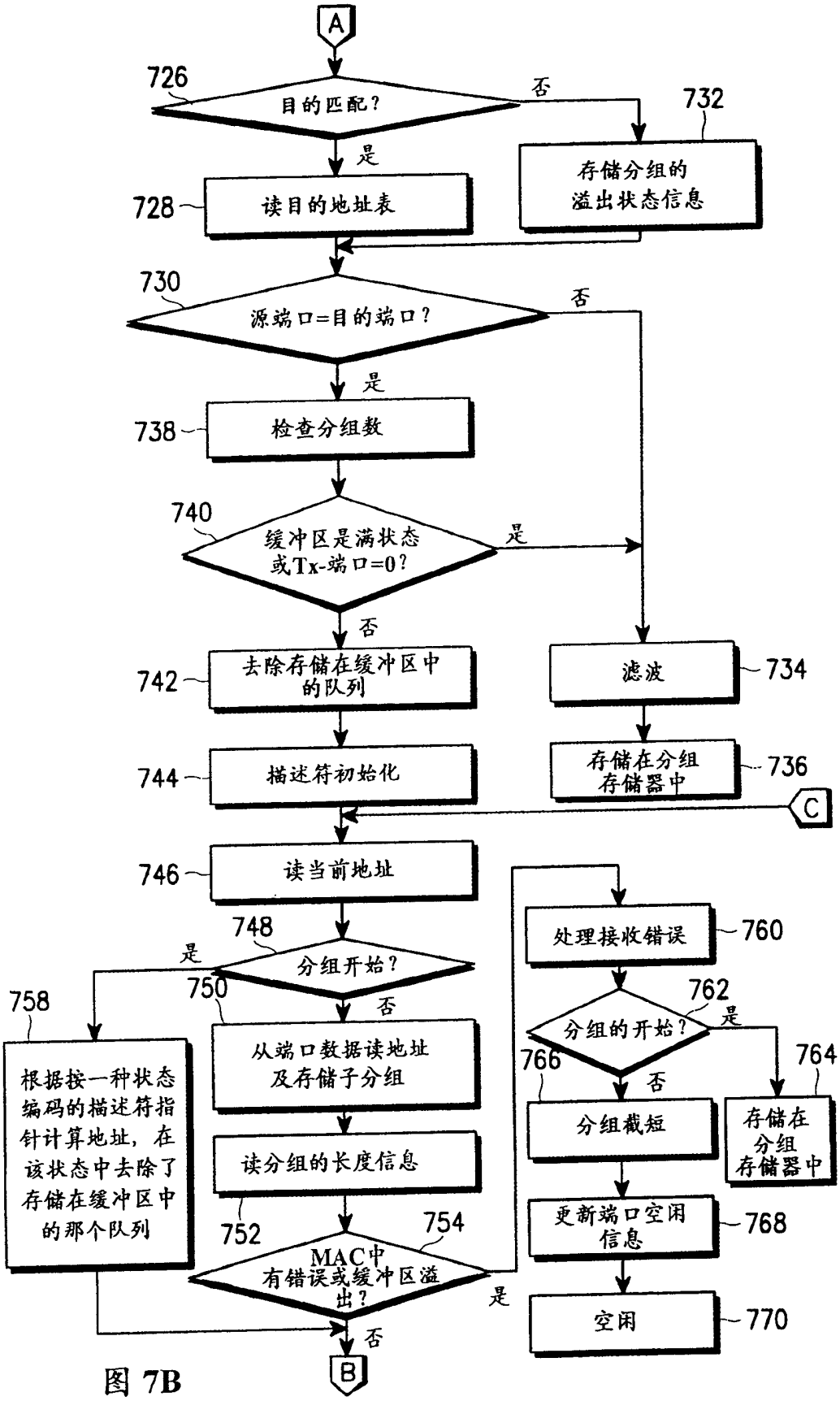

A
726
目的匹配？
否
732
存储分组的
溢出状态信息
是
728
读目的地址表
730
源端口=目的端口？
否
是
738
检查分组数
740
缓冲区是满状态
或Tx-端口=0？
是
否
742
去除存储在缓冲区中
的队列
滤波
734
744
描述符初始化
存储在分组
存储器中
736
C
746
读当前地址
748
是
分组开始？
750
否
758
根据按一种状态
编码的描述符指
针计算地址，在
该状态中去除了
存储在缓冲区中
的那个队列
从端口数据读地址
及存储子分组
读分组的长度信息
752
754
MAC中
有错误或缓冲区溢
出？
是
否
B
处理接收错误
760
762
分组的开始？
是
764
766
否
分组截短
存储在
分组
存储器中
更新端口空闲
信息
768
空闲
770

图 7B

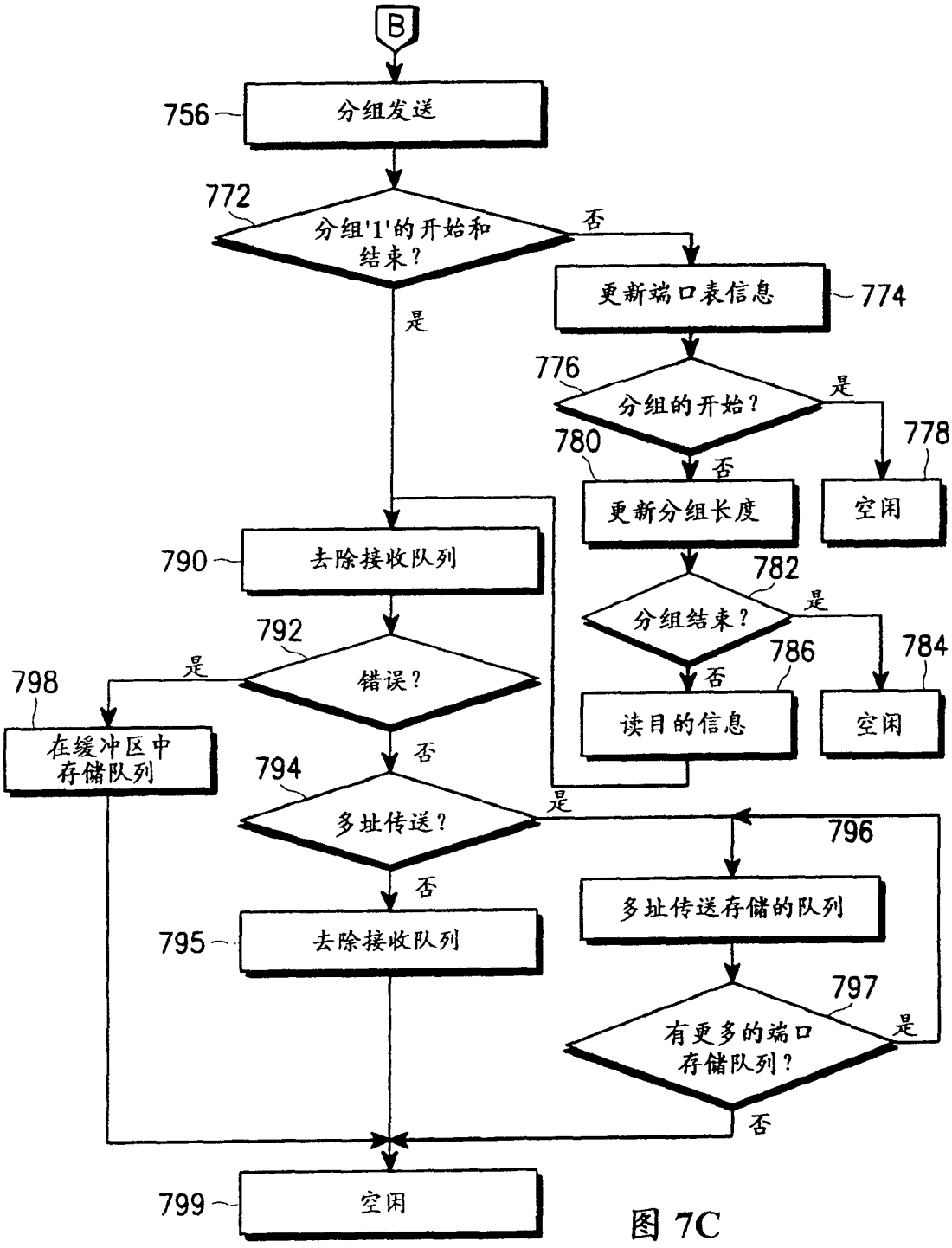

B
756
分组发送
772
分组'1'的开始和结束？
否
是
更新端口表信息
774
776
分组的开始？
是
778
780
否
更新分组长度
空闲
790
去除接收队列
782
分组结束？
是
784
792
错误？
是
786
否
读目的信息
空闲
798
在缓冲区中存储队列
否
794
多址传送？
是
796
否
多址传送存储的队列
795
去除接收队列
797
有更多的端口存储队列？
是
否
799
空闲

图 7C

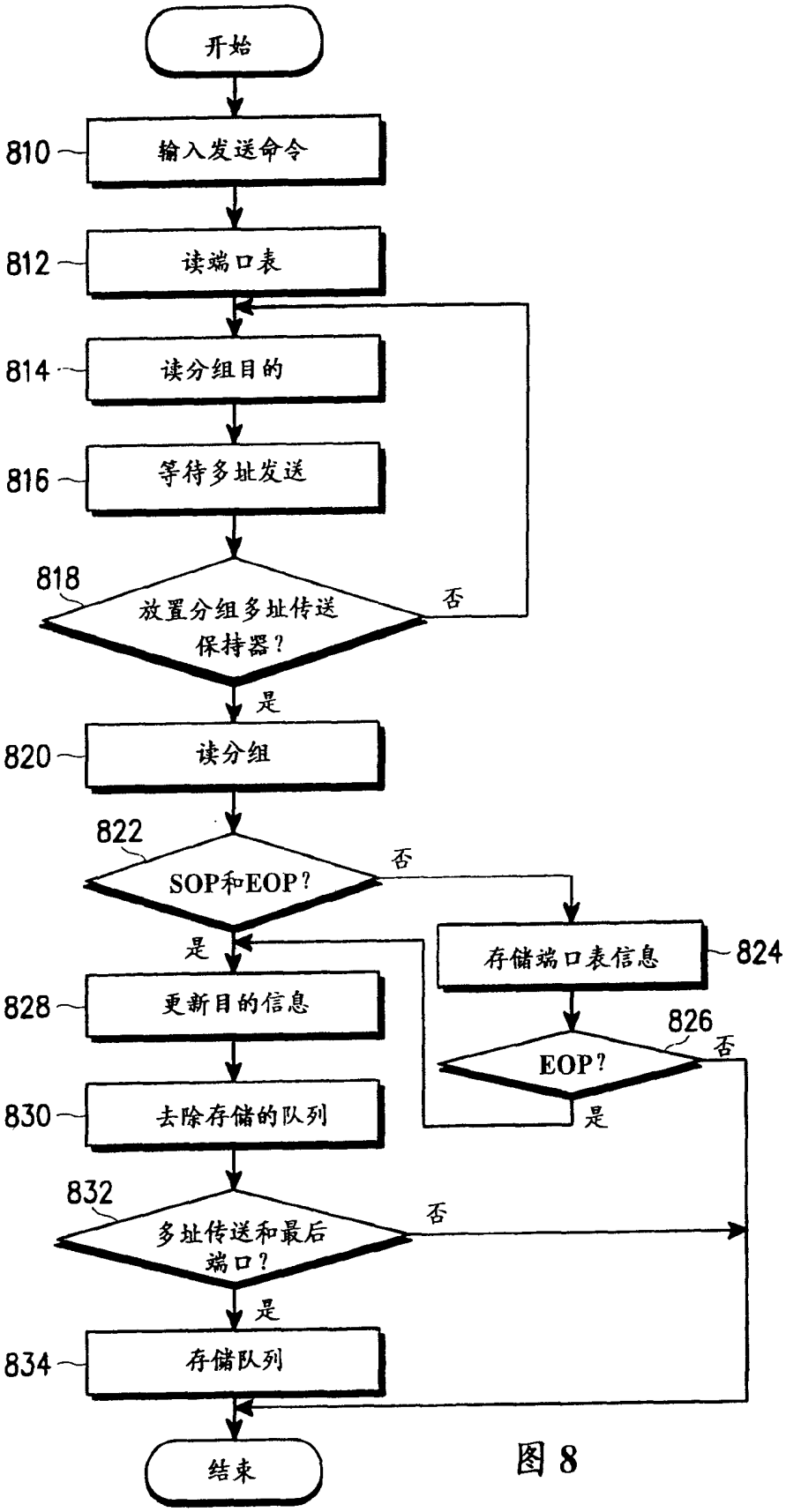

开始
810
输入发送命令
812
读端口表
814
读分组目的
816
等待多址发送
818
放置分组多址传送保持器?
否
是
820
读分组
822
SOP和EOP?
否
是
824
存储端口表信息
826
EOP?
否
是
828
更新目的信息
830
去除存储的队列
832
多址传送和最后端口?
否
是
834
存储队列
结束

图 8

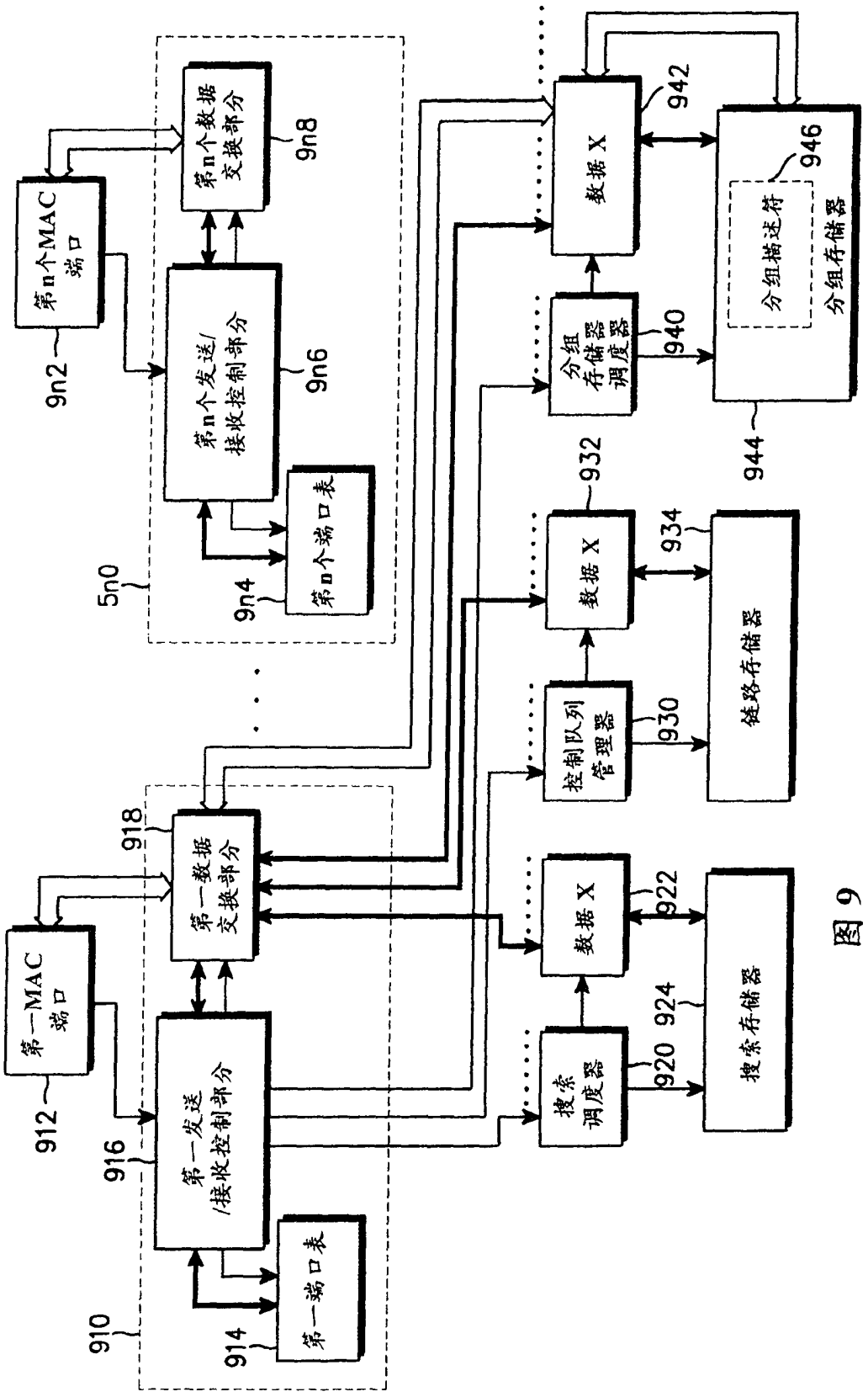

第一MAC端口
912
第一发送/接收控制部分
916
第一端口表
914
910
第一数据交换部分
918
第n个MAC端口
9n2
第n个发送/接收控制部分
9n6
第n个端口表
9n4
第n个数据交换部分
9n8
5n0
搜索调度器
920
搜索存储器
924
数据X
922
控制队列管理器
930
数据X
932
链路存储器
934
分组存储器调度器
940
数据X
942
分组描述符
946
分组存储器
944

图 9

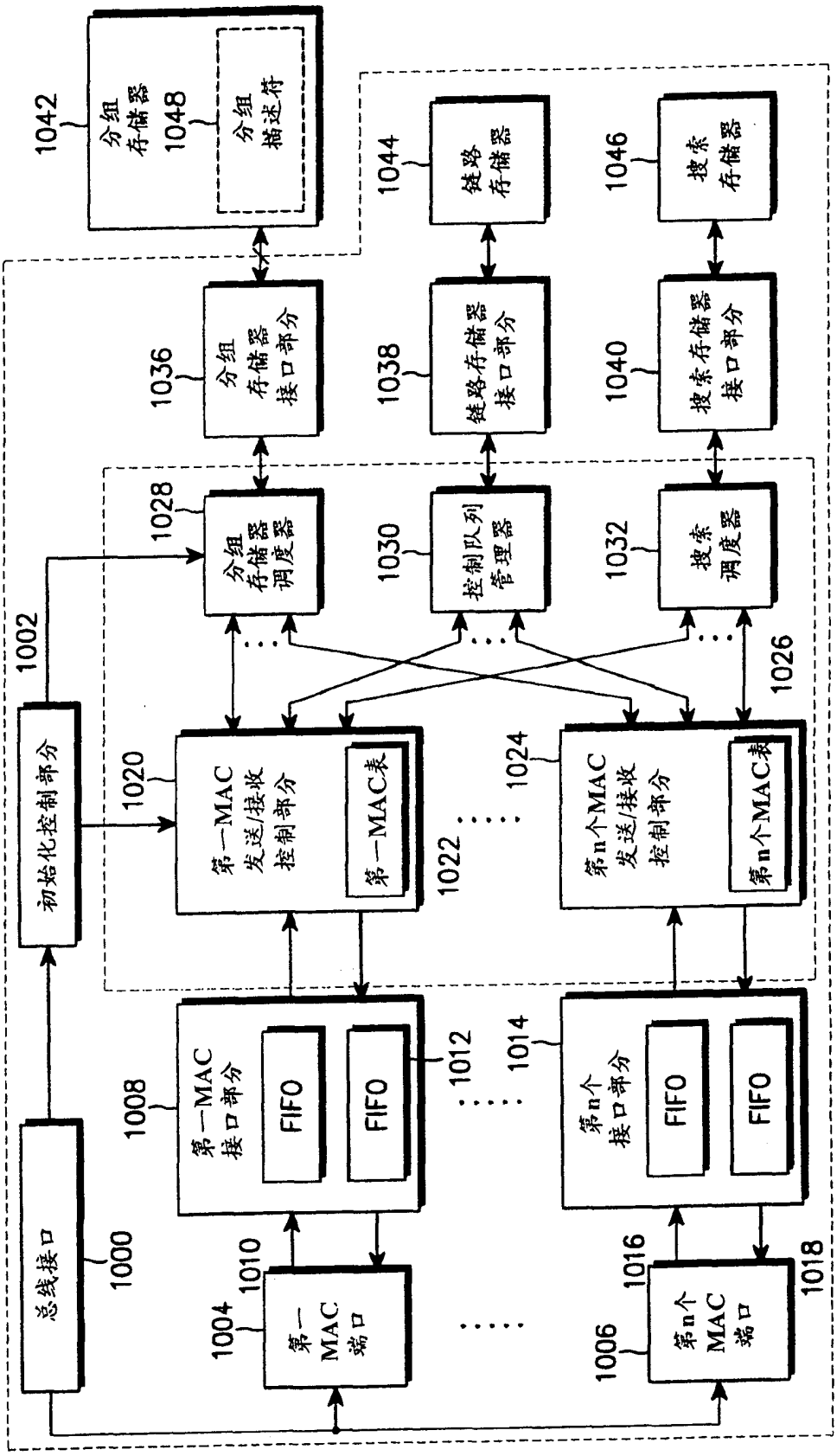

1042
分组存储器
1048
分组描述符
1044
链路存储器
1046
搜索存储器
1036
分组存储器接口部分
1038
链路存储器接口部分
1040
搜索存储器接口部分
1028
分组存储器调度器
1030
控制队列管理器
1032
搜索调度器
1002
初始化控制部分
1020
第一MAC发送/接收控制部分
第一MAC表
1022
1024
第n个MAC发送/接收控制部分
第n个MAC表
1026
1008
第一MAC接口部分
FIFO
FIFO
1012
1014
第n个接口部分
FIFO
FIFO
总线接口
1000
1004
第一MAC端口
1010
1006
第n个MAC端口
1016
1018

图 10

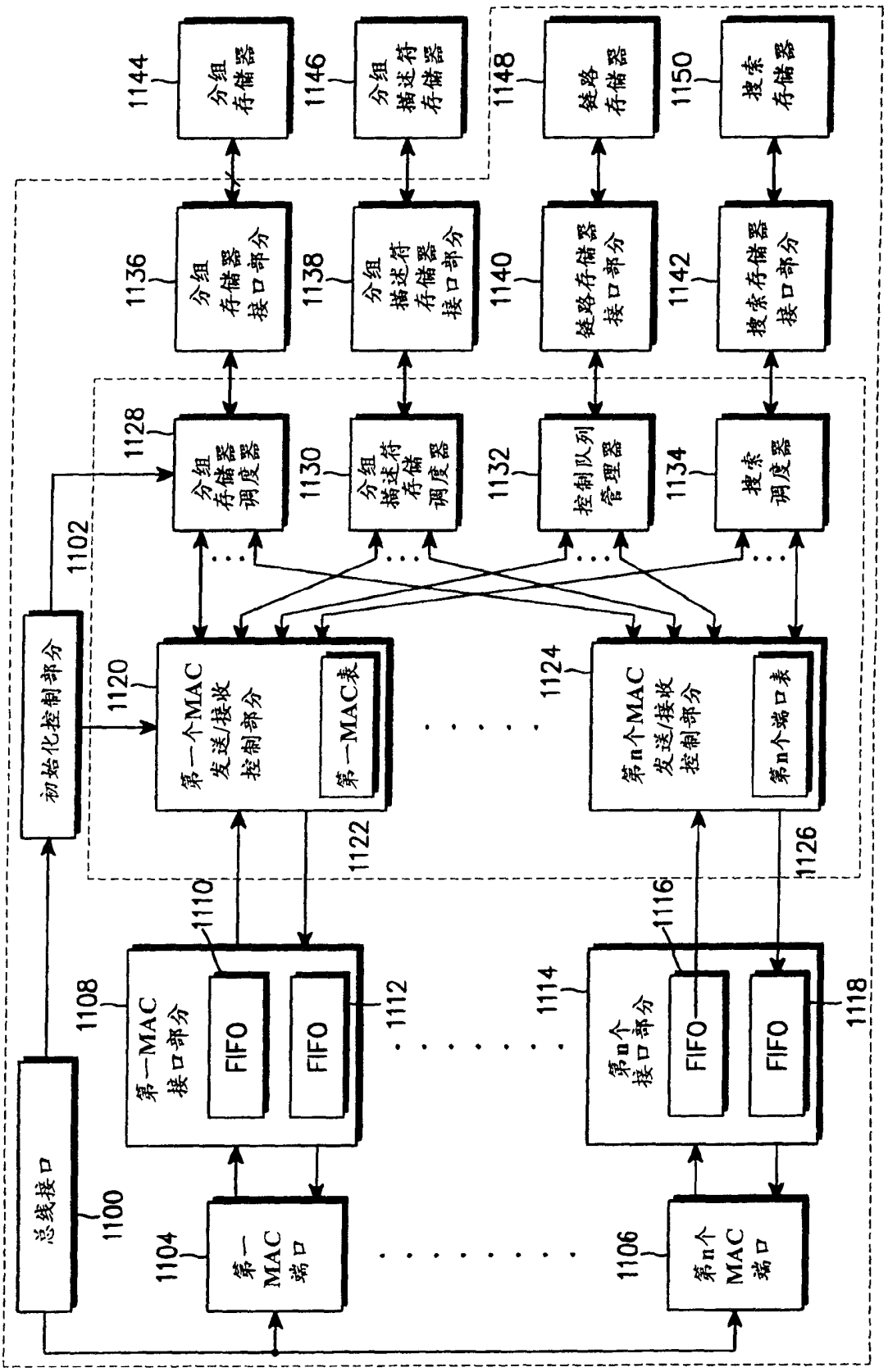

总线接口
1100
初始化控制部分
1102
第一MAC端口
1104
第n个MAC端口
1106
第一MAC接口部分
1108
第n个接口部分
1114
FIFO
1112
FIFO
1110
FIFO
1116
FIFO
1118
第一个MAC发送/接收控制部分
1120
第一MAC表
1122
第n个MAC发送/接收控制部分
1124
第n个端口表
1126
分组存储器调度器
1128
分组描述符存储调度器
1130
控制队列管理器
1132
搜索调度器
1134
分组存储器接口部分
1136
分组描述符存储器接口部分
1138
链路存储器接口部分
1140
搜索存储器接口部分
1142
分组存储器
1144
分组描述符存储器
1146
链路存储器
1148
搜索存储器
1150

图 11