(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 1940904 B
(45)授权公告日 2010.08.18

(21)申请号 200610106414.8
(22)申请日 2006.07.14
(30)优先权数据
11/236,458 2005.09.27 US
(73)专利权人 国际商业机器公司
地址 美国纽约
(72)发明人 威廉·J·斯塔克
本杰明·L·古德曼
普拉文·S·雷迪 维森特·E·丘恩
(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 邸万奎 黄小临
(51)Int.Cl.
G06F 15/163(2006.01)
(56)对比文件
US 6067603 A,2000.05.23, 全文.
US 5341504 A,1994.08.23, 全文.
US 5644716 A,1997.07.01, 全文.
US 20050034049 A1,2005.02.10, 全文.
US 5060141 A,1991.10.22, 全文.
审查员 丁文勋

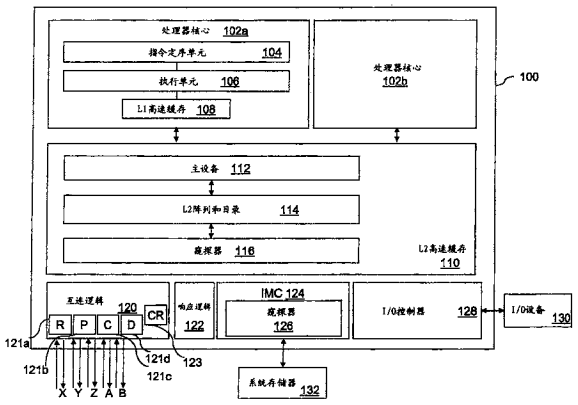
权利要求书 4 页 说明书 30 页 附图 28 页

(54)发明名称

数据处理系统和方法

(57)摘要

一种数据处理系统包括第一处理节点和第二处理节点。第一处理节点包括彼此耦接以便通信的第一多个处理单元，并且第二处理节点包括彼此耦接以便通信的第二多个处理单元。该第一多个处理单元中的每一个通过多个点到点链路中的相应一个耦接到第二处理节点中的该第二多个处理单元中的相应一个。



CN 1940904 B

1. 一种数据处理系统，包括：

第一处理节点和第二处理节点，其中：

所述第一处理节点包括彼此耦接以便通信的第一多个处理单元，并且所述第二处理节点包括彼此耦接以便通信的第二多个处理单元；

所述第一多个处理单元至少包括第一、第二和第三处理单元，并且所述第二多个处理单元至少包括第四、第五和第六处理单元；

所述第一和第四处理单元通过第一点对点链接而连接；

所述第二和第五处理单元通过第二点对点链接而连接；以及

所述第三和第六处理单元通过第三点对点链接而连接，

其中，

所述第一多个处理单元包括节点主设备处理单元和至少一个节点叶子处理单元；

所述第二多个处理单元包括远程中心处理单元和至少一个远程叶子处理单元；

所述节点主设备处理单元将请求广播到每个节点叶子处理单元和所述远程中心处理单元；

所述远程中心处理单元将所述请求广播到每个远程叶子处理单元；以及

所述节点主设备处理单元基于由所述节点主设备处理单元接收的对所述请求的部分响应，将对所述请求的组合响应广播到每个节点叶子处理单元、远程中心处理单元和远程叶子处理单元。

2. 如权利要求 1 所述的数据处理系统，其中：

所述第一、第二和第三点对点链接包括多个点对点第二层链接；以及

所述第一多个处理单元中的每一个通过多个点对点第一层链接中的相应一个耦接到所述第一多个处理单元中的每一个其它处理单元。

3. 如权利要求 2 所述的数据处理系统，其中，所述第一多个处理单元中的每一个都包括互连逻辑，该互连逻辑经由所有所述多个第一层链接以及所述多个第二层链接之一，将操作广播到所有所述第一多个处理单元以及所有所述第二多个处理单元。

4. 如权利要求 1 所述的数据处理系统，其中，所述第一多个处理单元中的至少一个包括配置寄存器，该配置寄存器包括一个或多个位，用于以第一模式配置第一处理单元以及以替换的第二模式配置第一处理单元，在第一模式中，所述第一多个处理单元中的每一个通过多个点对点链接中的相应一个耦接到所述第二处理节点中的所述第二多个处理单元中的相应一个，而在第二模式中，少于所有的所述第一多个处理单元通过所述多个点对点链接耦接到所述第二多个处理单元中的处理单元。

5. 如权利要求 1 所述的数据处理系统，其中：

所述第一处理节点中的第一多个处理单元和所述第二处理节点中的第二多个处理单元的操作按次序至少包括：请求阶段，其中广播请求；部分响应阶段，其中各个处理单元确定它们各自的对所述请求的响应；以及组合响应阶段，其中分发对所述请求的全系统范围的组合响应；以及

所述第一处理节点中的第一多个处理单元和所述第二处理节点中的第二多个处理单元以与所述请求相同的方向、经由所述请求穿过的每个链接路由所述组合响应，并且以与所述请求相反的方向、经由所述请求穿过的每个链接路由至少一个部分响应。

6. 一种用于数据处理系统的第一处理节点，该数据处理系统至少包括所述第一处理节点和第二处理节点，所述第一处理节点包括：

彼此耦接以便通信的第一多个处理单元，其中所述第一多个处理单元中的每一个都包括互连逻辑，通过该互连逻辑，每个所述第一处理单元可以通过多个点对点链接中的相应一个耦接到所述第二处理节点中的第二多个处理单元中的相应一个，使得：

所述第一处理节点中的第一处理单元通过第一点对点链接耦接到第二处理节点中的第四处理单元；

所述第一处理节点中的第二处理单元通过第二点对点链接耦接到第二处理节点中的第五处理单元；以及

所述第一处理节点中的第三处理单元通过第三点对点链接耦接到第二处理节点中的第六处理单元，

其中

所述第一多个处理单元包括节点主设备处理单元和至少一个节点叶子处理单元；

所述第二多个处理单元包括远程中心处理单元和至少一个远程叶子处理单元；

所述节点主设备处理单元将请求广播到每个节点叶子处理单元和所述远程中心处理单元；

所述节点主设备处理单元基于由所述节点主设备处理单元接收的对所述请求的部分响应、将对所述请求的组合响应广播到每个节点叶子处理单元、远程中心处理单元和远程叶子处理单元。

7. 如权利要求 6 所述的第一处理节点，其中：

所述多个点对点链接包括多个点对点第二层链接；以及

所述第一多个处理单元中的每一个通过多个点对点第一层链接中的相应一个耦接到所述第一多个处理单元中的每一个其它处理单元。

8. 如权利要求 7 所述的第一处理节点，其中，所述第一多个处理单元中的每一个都包括互连逻辑，该互连逻辑经由所有所述多个第一层链接以及所述多个第二层链接之一、将操作广播到所有所述第一多个处理单元以及所有所述第二多个处理单元。

9. 如权利要求 6 所述的第一处理节点，其中，所述第一多个处理单元中的至少一个包括配置寄存器，该配置寄存器包括一个或多个位，用于以第一模式配置第一处理单元以及以替换的第二模式配置第一处理单元，在第一模式中，所述第一多个处理单元中的每一个通过多个点对点链接中的相应一个耦接到所述第二处理节点中的所述第二多个处理单元中的相应一个，而在第二模式中，少于所有的所述第一多个处理单元通过所述多个点对点链接耦接到所述第二多个处理单元中的处理单元。

10. 如权利要求 6 所述的第一处理节点，其中：

所述第一多个处理单元的操作按次序至少包括：请求阶段，其中广播请求；部分响应阶段，其中各个处理单元确定它们各自的对所述请求的响应；以及组合响应阶段，其中分发对所述请求的全系统范围的组合响应；以及

所述第一多个处理单元中的节点主设备处理单元以与所述请求相同的方向、经由所述请求穿过的每个链接路由所述组合响应，并且所述第一多个处理单元中的每个节点叶子处理单元以与所述请求相反的方向、经由所述请求穿过的链接路由至少一个部分响应。

11. 一种数据处理系统中的数据处理方法，该数据处理系统包括第一处理节点和第二处理节点，其中所述第一处理节点包含第一多个处理单元，并且所述第二处理节点包含第二多个处理单元，所述方法包括：

彼此耦接所述第一多个处理单元；

彼此耦接所述第二多个处理单元；以及

耦接所述第一处理节点和所述第二处理节点，使得所述第一多个处理单元中的每一个通过多个点对点链接中的相应一个耦接到所述第二处理节点中的所述第二多个处理单元中的相应一个；其中所述耦接包括：

通过第一点对点链接将所述第一处理节点中的第一处理单元耦接到

所述第二处理节点中的第四处理单元；

通过第二点对点链接将所述第一处理节点中的第二处理单元耦接到

所述第二处理节点中的第五处理单元；以及

通过第三点对点链接将所述第一处理节点中的第三处理单元耦接到

所述第二处理节点中的第六处理单元，

其中，

所述第一多个处理单元包括节点主设备处理单元和至少一个节点叶子处理单元；

所述第二多个处理单元包括远程中心处理单元和至少一个远程叶子处理单元；

所述方法还包括：

所述节点主设备处理单元将请求广播到每个节点叶子处理单元和所述远程中心处理单元；

所述远程中心处理单元将所述请求广播到每个远程叶子处理单元；以及

所述节点主设备处理单元基于由所述节点主设备处理单元接收的对所述请求的部分响应，将对所述请求的组合响应广播到每个节点叶子处理单元、远程中心处理单元和远程叶子处理单元。

12. 如权利要求 11 所述的方法，其中，

所述多个点对点链接包括多个点对点第二层链接；以及

耦接所述第一多个处理单元包括，通过多个点对点第一层链接中的相应一个，将所述第一多个处理单元中的每一个耦接到所述第一多个处理单元中的每个其它处理单元。

13. 如权利要求 12 所述的方法，还包括：

所述第一多个处理单元之一内的互连逻辑，其经由所有所述多个第一层链接以及所述多个第二层链接之一将操作广播到所有所述第一多个处理单元以及所有所述第二多个处理单元。

14. 如权利要求 11 所述的方法，还包括：

响应于配置寄存器的第一设置，以第一模式传递操作，在第一模式中，所述第一多个处理单元中的每一个通过多个点对点链接中的相应一个与所述第二处理节点中的所述第二多个处理单元中的相应一个进行通信；以及

响应于配置寄存器的第二设置，以替换的第二模式传递操作，在第二模式中，少于所有的所述第一多个处理单元通过所述多个点对点链接与所述第二多个处理单元中的处理单元进行通信。

15. 如权利要求 11 所述的方法，其中：

所述第一处理节点中的第一多个处理单元和所述第二处理节点中的第二多个处理单元的操作按次序至少包括：请求阶段，其中广播请求；部分响应阶段，其中各个处理单元确定它们各自的对所述请求的响应；以及组合响应阶段，其中分布对所述请求的全系统范围的组合响应；以及

所述方法还包括，所述第一处理节点中的第一多个处理单元和所述第二处理节点中的第二多个处理单元以与所述请求相同的方向、经由所述请求穿过的每个链接路由所述组合响应，并且以与所述请求相反的方向、经由所述请求穿过的每个链接路由至少一个部分响应。

数据处理系统和方法

技术领域

[0001] 本发明一般涉及数据处理系统，尤其涉及改进的用于数据处理系统的互连构造。

[0002] 背景技术

[0003] 诸如服务器计算机系统之类的、传统的对称多处理器（SMP）计算机系统包括全部耦接到系统互连的多个处理单元，其中系统互连一般包括一条或多条地址、数据和控制总线。系统存储器耦接到系统互连，该系统存储器表示在多处理器计算机系统中最低等级的易失性存储器，并且其通常可由所有处理单元访问以用于读和写访问。为了减少对驻留在系统存储器中的指令和数据的访问延迟，每个处理单元通常还由相应的多级高速缓存分级结构所支持，该分级结构中的较低级别（一个或多个）可以由一个或多个处理器核心所共享。

[0004] 发明内容

[0005] 随着处理单元能够操作的时钟频率上升以及系统规模增加，在处理单元之间经由系统互连的通信延迟已经变为关键的性能问题。为了解决这个性能问题，已经提议和 / 或实现了旨在在传统的总线互连上提高性能和可伸缩性的各种互连设计。

[0006] 本发明提供了改进的数据处理系统、互连构造以及数据处理系统中的通信方法。在一个实施例中，数据处理系统包括第一处理节点和第二处理节点。第一处理节点包括彼此耦接以便通信的第一多个处理单元，而且第二处理节点包括彼此耦接以便通信的第二多个处理单元。该第一多个处理单元中的每一个通过多个点到点链接中的相应一个耦接到第二处理节点中的该第二多个处理单元中的相应一个。

[0007] 在以下详细编写的描述中，本发明的所有目的、特征、和优点将变得明显。

[0008] 附图说明

[0009] 在所附权利要求中阐述了认为是本发明的新颖特性的特征。然而，当结合附图阅读时，通过参考以下对说明性实施例的详细说明，本发明以及优选使用模式将得到最好的理解，在附图中：

[0010] 图 1 是根据本发明的处理单元的高级框图；

[0011] 图 2A 是根据本发明的数据处理系统的第一示例性实施例的高级框图；

[0012] 图 2B 是根据本发明、其中耦接多个节点以形成超级节点的数据处理系统的第二示例性实施例的高级框图；

[0013] 图 3 是包括请求阶段、部分响应阶段和组合响应阶段的示例性操作的时间 - 空间图；

[0014] 图 4A 是在图 2A 的数据处理系统内、全系统范围的示例性操作的时间 - 空间图；

[0015] 图 4B 是在图 2A 的数据处理系统内、仅仅节点范围的示例性操作的时间 - 空间图；

[0016] 图 4C 是图 2B 的数据处理系统内的示例性的超级节点广播操作的时间 - 空间图；

[0017] 图 5A-5C 描述了在图 4C 中描述的示例性超级节点广播操作的信息流；

[0018] 图 5D-5E 描述了根据本发明、用于示例性超级节点广播操作的示例性数据流；

[0019] 图 6 是示例性操作的时间 - 空间图，其说明了任意数据处理系统拓扑结构的定时

约束；

[0020] 图 7A-7B 说明了根据本发明、用于第一和第二层链接的示例性链接信息分配；

[0021] 图 7C 是包括在链接信息分配内的、用于写请求的部分响应字段的示例性实施例；

[0022] 图 8 是说明在操作的请求阶段利用的、图 1 中的互连逻辑部分的框图；

[0023] 图 9 是图 8 中的本地中心 (hub) 地址启动缓冲器的更详细框图；

[0024] 图 10 是图 8 中的标记 FIFO 队列的更详情框图；

[0025] 图 11 和 12 分别是图 8 中的本地中心部分响应 FIFO 队列和远程中心部分响应 FIFO 队列的更详细框图；

[0026] 图 13A-13D 是分别描述在本地主设备 (master)、本地中心、远程中心、和远程叶子处的操作中的请求阶段的流程图；

[0027] 图 13E 是根据本发明、在窥探器处生成部分响应的示例性方法的高级逻辑流程图；

[0028] 图 14 是说明在操作的部分响应阶段中利用的图 1 中的互连逻辑部分的框图；

[0029] 图 15A-15C 是分别描述在远程叶子、远程中心、本地中心、和本地主设备处的操作中的部分响应阶段的流程图；

[0030] 图 16 是说明在操作的组合响应阶段中利用的图 1 中的互连逻辑部分的框图；

[0031] 图 17A-7C 是分别描述在本地中心、远程中心、和远程叶子处的操作中的组合响应阶段的流程图；

[0032] 图 18 是图 2A 或者图 2B 中的数据处理系统的示例性窥探组件的更详细框图。

具体实施方式

[0033] 1. 处理单元和数据处理系统

[0034] 现在参见附图，并特别参见图 1，其中说明了根据本发明的处理单元 100 的示例性实施例的高级框图。在所描述的实施例中，处理单元 100 是包括用于独立地处理指令和数据的两个处理器核心 102a、102b 的单个集成电路。每个处理器核心 102 至少包括用于取出和排序指令以便执行的指令定序单元 (ISU)104 以及用于执行指令的一个或多个执行单元 106。由执行单元 106 执行的指令可以包括，例如，定点和浮点算术指令、逻辑指令、以及请求对存储块的读和写访问的指令。

[0035] 每个处理器核心 102a、102b 的操作由多级易失性存储器分级结构所支持，该结构在其最低级别具有一个或多个共享的系统存储器 132（图 1 中仅仅示出其中之一），并且在其较高级别具有一级或多级高速缓冲存储器。如同所示，处理单元 100 包括集成存储控制器 (IMC)，其响应于从处理器核心 102a、102b 接收的请求以及由窥探器在互连构造（如下所述）上窥探的操作，控制对系统存储器 132 的读和写访问。

[0036] 在该说明性实施例中，处理单元 100 的高速缓冲存储器分级结构包括在每个处理器核心 102a、102b 内的贯穿存储 (store-through) 第一级 (L1) 高速缓存 108，以及由处理单元 100 的所有处理器核心 102a、102b 共享的第二级 (L2) 高速缓存 110。L2 高速缓存 110 包括 L2 阵列和目录 114、主设备 112 和窥探器 116。主设备 112 启动在互连构造上的事务，并且响应于从相关联的处理器核心 102a、102b 接收的存储器访问（及其他）请求而访问 L2 阵列和目录 114。窥探器 116 检测在互连构造上的操作，提供适当的响应，并且执行操作所

需要的、对 L2 阵列和目录 114 的任何访问。虽然所说明的高速缓存分级结构仅仅包括二级高速缓存，但是本领域技术人员将理解，替换实施例可以包括另外级别 (L3、L4 等) 的片内或片外、内联或后备 (lookaside) 高速缓存，其可以完全包括、部分包括、或者不包括较高级别的高速缓存的内容。

[0037] 如图 1 中进一步所示，处理单元 100 包括集成的互连逻辑 120，处理单元 100 可以通过该互连逻辑 120 连接到互连构造，作为更大数据处理系统的一部分。在所描述的实施例中，互连逻辑 120 支持任意数目 t1 个“第一层”互连链接，其在这种情况下包括进入 (in-bound) 和外出 (out-bound)X、Y 和 Z 链接。互连逻辑 120 还支持任意数目 t2 个第二层链接，其在图 1 中指定为进入和外出 A 和 B 链接。利用这些第一和第二层链接，每个处理单元 100 可以双向通信地耦接到高达 t1/2+t2/2(在这种情况下，五个) 其它的处理单元 100。互连逻辑 120 包括用于在操作的不同阶段期间处理和转发信息的请求逻辑 121a、部分响应逻辑 121b、组合响应逻辑 121c 和数据逻辑 121d。此外，互连逻辑 120 包括配置寄存器 123，其包括用来配置处理单元 100 的多个模式位。如下面进一步所述，这些模式位优选地包括 :(1) 一个或多个模式位的第一集合，其选择期望的用于第一和第二层链接的链接信息分配 ;(2) 一个或多个模式位的第二集合，其指定处理单元 100 的第一和第二层链接中的哪些链接连接到其它处理单元 100 ;(3) 一个或多个模式位的第三集合，其确定保护窗口扩展的可编程持续时间 ;(4) 一个或多个模式位的第四集合，如在上面引用的美国专利申请第 11/055,305 号中所述，其在逐个操作的基础上，从仅仅节点的广播范围或者全系统的范围当中，预测性地选择用于由处理单元 100 发起的操作的广播范围 ;以及 (5) 一个或多个模式位的第五集合，其指示处理单元 100 是否属于以“超级节点”模式耦接到至少一个其它处理节点的处理节点，在“超级节点”模式中，广播操作跨越以下面参考图 2B 所述的方式耦接的多个物理处理节点。

[0038] 每个处理单元 100 还包括响应逻辑 122 的实例，其实现了分布式一致性信令机制部分，该机制维护在处理单元 100 的高速缓存分级结构和其它处理单元 100 中的高速缓存分级结构之间的高速缓存一致性。最后，每个处理单元 100 包括集成的 I/O(输入 / 输出) 控制器 128，其支持诸如 I/O 设备 130 之类的一个或多个 I/O 设备的附连。I/O 控制器 128 可以响应于 I/O 设备 130 的请求而在 X、Y、Z、A 和 B 链接上发出操作以及接收数据。

[0039] 现在参见图 2A，其中描述了根据本发明、由多个处理单元 100 形成的数据处理系统 200 的第一示例性实施例的框图。如同所示，数据处理系统 200 包括八个处理节点 202a0-202d0 和 202a1-202d1，在所描述的实施例中，其中每个都被实现为多芯片模块 (MCM)，该多芯片模块 (MCM) 包括包含四个处理单元 100 的封装。如同所示，在每个处理节点 202 内的处理单元 100 由处理单元的 X、Y、和 Z 链接耦接以用于点对点通信。每个处理单元 100 还由处理单元的 A 和 B 链接耦接到在两个不同的处理节点 202 中的处理单元 100 以用于点对点通信。虽然在图 2A 中用双头箭头进行了说明，但是应当理解，每对 X、Y、Z、A 和 B 链接优选地 (但不一定) 实现为两个单向链接而不是双向链接。

[0040] 用于形成图 2A 所示的拓扑结构的通用表示式可以如下所示给出 :

[0041] 对于所有 I ≠ J，节点 [I][K]. 芯片 [J]. 链接 [K] 连接到节点 [J][K]. 芯片 [I]. 链接 [K] ;以及

[0042] 节点 [I][K]. 芯片 [I]. 链接 [K] 连接到节点 [I][非 K]. 芯片 [I]. 链接 [非 K] ;

以及

[0043] 节点 [I][K]. 芯片 [I]. 链接 [非 K] 连接到 :

[0044] (1) 不进行连接,保留用于将来扩展 ;或者

[0045] (2) 节点 [额外][非 K]. 芯片 [I]. 链接 [K],在充分利用了所有链接 (即,形成 72 路系统的九个 8 路节点) 的情况下 ;以及

[0046] 其中 I 和 J 属于集合 {a,b,c,d},并且 K 属于集合 {A,B}。

[0047] 当然,可以定义替换的表示式以形成功能等效的拓扑结构。此外,应该理解,所述拓扑结构对于实施本发明的数据处理系统拓扑结构是代表性而非穷举性的,而且其它的拓扑结构是可能的。在这样的替换拓扑结构中,例如,耦接到每个处理单元 100 的第一层和第二层链接的数目可以是任意数目,而且在每一层内的处理节点 202 的数目 (即,I) 无需与每个处理节点 100 的处理单元 100 的数目 (即,J) 相等。

[0048] 即使以图 2A 所示的方式全连接,所有处理节点 202 也无需将每个操作传递到所有其它处理节点 202。具体而言,如上所述,处理单元 100 可能利用局限于它们的处理节点 202 的范围、或者利用诸如包括所有处理节点 202 的全系统范围之类的较大范围来广播操作。

[0049] 如图 18 所示,在数据处理系统 200 内的示例性窥探设备 1900,例如 L2(或者较低级别) 高速缓存的窥探器 116 或者 IMC 124 的窥探器 126,可以包括标识真实地址空间的一个或者多个区域的一个或多个基地址寄存器 (BAR)1902,该真实地址空间包括窥探设备 1900 所负责的真实地址。窥探设备 1900 还可以可选地包括哈希逻辑 1904,其在属于由 BAR 1902 所标识的真实地址空间的区域 (一个或多个) 内的真实地址上执行哈希函数,以进一步证明 (qualify) 该窥探设备 1900 是否对该地址负责。最后,窥探设备 1900 包括多个窥探器 1906a-1906m,其响应于窥探到的、指定由 BAR 1902 和哈希逻辑 1904 证明合格的请求地址的请求,访问资源 1910(例如,L2 高速缓存阵列和目录 114 或者系统存储器 132)。

[0050] 如同所示,资源 1910 可以具有存储体结构 (banked structure),其包括多个存储体 (bank)1912a-1912n,每个存储体与各自的真实地址集合相关联。如本领域技术人员所知的那样,经常采用这样的存储体设计,以通过有效地将资源 1910 细分为多个可独立访问资源,支持对资源 1910 的请求的较高到达速率。用这样的方式,即使窥探设备 1900 和 / 或资源 1910 的工作频率为窥探设备 1900 不能与访问资源 1910 的请求的最大到达速率那样快地为这些请求服务,但是只要在给定时间间隔内针对任何存储体 1912 接收的请求数目不超过该存储体 1912 可以在该时间间隔内进行服务的请求数目,则窥探设备 1900 可以为这些请求服务而无需重试。

[0051] 本领域的技术人员将会理解,SMP 数据处理系统 100 可以包括诸如互连桥接器、非易失性存储器、用于连接到网络或者附连设备的端口等之类的许多另外未说明的组件。因为这样的另外组件对于本发明的理解不是必需的,因此未在图 2A 中示出它们,或者在这里进一步讨论它们。

[0052] 图 2B 描述了根据本发明、其中耦接多个处理单元 100 以形成“超级节点”的数据处理系统 200 的示例性第二实施例的框图。如同所示,数据处理系统 220 包括两个处理节点 202a0 和 202b0,在所描述的实施例中,其中每个都被实现为多芯片模块 (MCM),该多芯片模块 (MCM) 包括包含四个依据图 1 的处理单元 100 的封装。如同所示,在每个处理节点 202 内的处理单元 100 由处理单元的 X、Y、和 Z 链接耦接以用于点对点通信。每个处理单元 100

还由处理单元的 A 和 / 或 B 链接耦接到在其它处理节点 202 中的相应处理单元 100 以用于点对点通信。虽然在图 2B 中用双头箭头进行了说明，但是应当理解，每对 X、Y、Z、A 和 B 链接优选（而不是必须）实现为两个单向链接而不是双向链接。

[0053] 用于形成图 2B 所示的拓扑结构的通用表示式可以给出如下：节点 [I]. 芯片 [J]. 链接 [L] 连接到节点 [非 I]. 芯片 [非 J]. 链接 [L]；以及

[0054] 节点 [I]. 芯片 [K]. 链接 [L] 连接到节点 [非 I]. 芯片 [非 K]. 链接 [L]，

[0055] 其中 I 属于 {a,b}，J 属于集合 {a,b}，K 属于集合 {c,d}，以及 L 属于集合 {A,B}。

[0056] 此外应该理解，所述拓扑结构是体现本发明的数据处理系统拓扑结构的代表而不是该拓扑结构的穷举，而且其它具有耦接特定节点对的多个链接的拓扑结构是可能的。如上参考图 2A 所述，在这样的替换拓扑结构中，耦接到每个处理单元 100 的第一层和第二层链接的数目可以是任意数目。此外，另外的处理节点 202 可以由另外的第二层链接耦接到处理节点 202a0 和 202b0。

[0057] 当期望最大化节点间通信的带宽时，可以采用诸如图 2B 中所述的拓扑结构。例如，如果在特定处理和它们的相关联数据之间的亲合力 (affinity) 不是大到足够让操作占优势地在单个处理节点 202 内得到服务，则可以采用图 2B 的拓扑结构来提高节点间的通信带宽（例如，在这种情况下，提高了高达 4 倍）。通过增加耦接特定节点对的第二层链接的数目来提高节点间带宽可以因此对特定的工作负载产生重要的性能益处。

[0058] II. 示例性操作

[0059] 现在参见图 3，其中描述了在图 2A 的数据处理系统 200 或者图 2B 的数据处理系统 220 的互连构造上的示例性操作的时间 - 空间图。当主设备 300（例如，L2 高速缓存 110 的主设备 112 或者 I/O 控制器 128 内的主设备）在互连构造上发出请求 302 时，开始该操作。请求 302 优选地至少包括指示期望访问类型的事务类型，以及指示要由该请求访问的资源的资源标识符（例如，真实地址）。常见的请求类型优选地包括下面在表格 I 中所阐述的那些。

[0060] 表格 I

[0061]

请求 | 描述

[0062]

READ	为查询目的，请求存储块的映像的副本
RWITM（带修改意图的读取）	带更新（修改）意图地请求存储块的映像的唯一副本，并且若有的话，要求破坏其它副本
DCLAIM（数据要求）	带更新（修改）意图地请求将存储块的现有仅仅查询副本提升为唯一副本的授权，并且若有的话，要求破坏其它副本
DCBZ（数据高速缓存块清零）	请求创建存储块的新唯一副本而不考虑其当前状态并且随后修改其内容的授权；若有的话，要求破坏其它副本
CASTOUT（逐出）	将存储块的映像从较高级别的存储器拷贝到较低级别的存储器，以准备破坏较高级别副本
WRITE	请求创建存储块的新唯一副本而不考虑其当前状态并且立即将存储块的映像从较高级别的存储器拷贝到较低级别的存储器以准备破坏较高级别副本的授权
PARTIAL WRITE	请求创建部分存储块的新唯一副本而不考虑其当前状态并且立即将部分存储块的映像从较高级别的存储器拷贝到较低级别的存储器以准备破坏较高级别副本的授权

[0063] 有关这些操作以及有助于这些操作的高效处理的示例性高速缓存一致性协议的更多细节可以在通过引用并入在此的、共同未决的美国专利申请第 11/055,305 号中找到。

[0064] 请求 302 由遍及数据处理系统 200 分布的窥探器 304 例如 L2 高速缓存 110 的窥探器 116 以及 IMC 124 的窥探器 126 所接收。一般而言，尽管有一些例外，在与请求 302 的主

设备 112 相同的 L2 高速缓存 110 中的窥探器 116 不窥探请求 302(即,一般不存在自我窥探),这是因为只有当请求 302 不能由处理单元 100 在内部进行服务时,才在互连构造上传送请求 302。接收和处理请求 302 的窥探器 304 每个都提供了各自的部分响应 306,其表示至少窥探器 304 对请求 302 的响应。IMC 124 内的窥探器 126 例如基于窥探器 126 是否对该请求地址负责以及它是否具有可用于为该请求服务的资源,确定要提供的部分响应 306。L2 高速缓存 110 中的窥探器 116 可以例如基于其 L2 高速缓存目录 114 的可用性、在窥探器 116 内处理该请求的窥探逻辑实例的可用性、以及与在 L2 高速缓存目录 114 中的请求地址相关联的一致性状态,确定它的部分响应 306。

[0065] 窥探器 304 的部分响应 306 由响应逻辑 122 的一个或多个实例分阶段地或同时地进行逻辑组合,以确定对请求 302 的组合响应 (CR)310。在一个下文中将要采用的优选实施例中,负责生成组合响应 310 的响应逻辑 122 的实例位于包含发出请求 302 的主设备 300 的处理单元 100 中。响应逻辑 122 经由互连构造向主设备 300 和窥探器 304 提供组合响应 310,以指示对请求 302 的响应(例如,成功、失败、重试等)。如果 CR 310 指示请求 302 的成功,则例如,CR 310 可以指示用于所请求存储块的数据源、其中所请求的存储块要由主设备 300 高速缓存的高速缓存状态、以及是否需要使一个或者多个 L2 高速缓存 110 中的所请求存储块无效的“清除”操作。

[0066] 响应于组合响应 310 的接收,一个或多个主设备 300 和窥探器 304 通常执行一个或多个操作以便为请求 302 服务。这些操作可以包括向主设备 300 提供数据、使在一个或多个 L2 高速缓存 110 中高速缓存的数据的一致性状态无效或者更新该状态、执行逐出 (castout) 操作、将数据写回到系统存储器 132 中等。如果请求 302 要求的话,可以在由响应逻辑 122 生成组合响应 310 之前或之后,向或从主设备 300 传送所请求或目标存储块。

[0067] 在以下的描述中,将参考窥探器相对于由请求所指定的请求地址是一致性最高点 (HPC)、一致性最低点 (LPC)、还是都不是,来描述窥探器 304 对请求 302 的部分响应 306 以及由窥探器 304 响应于请求 302 而执行的操作,和 / 或它的组合响应 310。此处将 LPC 定义为用作存储块的储存库 (repository) 的存储器设备或者 I/O 设备。在不存在用于存储块的 HPC 的情况下,LPC 保存存储块的真实映像,并且具有允许或者拒绝生成该存储块的附加高速缓存副本的请求的授权。对于在图 1 和 2 的数据处理系统实施例中的典型请求,LPC 将是用于保存所引用的存储块的系统存储器 132 的存储控制器 124。此处将 HPC 定义为对存储块的真实映像(其可能或者可能不与 LPC 处的对应存储块相一致)进行高速缓存的唯一标识的设备,并且具有允许或者拒绝修改该存储块的请求的授权。叙述性地,HPC 还可以响应于不修改存储块的操作而向请求者提供存储块的副本。因此,对于在图 1 和 2 的数据处理系统实施例中的典型请求,若有的话,HPC 将会是 L2 高速缓存 110。虽然可以利用其它指示符来指定用于存储块的 HPC,但是本发明的优选实施例利用在 L2 高速缓存 110 的 L2 高速缓存目录 114 内的选定高速缓存一致性状态(一个或多个),来指定用于存储块的 HPC(若有的话)。

[0068] 仍然参见图 3,用于在请求 302 中所引用的存储块的 HPC(若有的话),或者在不存在 HPC 时该存储块的 LPC,优选地具有响应于请求 302、在必要时保护存储块的所有权转移的责任。在图 3 所示的示例性场景中,在用于由请求 302 的请求地址所指定的存储块的 HPC(或者在不存在 HPC 时,LPC) 处的窥探器 304n 在保护窗口 312a 期间和随后的窗口扩展

312b 期间保护所请求的存储块的所有权到主设备 300 的转移，其中保护窗口 312a 从窥探器 304n 确定它的部分响应 306 的时间延伸到窥探器 304n 接收组合响应 310 为止，而窗口扩展 312b 从窥探器 304n 对组合响应 310 的接收开始外延一可编程时间。在保护窗口 312a 和窗口扩展 312b 期间，窥探器 304n 通过向其它指定相同请求地址的请求提供防止其它主设备获得所有权的部分响应 306（例如重试部分响应）直到所有权已经成功地被转移到主设备 300 为止，来保护所有权转移。主设备 300 同样也启动保护窗口 313，以在收到组合响应 310 之后保护它对在请求 302 中请求的存储块的所有权。

[0069] 因为窥探器 304 全都具有有限的资源用于处理上述 CPU 和 I/O 请求，几个不同级别的部分响应和对应的 CR 是可能的。例如，如果对所请求的存储块负责的存储控制器 124 内的窥探器 126 具有可用于处理请求的队列，则该窥探器 126 可以用指示它能够用作该请求的 LPC 的部分响应进行响应。相反，如果窥探器 126 没有可用于处理该请求的队列，则窥探器 126 可以用指示它是该存储块的 LPC、但是当前不能为该请求服务的部分响应进行响应。类似地，L2 高速缓存 110 中的窥探器 116 可能需要窥探逻辑的可用实例以及对 L2 高速缓存目录 114 的访问，以便处理请求。缺少对这些资源中的任何一个（或者两者）的访问导致通知由于缺少所需要的资源而不能为该请求服务的部分响应（以及对应的 CR）。

[0070] III. 示例性操作的广播流程

[0071] 现在参见图 4A，其中说明了在图 2A 的数据处理系统 200 中、全系统范围的操作的示例性操作流程的时间－空间图。在这些附图中，数据处理系统 200 内的各个处理单元 100 用两个位置标识符标记，第一个标识处理单元 100 所属的处理节点 202，而且第二个标识在处理节点 202 内的特定处理单元 100。因此，例如，处理单元 100a0c 是指处理节点 202a0 的处理单元 100c。此外，每一处理单元 100 用功能标识符标记，该功能标识符指示其相对于参与操作的其它处理单元 100 的功能。这些功能标识符包括：(1) 本地主设备 (LM)，其指定发起操作的处理单元 100；(2) 本地中心 (LH)，其指定在与本地主设备相同的处理节点 202 中并且负责将该操作传送到其它处理节点 202 的处理单元 100（本地主设备也可以是本地中心）；(3) 远程中心 (RH)，其指定在与本地主设备不同的处理节点 202 中、并且负责将该操作分发到它的处理节点 202 中的其它处理单元 100 的处理单元 100；以及 (4) 远程叶子 (RL)，其指定在与本地主设备不同的处理节点 202 中、并且不是远程中心的处理单元 100。

[0072] 如图 4A 所示，示例性操作至少具有上面参考图 3 所述的三个阶段，即请求（或者寻址）阶段、部分响应 (Presp) 阶段、和组合响应 (Cresp) 阶段。这三个阶段优选地以上述次序出现并且不重叠。该操作另外可以具有数据阶段，其可以可选地与请求、部分响应和组合响应阶段中的任意一个相重叠。

[0073] 仍然参见图 4A，当本地主设备 100a0c（即，处理节点 202a0 的处理单元 100c）向在它的处理节点 202a0 内的每个本地中心 100a0a、100a0b、100a0c 和 100a0d 执行例如读取请求的请求的同步广播时，请求阶段开始。应当注意到，本地中心列表包括本地中心 100a0c，其也是本地主设备。如下面进一步所述，有利地采用该内部传送来同步本地中心 100a0c 和本地中心 100a0a、100a0b 和 100a0d 的操作，以便可以更容易地满足下面讨论的定时约束。

[0074] 响应于请求的接收，每个通过其 A 或 B 链接耦接到远程中心 100 的本地中心 100 将该操作传送到其远程中心（一个或多个）100。因此，本地中心 100a0a 不在其外出 A 链接上进行操作的传送，而是经由其外出 B 链接将操作传送到处理节点 202a1 内的远程中心。本

地中心 100a0b、100a0c 和 100a0d 经由其各自的外出 A 和 B 链接分别将操作传送到处理节点 202b0 和 202b1、处理节点 202c0 和 202c1、以及处理节点 202d0 和 202d1 中的远程中心。每个接收操作的远程中心 100 又将该操作传送到其处理节点 202 中的每个远程叶子 100。因此，例如，本地中心 100b0a 将操作传送到远程叶子 100b0b、100b0c 和 100b0d。用这样的方式，利用在不超过三个链接上的传送，将操作高效地广播到数据处理系统 200 内的所有处理单元 100。

[0075] 如图 4A 所示，在请求阶段之后，发生部分响应 (Presp) 阶段。在部分响应阶段，每个远程叶子 100 计算 (evaluate) 该操作，并且将其对该操作的部分响应提供给其各自的远程中心 100。例如，远程叶子 100b0b、100b0c 和 100b0d 将其各自的部分响应传送到远程中心 100b0a。每个远程中心 100 又将这些部分响应、以及它自己的部分响应传送到本地中心 100a0a、100a0b、100a0c 和 100a0d 中的各个。本地中心 100a0a、100a0b、100a0c 和 100a0d 然后将这些部分响应、以及它们自己的部分响应广播到处理节点 202a0 中的每个本地中心 100。应当注意，由处理节点 202a0 内的本地中心进行的部分响应广播包括由于定时的原因，由每个本地中心 100 对它自己的部分响应的自我广播。

[0076] 应当理解，以所示方式收集部分响应可以以多种不同的方式实现。例如，有可能从每个其它本地中心、远程中心和远程叶子将单独的部分响应传递回到每个本地中心。可替换地，为了更大的效率，可能期望在将部分响应传递回到本地中心时累积这些部分响应。为了确保将每个部分响应的效果准确地传递回到本地中心 100，优选地，例如利用逻辑或函数以及其中当经受这样的函数时不丢失相关信息的编码（例如，“一热 (one-hot)”编码），以非破坏性方式累积部分响应（若有的话）。

[0077] 如在图 4A 中进一步所示，在处理节点 202a0 内的每个本地中心 100 处的响应逻辑 122 汇集其它处理单元 100 的部分响应以获得代表对该请求的全系统范围响应的组合响应。然后，本地中心 100a0a-100a0d 沿着与请求阶段所采用的路径相同的分发路径将该组合响应广播到所有处理单元 100。因此，该组合响应首先被广播到远程中心 100，其又将该组合响应传送到在其各自的处理节点 202 内的每个远程叶子 100。例如，远程中心 100a0b 将该组合响应传送到远程中心 100b0a，其又将该组合响应传送到远程叶子 100b0b、100b0c 和 100b0d。

[0078] 如上所述，为操作服务可能需要另外的数据阶段。例如，如果操作是诸如读取或 RWITM 操作之类的读取类型的操作，则远程叶子 100b0d 可以经由将远程叶子 100b0d 连接到远程中心 100b0a 的链接、将远程中心 100b0a 连接到本地中心 100a0b 的链接、以及将本地中心 100a0b 连接到本地主设备 100a0c 的链接，将所请求的存储块供应 (source) 到本地主设备 100a0c。相反，如果操作是写入类型的操作，例如，将修改的存储块写回到远程叶子 100b0b 的系统存储器 132 中的高速缓存逐出操作，则经由将本地主设备 100a0c 连接到本地中心 100a0b 的链接、将本地中心 100a0b 连接到远程中心 100b0a 的链接、以及将远程中心 100b0a 连接到远程叶子 100b0b 的链接传送该存储块。

[0079] 现在参见图 4B，其中说明了在图 2A 的数据处理系统 200 中、仅仅节点范围的操作的示例性操作流程的时间 - 空间图。在这些附图中，数据处理系统 200 内的各个处理单元 100 用两个位置标识符标记，第一个标识处理单元 100 所属的处理节点 202，而且第二个标识处理节点 202 内的特定处理单元 100。因此，例如，处理单元 100b0a 是指处理节点 202b0

的处理单元 100b。此外，每一处理单元 100 用功能标识符标记，该功能标识符指示其相对于参与操作的其它处理单元 100 的功能。这些功能标识符包括：(1) 节点主设备 (NM)，其指定发起仅仅节点范围的操作的处理单元 100，以及 (2) 节点叶子 (NL)，其指定在与节点主设备相同的处理节点 202 中并且不是节点主设备的处理单元 100。

[0080] 如图 4B 所示，示例性的仅仅节点的操作至少具有如上所述的三个阶段，即请求（或者寻址）阶段、部分响应 (Presp) 阶段、和组合响应 (Cresp) 阶段。同样，这三个阶段优选地以上述次序出现并且不重叠。该操作另外可以具有数据阶段，其可以可选地与请求、部分响应和组合响应阶段中的任意一个相重叠。

[0081] 仍然参见图 4B，当非常类似于图 4A 的操作场景中的远程中心而工作的节点主设备 100b0a（即，处理节点 202b0 的处理单元 100a) 向其处理节点 202b0 内的每个节点叶子 100b0b、100b0c、和 100b0d 执行例如读取请求的请求的同步广播时，请求阶段开始。应当注意，因为广播传送的范围限于单个节点，因此没有采用在节点主设备 100b0a 内的请求的内部传送来同步该请求的节点外传送。

[0082] 如图 4B 所示，在请求阶段之后，发生部分响应 (Presp) 阶段。在部分响应阶段，节点叶子 100b0b、100b0c 和 100b0d 中的每一个计算该操作并且向节点主设备 100b0a 提供它对该操作的部分响应。接下来，如在图 4B 中进一步所示，在处理节点 202b0 内的节点主设备 100b0a 处的响应逻辑 122 汇集其它处理单元 100 的部分响应，以获得代表对该请求的节点范围响应的组合响应。节点主设备 100b0a 然后利用节点主设备 100b0a 的 X、Y 和 Z 链接，将该组合响应广播到所有节点叶子 100b0b、100b0c 和 100b0d。

[0083] 如上所述，为操作服务可能需要另外的数据阶段。例如，如果操作是诸如读取或者 RWITM 操作之类的读取类型操作，则节点叶子 100b0d 可以经由将节点叶子 100b0d 连接到节点主设备 100b0a 的 Z 链接将所请求的存储块供应到节点主设备 100b0a。相反，如果操作是写入类型操作，例如将修改的存储块写回到远程叶子 100b0b 的系统存储器 132 中的高速缓存逐出操作，则经由将节点主设备 100b0a 连接到节点叶子 100b0b 的 X 链接传送该存储块。

[0084] 现在参见图 4C，将结合图 5A-5E 对其进行描述，其说明了在图 2B 的数据处理系统 220 中的操作的示例性操作流程的时间 - 空间图。在这些图中，在数据处理系统 220 内的各个处理单元 100 利用与上述相同的两个位置标识符标记。此外，每个处理单元 100 用功能标识符标记，该功能标识符指示其相对于参与操作的其它处理单元 100 的功能。这些功能标识符包括：(1) 节点主设备 (NM)，其指定发起操作的处理单元 100；(2) 节点叶子 (NL)，其指定在与节点主设备相同的处理节点 202 中但不是该节点主设备的处理单元 100；(3) 远程中心 (RH)，其指定在与本地主设备不同的处理节点 202 中、并且负责将操作分发到其处理节点 202 中的其它处理单元 100 的处理单元 100；以及 (4) 远程叶子 (RL)，其指定在与本地主设备不同的处理节点 202 中、并且不是远程中心的处理单元 100。

[0085] 如图 4C 所示，示例性操作至少具有上面参考图 3 所述的三个阶段，即请求（或者寻址）阶段、部分响应 (Presp) 阶段、和组合响应 (Cresp) 阶段。这三个阶段优选地以上述次序出现并且不重叠。该操作另外可以具有数据阶段，其可以可选地与请求、部分响应和组合响应阶段中的任意一个相重叠。

[0086] 仍然参见图 4C 并且另外参见图 5A，当节点主设备 (NM)100a0c（即，处理节点 202a0 的处理单元 100c) 向其处理节点 202a0 内的每个节点叶子 100a0a、100a0b、和 100a0d

以及向处理节点 202b0 中的远程中心 100b0d 执行例如读取请求的请求的同步广播时，请求阶段开始。远程中心 100b0d 又将该操作传送到每个远程叶子 100b0a、100b0b 和 100b0c。用这样的方式，利用在不超过两个链接上的传送，将该操作高效地广播到在数据处理系统 200 内的所有处理单元 100。

[0087] 如图 4A 和 5B 所示，在请求阶段之后，发生部分响应 (Presp) 阶段。在部分响应阶段，每个远程叶子 100 计算该操作并且将其各自的对该操作的部分响应提供给其各自的远程中心 100。例如，远程叶子 100b0a、100b0c 和 100b0c 将它们各自的部分响应传送到远程中心 100b0d。每个远程中心 100 又将这些部分响应、以及它自己的部分响应传送到节点主设备 100a0c。每一个节点叶子 100a0a、100a0b 和 100a0d 类似地计算该请求，并且将其各自的部分响应传送到节点主设备 100a0c。

[0088] 可以理解，以所示方式收集部分响应可以以多种不同的方式实现。例如，有可能从每个其它节点叶子、远程中心和远程叶子将单独的部分响应传递回到节点主设备。作为替换，为了更大的效率，可能期望在将部分响应传递回到发起处理节点时累积这些部分响应。为了确保将每个部分响应的效果准确地传递回到节点主设备 100，优选地，例如，利用逻辑或函数以及其中当经受这样的函数时不丢失相关信息的编码（例如，“一热”编码），以非破坏性方式累积部分响应（若有的话）。

[0089] 如图 4A 和图 5C 中进一步所示，在节点主设备 100a0c 处的响应逻辑 122 汇集其它处理单元 100 的部分响应，以获得代表对该请求的全系统范围响应的组合响应。节点主设备 100a0c 然后沿着与请求阶段所采用的路径相同的分发路径将该组合响应广播到所有处理单元 100。因此，该组合响应首先被广播到节点叶子 100a0a、100a0b 和 100a0d 以及远程中心 100b0d。远程中心 100b0d 又将该组合响应传送到每个远程叶子 100b0a、100b0b 和 100b0c。

[0090] 如上所述，如图 5D 或 5E 所示，为操作服务可能需要另外的数据阶段。例如，如图 5D 所示，如果操作是诸如读取或 RWITM 操作之类的读取类型的操作，则远程叶子 100b0b 可以经由将远程叶子 100b0b 连接到节点叶子 100a0a 的链接、以及将节点叶子 100a0a 连接到本地主设备 100a0c 的链接，将所请求的存储块供应到节点主设备 100a0c。相反，如果操作是写入类型的操作，例如将修改的存储块写回到远程叶子 100b0b 的系统存储器 132 中的高速缓存逐出操作，则如图 5E 所示，经由将节点主设备 100a0c 连接到远程中心 100b0d 的链接传送该存储块。

[0091] 当然，对于可能在诸如数据处理系统 200 或数据处理系统 220 之类的多处理器数据处理系统中同时发生的多种可能操作，图 4A-4C 中描述的操作仅仅是示例性的。

[0092] IV. 定时考虑

[0093] 如上面参考图 3 所述，通过保护窗口 312a、窗口扩展 312b、和保护窗口 313，在可能存在竞争相同存储块的所有权的其它主设备的情况下，在存储块的一致性所有权从窥探器 304n “移交”到请求主设备 300 的期间保持一致性。例如，如图 6 所示，保护窗口 312a 和窗口扩展 312b 一起必须具有足够的持续时间，以在存在竞争主设备 (CM)320 的竞争请求 322 的情况下，保护所请求的存储块的一致性所有权从窥探器 304n 到获胜主设备 (WM)300 的转移。为了确保保护窗口 312a 和窗口扩展 312b 具有足够的持续时间来保护所请求的存储块的所有权从窥探器 304n 转移到获胜主设备 300，优选地约束依据图 4A、4B 和 4C 的处理单元

100 之间的通信的延迟，使得满足以下条件：

[0094] $A\_lat(CM\_S) \leq A\_lat(CM\_WM)+C\_lat(WM\_S)+\varepsilon$，

[0095] 其中 A_lat(CM_S) 是任何竞争主设备 (CM)320 到拥有所请求存储块的一致性的窥探器 (S)304n 的寻址延迟，A_lat(CM_WM) 是任何竞争主设备 (CM)320 到由窥探器 304n 给予一致性所有权的“获胜”主设备 (WM)300 的寻址延迟，C_lat(WM_S) 是从由获胜主设备 (WM)300 收到组合响应的时间到由拥有所请求存储块的窥探器 (S)304n 收到组合响应的时间的组合响应延迟，以及 ε 是窗口扩展 312b 的持续时间。

[0096] 如果不满足可应用于任意拓扑结构的系统的前述定时约束，则可以 (1) 在获胜主设备 300 取得一致性所有权并且启动保护窗口 312b 之前由获胜主设备 300、以及 (2) 在保护窗口 312a 和窗口扩展 312b 结束之后由窥探器 304n，接收竞争主设备 320 的请求 322。在这样的情况下，获胜主设备 300 和窥探器 304n 都将不会对竞争请求 322 提供防止竞争主设备 320 取得存储块的一致性所有权并且从存储器中读取非一致数据的部分响应。然而，为了避免这个一致性错误，可以将窗口扩展 312b 可编程地（例如，通过适当地设置配置寄存器 123）设置为任意长度（ε），以补偿延迟差异或者物理实现的欠缺，其否则可能不能满足对于保持一致性必须满足的定时约束。因此，通过求解上述方程以获得 ε，可以确定用于任何实现的窗口扩展 312b 的理想长度。对于图 2A 和 2B 的数据处理系统实施例，如果对于具有包括多个处理节点 202 的范围的广播操作，ε 具有等于一个第一层链接芯片跳跃 (chip-hop) 的延迟的持续时间，并且对于仅仅节点范围的操作，ε 具有零持续时间，则这是优选的。

[0097] 可以进行有关上述定时约束的几个观察 (observation)。首先，从竞争主设备 320 到拥有窥探器 304a 的寻址延迟不必具有下限，但是必须具有上限。通过在尤其给定最大可能的振荡器漂移、耦接处理单元 100 的最长链接、累积停顿的最大数目、以及所保证的最坏情况吞吐量的情况下确定可得到的最坏情况延迟，来设计上限。为了确保观察到上限，互连构造必须确保非阻塞的行为。

[0098] 其次，从竞争主设备 320 到获胜主设备 300 的寻址延迟不必具有上限，但是必须具有下限。通过在尤其给定不存在停顿、处理单元 100 之间的最短可能链接、以及给定特定静态配置时的最慢振荡器漂移的情况下可得到的最佳情况延迟，来确定该下限。

[0099] 虽然对于给定操作，获胜主设备 300 与竞争主设备 320 中的每一个都仅仅具有一个定时限制用于其各自的请求，但是应当理解，在操作过程中，任何处理单元 100 对于一些操作可能是获胜主设备，而对于其它操作可能是竞争（以及失败）主设备。因此，每个处理单元 100 有效地具有用于其寻址延迟的上限和下限。

[0100] 第三，从生成组合响应的时间到由获胜主设备 300 观察到该组合响应的时间的组合响应延迟不必具有下限（该组合响应可能在任意的较早时间到达获胜主设备 300），但是必须具有上限。相反，从生成组合响应的时间直到由窥探器 304n 收到该组合响应为止的组合响应延迟具有下限，但是不必具有上限（虽然可以任意地施加上限来限制同时进行的操作数目）。

[0101] 第四，对部分响应延迟没有约束。也就是说，因为上面列举的定时约束的所有项都与请求 / 寻址延迟以及组合响应延迟有关，所以窥探器 304 和竞争主设备 320 到获胜主设备 300 的部分响应延迟不必具有上限或下限。

[0102] V. 示例性链接信息分配

[0103] 连接处理单元 100 的第一层和第二层链接可以各种方式实现以获得在图 2A 和 2B 中描述的拓扑结构以及满足图 6 所述的定时约束。在一个优选实施例中，每个进入和外出第一层 (X，Y 和 Z) 链接和每个进入和外出第二层 (A 和 B) 链接被实现为单向 8 字节总线，其包含多个不同的虚拟通道或者占有期 (tenure) 来传达地址、数据、控制和一致性信息。

[0104] 现在参见图 7A-7B，其中说明了用于第一层 X、Y 和 Z 链接以及第二层 A 和 B 链接的第一示例性时间分割信息分配。如同所示，在这个第一实施例中，信息以重复的 8 个周期帧分配在第一层和第二层链接上，其中前 4 个周期包含传输地址、一致性和控制信息的两个地址占有期，并且后 4 个周期专用于提供数据传输的数据占有期。

[0105] 首先参见图 7A，其说明了用于第一层链接的链接信息分配。在周期数对 8 取模的余数为 0 的每个周期中，字节 0 传递第一操作的事务类型 700a（例如，读取），字节 1-5 提供了第一操作的请求地址的较低地址字节 702a1，而且字节 6-7 形成保留字段 704。在下一个周期（即，周期数对 8 取模的余数是 1 的周期）中，字节 0-1 传递标识第一操作的主设备 300 的主设备标记 706a（例如，L2 高速缓存主设备 112 或者 I/O 控制器 128 内的主设备之一），而且字节 2 传达第一操作的请求地址的高地址字节 702a2。连同这个关于第一操作的信息一起传递的高达关于不同操作的三个另外的字段，即旨在用于在同一处理节点 202 中的本地主设备的本地部分响应 708a（字节 3-4），在字节 5 中的组合响应 710a，以及旨在用于在不同的处理节点 202 中的本地主设备的远程部分响应 712a（或者在仅仅节点范围的广播的情况下，从节点叶子 100 传递到节点主设备 100 的部分响应）（字节 6-7）。如上所述，这前两个周期形成了这里所谓的地址占有期。

[0106] 如图 7A 中进一步说明的那样，除了保留字段 704 被替换为形成数据占有期的一部分的数据标记 714 和数据令牌 715 之外，接下来的两个周期（即，周期数对 8 取模的余数为 2 和 3 的周期）形成了具有与第一地址占有期相同的基本模式的第二地址占有期。具体地说，数据标记 714 标识出现在周期 4-7 中的 32 个字节的数据有效载荷 716a-716d 被定向的目的数据宿。它在地址占有期内的位置紧邻在有效载荷数据之前，这有利地允许在接收有效载荷数据之前进行下游操纵的配置，并由此高效地将数据路由到指定的数据宿。数据令牌 715 提供了已被释放的下游队列条目的指示，并且因此，另外的数据可以在成对的 X、Y、Z 或者 A 链接上传送而没有过载的危险。再次应当注意，事务类型 700b、主设备标记 706b、低地址字节 702b1、和高地址字节 702b2 全都属于第二操作，而数据标记 714、本地部分响应 708b、组合响应 710b 以及远程部分响应 712b 全都涉及不同于第二操作的一个或多个操作。

[0107] 每个事务类型字段 700 和组合响应字段 710 优选地包括范围指示符 730，其能够指示它所属的操作具有仅仅节点（本地）的范围还是具有全系统（全局）的范围。当配置寄存器 123 被设置为以超级节点模式配置处理单元 100 时，不使用范围指示符 730，并且该范围指示符具有“无关”值。如在上面通过引用并入在此的、交叉引用的美国专利申请第 11/055,305 号中更详细描述的那样，数据标记 714 还包括域指示符 732，其可以由 LPC 设置来指示是否可能存在包含在数据有效载荷 716a-716d 内的数据的远程副本。优选地，当配置寄存器 123 被设置为以超级节点模式配置处理单元 100 时，也不使用域指示符 732，并且该域指示符具有“无关”值。

[0108] 图 7B 描述了用于第二层 A 和 B 链接的链接信息分配。通过与图 7A 的比较可以看

出，除了本地部分响应字段 708a、708b 被替换为保留字段 718a、718b 之外，在第二层 A 和 B 链接上的链接信息分配与在图 7A 中给出的用于第一层链接的分配相同。因为简单的原因，即作为第二层链接，不需要传递本地部分响应，所以进行这个替换。

[0109] 图 7C 说明了写入请求部分响应 720 的示例性实施例，其可以响应于写入请求，在本地部分响应字段 708a、708b 或者远程部分响应字段 712a、712b 内传输。如同所示，写入请求部分响应 720 在长度上为两个字节，并且包括 15 位目的地标记字段 724 和 1 位有效 (V) 标记 722，其中目的地标记字段 724 用于指定作为写入数据的目的地的窥探器（例如，IMC 窥探器 126）的标记，而有效标记 722 用于指示目的地标记字段 724 的有效性。

[0110] VI. 请求阶段结构和操作

[0111] 现在参见图 8，其描述了这样的框图，该框图说明了在操作的请求阶段处理中利用的、在图 1 的互连逻辑 120 内的请求逻辑 121a。如同所示，请求逻辑 121a 包括主设备多路复用器 900，其被耦接以由处理单元 100 的主设备 300（例如，L2 高速缓存 110 内的主设备 112 以及 I/O 控制器 128 内的主设备）接收请求。主设备多路复用器 900 的输出形成请求多路复用器 904 的一个输入。请求多路复用器 904 的第二输入端耦接到远程中心多路复用器 903 的输出端，该远程中心多路复用器 903 的输入端与保持缓冲器 902a、902b 的输出端相耦接，该保持缓冲器 902a、902b 又被耦接以分别在进入 A 和 B 链接上接收和缓冲请求。远程中心多路复用器 903 实现了下面详细描述的公平分配策略，其从在保持缓冲器 902a-902b 中缓冲的、从进入 A 和 B 链接接收的请求当中公平地进行选择。如果存在的话，由远程中心多路复用器 903 向请求多路复用器 904 提供的请求总是由请求多路复用器 904 给予优先权。请求多路复用器 904 的输出驱动请求总线 905，该请求总线 905 与输出 X、Y 和 Z 链接、节点主设备 / 远程中心 (NM/RH) 保持缓冲器 906、以及本地中心 (LH) 地址启动缓冲器 910 中的每一个相耦接。也耦接到请求总线 905 的先前请求 FIFO 缓冲器 907 优选地为多个先前地址占有期中的每一个保持少量地址相关的信息，以允许确定在该地址占有期中传递的地址（若有的话）哈希 (hash) 到其的地址片段或者资源存储体 1912。例如，在一个实施例中，先前请求 FIFO 缓冲器 907 的每个条目包含“一热”编码，其标识相关联请求的请求地址哈希到其的存储体 1912a-1912n 中的特定一个。对于其中没有在请求总线 905 上传送请求的地址占有期，一热编码将是全 0。

[0112] 进入第一层 (X,Y 和 Z) 链接每个都耦接到 LH 地址启动缓冲器 910，以及节点叶子 / 远程叶子 (NL/RL) 保持缓冲器 914a-914c 中的各个。NM/RH 保持缓冲器 906、LH 地址启动缓冲器 910、和 NL/RL 保持缓冲器 914a-914c 的输出全都形成窥探多路复用器 920 的输入。另一个先前缓冲器 911 也耦接到 LH 地址启动缓冲器 910 的输出，该先前缓冲器 911 优选地也类似于先前的请求 FIFO 缓冲器 907 进行构造。窥探多路复用器 920 的输出驱动窥探总线 922，其中标记 FIFO 队列 924、处理单元 100 的窥探器 304（例如，L2 高速缓存 110 的窥探器 116 和 IMC 124 的窥探器 126）、以及外出 A 和 B 链接耦接到该窥探总线 922。窥探器 304 还耦接到本地中心 (LH) 部分响应 FIFO 队列 930 和节点主设备 / 远程中心 (NM/RH) 部分响应 FIFO 队列 940，并且由它们支持。

[0113] 虽然其它实施例是可能的，但是如果缓冲器 902、906、和 914a-914c 保持简短以便最小化通信延迟，则这是优选的。在一个优选实施例中，控制每个缓冲器 902、906、和 914a-914c 的大小以仅仅保持选定链接信息分配的单个帧的地址占有期（一个或多个）。

[0114] 现在参见图 9，其中说明了图 8 的本地中心 (LH) 地址启动缓冲器 910 的更详细框图。如同所述，LH 地址启动缓冲器 910 的本地和进入 X、Y 和 Z 链接输入形成了映射逻辑 1010 的输入，该映射逻辑 1010 将在每个特定输入端上接收的请求放置到各自对应的位置相关 FIFO 队列 1020a-1020d 中。在所描述的命名法中，在处理节点 /MCM 202 的左上角的处理单元 100a 是“S”芯片；在处理节点 /MCM 202 的右上角的处理单元 100b 是“T”芯片；在处理节点 /MCM 202 的左下角的处理单元 100c 是“U”芯片；以及在处理节点 202 的右下角的处理单元 100d 是“V”芯片。因此，例如，对于本地主设备 / 本地中心 100ac，在本地输入端上接收的请求由映射逻辑 1010 放置在 U FIFO 队列 1020c 中，在进入 Y 链接上接收的请求由映射逻辑 1010 放置在 S FIFO 队列 1020a 中。采用映射逻辑 1010 来标准化输入流，以便同步所有本地中心 100 中的判优逻辑 1032（如下所述）来同一地处理请求而无需采用任何显式的相互通信。

[0115] 虽然被放置在位置相关 FIFO 队列 1020a-1020d 中，但是请求未被立即标记为有效并且可用于分派。相反，在每个位置相关 FIFO 队列 1020a-1020d 中的请求的验证经受可编程延迟 1000a-1000d 中的各自一个，以便同步在每个地址占有期期间在四个输入端上接收的请求。因此，与本地输入端相关联的可编程延迟 1000a 通常比与其它输入端相关联的可编程延迟明显更长，其中该本地输入端在本地主设备 / 本地中心 100 处接收请求的自我广播。为了确保验证适当的请求，由可编程延迟装置 1000a-1000d 生成的验证信号经受与底层请求相同的、由映射逻辑 1010 进行的映射。

[0116] 位置相关 FIFO 队列 1020a-1020d 的输出形成本地中心请求多路复用器 1030 的输入，该本地中心请求多路复用器 1030 响应于由判优器 1032 生成的选择信号，从位置相关 FIFO 队列 1020a-1020d 中选择一个请求以便提供给窥探多路复用器 920。判优器 1032 实现了公平判优策略，其在选择时与在给定处理节点 202 内的所有其它本地中心 100 的判优器 1032 同步，以便如图 4 和 5A 中所述，相同的请求由处理节点 202 中的所有本地中心 100 同时在外出 A 链接上广播。因此，给定图 7B 和 8B 所示的示例性链接信息分配中的任何一个，本地中心请求多路复用器 1030 的输出与外出 A 链接请求帧的地址占有期（一个或多个）时间片对齐。

[0117] 因为 LH 地址启动缓冲器 910 的输入带宽是其输出带宽的四倍，所以位置相关 FIFO 队列 1020a-1020d 的过载是设计考虑。在优选实施例中，通过为每个位置相关 FIFO 队列 1020 实现在大小上等于相关联的位置相关 FIFO 队列 1020 的深度的本地中心令牌池，来防止队列过载。需要空闲的本地中心令牌来让本地主设备发送请求到本地中心并且保证本地中心可以将该请求排入队列。因此，当由本地主设备 100 向本地中心 100 中的位置相关 FIFO 队列 1020 发出请求时分配本地中心令牌，并且当判优器 1032 从位置相关 FIFO 队列 1020 中发出条目 (entry) 时，释放该令牌以便重用。

[0118] 现在参见图 10，其中描述了图 8 的标记 FIFO 队列 924 的更详细框图。如同所示，标记 FIFO 队列 924 包括本地中心 (LH) 标记 FIFO 队列 924a、远程中心 (RH) 标记 FIFO 队列 924b0-924b1、节点主设备 (NM) 标记 FIFO 队列 924b2、远程叶子 (RL) 标记 FIFO 队列 924c0-924c1、924d0-924d1 和 924e0-924e1、以及节点叶子 (NL) 标记 FIFO 队列 924c2、924d2 和 924e2。当在以针对特定请求的这些给定角色 (LH、RH、和 RL) 中的每一个服务的处理单元（一个或多个）100 处接收到该请求时，全系统范围的操作的请求的主设备标记被

存放在标记 FIFO 队列 924a、924b0-924b1、924c0-924c1、924d0-924d1 和 924e0-924e1 的每一个中。类似地，当在以针对特定请求的这些给定角色 (NM 和 NL) 中的每一个服务的处理单元 (一个或多个)100 处接收到该请求时，仅仅节点范围的操作的请求的的主设备标记被存放在标记 FIFO 队列 924b2、924c2、924d2 和 924e2 的每一个中。当在相关联的处理单元 100 处接收到组合响应时，从标记 FIFO 队列 924 的每一个中检索主设备标记。因此，不是与组合响应一起传送主设备标记，而是根据需要由处理单元 100 从其标记 FIFO 队列 924 中检索主设备标记，从而导致在第一和第二层链接上的带宽节省。假定在各个处理单元 100 处接收组合响应的次序与接收相关联的请求的次序相同，则可以有利地采用用于分配和检索主设备标记的 FIFO 策略。

[0119] LH 标记 FIFO 队列 924a 包括多个条目，每个条目包括主设备标记字段 1100，用于存储由判优器 1032 启动的请求的主设备标记。标记 FIFO 队列 924b0-924b1 中的每一个类似地包括多个条目，每个至少包括主设备标记字段 1100，用于存储由远程中心 100 经由进入 A 和 B 链接中的相应一个接收的、全系统范围的请求的主设备标记。类似地构造标记 FIFO 队列 924c0-924c1、924d0-924d1 和 924e0-924e1，并且这些队列中的每个都保持由远程叶子 100 经由进入第一和第二层链接的唯一对接收的、全系统范围的请求的主设备标记。对于仅仅节点广播范围的请求，NM 标记 FIFO 队列 924b2 保持由节点主设备 100 发起的请求的主设备标记，并且 NL 标记 FIFO 队列 924c2、924d2 和 924e2 中的每一个提供了对由节点叶子 100 在第一层 X、Y 和 Z 链接的相应一个上接收的请求的主设备标记的存储。

[0120] 在 LH 标记 FIFO 队列 924a 内的条目具有全系统范围广播操作的最长占有期，并且 NM 标记 FIFO 队列 924b2 具有仅仅节点范围广播操作的最长占有期。因此，LH 标记 FIFO 队列 924a 和 NM 标记 FIFO 队列 924b2 的深度分别限制了处理节点 202 可以在互连构造上发出的、全系统范围的并行操作的数目，以及给定处理单元 100 可以在互连构造上发出的、仅仅节点范围的并行操作的数目。这些深度不一定有关系并且可以是不同的。然而，优选地将标记 FIFO 队列 924b0-924b1、924c0-924c1、924d0-924d1 和 924e0-924e1 的深度设计为等于 LH 标记 FIFO 队列 924a 的深度，以及优选地将标记 FIFO 队列 924c2、924d2 和 924e2 的深度设计为等于 NM 标记 FIFO 队列 924b2 的深度。

[0121] 现在参见图 11 和 12，其中说明了图 8 的本地中心 (LH) 部分响应 FIFO 队列 930 和节点主设备 / 远程中心 (NM/RH) 部分响应 FIFO 队列 940 的示例性实施例的更详细框图。如同所示，LH 部分响应 FIFO 队列 930 包括多个条目 1200，其中每个条目包括部分响应字段 1202 和响应标记阵列 1204，该部分响应字段 1202 用于为请求存储累积的部分响应，响应标记阵列 1204 具有相应的标记用于 6 个可能来源中的每一个，其中本地中心 100 可以在不同的时间或者可能同时从这些来源中接收部分响应 (即，本地 (L)、第一层 X、Y、Z 链接、以及第二层 A 和 B 链接)。在 LH 部分响应 FIFO 队列 930 内的条目 1200 经由分配指针 1210 分配，并且经由释放指针释放。利用 A 指针 1214、B 指针 1215、X 指针 1216、Y 指针 1218、和 Z 指针 1220 访问组成响应标记阵列 1204 的各个标记。

[0122] 如下面进一步描述的那样，当由本地中心 100 处的部分响应逻辑 121b 接收到对特定请求的部分响应时，在部分响应字段 1202 内累积该部分响应，并且通过在响应标记阵列 1204 内设置对应标记来记录从其接收到该部分响应的链接。指针 1214、1215、1216、1218 和 1220 中的对应一个然后前进到随后的条目 1200。

[0123] 当然，如上所述，每个处理单元 100 无需通过其 5 个进入 (X，Y，Z，A 和 B) 链接中的每一个全耦接到其它处理单元 100。因此，忽略在响应标记阵列 1204 内与未连接的链接相关联的标记。例如，可以通过在配置寄存器 123 中指示的配置来指示每个处理单元 100 中的未连接的链接（若有的话），该配置寄存器 123 可以例如在系统启动时由引导代码或者在对数据处理系统 200 进行分区时由操作系统进行设置。

[0124] 通过图 12 和图 11 的比较可以看出，NM/RH 部分响应 FIFO 队列 940 类似于 LH 部分响应 FIFO 队列 930 进行构造。NM/RH 部分响应 FIFO 队列 940 包括多个条目 1230，其中每个条目包括部分响应字段 1202 和响应标记阵列 1234，该部分响应字段 1202 用于存储累积的部分响应，而且响应标记阵列 1234 具有相应的标记用于高达 4 个可能来源中的每一个，其中节点主设备或者远程中心 100 可以从这些来源接收部分响应（即，节点主设备 (NM)/远程 (R)、以及第一层 X、Y、和 Z 链接）。此外，每个条目 1230 包括路由字段 1236，其标识操作是仅仅节点范围的广播操作还是全系统范围的广播操作，以及对于全系统范围的广播操作，该请求在哪个进入第二层链接上接收（以及因此累积的部分响应将在哪个外出第二层链接上传送）。在 NM/RH 部分响应 FIFO 队列 940 内的条目 1230 经由分配指针 1210 分配，并且经由释放指针 1212 释放。利用 X 指针 1216、Y 指针 1218、和 Z 指针 1220 访问和更新组成响应标记阵列 1234 的各个标记。

[0125] 如上面相对于图 11 所述，每个处理单元 100 无需通过其第一层 X、Y、和 Z 链接中的每一个全耦接到其它处理单元 100。因此，忽略在响应标记阵列 1204 内与未连接的链接相关联的标记。可以例如通过在配置寄存器 123 中指示的配置来指示每个处理单元 100 中的未连接链接（若有的话）。

[0126] 现在参见图 13A-13D，分别给出了描述根据本发明的示例性实施例、在请求阶段期间在本地主设备（或节点主设备）、本地中心、远程中心（或节点主设备）、以及远程叶子（或节点叶子）处的操作的示例性处理的流程图。现在具体参见图 13A，在本地主设备（或如果是仅仅节点或者超级节点广播的话，节点主设备）100 处的请求阶段处理从块 1400 开始，其中由本地（或节点）主设备 100 内的特定主设备 300（例如，L2 高速缓存 110 内的主设备 112 或者 I/O 控制器 128 内的主设备之一）生成请求。在块 1400 之后，处理继续到块 1402、1404、1406、和 1408，这些块中的每个都代表有关由特定主设备 300 发布请求的条件。在块 1402 和 1404 处说明的条件代表主设备多路复用器 900 的操作，并且在块 1406 和 1408 处说明的条件代表请求多路复用器 904 的操作。

[0127] 首先转向块 1402 和 1404，如果支配主设备多路复用器 900 的公平判优策略从（可能的）多个竞争主设备 300 的请求中选择特定主设备 300 的请求（块 1402），并且如果该请求是全系统范围的广播，如果存在可分配给该请求的本地中心令牌（块 1404），则主设备多路复用器 900 输出特定主设备 300 的请求。如块 1415 所示，如果主设备 300 将其请求范围选择为具有仅仅节点或者超级节点范围（例如，如在上面引用的美国专利申请第 11/055,305 号中所述的那样，通过参考配置寄存器 123 的设置和 / 或范围预测机制），则不需要本地中心令牌，并且省略在块 1404 处说明的条件。

[0128] 假定特定主设备 300 的请求通过主设备多路复用器 900 前进到请求多路复用器 904，则只有当在外出第一层链接信息分配中有地址占有期可用于请求时，请求多路复用器 904 才在请求总线 905 上发出该请求（块 1406）。也就是说，请求多路复用器 904 的输出与

选定链接信息分配进行时间片对齐，并且将仅仅在被设计成承载请求的周期（例如，图 7A 的实施例中的周期 0 或者 2）期间生成输出。如在块 1408 处进一步说明的那样，只有当不存在由远程中心多路复用器 903 提供的、来自进入的第二层 A 和 B 链接的请求（总是向该请求给予优先权）（块 1406）时，请求多路复用器 904 才会发布请求。因此，保证相对于进入请求、第二层链接不被阻塞。即使利用这样的不阻塞策略，也可以通过在上游中心的判优器 1032 中的适当策略的实现来防止主设备 300 发出的请求的“挨饿 (starving)”，其中上游中心的判优器 1032 中的适当策略防止在下游中心的进入 A 和 B 链接上的多个连续地址占有期期间“挡住 (brickwalling)”请求。

[0129] 如果在任何块 1402-1408 处做出了否定的确定，则如块 1410 处所指示的那样，延迟该请求直到其中在块 1402-1408 处说明的所有确定都是肯定的后续周期为止。相反，如果在所有块 1402-1408 上都做出了肯定的确定，则处理继续到块 1417。块 1417 代表仅仅节点范围（由 Ttype 字段 700 的范围指示符 730 所指示）或者超级节点范围（如由配置寄存器 123 所指示）的请求经受附加的条件。

[0130] 首先，如块 1419 所示，如果该请求是仅仅节点或者超级节点的广播请求，则只有当 NM 标记 FIFO 队列 924b2 中存在可分配给该请求的条目时，请求多路复用器 904 才将发布该请求。如果不存在的话，则该处理从块 1419 传递到已经描述的块 1410。

[0131] 其次，如块 1423 所述，在所描述的实施例中，只有当请求地址不哈希到与在先前请求 FIFO 缓冲器 907 内缓冲的、任意选定数目的先前请求相同的存储体资源 1910 中的存储体 1912 时，请求多路复用器 904 才将发布仅仅节点或者超级节点范围的请求。例如，假定窥探设备 1900 和它相关联的资源 1910 被构造成窥探设备 1900 不能以最大请求到达速率为请求服务，而是可以被表达为 1/R 的最大到达速率的一部分来为请求服务，则先前请求的选定数目优选地为 R-1，其中将为由请求多路复用器 904 启动而竞争的当前仅仅节点范围请求与这些先前请求进行比较，以确定它是否在同一地址片段中。如果要以这样的方式保护多个不同的窥探设备 1900 以防止请求过载，则优选地将请求的选定数目 R-1 设置成为各个窥探设备 1900 计算的量 R-1 的集合中的最大值。因为处理单元 100 优选地不协调它们对用于广播的请求的选择，所以以在块 1423 处所述的方式进行请求调速 (throttle) 不保证在特定窥探设备 1900 处的请求到达速率不会超过窥探设备 1900 的服务速率。然而，以所示方式对仅仅节点范围的广播请求进行调速将限制在给定数目的周期中可到达的请求数目，其可以表示为：

[0132] throttled_arr_rate ＝每 R 个周期的 PU 请求数

[0133] 其中 PU 是每个处理节点 202 的处理单元 100 的数目。窥探设备 1900 优选地被设计为处理以这样已调速的到达速率到达的、仅仅节点范围的广播请求而不重试。

[0134] 如果不满足在块 1423 处所示的条件，则处理从块 1423 传递到已经描述的块 1410。如果满足在块 1419 和 1423 处说明的条件，则如果请求为仅仅节点范围，则请求多路复用器 904 在请求总线 905 上发布请求，并且该处理通过页面连接符 1425 传递到图 13C 的块 1427。相反，如果如在块 1401 处确定该请求为超级节点范围，则只有当请求多路复用器 904 确定了它在连续的地址占有期中没有输出太多请求时，它才发布该请求。具体地说，如块 1403 所示，为了避免使在 A 和 / 或 B 链接上的进入请求挨饿，请求多路复用器 904 在不超过可用地址占有期一半（即，1/t2）的期间启动主设备 300 的请求。如果满足了在块 1401 处所描

述的条件，则请求多路复用器 904 在请求总线 905 上发布超级节点请求，并且该处理通过页面连接符 1425 传递到图 13C 的块 1427。如果不满足在块 1401 处所描述的条件，则该处理从块 1401 传递到已经描述的块 1410。

[0135] 再次返回到块 1417，如果该请求是全系统范围的广播请求而不是仅仅节点范围或者超级节点广播请求，则该处理继续到块 1412。块 1412 描述了请求多路复用器 904 在请求总线 905 上将该请求广播到外出 X、Y 和 Z 链接的每一个上以及广播到本地中心地址启动缓冲器 910。此后，该处理分叉并且通过页面连接符 1414 和 1416 传递到图 13B，该图 13B 说明了在每个本地中心 100 处对请求的处理。

[0136] 现在参见图 13B，在也是本地主设备 100 的本地中心 100 处对全系统范围请求的处理被示出为从块 1416 开始，并且在与本地主设备 100 相同的处理节点 202 中的每个其它本地中心 100 处对请求的处理被描述为从块 1414 开始。首先转向块 1414，由本地中心 100 在进入 X、Y 和 Z 链接上接收的请求由 LH 地址启动缓冲器 910 所接收。如块 1420 和图 9 中所述，映射逻辑 1010 将每个 X、Y 和 Z 请求映射到位置相关 FIFO 队列 1020a-1020d 中的适当一个以便缓冲。如上所述，在 X、Y 和 Z 链接上接收并且放置在位置相关队列 1020a-1020d 内的请求不被立即验证。相反，这些请求经受调整延迟装置 1000a-1000d 中的相应一个，这些延迟同步对给定本地中心 100 上的 X、Y 和 Z 请求以及本地请求的处理和对同一处理节点 202 中的其它本地中心 100 处的对应请求的处理（块 1422）。此后，如块 1430 所示，调整延迟装置 1000 验证在位置相关 FIFO 队列 1020a-1020d 中的它们各自的请求。

[0137] 现在参见块 1416，在本地主设备 / 本地中心 100 处，请求总线 905 上的请求被直接馈送到 LH 地址启动缓冲器 910 中。因为没有穿越芯片间链接，所以这个本地请求到达 LH 地址启动 FIFO 910 早于在相同周期中发布的请求到达进入 X、Y 和 Z 链接。因此，在由块 1424 所述的、由映射逻辑 1010 进行的映射之后，调整延迟装置 1000a-1000d 之一将长延迟应用到该本地请求以同步其验证与在进入 X、Y 和 Z 链接上接收的请求的验证（块 1426）。在这个延迟间隔之后，如块 1430 所示，相关调整延迟装置 1000 验证该本地请求。

[0138] 在块 1430 处对在 LH 地址启动缓冲器 910 内排队的请求进行验证之后，该处理然后继续到块 1434-1440，其中每个块都代表由判优器 1032 强制的、从 LH 地址启动缓冲器 910 中发布请求的条件。如上所述，同步在所有处理单元 100 内的判优器 1032，以便由所有本地中心 100 进行相同的判定而无需相互通信。如块 1434 所述，只有当在外出第二层链接信息分配中存在可用于请求的地址占有期时，判优器 1032 才允许本地中心请求多路复用器 1030 输出该请求。因此，例如，判优器 1032 导致本地中心请求多路复用器 1030 仅仅在图 7B 的实施例的周期 0 或 2 期间才启动请求的传送。此外，如果由判优器 1032 实现的公平判优策略确定请求属于接下来应该被服务的位置相关 FIFO 队列 1020a-1020d，则由本地中心请求多路复用器 1030 输出该请求（块 1436）。

[0139] 如在块 1437 和 1438 处进一步描述的那样，判优器 1032 导致只有当本地中心请求多路复用器 1030 确定它没有在连续的地址占有期中输出太多请求时，它才输出请求。具体地说，如块 1437 所示，为了避免过度驱动连接到外出 A 和 B 链接的中心 100 的请求总线 905，判优器 1032 假定最坏情况（即，连接到下游中心 100 的另一个第二层链接的上游中心 100 在相同的周期内传送请求），并且在不超过可用地址占有期一半（即，1/t2）的期间启动请求。此外，如块 1438 所述，判优器 1032 还将请求的启动限制为低于通信量（traffic）在第

二层链接上的公平分配，以避免可能使耦接到其外出 A 和 B 链接的处理单元 100 中的主设备 300“挨饿”。

[0140] 例如，给定图 2A 的实施例，其中每个处理节点有 2 对第二层链接和 4 个处理单元 100，在下游中心 100 的请求总线 905 上的通信量经受高达 9 个处理单元 100 的竞争，即在通过第二层链接耦接到下游中心 100 的 2 个处理节点 202 的每一个中的 4 个处理单元 100，以及下游中心 100 本身。因此，将请求总线 905 的带宽在可能的请求源中平均分配的示例性公平分配策略向进入 A 和 B 链接中的每一个分配 4/9 的带宽，并且向本地主设备 300 分配 1/9 的带宽。针对任意数目的第一和第二层链接而推广，由判优器 1032 所采用的示例性公平分配策略所消耗的、分配的可用地址帧的部分可以表示为：

[0141] 部分＝ (t1/2+1)/(t2/2*(t1/2+1)+1)

[0142] 其中 t1 和 t2 代表处理单元 100 可以耦接到其的第一和第二层链接的总数，量“t1/2+1”代表每个处理节点 202 的处理单元 100 数目，量“t2/2”代表下游中心 100 可以耦接到其的处理节点 202 的数目，并且常量“1”代表分配给下游中心 100 的部分带宽。

[0143] 如块 1439 所示，判优器 1032 还对全系统范围广播请求的传送进行调速，这是通过只有当请求地址不哈希到与在先前请求 FIFO 缓冲器 911 内缓冲的 R-1 个先前请求中的任何一个请求相同的存储体资源 1910 的存储体 1912 时才发布全系统范围的广播请求，其中 1/R 是最大到达速率的一部分，并且最慢的受保护窥探设备 1900 可以该到达速率为请求服务。因此，以所示方式对全系统范围的广播请求进行调速将限制在给定数目的周期内可到达给定窥探设备 1900 的请求的数目，其可以表示为：

[0144] throttled_arr_rate ＝每 R 个周期的 N 个请求

[0145] 其中 N 是处理节点 202 的数目。窥探设备 1900 优选地被设计为处理以这样已调速的到达速率到达的请求而不重试。

[0146] 最后参见块 1440 所示的条件，只有当 LH 标记 FIFO 队列 924a 中有可用于分配的条目时，判优器 1032 才允许由本地中心请求多路复用器 1030 输出请求（块 1440）。

[0147] 如果在任何块 1434-1440 处做出了否定的确定，则如块 1442 所指示的那样，延迟该请求直到其中在块 1434-1440 处说明的所有确定都是肯定的后续周期为止。相反，如果在所有块 1434-1440 处进行了肯定的确定，则判优器 1032 向本地中心请求多路复用器 1030 通知将选定请求输出到多路复用器 920 的输入端，该多路复用器 920 总是向由 LH 地址启动缓冲器 910 提供的请求（若有的话）给予优先权。因此，多路复用器 920 在窥探总线 922 上发布该请求。应当注意，多路复用器 920 的其它端口（例如，RH、RLX、RLY、和 RLZ）可以与 LH 地址启动缓冲器 910 一起同时提供请求，这意指窥探总线 922 的最大带宽必须等于外出 A 和 B 链接的带宽的 10/8（假定图 7B 的实施例），以便跟得上最大到达速率。

[0148] 还应当观察到，仅仅在本地中心地址启动缓冲器 910 内缓冲的请求在外出 A 和 B 链接上传送，并且要求与链接信息分配内的地址占有期对齐。因为所有其它竞争由多路复用器 920 发布的请求仅仅以本地窥探器 304 以及它们相应的 FIFO 队列而不是外出 A 和 B 链接作为目标，所以可以在信息帧的剩余周期内发布这样的请求。因此，与由多路复用器 920 采用的特定判优方案无关，保证所有同时向多路复用器 920 提供的请求在单个信息帧的延迟之内传送。

[0149] 如块 1444 所示，响应于请求在窥探总线 922 上的发布，LH 标记 FIFO 队列 924a 在

下一个可用条目的主设备标记字段 1100 中记录在该请求中指定的主设备标记。然后，如块 1446 所示，将该请求路由到外出 A 和 B 链接。该处理然后通过页面连接符 1448 传递到图 13B，其描述了在请求阶段期间在每个远程中心处对请求的处理。

[0150] 图 13B 中描述的处理还从块 1446 继续到块 1450，其说明了响应于从 LH 地址启动缓冲器 910 中移除请求，本地中心 100 释放分配给该请求的本地中心令牌。如块 1452 所示，该请求被进一步路由到本地中心 100 中的窥探器 304。响应于该请求的接收，窥探器 304 生成部分响应（块 1454），其被记录在 LH 部分响应 FIFO 队列 930 内（块 1456）。具体而言，在块 1456 处，通过参考分配指针 1210 将 LH 部分响应 FIFO 队列 930 中的条目 1200 分配给该请求，递增分配指针 1210，在所分配条目的部分响应字段 1202 内放置本地中心的部分响应，并且在响应标记字段 1204 中设置本地 (L) 标记。此后，在块 1458 处，结束在本地中心 100 处的请求阶段处理。

[0151] 现在参见图 13C，其中描述了根据本发明、在远程中心（或者对于仅仅节点范围的广播请求，节点主设备）100 处的请求处理的示例性方法的高级逻辑流程图。如同所述，对于全系统范围或者超级节点广播请求，当在远程中心 100 处在其进入 A 和 B 链接之一上收到请求时，该处理从页面连接符 1448 开始。如上所述，在如块 1460 所示将请求锁存到保持缓冲器 902a-902b 的相应一个中之后，如块 1464 和 1465 所述，由远程中心多路复用器 903 和请求多路复用器 904 计算该请求以便在请求总线 905 上传送该请求。具体而言，在块 1464 处，远程中心多路复用器 903 根据将地址占有期平均地分配给在进入第二层链接上接收的请求的公平分配策略，确定是否输出全系统范围的广播请求（超级节点请求总是“获胜”的请求，这是因为没有由节点主设备 100 在其它第二层链接上同时供应的竞争请求）。此外，如块 1465 所述，只有当地址占有期可用时，与第一层链接信息分配时间片对齐的请求多路复用器 904 才会输出请求。因此，如块 1466 所示，如果请求不是多路复用器 903 的公平分配策略下的获胜请求，如果适用，或者如果没有地址占有期是可用的，则多路复用器 904 等待下一地址占有期。然而应当理解，即使延迟了在进入第二层链接上接收的请求，该延迟也将不会超过第一层链接信息分配的一帧。

[0152] 如果满足在块 1464 和 1465 处描述的条件，则多路复用器 904 在请求总线 905 上启动该请求，并且该处理从块 1465 继续到块 1468。如同所示，从图 13A 的块 1421 开始、在块 1423 处继续的节点主设备 100 处的请求阶段处理也传递到块 1468。块 1468 说明了在请求总线 905 上发布的请求到外出 X、Y 和 Z 链接以及 NM/RH 保持缓冲器 906 的路由。在块 1468 之后，该处理分叉。第一路径通过页面连接符 1470 传递到图 13D，其说明了在远程（或节点）叶子 100 处的请求处理的示例性方法。第二路径从块 1468 继续到块 1474，其说明了窥探多路复用器 920 确定要将在其输入端提供的哪个请求在窥探总线 922 上输出。如同所示，窥探多路复用器 920 使本地中心请求优先于远程中心请求，远程中心请求又优先于在 NL/RL 保持缓冲器 914a-914c 中缓冲的请求。因此，如果由 LH 地址启动缓冲器 910 提供了本地中心请求以便选择，则如块 1476 所示，延迟在 NM/RH 保持缓冲器 906 内缓冲的请求。然而，如果没有由 LH 地址启动缓冲器 910 提供请求，则窥探多路复用器 920 将来自 NM/RH 保持缓冲器 906 的请求发布在窥探总线 922 上（在超级节点请求的情况下，没有由 LH 地址启动缓冲器 910 提供竞争请求，并且在块 1474 处描述的确定将总是具有否定的结果）。

[0153] 响应于在窥探总线 922 上检测到请求，标记 FIFO 队列 924b 中的适当一个（即，在

节点主设备处，NM 标记 FIFO 队列 924b2，或者在远程中心处，与在其上接收了请求的进入第二层链接相关联的 RH 标记 FIFO 队列 924b0 和 924b1 之一）将由该请求指定的主设备标记放置到其下一个可用条目的主设备标记字段 1100 中（块 1478）。如上所述，仅仅节点范围的广播请求和全系统范围的广播请求通过该请求的 Ttype 字段 700 内的范围指示符 730 来区分，而超级节点模式由配置寄存器 123 所指示。如块 1480 所示，该请求被进一步路由到节点主设备 100 或远程中心 100 中的窥探器 304。此后，该处理分叉并且继续到块 1482 和 1479 中的每一个。

[0154] 首先参见块 1482，窥探器 304 响应于请求的接收而生成部分响应，并且将该部分响应记录在 NM/RH 部分响应 FIFO 队列 940 内（块 1484）。具体而言，通过参考分配指针 1210 将 NM/RH 部分响应 FIFO 队列 940 中的条目 1230 分配给该请求，递增该分配指针 1210，在部分响应字段 1202 内放置远程中心的部分响应，并且在响应标记字段 1234 中设置节点主设备 / 远程标志 (NM/R)。应当注意到，NM/RH 部分响应 FIFO 队列 940 因此以相同的数据结构缓冲不同范围的操作的部分响应。此外，如块 1483 和 1485 所示，如果该请求是在节点主设备 100 处的超级节点请求，则还在 LH 部分响应 FIFO 队列 930 的条目 1200 内屏蔽处理器 100 的部分响应，并且设置响应标记阵列 1204 内的本地标志。在块 1483 或者块 1485 之后，在节点主设备 100 或者远程中心 100 处的请求阶段处理在块 1486 结束。

[0155] 现在转向块 1479，如果配置寄存器 123 指示超级节点模式并且该处理器是节点主设备 100，则进一步将该请求路由到第二层链接中的预定一个（例如，链接 A）。然后，该处理通过块 1477 传递到块 1448，其中块 1448 表示在远程中心 100 处对该请求的请求阶段处理。相反，如果在块 1479 处做出了否定的确定，则该处理简单地在块 1481 终止。

[0156] 现在参见图 13D，其中说明了根据本发明、在远程叶子（或节点叶子）100 处的请求处理的示例性方法的高级逻辑流程图。如同所示，当在远程叶子或节点叶子 100 处在其进入 X、Y 和 Z 链接之一上接收到请求时，该处理从页面连接符 1470 开始。如块 1490 所示，响应于请求的接收，该请求被锁存到与在其上接收了请求的第一层链接相关联的 NL/RL 保持缓冲器 914a-914c 中的特定一个中。接下来，如块 1491 所述，该请求连同在窥探多路复用器 920 的输入端提供的其它请求一起由窥探多路复用器 920 计算。如上所述，窥探多路复用器 920 使本地中心请求优先于远程中心请求，远程中心请求又优先于在 NL/RL 保持缓冲器 914a-914c 中缓冲的请求。因此，如果提供了本地中心或者远程中心请求以便选择，则如块 1492 所示，延迟在 NL/RL 保持缓冲器 914 内缓冲的请求。然而，如果没有向窥探多路复用器 920 提供较高优先级的请求，则窥探多路复用器 920 通过在 X、Y 和 Z 请求之间进行公平的选择，在窥探总线 922 上发布来自 NL/RL 保持缓冲器 914 的请求。

[0157] 响应于在窥探总线 922 上检测到请求，与请求范围以及接收该请求的路由相关联的标记 FIFO 队列 924c0-924c2、924d0-924c2 和 924e0-924e2 中的特定一个将由该请求指定的主设备标记放置到其下一个可用条目的主设备标记字段 1100 中（块 1493）。也就是说，在该请求的 Ttype 字段 700 内的范围指示符 730 用来确定该请求是仅仅节点范围还是全系统范围，而配置寄存器 123 的设置用来指示超级节点模式。对于仅仅节点范围以及超级节点广播请求，与在其上接收了请求的进入第一层链接相关联的 NL 标记 FIFO 队列 924c2、924d2 和 924e2 中的特定一个缓冲该主设备标记。对于全系统范围和超级节点广播请求，将主设备标记放置在与在其上接收了请求的进入第一和第二层链接的组合相对应的远程节

点中的 RL 标记 FIFO 队列 924c0-924c1、924d0-924d1 和 924e0-924e1 中的特定一个中。如块 1494 所示，该请求被进一步路由到远程叶子 100 中的窥探器 304。响应于请求的接收，窥探器 304 处理该请求，生成它们各自的部分响应，并且累积该部分响应以获得该处理单元 100 的部分响应（块 1495）。如页面连接符 1497 所示，远程叶子或节点叶子 100 的窥探器 304 的部分响应根据下面所述的图 15A 进行处理。

[0158] 图 13E 是示例性方法的高级逻辑流程图，窥探器 304 利用该方法例如在图 13B-13D 的块 1454、1482 和 1495 处生成对请求的部分响应。响应于由窥探器 304（例如，IMC 窥探器 126、L2 高速缓存窥探器 116 或者 I/O 控制器 128 内的窥探器）接收到请求，该处理从块 1401 开始。响应于请求的接收，窥探器 304 通过参考由该请求指定的事务类型，确定该请求是否是诸如逐出请求、写入请求、或者部分写入请求之类的写入类型的请求。响应于窥探器 304 在块 1403 处确定该请求不是写入类型的请求（例如，读取或者 RWITM 请求），该处理继续到块 1405，其中说明了如果需要的话，窥探器 304 通过常规处理生成对该请求的部分响应。然而，如果窥探器 304 确定该请求是写入类型的请求，则该处理继续到块 1407。

[0159] 块 1407 描述了窥探器 304 确定它是否是用于由该写入类型请求所指定的请求地址的 LPC。例如，窥探器 304 可以通过参考一个或多个基地址寄存器 (BAR) 和 / 或指定窥探器 304 所负责的地址范围（即，LPC) 的地址哈希函数，来进行所述确定。如果窥探器 304 确定它不是用于该请求地址的 LPC，则该处理传递到块 1409。块 1409 说明了窥探器 304 生成写入请求部分响应 720（图 7C)，其中有效字段 722 和目的地标记字段 724 由全‘0’形成，由此表示该窥探器 304 不是用于该请求地址的 LPC。然而，如果窥探器 304 在块 1407 处确定它是用于该请求地址的 LPC，则该处理传递到块 1411，其中描述了窥探器 304 生成写入请求部分响应 720，其中有效字段 722 被设置为‘1’，并且目的地标记字段 724 指定了唯一地标识窥探器 304 在数据处理系统 200 内的位置的目的地标记或者路由。在块 1409 或 1411 中的任一个之后，图 13E 所示的处理在块 1413 结束。

[0160] VII. 部分响应阶段结构和操作

[0161] 现在参见图 14，其中描述了说明在图 1 的互连逻辑 120 内的部分响应逻辑 121b 的示例性实施例的框图。如同所示，部分响应逻辑 121b 包括路由逻辑 1500，其将由远程叶子（或节点叶子）100 处的窥探器 304 生成的远程部分响应路由回到远程中心（或节点主设备）100，其中经由外出第一层 X、Y 和 Z 链接中的适当一个从该远程中心（或节点主设备）100 接收了该请求。此外，部分响应逻辑 121b 包括组合逻辑 1502 和路由逻辑 1504。组合逻辑 1502 累积从远程（或节点）叶子 100 接收的部分响应以及在 NM/RH 部分响应 FIFO 队列 940 内缓冲的、对相同请求的其它部分响应（一个或多个）。对于仅仅节点范围的广播操作，节点主设备 100 的组合逻辑 1502 直接将累积的部分响应提供给响应逻辑 122。对于全系统范围或者超级节点广播操作，组合逻辑 1502 将累积的部分响应提供给路由逻辑 1504，其经由外出 A 和 B 链接之一将累积的部分响应路由到本地中心 100。

[0162] 部分响应逻辑 121b 还包括：保持缓冲器 1506a-1506b，其从远程中心 100 接收和缓冲部分响应；多路复用器 1507，其应用公平判优策略以从在保持缓冲器 1506a-1506b 内缓冲的部分响应当中进行选择；以及广播逻辑 1508，其将由多路复用器 1507 选择的部分响应广播到其处理节点 202 内的每个其它处理单元 100。如将多路复用器 1507 的输出端耦接到可编程延迟装置 1509 的路径进一步所示，多路复用器 1507 执行由可编程延迟装置 1509

延迟大约一个第一层链接延迟时间的部分响应的本地广播，以便与在进入 X、Y 和 Z 链接上从其它处理单元 100 接收的部分响应大致同时地由组合逻辑 1510 接收该本地广播部分响应。组合逻辑 1510 累积在进入 X、Y 和 Z 链接上接收的部分响应和从进入第二层链接接收的本地广播部分响应以及本地生成的部分响应（其在 LH 部分响应 FIFO 队列 930 内缓冲），并且当不处于超级节点模式时，将所累积的部分响应传送到响应逻辑 122，以便生成该请求的组合响应。

[0163] 现在参见图 15A-15C，其中说明了分别描述了在远程叶子（和节点叶子）、远程中心（和对于非超级节点模式操作的节点主设备）、以及本地中心（或者对于超级节点模式操作的节点主设备）处的操作的部分响应阶段期间的示例性处理。在这些附图中，部分响应的传送可能经受未显式示出的各种延迟。然而，因为如上所述、在部分响应延迟上没有定时约束，所以这样的延迟（如果存在的话）将不会导致操作出错，并因此不在此处进行进一步的描述。

[0164] 现在具体参见图 15A，当远程叶子（或节点叶子）100 的窥探器 304 生成请求的部分响应时，在远程叶子（或节点叶子）100 处的部分响应阶段处理从块 1600 开始。如块 1602 所示，路由逻辑 1500 然后使用链接信息分配的远程部分响应字段 712，经由与在其上接收了请求的进入第一层链接相对应的外出 X、Y 或 Z 链接，将该部分响应路由到用于该请求的远程中心 100。如上所示，在其上接收了请求的进入第一层链接通过标记 FIFO 队列 924c0-924c2、924d0-924d2 和 924e0-924e2 中的哪一个保持用于该请求的主设备标记来指示。此后，如页面连接符 1604 所示以及如下面参考图 15B 所述，在远程中心（或节点主设备）100 处继续部分响应处理。

[0165] 现在参见图 15B，其中说明了根据本发明、在远程中心（以及对于非超级节点模式操作，在节点主设备）处的部分响应处理的方法的示例性实施例的高级逻辑流程图。响应于接收到通过第一层 X、Y 和 Z 链接之一耦接到远程中心（或节点主设备）100 的远程叶子（或节点叶子）100 之一的部分响应，所述处理从页面连接符 1604 开始。响应于该部分响应的接收，组合逻辑 1502 读出分配给该操作的、在 NM/RH 部分响应 FIFO 队列 940 内的条目 1230。如与在其上接收了部分响应的链接相关联的 X、Y 或 Z 指针 1216-1220 所示，由在 NM/RH 部分响应 FIFO 队列 940 内所观察到的 FIFO 次序标识该条目。组合逻辑 1502 然后累积远程（或节点）叶子 100 的部分响应与所读取条目 1230 的部分响应字段 1202 的内容。如上所述，累积操作优选地是诸如逻辑或操作之类的非破坏性操作。如块 1605 和 1607 所示，对于超级节点模式中在节点主设备 100 处的请求，还在 LH 部分响应 FIFO 队列 930 的条目 1200 内屏蔽所累积的部分响应，并且设置响应标志阵列 1204 内的适当标志。在块 1605 或块 1607 之后，该处理继续到块 1614。在块 1614 处，组合逻辑 1502 通过参考 NM/RH 部分响应 FIFO 队列 940 中的条目 1230 的响应标志阵列 1234，利用在块 1604 接收的部分响应来确定所有的远程（或节点）叶子 100 是否已经报告了它们各自的部分响应。如果否，则该处理继续到块 1616，其中说明了组合逻辑 1502 用所累积的部分响应更新分配给该操作的条目 1230 的部分响应字段 1202，在响应标志阵列 1234 中设置适当的标志来指示哪个远程（或节点）叶子 100 提供了部分响应，以及使指针 1216-1220 中相关联的一个前进。此后，在块 1618 处结束该处理。

[0166] 再次参见块 1614，响应于由组合逻辑 1502 确定所有远程（或节点）叶子 100 都已

经报告了它们各自的对该操作的部分响应，组合逻辑 1502 通过参考释放指针 1212，从 NM/RH 部分响应 FIFO 队列 940 中释放用于该操作的条目 1230（块 1620）。接下来，如块 1621 所述，组合逻辑 1502 检查已释放条目的路由字段 1236，以确定该操作的范围。如果已释放条目的路由字段 1236 指示该操作在远程节点处处理，则如块 1622 所述，组合逻辑 1502 利用链接分配信息中的远程部分响应字段 712，将所累积的部分响应路由到由路由字段 1236 的内容所指示的外出 A 和 B 链接中的特定一个。（超级节点模式中的操作的部分响应优选地在第二层链接的预定一个（例如，链接 A）上传送。）此后，该处理通过页面连接符 1624 传递到图 15C。再次参见块 1621，如果该条目的路由字段 1236 指示该操作在节点主设备 100 处处理，则如果配置寄存器 123 未指示超级节点模式，则组合逻辑 1502 直接将所累积的部分响应提供给响应逻辑 122（块 1617）。此后，该处理通过页面连接符 1625 传递到下面所述的图 17A。然而，如果组合逻辑 1502 在块 1617 确定配置寄存器 123 指示超级节点模式，则该处理简单地在块 1619 结束而不用让组合逻辑 1502 将从 NM/RH 部分响应 FIFO 队列 940 中释放的部分响应路由到响应逻辑 122。不需要这样的路由是因为如下面参考图 15C 所述，用于这样操作的组合响应从由 LH 部分响应 FIFO 队列 930 所维护的屏蔽副本中生成。

[0167] 现在参见图 15C，其中描述了根据本发明的实施例、在本地中心 100（对于超级节点模式，包括本地主设备 100 或者节点主设备 100）处的部分响应处理的示例性方法的高级逻辑流程图。响应于在本地中心 100 处经由进入 A 和 B 链接之一从远程中心 100 接收到部分响应，该处理从块 1624 开始。在接收时，该部分响应被放置在耦接到在其上接收了部分响应的进入第二层链接的保持缓冲器 1506a、1506b 内（块 1626）。如块 1627 所示，如果配置寄存器 123 未指示超级节点模式，则多路复用器 1507 应用公平判优策略，以从在保持缓冲器 1506a-1506b 内缓冲的部分响应当中进行选择。因此，如果通过公平判优策略未选择部分响应，则如块 1628 所示，延迟部分响应的广播。必要时，一旦通过公平判优策略（可能在延迟之后）选择了部分响应，则多路复用器 1507 将部分响应输出到广播逻辑 1508 和可编程延迟装置 1509。因为部分响应的到达速率由请求启动的速率所限制，所以多路复用器 1507 的输出总线将不会因部分响应而变得过载。如块 1625 所示，如果配置寄存器 123 未指示超级节点模式，则该处理接下来继续到块 1629，否则省略块 1629 并且直接继续到块 1630。

[0168] 块 1629 描述了广播逻辑 1508 经由第一层 X、Y 和 Z 链接将由多路复用器 1507 选择的部分响应广播到其处理节点 202 中的每个其它处理单元 100，并且多路复用器 1507 通过将部分响应输出到可编程延迟装置 1509 来执行部分响应的本地广播。此后，该处理分叉并且继续到块 1631 和块 1630 中的每一个，其中块 1631 说明了在其它本地中心 100 处的部分响应阶段处理的继续。如块 1630 所示，如果配置寄存器 123 未指示超级节点模式，则当前本地中心 100 内的部分响应广播由有选择应用的可编程延迟装置 1509 延迟大约第一层链接的传送延迟时间，以便该本地广播部分响应与在进入 X、Y 和 Z 链接上、从其它处理单元 100 接收的部分响应（一个或多个）大约同时地由组合逻辑 1510 接收。如块 1640 所示，组合逻辑 1510 将远程中心 100 的本地广播部分响应与从进入第一层链接（一个或多个）接收的部分响应（一个或多个），以及与在 LH 部分响应 FIFO 队列 930 内缓冲的本地生成的部分响应一起累积。

[0169] 为了累积部分响应，组合逻辑 1510 首先读出 LH 部分响应 FIFO 队列 930 内分配给

该操作的条目1200。如与在其上接收了本地广播部分响应的链接相对应的指针1214、1215中的特定一个所示，该条目由LH部分响应FIFO队列930内所观察到的FIFO次序所标识。组合逻辑1510然后累积远程中心100的本地广播部分响应与所读取条目1200的部分响应字段1202的内容。接下来，如块1642所示，组合逻辑1510还通过参考条目1200的响应标志阵列1204，利用当前接收的部分响应（一个或多个），确定是否已经从被期望部分响应的每个处理单元中接收了部分响应。如果否，则该处理传递到块1644，其描述了组合逻辑1510用新累积的部分响应更新从LH部分响应FIFO队列930中读取的条目1200。此后，在块1646结束该处理。

[0170] 回到块1642，如果组合逻辑1510确定所有被期望部分响应的处理单元100都已经报告了它们的部分响应，则该处理继续到块1650。块1650描述了组合逻辑1510通过参考释放指针1212从LH部分响应FIFO队列930中释放分配给该操作的条目1200。然后如块1652所述，组合逻辑1510将所累积的部分响应传递到响应逻辑122，以便生成组合响应。此后，该处理通过页面连接符1654传递到图17A，其中说明了在本地中心（或节点主设备）100处的组合响应处理。

[0171] 现在参见块1632，当由组合逻辑1510接收部分响应（一个或多个）时，在非超级节点模式中在一个或多个第一层链接上由本地中心100接收的部分响应（一个或多个）的处理开始。如块1634所示，组合逻辑1510可以向在进入第一层链接上接收的部分响应（一个或多个）应用小的调整延迟，以便使部分响应（一个或多个）的处理彼此同步以及与本地广播部分响应同步。此后，如在已经描述过的块1640以及后续块处所描述的那样，处理该部分响应（一个或多个）。

[0172] VIII. 组合响应阶段结构和操作

[0173] 现在参见图16，其中描述了根据本发明、在图1的互连逻辑120内的组合响应逻辑121c的示例性实施例的框图。如同所示，组合响应逻辑121c包括保持缓冲器1702a-1702b，每个保持缓冲器从通过进入A和B链接中的相应一个耦接到本地中心100的远程中心100接收并缓冲组合响应。保持缓冲器1702a-1702b的输出形成第一多路复用器1704的两个输入，第一多路复用器1704应用公平判优策略以从由保持缓冲器1702a-1702b缓冲的组合响应（若有的话）当中进行选择，以便启动到信息帧的组合响应字段710内的第一总线1705上。

[0174] 第一多路复用器1704具有第三输入端，通过该第三输入端，由响应逻辑122提供仅仅节点范围的广播操作的组合响应，以便在保持缓冲器1702a-1702b中缺少任何组合响应的情况下，选择和启动到信息帧的组合响应字段710内的第一总线1705上。因为第一多路复用器1704总是给予从远程中心100接收的全系统范围的广播操作的组合响应优先于仅仅节点范围的广播操作的本地生成组合响应的优先权，所以在某些操作条件下，响应逻辑122可能必须等待相当长的时段以便让第一多路复用器1704选择它提供的组合响应。因此，在最坏情况下，响应逻辑122必须能够让数目等于NM标记FIFO队列924b2中的条目数的组合响应和部分响应对排队，该条目数确定了给定处理单元100在任一时刻可以同时具有的仅仅节点范围的广播操作的最大数目。即使组合响应被延迟了相当长的时段，由主设备300和窥探器304对该组合响应的观察也将被延迟相同的时间量。因此，组合响应的延迟启动不会有违背上述定时约束的危险，这是因为在由获胜主设备300对该组合响应的观

察和由拥有窥探器 304 对该组合响应的观察之间的时间没有因此减少。

[0175] 第一总线 1705 耦接到外出 X、Y 和 Z 链接以及节点主设备 / 远程中心 (NM/RH) 缓冲器 1706 中的每一个。对于仅仅节点范围的广播操作，NM/RH 缓冲器 1706 对由在这个节点主设备 100 处的响应逻辑 122 提供的组合响应和累积部分响应（即，目的地标记）进行缓冲。

[0176] 进入第一层 X、Y 和 Z 链接每个都耦接到远程叶子 (RL) 缓冲器 1714a-1714c 中的相应一个。NM/RH 缓冲器 1706 和 RL 缓冲器 1714a-1714c 的输出形成第二多路复用器 1720 的 4 个输入。第二多路复用器 1720 具有另外的第五输入端，其耦接到本地中心 (LH) 保持缓冲器 1710 的输出端，该本地中心 (LH) 保持缓冲器 1710 为全系统范围的广播操作缓冲由在这个本地中心 100 处的响应逻辑 122 提供的组合响应和累积的部分响应（即，目的地标记）。第二多路复用器 1720 的输出将组合响应驱动到第二总线 1722 上，其中标记 FIFO 队列 924 和外出第二层链接耦接到该第二总线 1722。如同所述，标记 FIFO 队列 924 还被耦接以经由附加信道接收在 LH 保持缓冲器 1710 或 NM/RH 缓冲器 1706 中缓冲的累积的部分响应（即，目的地标记）。主设备 300 和窥探器 304 还耦接到标记 FIFO 队列 924。与标记 FIFO 队列 924 的连接允许窥探器 304 观察组合响应并且允许相关的主设备 300 接收组合响应和目的地标记（若有的话）。

[0177] 在没有上述窗口扩展 312b 的情况下，由主设备 300 和窥探器 304 在基本上相同的时间观察组合响应，在一些操作情况下，可能导致关于从获胜主设备 300 到窥探器 304n 的组合响应延迟的定时约束项（即，C_lat(WM_S)）接近于零，这违反了定时约束。然而，因为窗口扩展 312b 具有大致为第一层链接传送延迟的持续时间，所以尽管主设备 300 和窥探器 304 基本上同时观察组合响应，但是还可以满足上述定时约束。

[0178] 现在参见图 17A-17C，其中描述了分别描述根据本发明的示例性实施例、在本地中心（或节点主设备）、远程中心（或节点主设备）、以及远程叶子（或节点叶子）处的示例性组合响应阶段处理的高级逻辑流程图。现在具体参见图 17A，在本地中心（或节点主设备）100 处的组合响应阶段处理从块 1800 开始，然后继续到块 1802，其描述了响应逻辑 122 基于请求类型和累积的部分响应而生成对操作的组合响应。如块 1803-1805 所示，如果组合响应 710 内的范围指示符 730 指示该操作是仅仅节点范围的广播操作或者配置寄存器 123 指示超级节点模式，则在节点主设备 100 处的组合响应阶段处理在图 17B 的块 1863 处继续。然而，如果范围指示符 730 指示该操作是全系统范围的广播操作，则如块 1804 所示，远程中心 100 的响应逻辑 122 将该组合响应和累积的部分响应放置到 LH 保持缓冲器 1710 中。借助于利用或操作累积部分响应，对于写入类型的请求，累积的部分响应将包含被设置为‘1’的有效字段 722，以表示在伴随的目的地标记字段 724 中存在有效的目的地标记。对于其它类型的请求，累积的部分响应中的位 0 将被设置为‘0’，以指示不存在这样的目的地标记。

[0179] 如块 1844 所述，第二多路复用器 1720 与所选择的第二层链接信息分配时间片对齐，并且只有当外出第二层链接信息分配中有可用于组合响应的地址占有期时，才从 LH 保持缓冲器 1710 中选择组合响应和累积的部分响应，以便启动。因此，例如，仅仅在图 7B 的实施例的周期 1 或 3 期间，第二多路复用器 1720 才从 LH 保持缓冲器 1710 输出组合响应和累积的部分响应。如果在块 1844 做出了否定的确定，则如块 1846 所示，延迟 LH 保持缓冲

器 1710 内的组合响应的启动，直到其中地址占有期可用的后续周期为止。相反，如果在块 1844 处做出了肯定的确定，则第二多路复用器 1720 优先于其它输入而选择 LH 保持缓冲器 1710 内的组合响应，以便启动到第二总线 1722 上并且随后在外出第二层链接上传送。

[0180] 还应当注意，第二多路复用器 1720 的其它端口（例如，RH、RLX、RLY、和 RLZ）也可以与 LH 保持缓冲器 1710 同时地提供请求，这意味着第二总线 1722 的最大带宽必须等于外出第二层链接的带宽的 10/8（假定图 7B 的实施例），以便跟得上最大到达速率。此外应当观察到，仅仅在 LH 保持缓冲器 1710 内缓冲的组合响应在外出第二层链接上传送，并且要求与链接信息分配内的地址占有期对齐。因为竞争由第二多路复用器 1720 发布的所有其它组合响应仅仅以本地主设备 300、窥探器 304 以及它们各自的 FIFO 队列而不是外出第二层链接作为目标，所以可以在信息帧的剩余周期内发布这样的组合响应。因此，与由第二多路复用器 1720 采用的特定判优方案无关，保证同时向第二多路复用器 1720 提供的所有组合响应在单个信息帧的延迟之内传送。

[0181] 在第二总线 1722 上发布组合响应之后，该处理分叉并且继续到块 1848 和 1852 中的每一个。块 1848 描述了将启动到第二总线 1722 上的组合响应路由到外出第二层链接，以便传送到远程中心 100。此后，该处理通过页面连接符 1850 继续到图 17C，其中图 17C 描述了在远程中心 100 处的组合响应处理的示例性方法。

[0182] 现在参见块 1852，还利用在第二总线 1722 上发布的组合响应来查询 LH 标记 FIFO 队列 924a，以从其中的最旧条目中获得主设备标记。此后，LH 标记 FIFO 队列 924a 释放分配给操作的条目（块 1854）。在块 1854 之后，该处理分叉并且继续到块 1810 和 1856 中的每一个。在块 1810，LH 标记 FIFO 队列 924a 确定该主设备标记是否指示发起了与组合响应相关联的请求的主设备 300 驻留在这个本地中心 100 中。如果不是的话，则这个路径中的处理在块 1816 结束。然而，如果该主设备标记指示发起的主设备 300 驻留在当前本地中心 100 中，则 LH 标记 FIFO 队列 924a 将该主设备标记、组合响应和累积的部分响应路由到由该主设备标记标识的发起主设备 300（块 1812）。响应于组合响应和主设备标记的接收，发起主设备 300 处理该组合响应，如果对应的请求是写入类型请求，则还处理累积的部分响应（块 1814）。

[0183] 例如，如果组合响应指示“成功”并且对应的请求是读取类型的请求（例如，读取、DClaim 或 RWITM 请求），则发起主设备 300 可以更新或准备接收所请求的存储块。在这种情况下，丢弃累积的部分响应。如果组合响应指示“成功”而且对应的请求是写入类型的请求（例如，逐出、写入或部分写入请求），则发起主设备 300 从累积的部分响应中提取目的地标记字段 724，并且利用其内容作为数据标记 714，该数据标记 714 用于将该操作的后续数据阶段路由到其目的地。如果“成功”的组合响应指示或暗示用于发起主设备 300 的 HPC 状态的准予，则如标号 313 所述，发起主设备 300 将另外开始保护它对存储块的所有权。然而，如果在块 1814 处接收的组合响应指示诸如“重试”之类的其它结果，则可能要求发起主设备 300 或许以不同的范围（例如，全局而非本地）重新发布请求。此后，该处理在块 1816 结束。

[0184] 现在参见块 1856，LH 标记 FIFO 队列 924a 还将组合响应和相关联的主设备标记路由到本地中心 100 内的窥探器 304。响应于组合响应的接收，窥探器 304 处理该组合响应，并且执行响应于其所需的任何操作（块 1857）。例如，窥探器 304 可以将所请求的存储块供

应到请求的发起主设备300，使所请求的存储块的高速缓存的副本无效等。如果该组合响应包括窥探器304要将存储块的所有权转移到请求主设备300的指示，则窥探器304在其保护窗口312a的末尾附加一可编程长度的窗口扩展312b，对于所说明的拓扑结构，其优选地具有大致为在第一层链接上的一个芯片跳跃的延迟的持续时间（块1858）。当然，对于其它数据处理系统拓扑结构和互连逻辑120的不同实现，可以有利地将可编程窗口扩展312b设置为其它长度，以补偿链接延迟差异（例如，耦接不同处理节点202的不同长度的电缆）、拓扑结构或者物理约束、电路设计约束、或者各个操作阶段的有界延迟的较大变化性。此后，在本地中心100处的组合响应阶段处理在块1859结束。

[0185] 现在参见图17B，其中描述了根据本发明、在远程中心（或节点主设备）100处的组合响应阶段处理的示例性方法的高级逻辑流程图。如同所述，对于在远程中心100处的组合响应阶段处理，当在远程中心100处在其进入A或B链接之一上接收到组合响应时，该处理从页面连接符1860开始。然后如块1862所示，在保持缓冲器1702a-1702b中相关联的一个内缓冲该组合响应。一旦都满足了在块1864和1865处描述的条件，则由第一多路复用器1704在第一总线1705上传送缓冲的组合响应。特别地，在第一层链接信息分配中地址占有期必须可用（块1864），并且由第一多路复用器1704实现的公平分配策略必须选择其中缓冲了组合响应的保持缓冲器1702a、1702b（块1865）。如前所述，在超级节点模式中，缓冲组合响应的保持缓冲器1702a总是第一多路复用器1704的公平分配策略的获胜者，这是因为在其它第二层链接（一个或多个）上没有为访问第一总线1705而竞争的操作。

[0186] 如块1864所示，如果不满足这些条件中的任何一个，则在块1866延迟由第一多路复用器1704将该组合响应启动到第一总线1705上直到下一个地址占有期为止。然而，如果满足块1864和1865所述的两个条件，则该处理从块1865继续到块1868，其说明了第一多路复用器1704在第一总线1705上将组合响应广播到外出X、Y和Z链接以及组合响应字段710内的NM/RH保持缓冲器1706。如包含块1863和1867的路径到块1868的连接所示，对于仅仅节点范围和超级节点广播操作，只有当没有由保持缓冲器1702a-1702b提供竞争的组合响应时，第一多路复用器1704才将由响应逻辑122提供的组合响应发布到第一总线1705上，以便路由到外出X、Y和Z链接以及NM/RH保持缓冲器1706。如果经由进入第二层链接之一从远程中心100接收了对全系统范围的广播操作的任何竞争组合响应，则如块1867所示，延迟仅仅节点范围广播操作的本地生成的组合响应。当第一多路复用器1704最终选择仅仅节点范围广播操作的本地生成的组合响应时，响应逻辑122直接将相关联的累积部分响应放置到NM/RH保持缓冲器1706中。

[0187] 在块1868之后，该处理分叉。第一路径通过页面连接符1870传递到图17C，其说明了在远程叶子（或节点叶子）100处的组合响应阶段处理的示例性方法。第二路径从块1868继续到块1874，其说明了第二多路复用器1720确定在其输入端提供的哪个组合响应要在第二总线1722上输出。如同所示，第二多路复用器1720使本地中心组合响应优先于远程中心组合响应，远程中心组合响应又优先于在远程叶子缓冲器1714a-1714c中缓冲的组合响应。因此，如果由LH保持缓冲器1710提供了本地中心组合响应以便选择，则如块1876所示，延迟在远程中心缓冲器1706内缓冲的组合响应。然而，如果没有由LH保持缓冲器1710给出组合响应（在超级节点模式的情况下总是这样），则第二多路复用器1720将来自NM/RH缓冲器1706的组合响应发布到第二总线1722上。

[0188] 响应于第二总线 1722 上的组合响应检测，如块 1878 所述，与在其上接收组合响应的第二层链接相关联的标记 FIFO 队列 924b0 和 924b1 中的特定一个（或者，对于仅仅节点或者超级节点广播操作，NM 标记 FIFO 队列 924b2）从其最老条目的主设备标记字段 1100 中读出由相关的请求指定的主设备标记，然后释放该条目（块 1880）。该处理然后分成三叉并且继续到块 1882、1881、和 1861 中的每一个。块 1882 描述了标记 FIFO 队列 924b 中的相关一个将组合响应和主设备标记路由到远程中心（或节点主设备）100 中的窥探器 304。响应于组合响应的接收，窥探器 304 处理该组合响应（块 1884）并且如上所述，执行任何需要的操作。如果该操作是全系统范围的或者超级节点广播操作并且如果组合响应包括窥探器 304 要将存储块的一致性所有权转移到进行请求的主设备 300 的指示，则如块 1885 所示，窥探器 304 向其保护窗口 312a 附加窗口扩展 312b。此后，在块 1886 处结束在远程中心 100 的组合响应阶段处理。

[0189] 现在参见块 1881，如果组合响应字段 710 内的范围指示符 730 和配置寄存器 123 的设置指示该操作不是仅仅节点范围或超级节点广播操作，而是全系统范围广播操作，则不在远程中心 100 处执行进一步的处理，并且在块 1886 结束该处理。然而，如果范围指示符 730 指示该操作是仅仅节点范围广播操作，或者配置寄存器 123 指示超级节点模式并且当前的处理器 100 是节点主设备 100，则该处理传递到块 1883，其说明了 NM 标记 FIFO 队列 924b2 将主设备标记、组合响应和累积的部分响应路由到由主设备标记所识别的发起主设备 300。响应于组合响应和主设备标记的接收，发起主设备 300 处理该组合响应，并且如果对应的请求是写入类型的请求，则还处理累积的部分响应（块 1887）。

[0190] 例如，如果组合响应指示“成功”并且对应的请求是读取类型的请求（例如，读取、DClaim 或 RWITM 请求），则发起主设备 300 可以更新或者准备接收所请求的存储块。在这种情况下，丢弃累积的部分响应。如果组合响应指示“成功”并且对应的请求是写入类型的请求（例如，逐出、写入或部分写入请求），则发起主设备 300 从累积的部分响应中提取目的地标记字段 724，并且使用其内容作为数据标记 714，该数据标记 714 用于将该操作的后续数据阶段路由到其目的地。如果“成功”的组合响应指示或暗示用于发起主设备 300 的 HPC 状态的准予，则如标号 313 所述，发起主设备 300 将另外开始保护它对存储块的所有权。然而，如果在块 1814 接收的组合响应指示诸如“重试”之类的其它结果，则可能要求发起主设备 300 重新发布该请求。此后，在块 1886 结束该处理。

[0191] 现在转向块 1861，如果处理该组合响应的处理单元 100 是节点主设备 100 并且配置寄存器 123 指示超级节点模式，则如块 1874 所示，第二多路复用器 1720 另外将该组合响应路由到第二层链接中的选定一个（例如，链接 A）。此后，该处理通过页面连接符 1860，并且组合响应的处理在远程中心 100 处继续。

[0192] 现在参见图 17C，其中说明了根据本发明、在远程叶子（或节点）叶子 100 处的组合响应阶段处理的示例性方法的高级逻辑流程图。如同所示，当在远程（或节点）叶子 100 处在其进入 X、Y 和 Z 链接之一上接收到组合响应时，该处理从页面连接符 1888 开始。如块 1890 所示，该组合响应被锁存到 NL/RL 保持缓冲器 1714a-1714c 之一中。接下来，如块 1891 所述，由第二多路复用器 1720 计算该组合响应以及在其输入端提供的其它组合响应。如上所述，第二多路复用器 1720 使本地中心组合响应优先于远程中心组合响应，远程中心组合响应又优先于在 NL/RL 保持缓冲器 1714a-1714c 中缓冲的组合响应。因此，如果提供了本

地中心或远程中心组合响应以便选择，则如块 1892 所示，延迟在 NL/RL 保持缓冲器 1714 内缓冲的组合响应。然而，如果没有向第二多路复用器 1720 提供较高优先级的组合响应，则第二多路复用器 920 将组合响应从 NL/RL 保持缓冲器 1714 发布到第二总线 1722 上。

[0193] 响应于在第二总线 1722 上检测到组合响应，如块 1893 所述，与操作的范围以及接收该组合响应的路由相关联的标记 FIFO 队列 924c0-924c2、924d0-924d2、和 924e0-924e2 中的特定一个从其最旧条目的主设备标记字段 1100 中读出由相关联的请求指定的主设备标记。也就是说，利用配置寄存器 123 的设置或者组合响应字段 710 内的范围指示符 730 来确定是否在超级节点模式中进行请求，或者如果不是的话，该请求是仅仅节点范围还是全系统范围。对于仅仅节点范围和超级节点广播请求，与在其上接收了组合响应的进入第一层链接相关联的 NL 标记 FIFO 队列 924c2、924d2 和 924e2 中的特定一个缓冲主设备标记。对于全系统范围广播请求，从与在其上接收了组合响应的进入第一和第二层链接的组合相对应的、RL 标记 FIFO 队列 924c0-924c1、924d0-924d1 和 924e0-924e1 中的特定一个中检索主设备标记。

[0194] 一旦相关的标记 FIFO 队列 924 标识了用于该操作的适当条目，标记 FIFO 队列 924 就释放该条目（块 1894）。如块 1895 所示，组合响应和主设备标记被进一步路由到远程（或节点）叶子 100 中的窥探器 304。响应于组合响应的接收，窥探器 304 处理该组合响应（块 1896）并且如上所述，执行任何所需操作。如果该操作不是仅仅节点范围的操作并且如果组合响应包括窥探器 304 要将存储块的一致性所有权转移到请求主设备 300 的指示，则如上所述并且如块 1897 所示，窥探器 304 向其保护窗口 312a 的末尾附加窗口扩展 312b。此后，在块 1898 结束在远程叶子 100 处的组合响应阶段处理。

[0195] IX. 数据阶段结构和操作

[0196] 数据逻辑 121d 以及其对数据递送的处理可以各种方式实现。在一个优选实施例中，数据逻辑 121d 及其操作如上面通过引用并入的、共同未决的美国专利申请中详细描述的那样实现。当然，另外的未被请求和响应流使用的第二层链接（一个或多个）（例如，B 链接）可以用于数据递送以增强数据带宽。

[0197] X. 结论

[0198] 如已经描述的那样，本发明提供了改进的处理单元、数据处理系统以及用于数据处理系统的互连构造。此处公开的创造性数据处理系统拓扑结构通过在处理节点的多个处理单元之间点对点节点间链接的实现，提供了在不同处理节点的处理单元之间的高带宽通信。此外，因为例如，如图 2A-2B 所示，相同的互连逻辑可能支持各种互连构造拓扑结构，所以此处公开的处理单元和处理节点显示出非常高的灵活性，并因此允许数据处理系统的处理节点以最适于预期工作负载的方式互连。

[0199] 虽然如参考优选实施例所述的那样已经具体示出了本发明，但是本领域的技术人员应当理解，可以在其中进行形式和细节上的各种改变而不背离本发明的精神和范围。例如，虽然本发明公开了利用 FIFO 队列来对操作相关标记和部分响应进行排序的优选实施例，但是本领域技术人员将会理解，可以采用其它有序的数据结构来以所描述的方式维持在操作的各个标记和部分响应之间的次序。此外，虽然本发明的优选实施例采用单向通信链接，但是本领域的技术人员通过参考前文将会理解，可以可选地采用双向通信链接。此外，虽然已经参考特定的示例性互连构造拓扑结构描述了本发明，但是本发明不局限于此处具体描述的那些互连构造拓扑结构，而是可以广泛地应用于多种不同的互连构造拓扑结构。

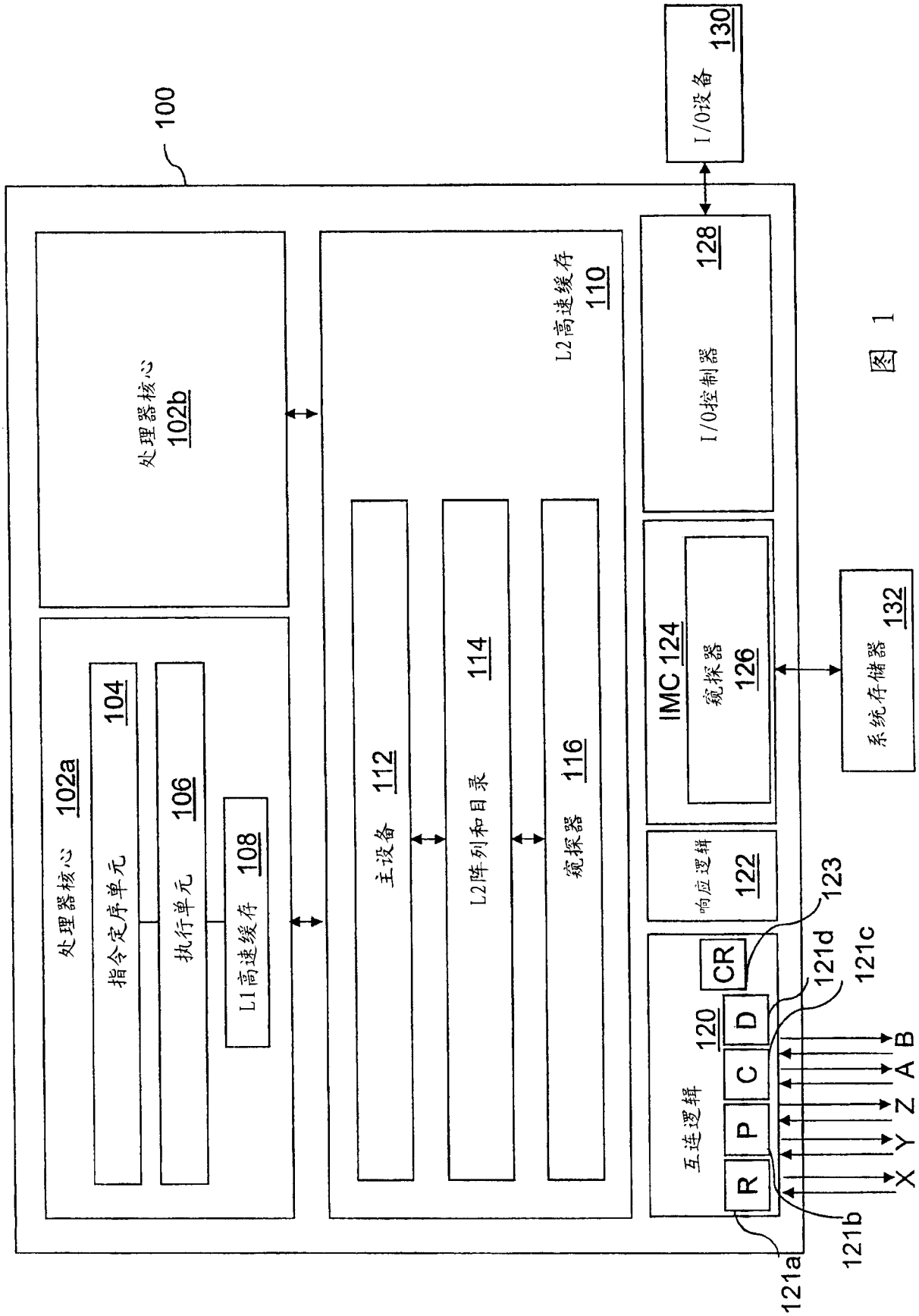

100
处理器核心 102a
指令定序单元 104
执行单元 106
L1高速缓存 108
处理器核心 102b
L2高速缓存 110
主设备 112
L2阵列和目录 114
窥探器 116
互连逻辑 120
R
P
C
D
CR
响应逻辑 122
IMC 124
窥探器 126
I/O控制器 128
I/O设备 130
系统存储器 132
121a
121b
121c
121d
123
X
Y
Z
A
B

图 1

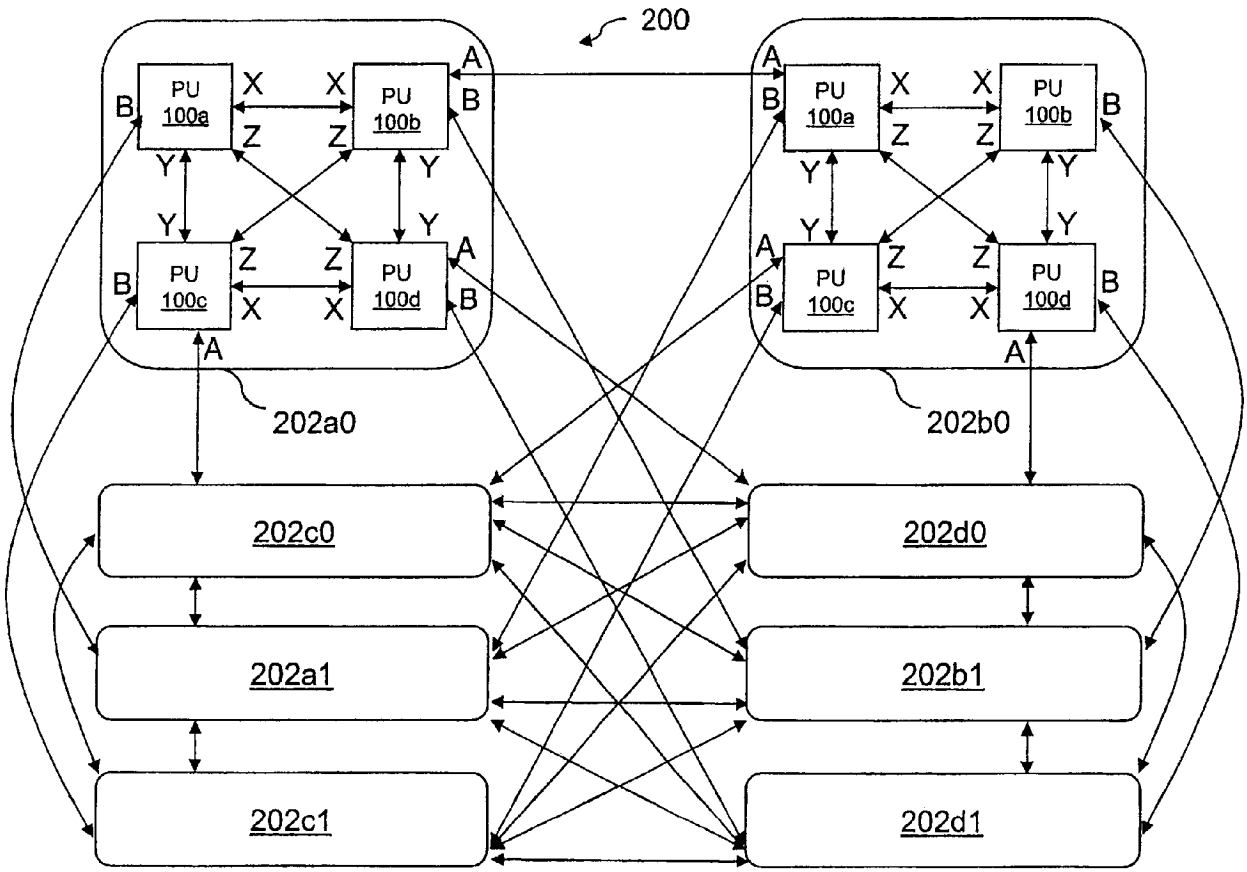

200
PU 100a
PU 100b
PU 100c
PU 100d
202a0
202b0
202c0
202d0
202a1
202b1
202c1
202d1

图 2A

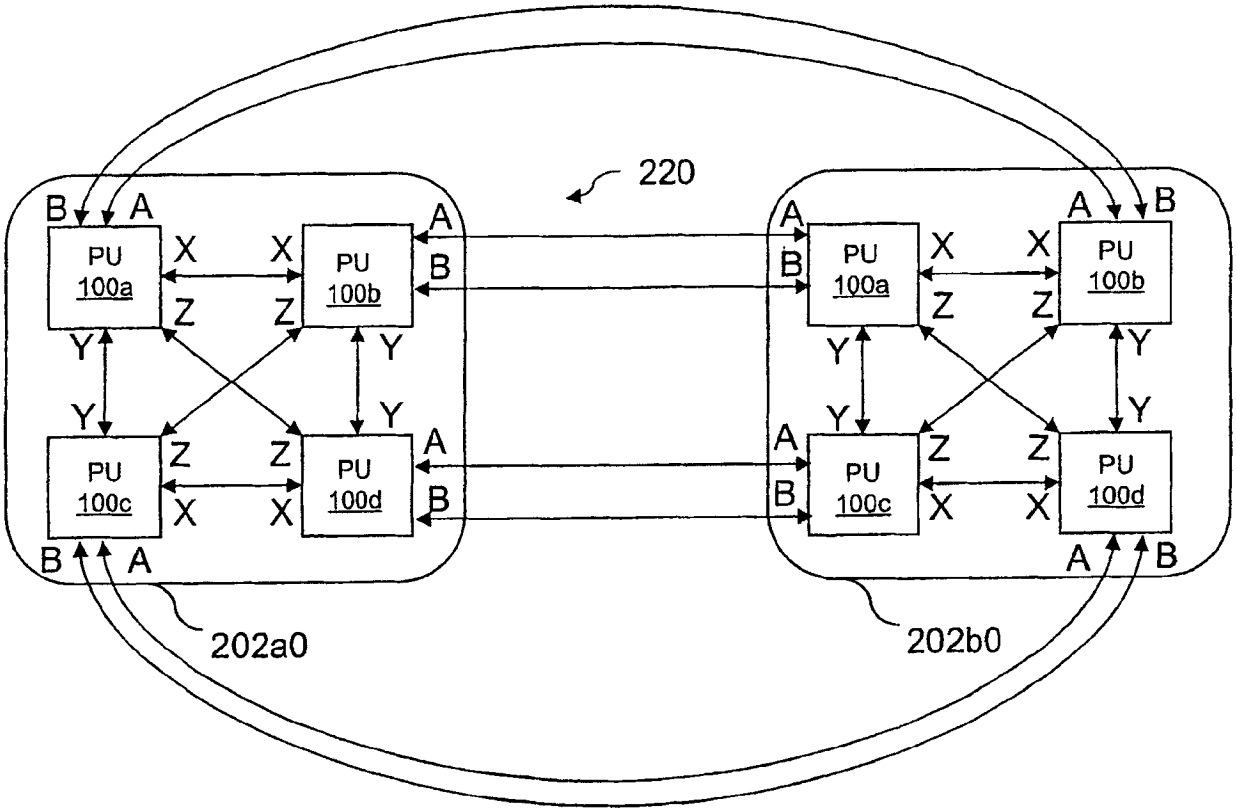


图 2B

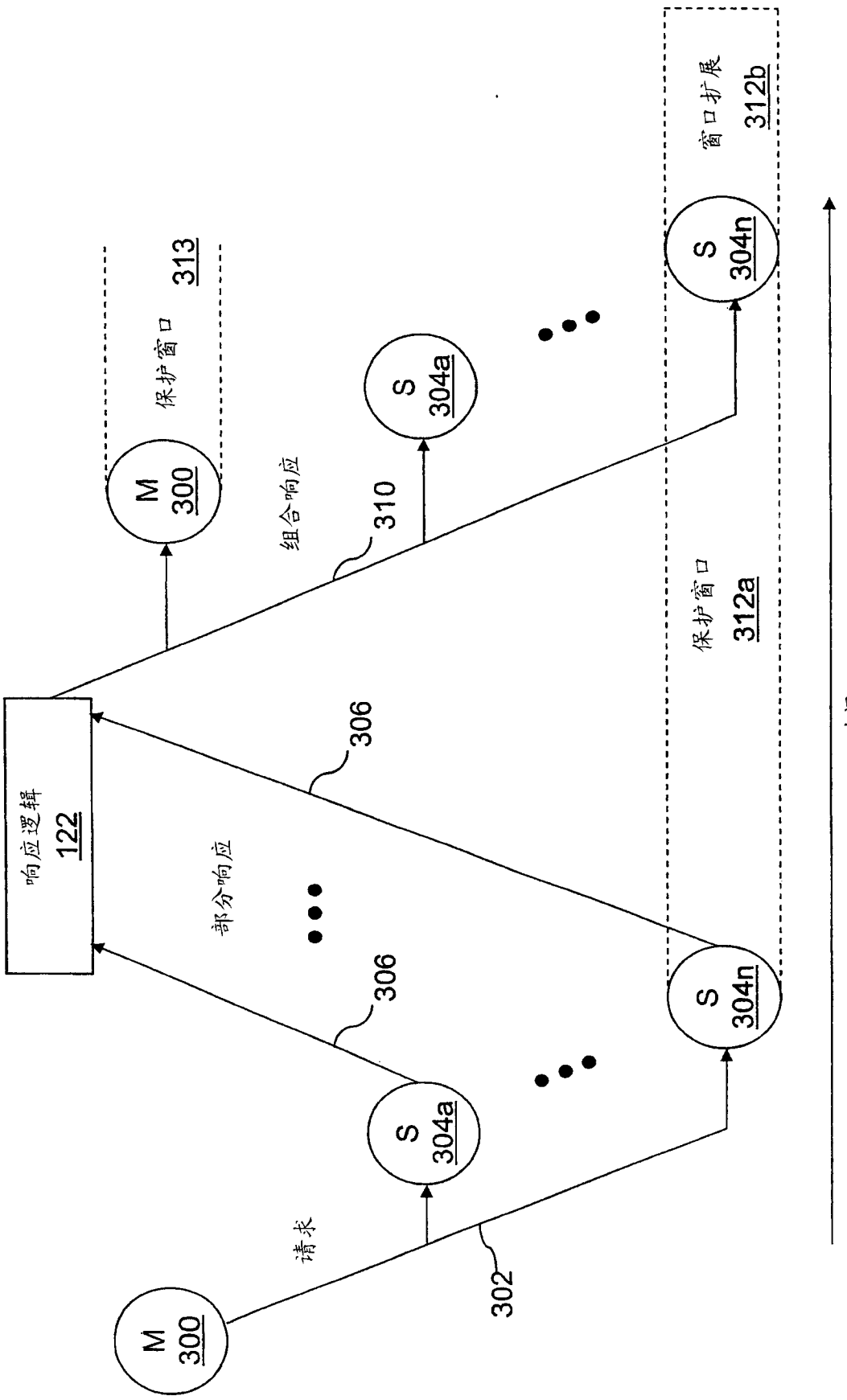

M 300
请求
302
S 304a
306
部分响应
S 304n
响应逻辑 122
306
组合响应
310
M 300
保护窗口
313
S 304a
S 304n
保护窗口 312a
窗口扩展 312b
时间

图 3

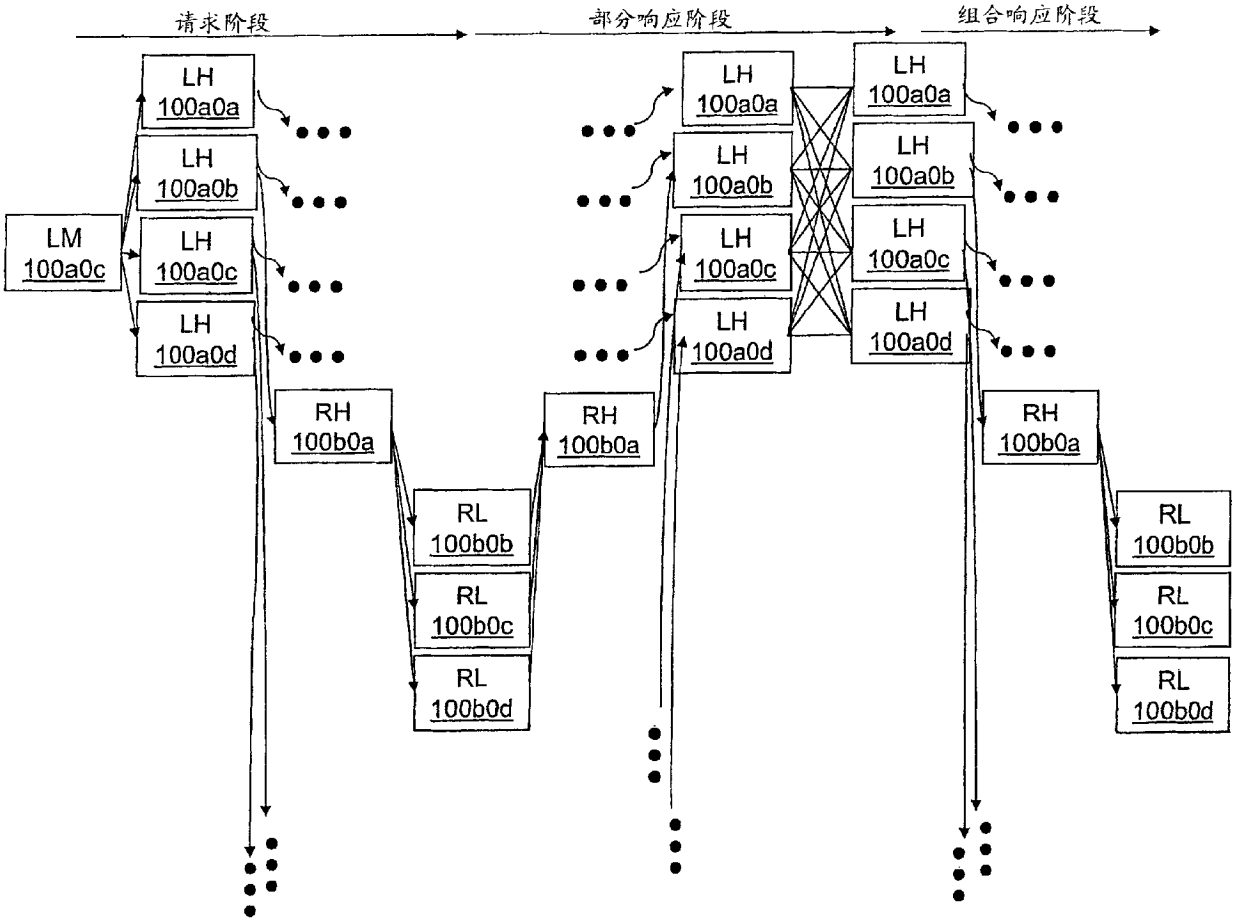

请求阶段
部分响应阶段
组合响应阶段
LM 100a0c
LH 100a0a
LH 100a0b
LH 100a0c
LH 100a0d
RH 100b0a
RL 100b0b
RL 100b0c
RL 100b0d

图 4A

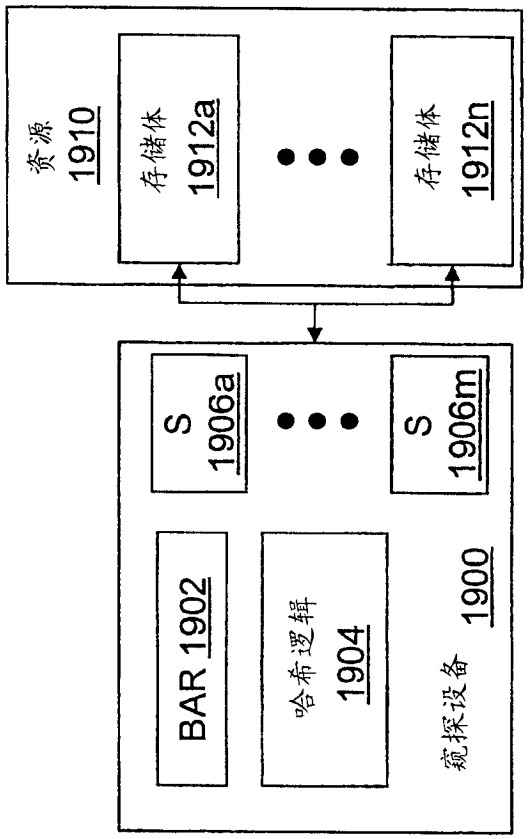

资源
1910
存储体
1912a
存储体
1912n
S
1906a
S
1906m
BAR 1902
哈希逻辑
1904
窥探设备
1900

图 18

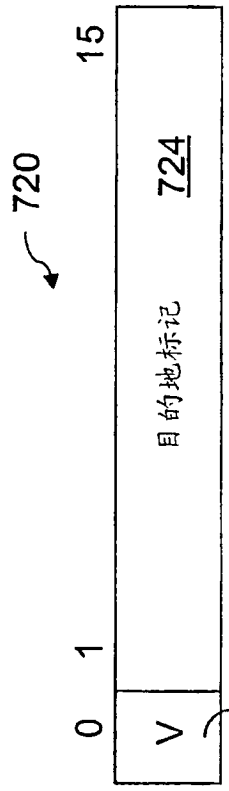

720
15
1
0
目的地标记
724
V
722

图 7C

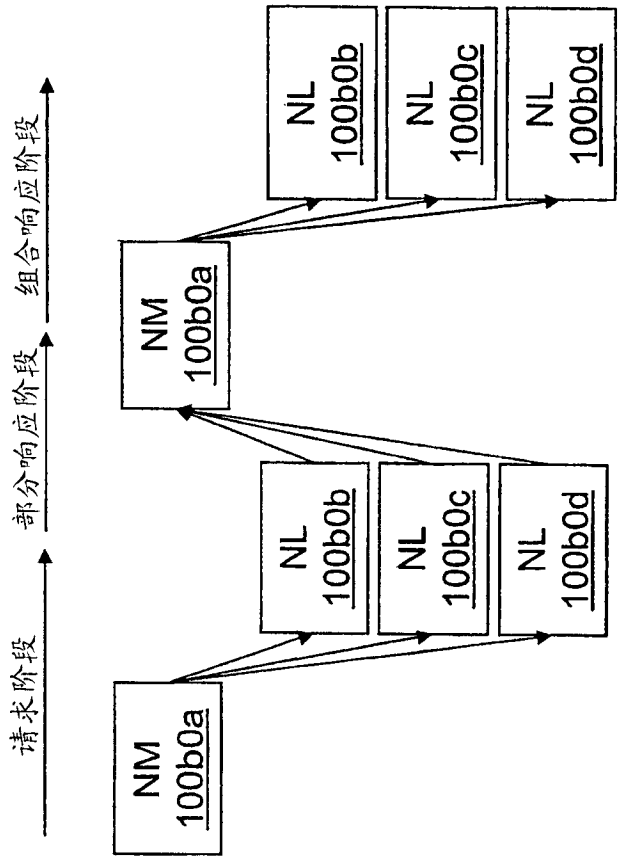

请求阶段
部分响应阶段
组合响应阶段
NM
100b0a
NL
100b0b
NL
100b0c
NL
100b0d
NM
100b0a
NL
100b0b
NL
100b0c
NL
100b0d

图 4B

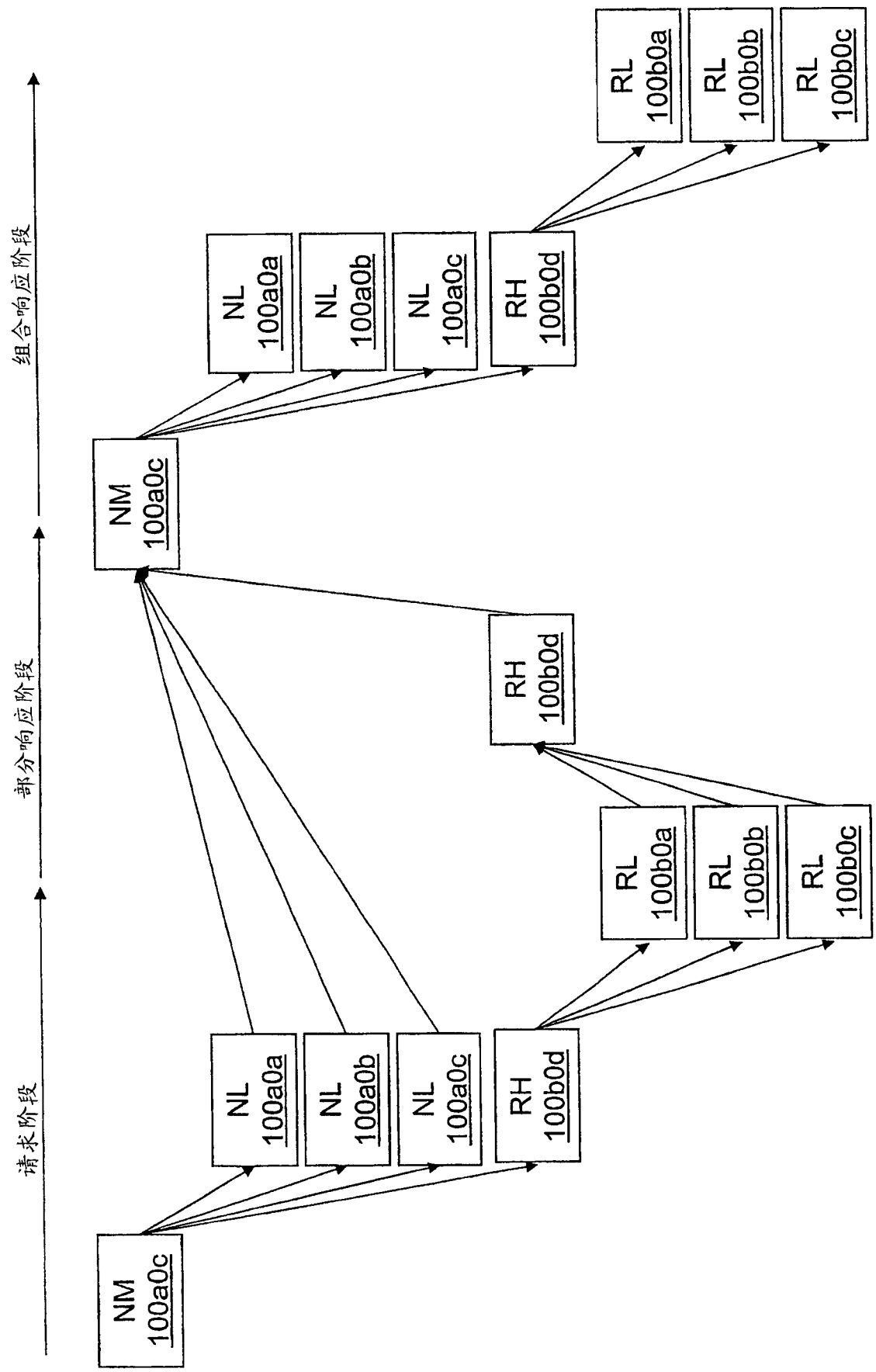

请求阶段
部分响应阶段
组合响应阶段
NM 100a0c
NL 100a0a
NL 100a0b
NL 100a0c
RH 100b0d
RL 100b0a
RL 100b0b
RL 100b0c
NM 100a0c
NL 100a0a
NL 100a0b
NL 100a0c
RH 100b0d
RL 100b0a
RL 100b0b
RL 100b0c
RH 100b0d

图 4C

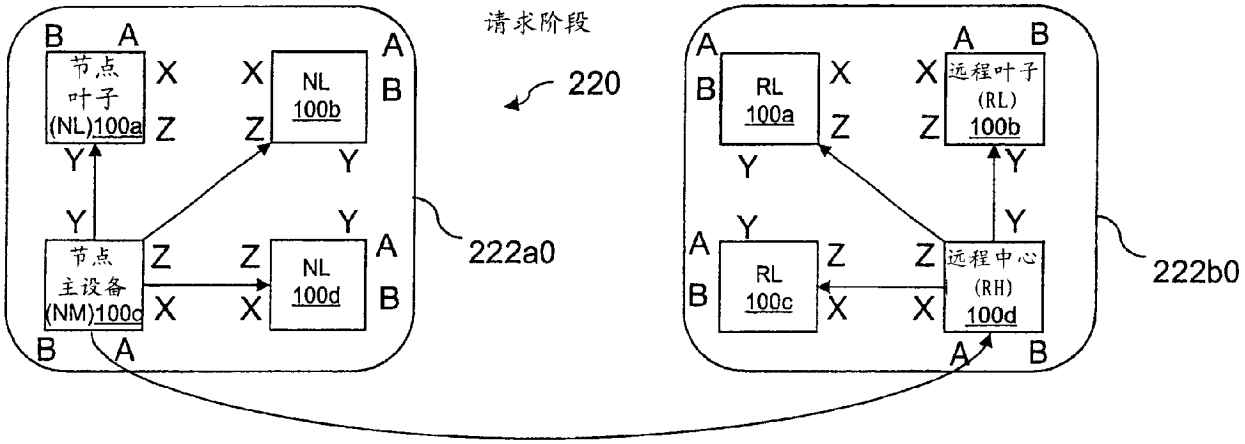

请求阶段
220
节点
叶子
(NL)100a
NL
100b
节点
主设备
(NM)100c
NL
100d
222a0
RL
100a
远程叶子
(RL)
100b
RL
100c
远程中心
(RH)
100d
222b0

图 5A

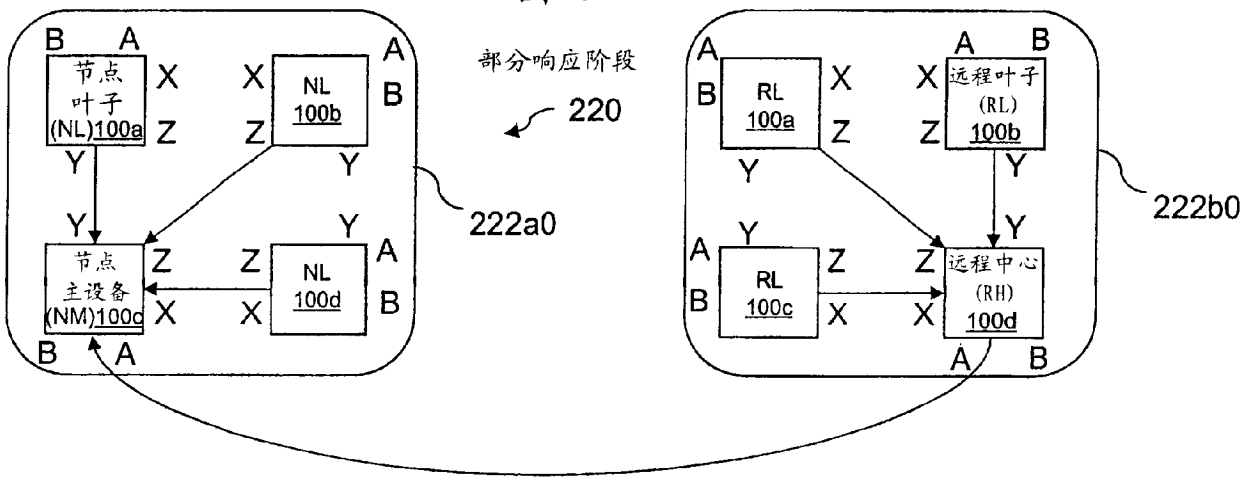

部分响应阶段
220
节点
叶子
(NL)100a
NL
100b
节点
主设备
(NM)100c
NL
100d
222a0
RL
100a
远程叶子
(RL)
100b
RL
100c
远程中心
(RH)
100d
222b0

图 5B

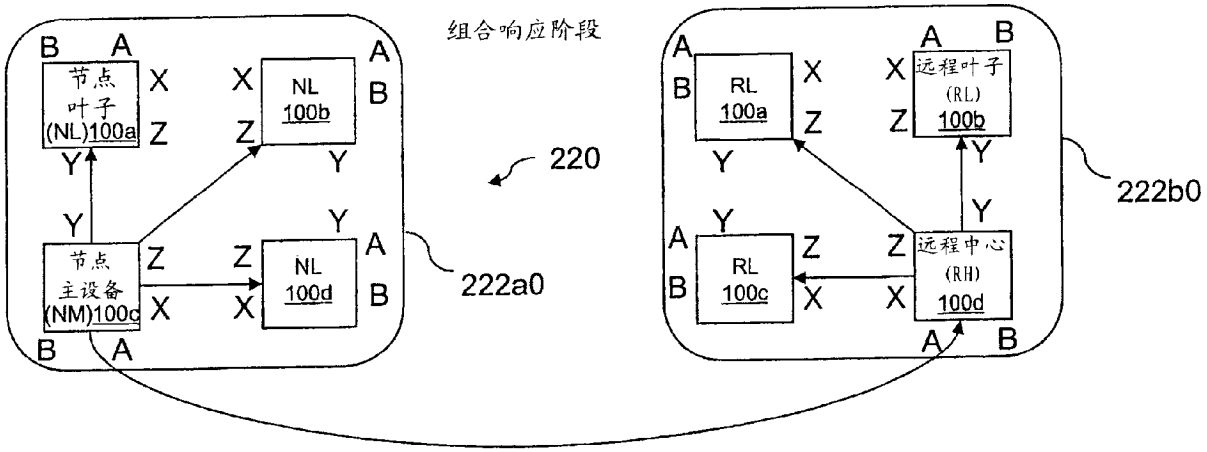

组合响应阶段
节点叶子(NL)100a
NL 100b
节点主设备(NM)100c
NL 100d
220
222a0
RL 100a
远程叶子(RL) 100b
RL 100c
远程中心(RH) 100d
222b0

图 5C

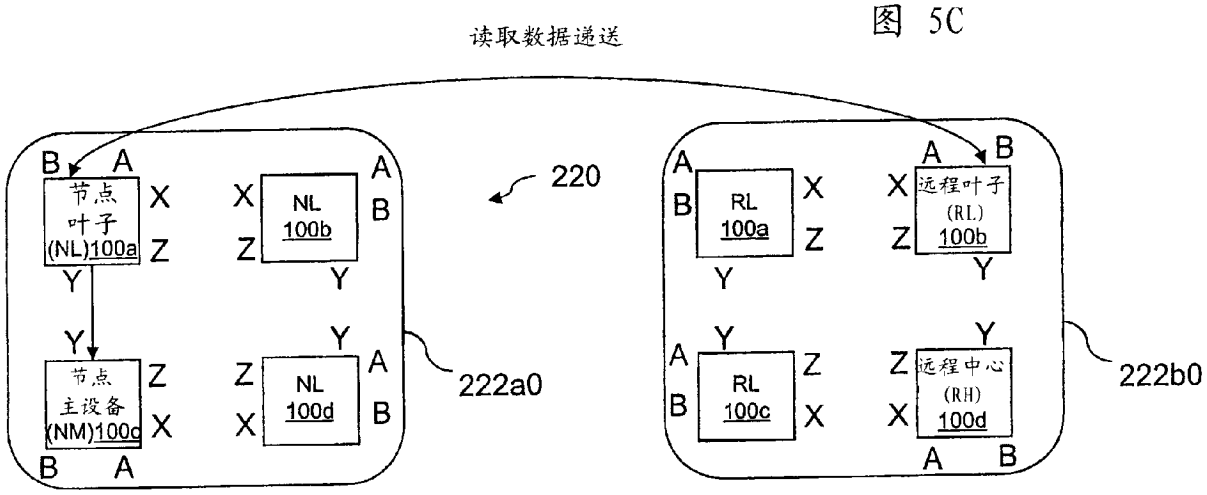

读取数据递送
节点叶子(NL)100a
NL 100b
节点主设备(NM)100c
NL 100d
220
222a0
RL 100a
远程叶子(RL) 100b
RL 100c
远程中心(RH) 100d
222b0

图 5D

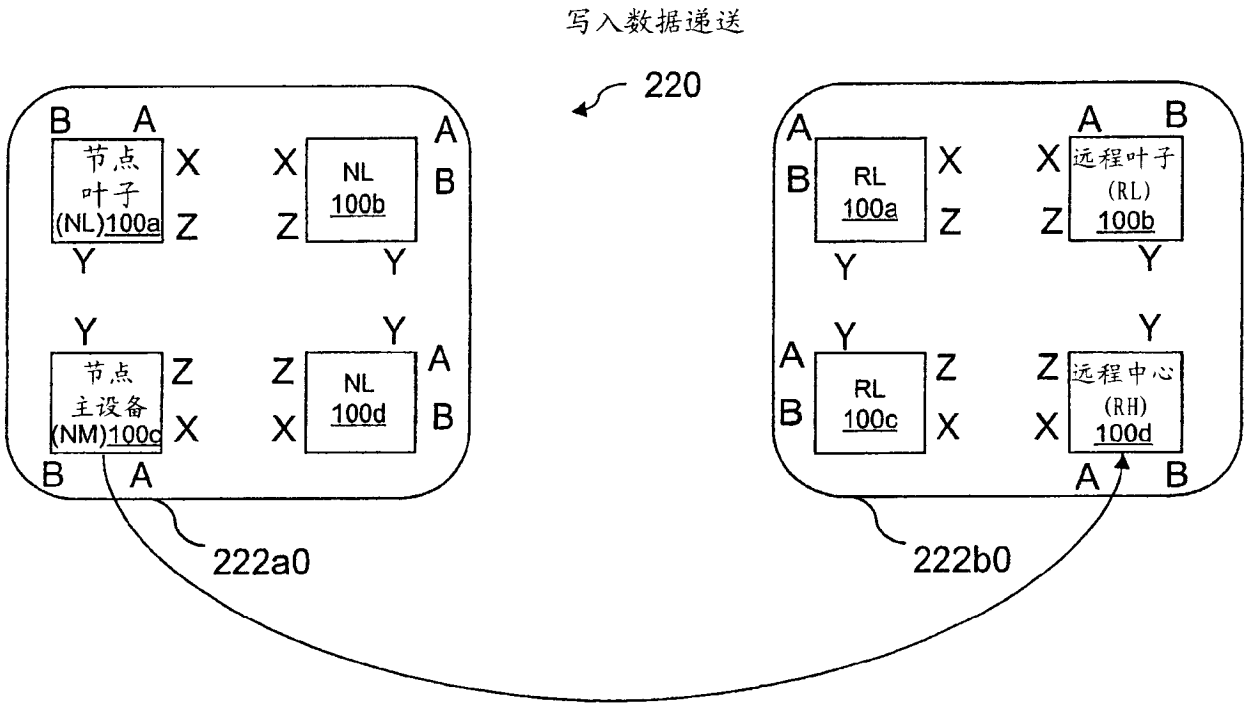

写入数据递送
220
节点
叶子
(NL)100a
NL
100b
节点
主设备
(NM)100c
NL
100d
RL
100a
远程叶子
(RL)
100b
RL
100c
远程中心
(RH)
100d
222a0
222b0

图 5E

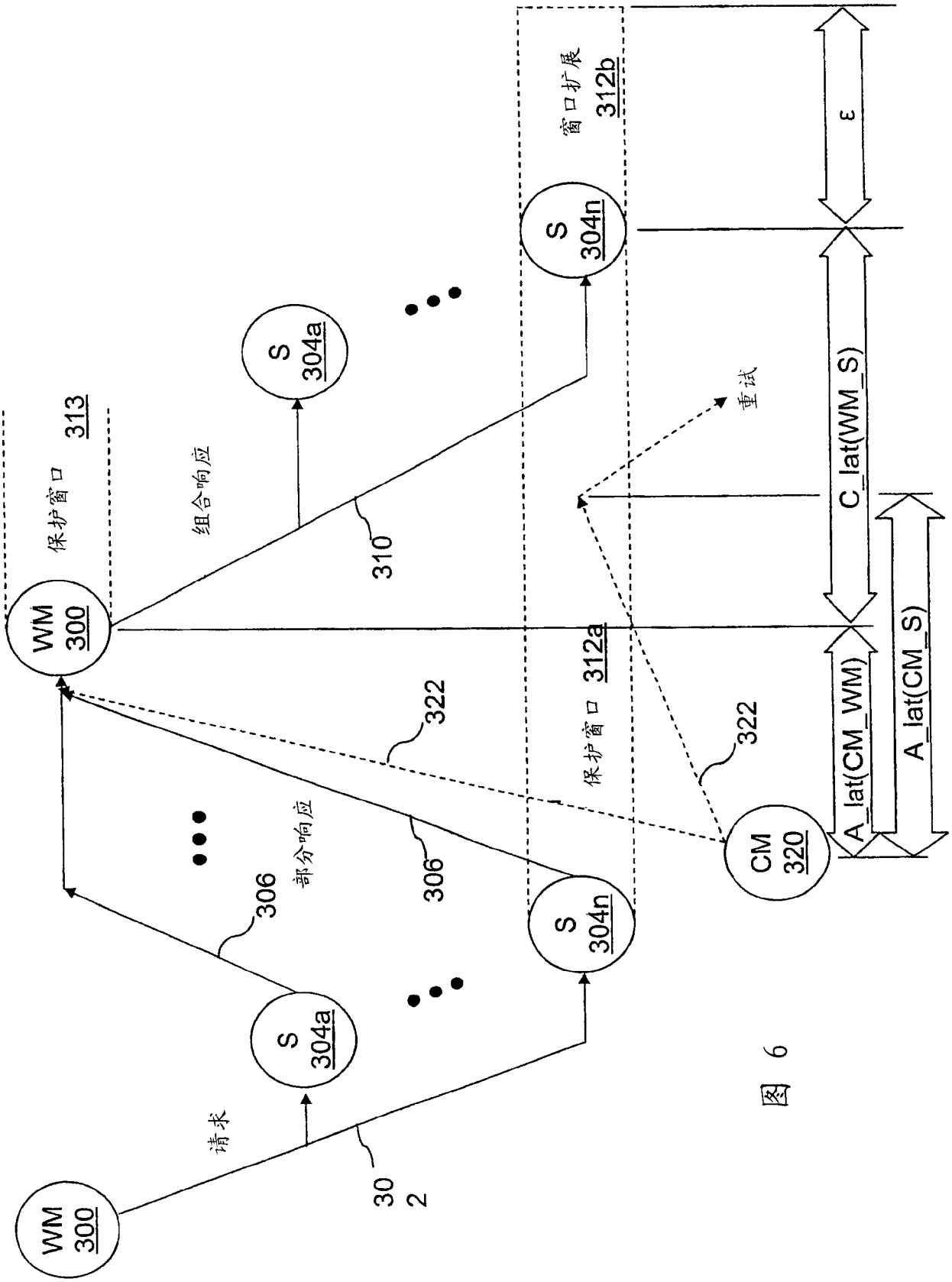

WM 300
请求
302
S 304a
S 304n
部分响应
306
306
保护窗口 312a
322
322
CM 320
A_lat(CM_WM)
A_lat(CM_S)
C_lat(WM_S)
重试
WM 300
保护窗口 313
组合响应
310
S 304a
S 304n
窗口扩展 312b
ε

图 6

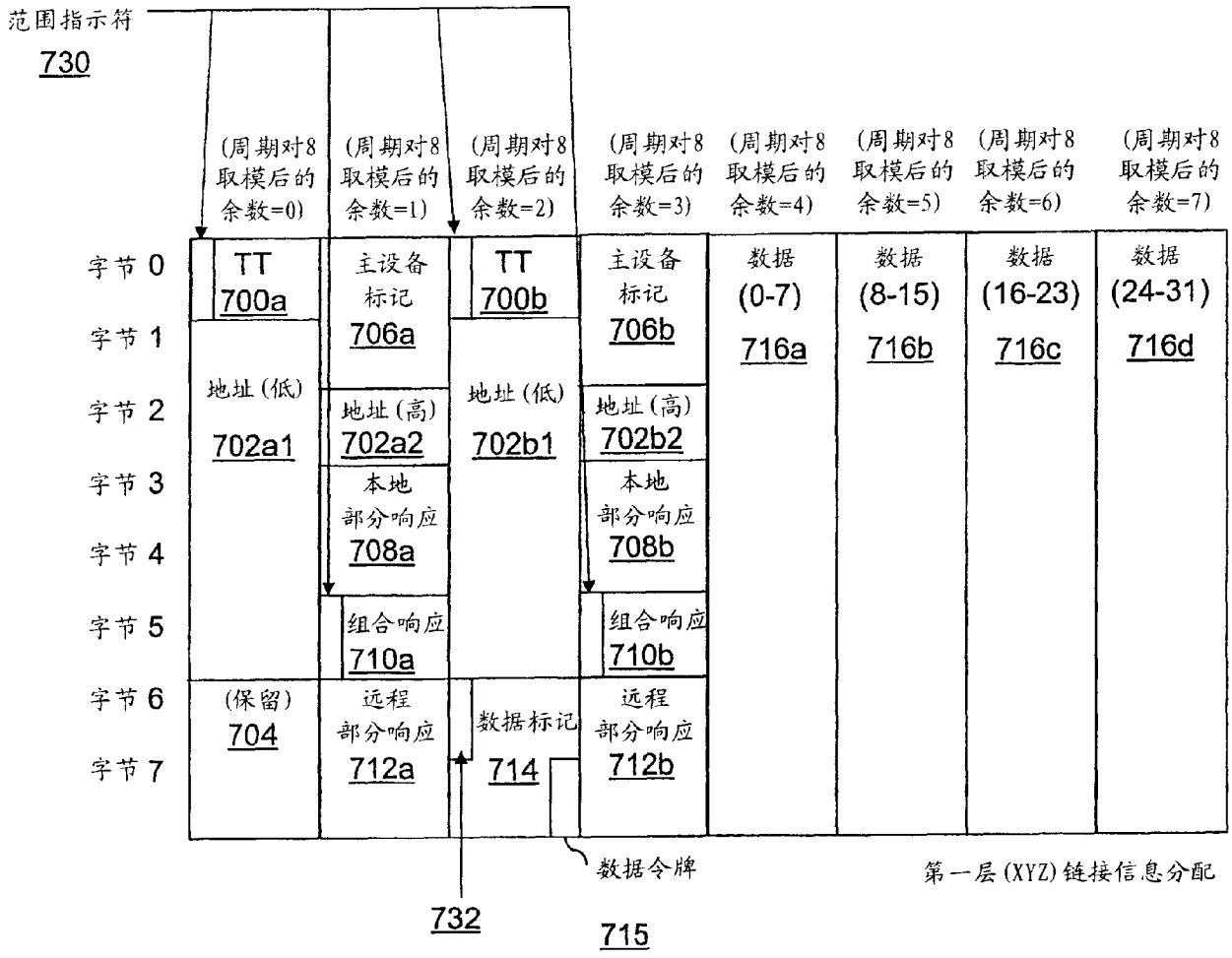

范围指示符
730
(周期对8取模后的余数=0)
(周期对8取模后的余数=1)
(周期对8取模后的余数=2)
(周期对8取模后的余数=3)
(周期对8取模后的余数=4)
(周期对8取模后的余数=5)
(周期对8取模后的余数=6)
(周期对8取模后的余数=7)
字节 0
字节 1
字节 2
字节 3
字节 4
字节 5
字节 6
字节 7
TT 700a
主设备标记 706a
TT 700b
主设备标记 706b
数据 (0-7) 716a
数据 (8-15) 716b
数据 (16-23) 716c
数据 (24-31) 716d
地址(低) 702a1
地址(高) 702a2
地址(低) 702b1
地址(高) 702b2
本地部分响应 708a
本地部分响应 708b
组合响应 710a
组合响应 710b
(保留) 704
远程部分响应 712a
数据标记 714
远程部分响应 712b
数据令牌 715
732
第一层(XYZ)链接信息分配

图 7A

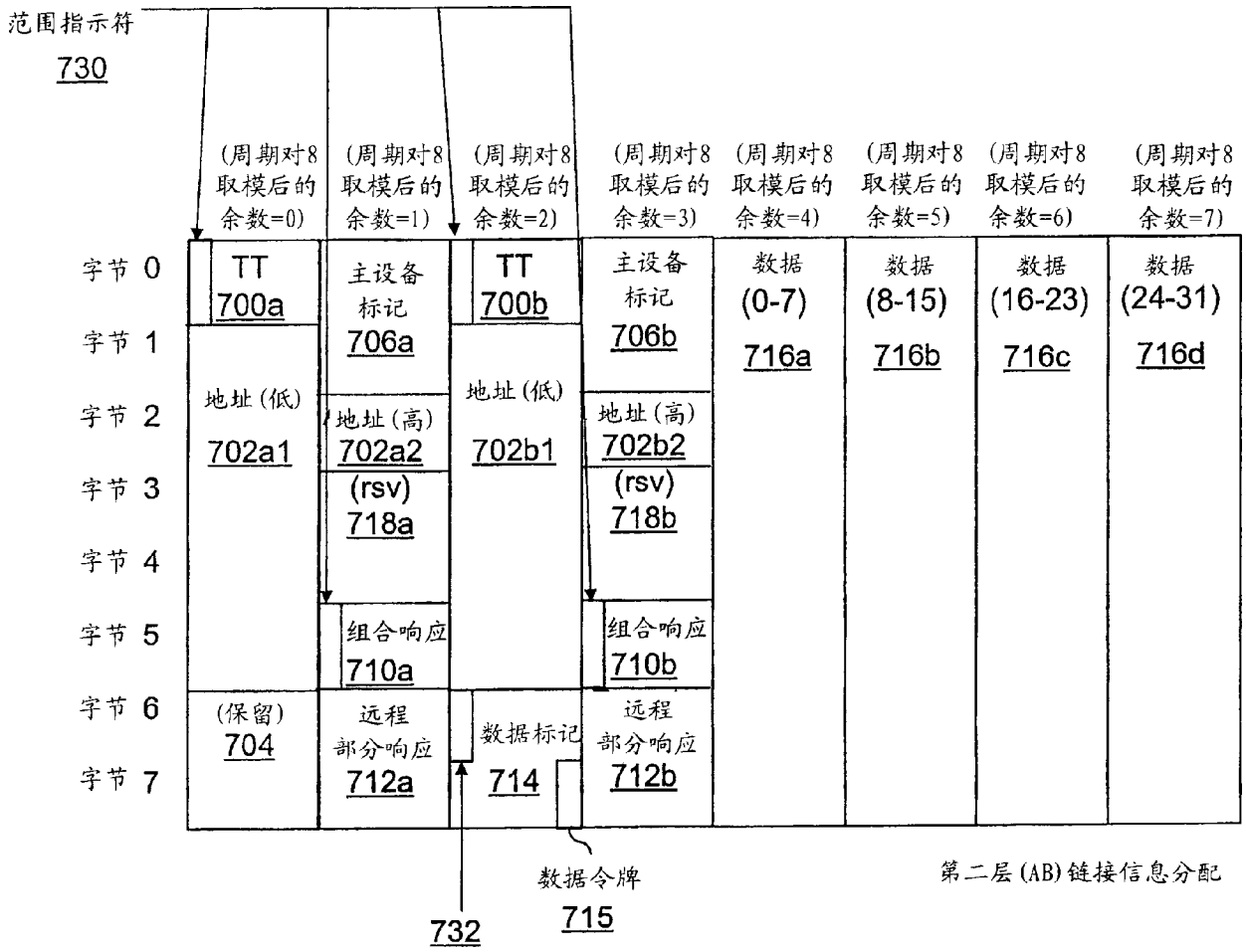

范围指示符
730
(周期对8取模后的余数=0)
(周期对8取模后的余数=1)
(周期对8取模后的余数=2)
(周期对8取模后的余数=3)
(周期对8取模后的余数=4)
(周期对8取模后的余数=5)
(周期对8取模后的余数=6)
(周期对8取模后的余数=7)
字节 0
字节 1
字节 2
字节 3
字节 4
字节 5
字节 6
字节 7
TT 700a
地址(低) 702a1
(保留) 704
主设备标记 706a
地址(高) 702a2
(rsv) 718a
组合响应 710a
远程部分响应 712a
TT 700b
地址(低) 702b1
数据标记 714
主设备标记 706b
地址(高) 702b2
(rsv) 718b
组合响应 710b
远程部分响应 712b
数据(0-7) 716a
数据(8-15) 716b
数据(16-23) 716c
数据(24-31) 716d
732
数据令牌 715
第二层(AB)链接信息分配

图 7B

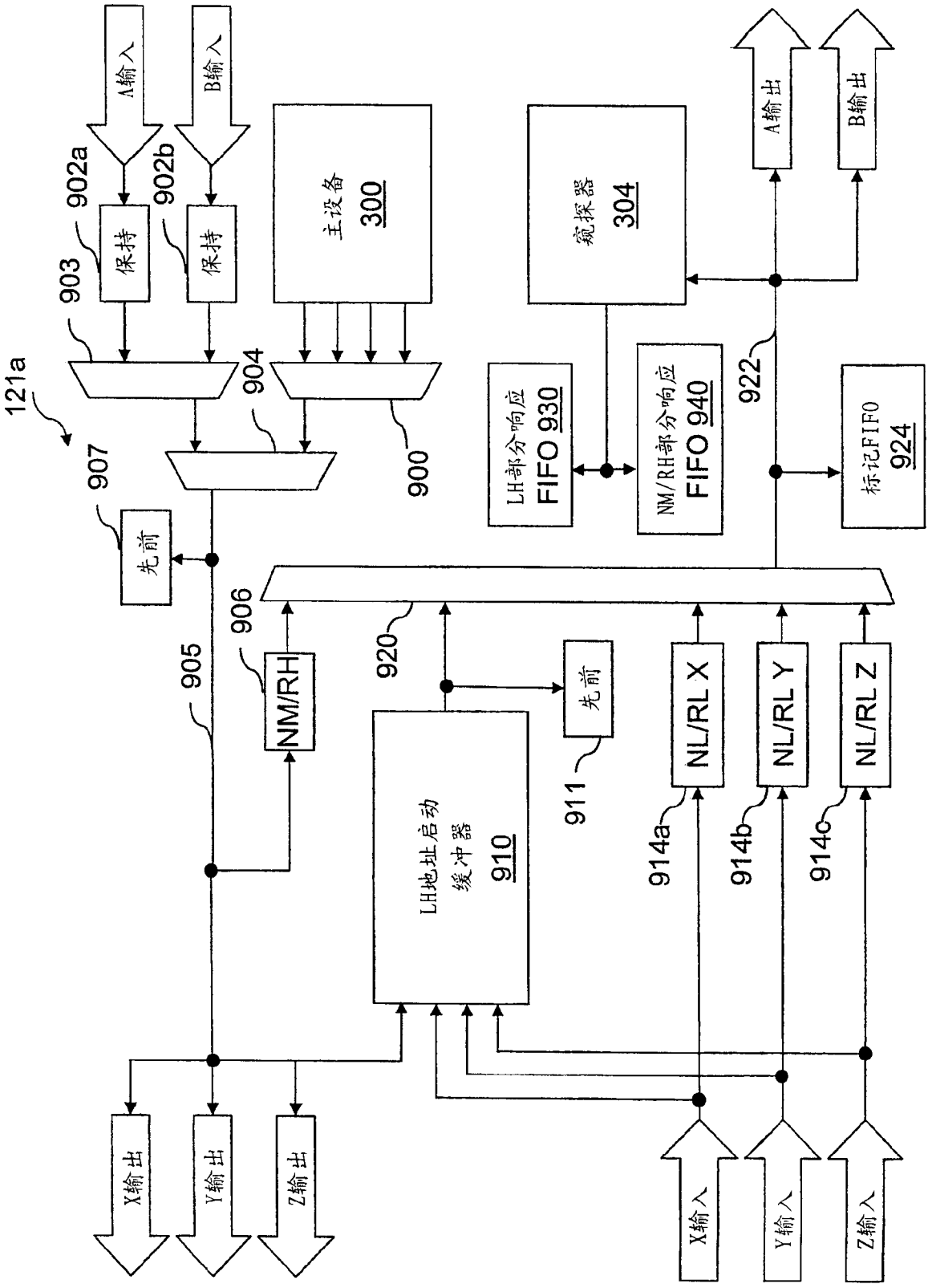

121a
A输入
B输入
902a
902b
903
保持
保持
主设备
300
窥探器
304
A输出
B输出
904
907
900
922
LH部分响应
FIFO 930
NM/RH部分响应
FIFO 940
标记FIFO
924
先前
906
905
NM/RH
920
911
先前
LH地址启动
缓冲器
910
NL/RL X
NL/RL Y
NL/RL Z
914a
914b
914c
X输出
Y输出
Z输出
X输入
Y输入
Z输入

图 8

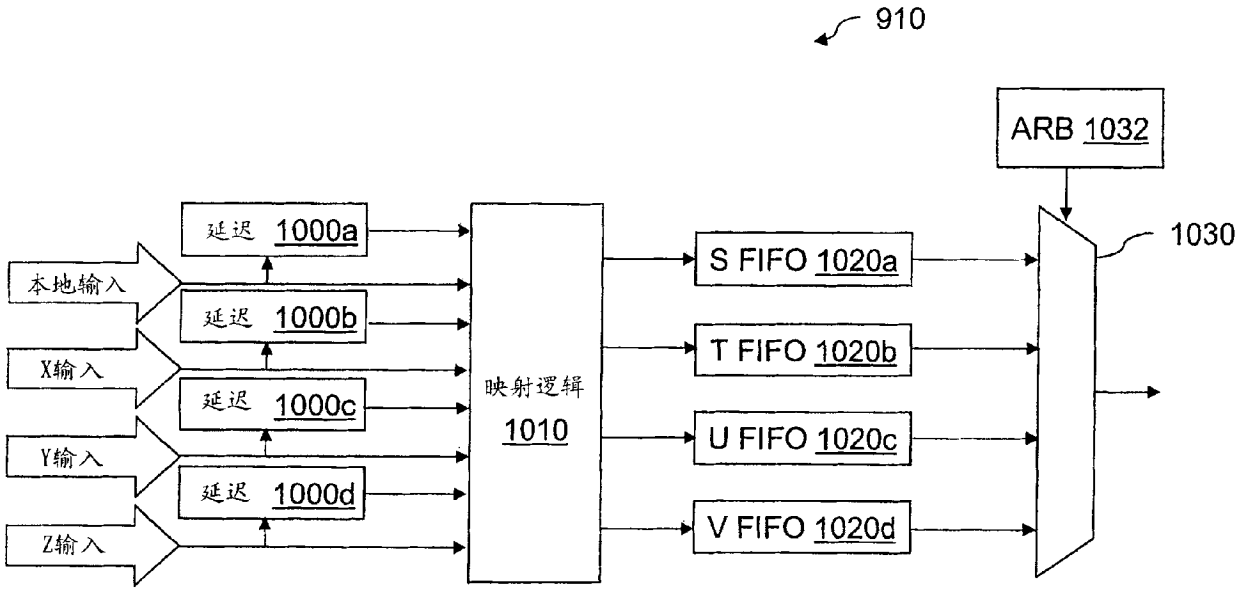

910
ARB 1032
1030
延迟 1000a
延迟 1000b
延迟 1000c
延迟 1000d
本地输入
X输入
Y输入
Z输入
映射逻辑
1010
S FIFO 1020a
T FIFO 1020b
U FIFO 1020c
V FIFO 1020d

图 9

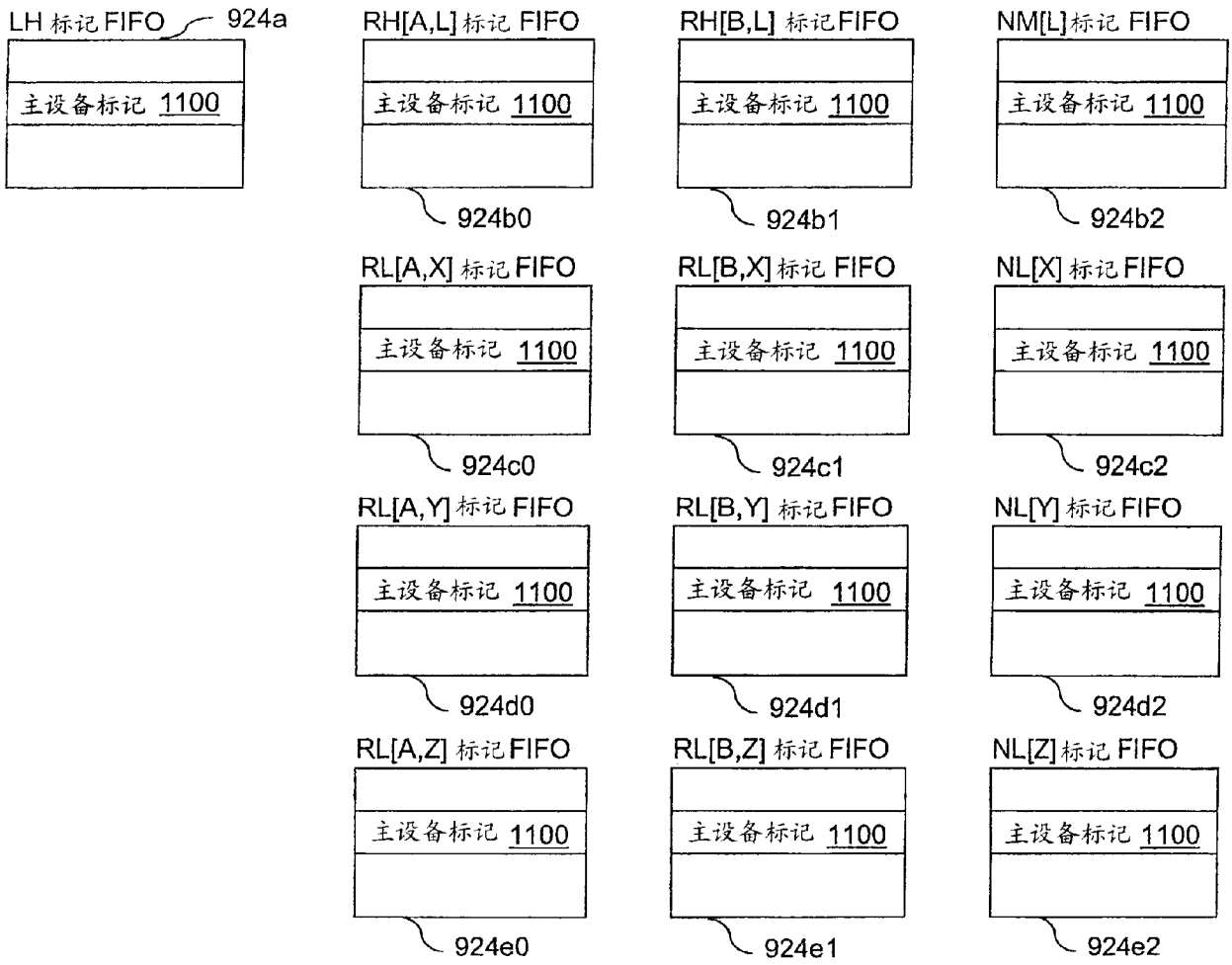

LH 标记 FIFO
924a
主设备标记 1100
RH[A,L] 标记 FIFO
主设备标记 1100
924b0
RH[B,L] 标记 FIFO
主设备标记 1100
924b1
NM[L] 标记 FIFO
主设备标记 1100
924b2
RL[A,X] 标记 FIFO
主设备标记 1100
924c0
RL[B,X] 标记 FIFO
主设备标记 1100
924c1
NL[X] 标记 FIFO
主设备标记 1100
924c2
RL[A,Y] 标记 FIFO
主设备标记 1100
924d0
RL[B,Y] 标记 FIFO
主设备标记 1100
924d1
NL[Y] 标记 FIFO
主设备标记 1100
924d2
RL[A,Z] 标记 FIFO
主设备标记 1100
924e0
RL[B,Z] 标记 FIFO
主设备标记 1100
924e1
NL[Z] 标记 FIFO
主设备标记 1100
924e2

图 10

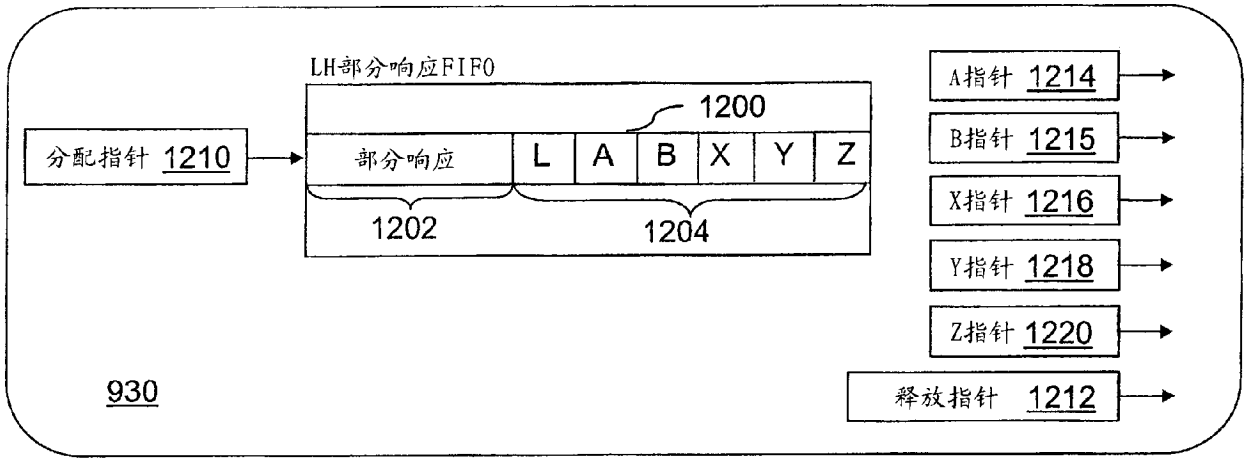

LH部分响应FIFO
1200
分配指针 1210
部分响应
L
A
B
X
Y
Z
1202
1204
A指针 1214
B指针 1215
X指针 1216
Y指针 1218
Z指针 1220
释放指针 1212
930

图 11

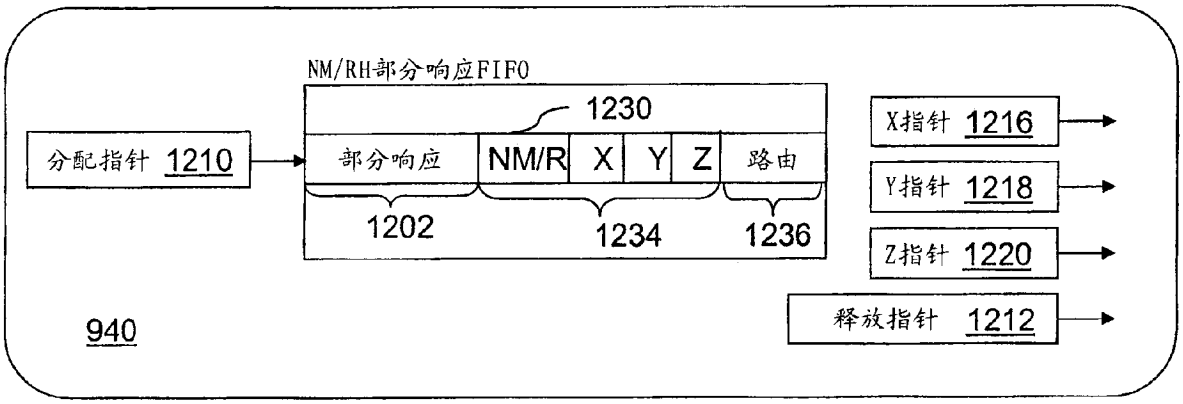

NM/RH部分响应FIFO
1230
分配指针 1210
部分响应
NM/R
X
Y
Z
路由
1202
1234
1236
X指针 1216
Y指针 1218
Z指针 1220
释放指针 1212
940

图 12

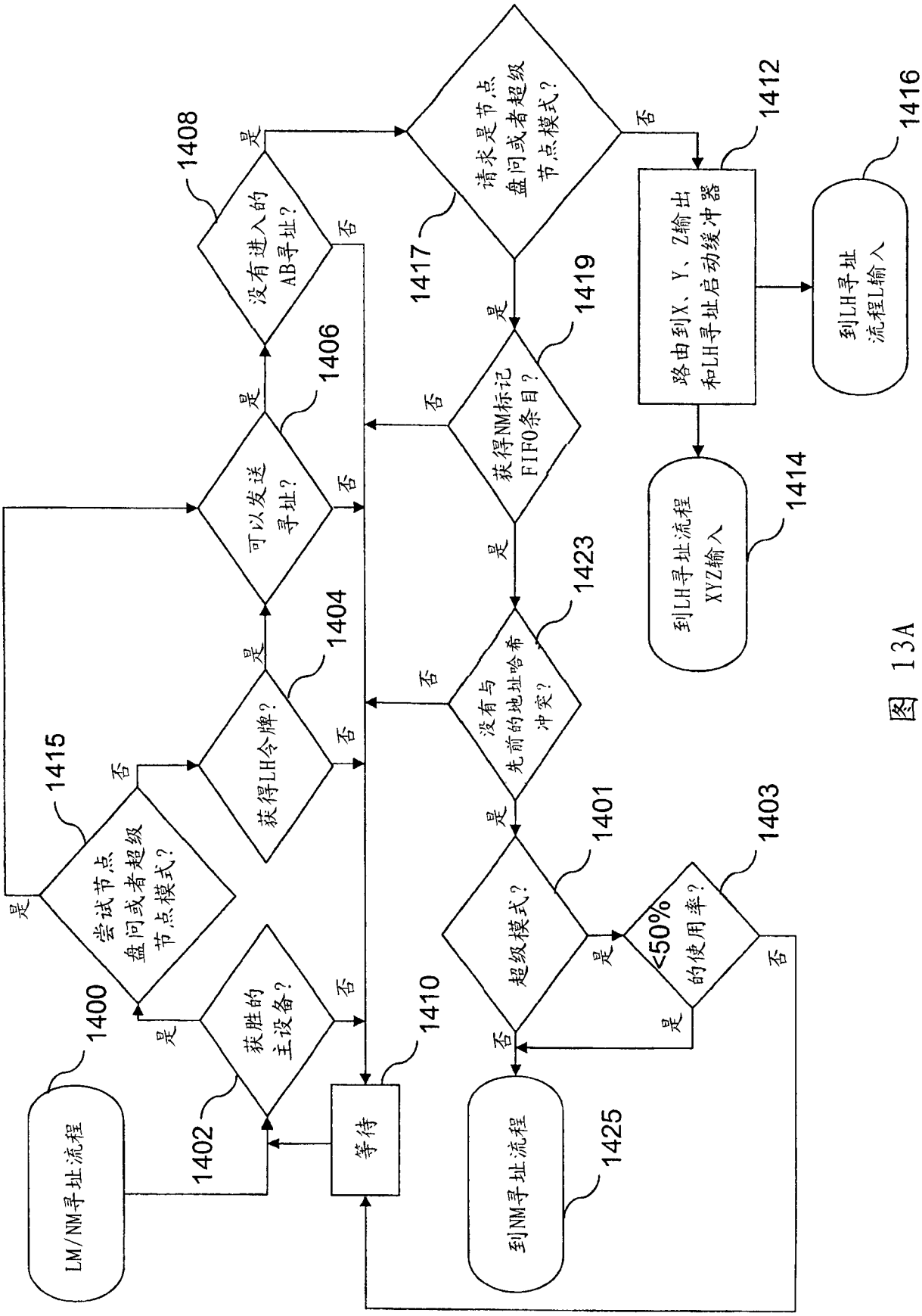

LM/NM寻址流程
1400
尝试节点盘问或者超级节点模式?
1415
获胜的主设备?
1402
获得LH令牌?
1404
可以发送寻址?
1406
没有进入的AB寻址?
1408
请求是节点盘问或者超级节点模式?
1417
获得NM标记FIFO条目?
1419
没有与先前的地址哈希冲突?
1423
超级模式?
1401
<50%的使用率?
1403
等待
1410
到NM寻址流程
1425
路由到X、Y、Z输出和LH寻址启动缓冲器
1412
到LH寻址流程XYZ输入
1414
到LH寻址流程L输入
1416
是
否

图 13A

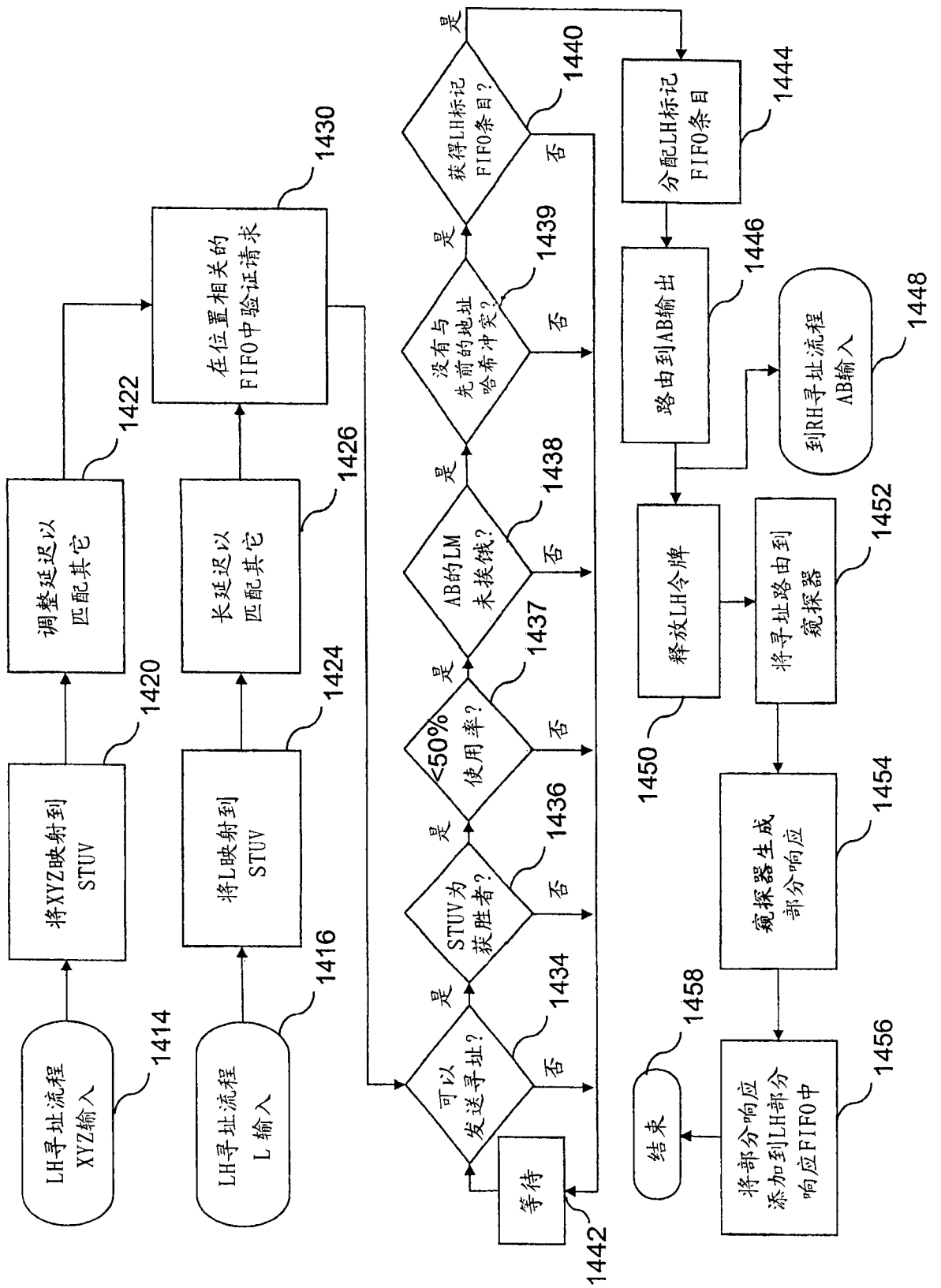

LH寻址流程 XYZ输入
1414
将XYZ映射到 STUV
1420
调整延迟以 匹配其它
1422
在位置相关的 FIFO中验证请求
1430
LH寻址流程 L 输入
1416
将L映射到 STUV
1424
长延迟以 匹配其它
1426
可以 发送寻址?
1434
是
否
STUV为 获胜者?
1436
是
否
<50% 使用率?
1437
是
否
AB的LM 未挨饿?
1438
是
否
没有与 先前的地址 哈希冲突?
1439
是
否
获得LH标记 FIFO条目?
1440
是
否
等待
1442
结束
1458
将部分响应 添加到LH部分 响应FIFO中
1456
窥探器生成 部分响应
1454
释放LH令牌
1450
将寻址路由到 窥探器
1452
路由到AB输出
1446
分配LH标记 FIFO条目
1444
到RH寻址流程 AB输入
1448

图 13B

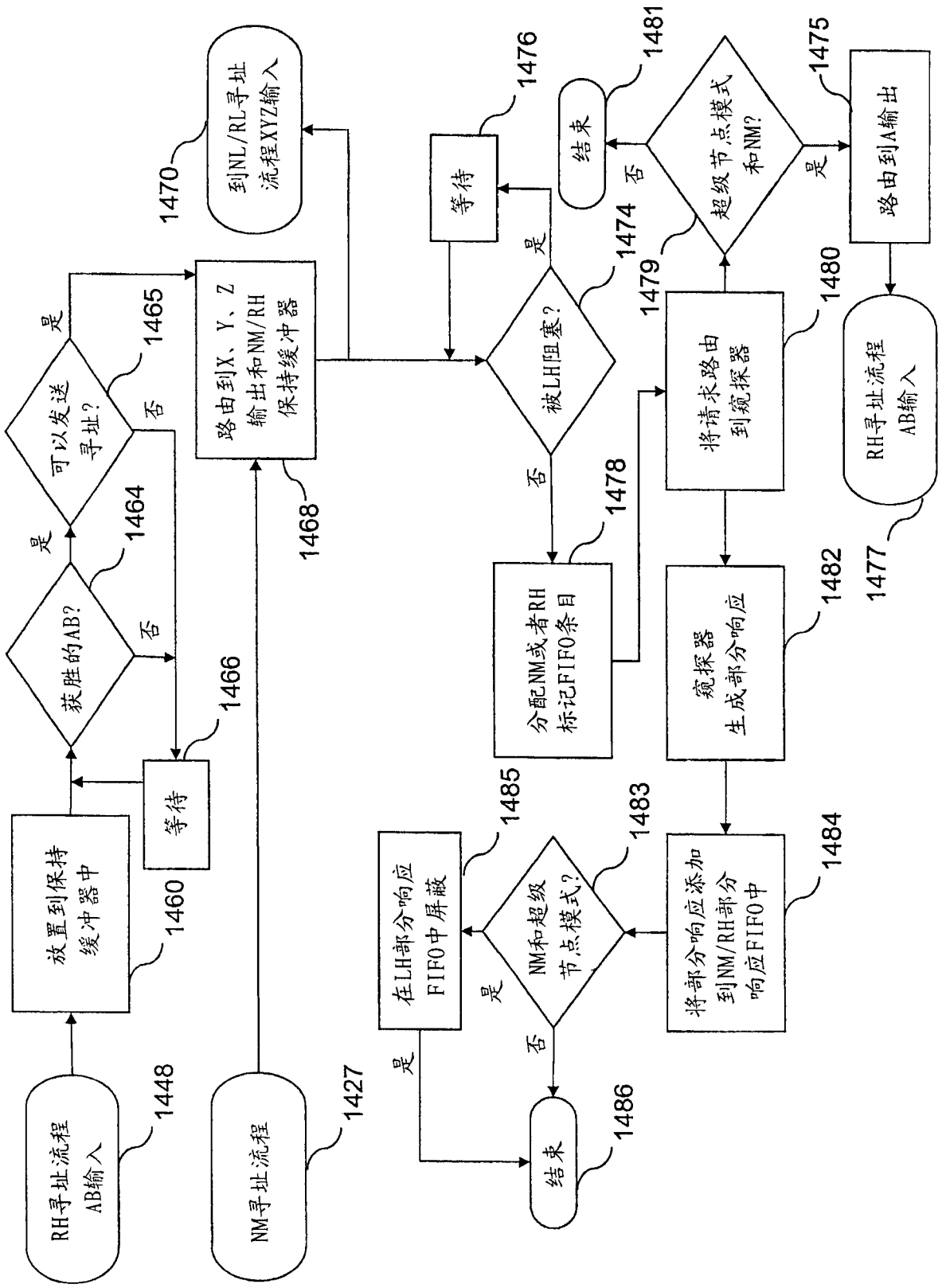

RH寻址流程 AB输入
1448
放置到保持缓冲器中
1460
获胜的AB?
1464
可以发送寻址?
1465
等待
1466
NM寻址流程
1427
路由到X、Y、Z输出和NM/RH保持缓冲器
1468
到NL/RL寻址流程XYZ输入
1470
等待
1476
被LH阻塞?
1474
分配NM或者RH标记FIFO条目
1478
结束
1481
超级节点模式和NM?
1475
路由到A输出
RH寻址流程 AB输入
1477
将请求路由到窥探器
1479
1480
窥探器生成部分响应
1482
将部分响应添加到NM/RH部分响应FIFO中
1484
NM和超级节点模式?
1483
在LH部分响应FIFO中屏蔽
1485
结束
1486
是
否

图 13C

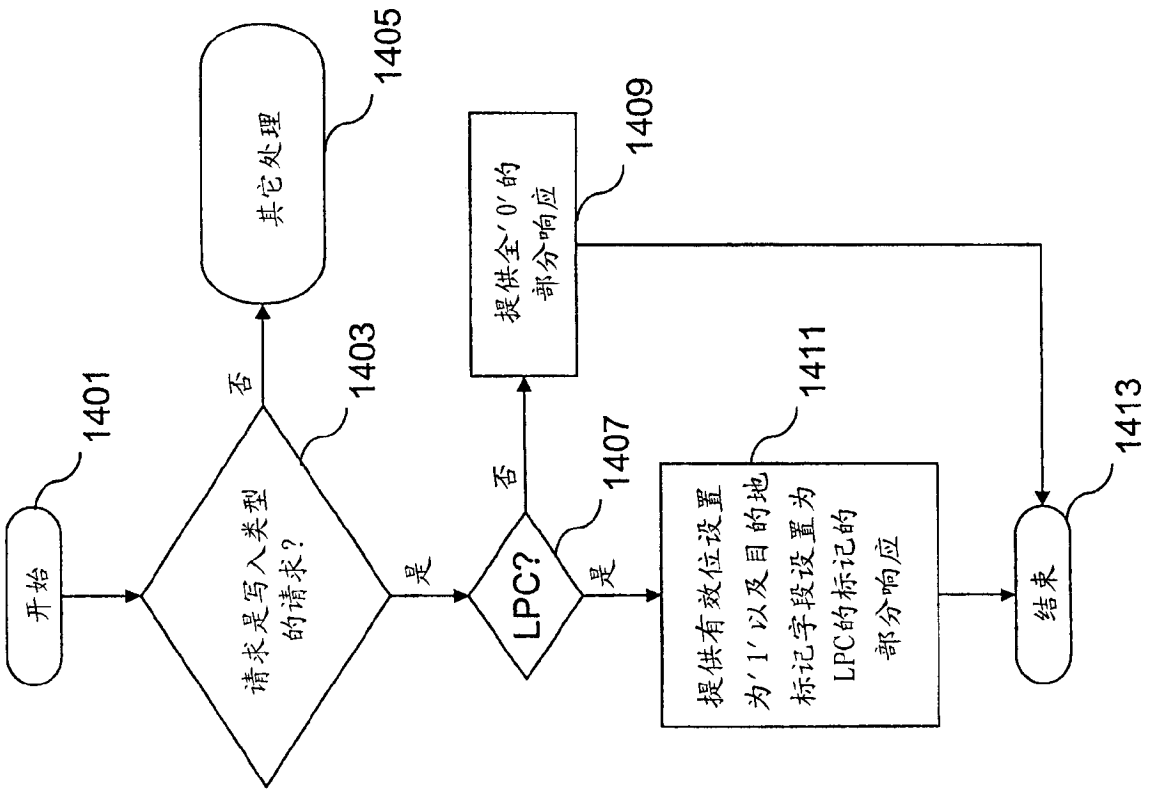

开始
1401
请求是写入类型的请求?
1403
否
其它处理
1405
是
LPC?
1407
否
提供全'0'的部分响应
1409
是
提供有效位设置为'1'以及目的地标记字段设置为LPC的标记的部分响应
1411
结束
1413

图 13E

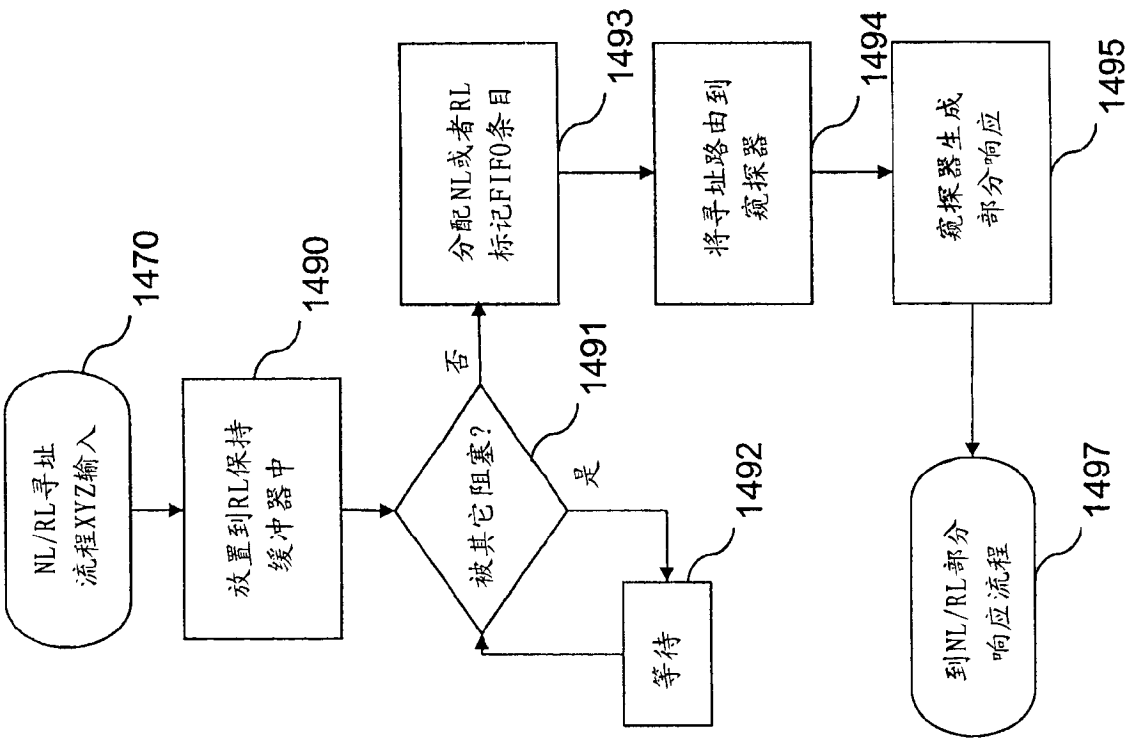

NL/RL寻址流程XYZ输入
1470
放置到RL保持缓冲器中
1490
被其它阻塞?
1491
否
是
等待
1492
分配NL或者RL标记FIFO条目
1493
将寻址路由到窥探器
1494
窥探器生成部分响应
1495
到NL/RL部分响应流程
1497

图 13D

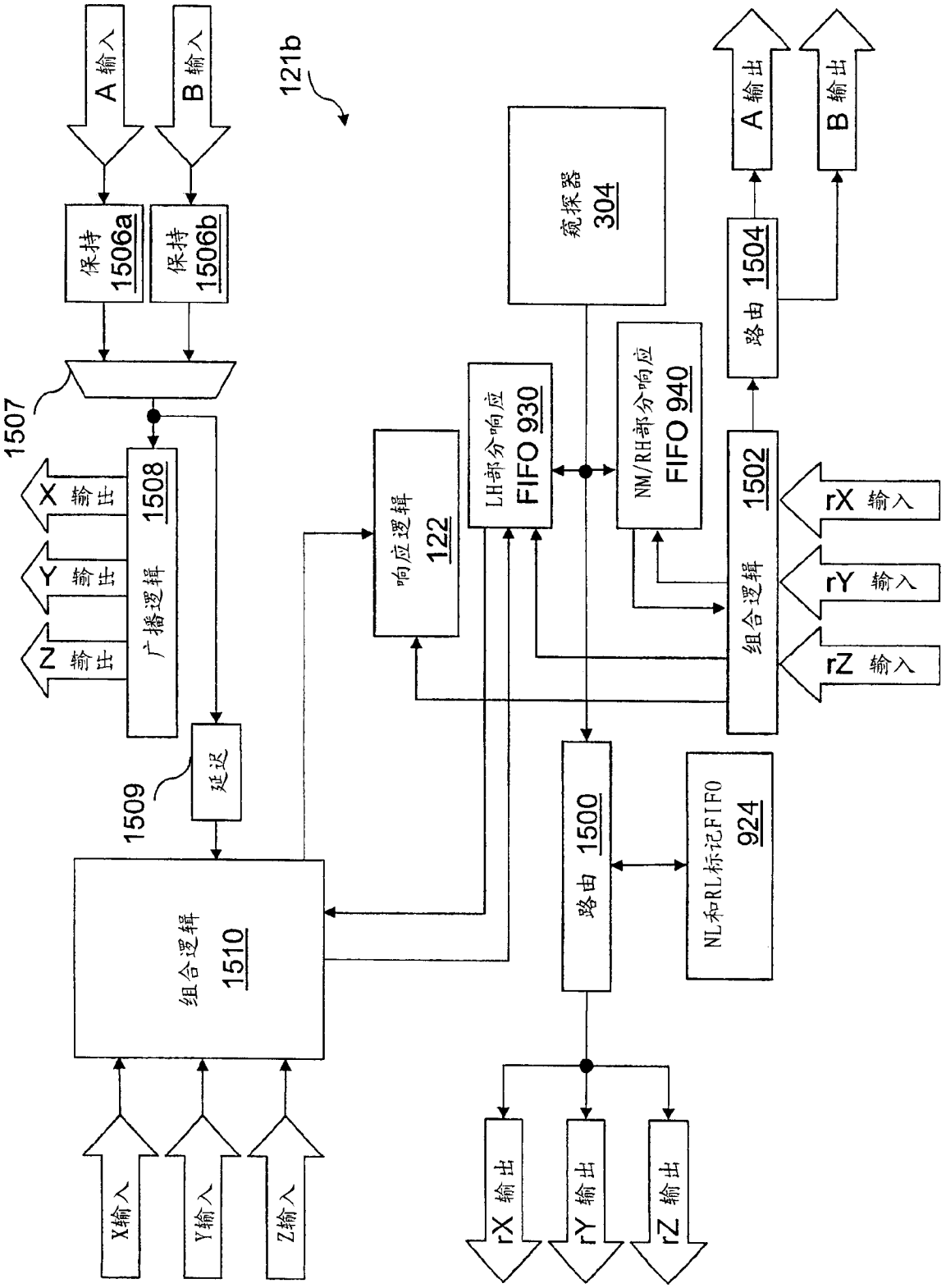

121b
A 输入
B 输入
保持 1506a
保持 1506b
1507
1508
广播逻辑
X 输出
Y 输出
Z 输出
1509
延迟
组合逻辑 1510
X输入
Y输入
Z输入
响应逻辑 122
LH部分响应 FIFO 930
NM/RH部分响应 FIFO 940
窥探器 304
路由 1500
NL和RL标记FIFO 924
rX 输出
rY 输出
rZ 输出
路由 1504
A 输出
B 输出
1502
组合逻辑
rX 输入
rY 输入
rZ 输入

图 14

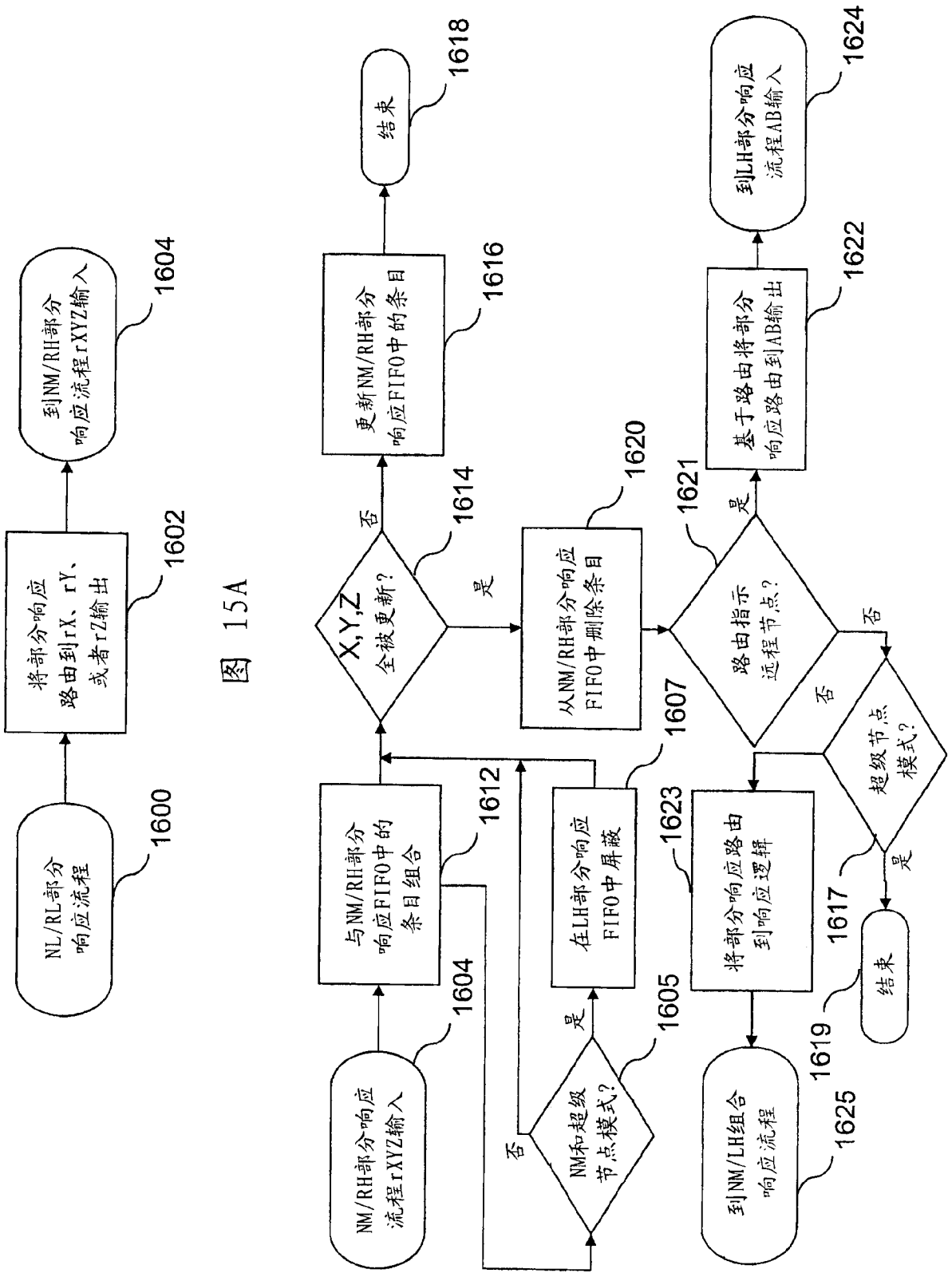

NL/RL部分响应流程
1600
将部分响应路由到rX、rY、或者rZ输出
1602
到NM/RH部分响应流程rXYZ输入
1604
NM/RH部分响应流程rXYZ输入
1604
与NM/RH部分响应FIFO中的条目组合
1612
在LH部分响应FIFO中屏蔽
1605
NM和超级节点模式?
1607
是
否
X,Y,Z全被更新?
1614
否
是
更新NM/RH部分响应FIFO中的条目
1616
结束
1618
从NM/RH部分响应FIFO中删除条目
1620
路由指示远程节点?
1621
是
否
基于路由将部分响应路由到AB输出
1622
到LH部分响应流程AB输入
1624
超级节点模式?
1617
是
否
结束
1619
将部分响应路由到响应逻辑
1623
到NM/LH组合响应流程
1625

图 15A

图 15B

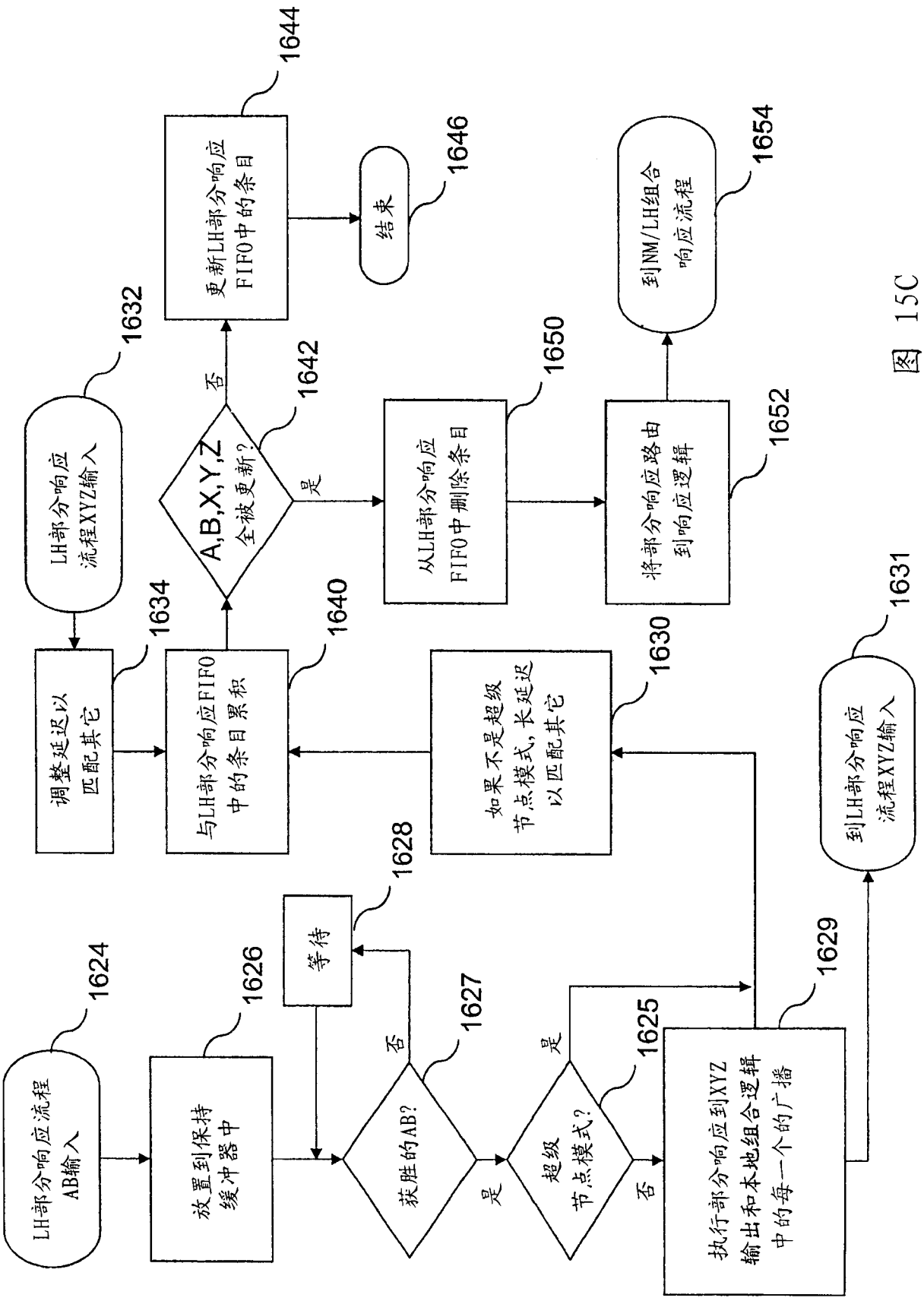

LH部分响应流程AB输入
1624
放置到保持缓冲器中
1626
等待
1628
获胜的AB?
1627
否
是
超级节点模式?
1625
是
否
执行部分响应到XYZ输出和本地组合逻辑中的每一个的广播
1629
到LH部分响应流程XYZ输入
1631
如果不是超级节点模式，长延迟以匹配其它
1630
与LH部分响应FIFO中的条目累积
1640
调整延迟以匹配其它
1634
LH部分响应流程XYZ输入
1632
A,B,X,Y,Z全被更新?
1642
否
是
更新LH部分响应FIFO中的条目
1644
结束
1646
从LH部分响应FIFO中删除条目
1650
将部分响应路由到响应逻辑
1652
到NM/LH组合响应流程
1654

图 15C

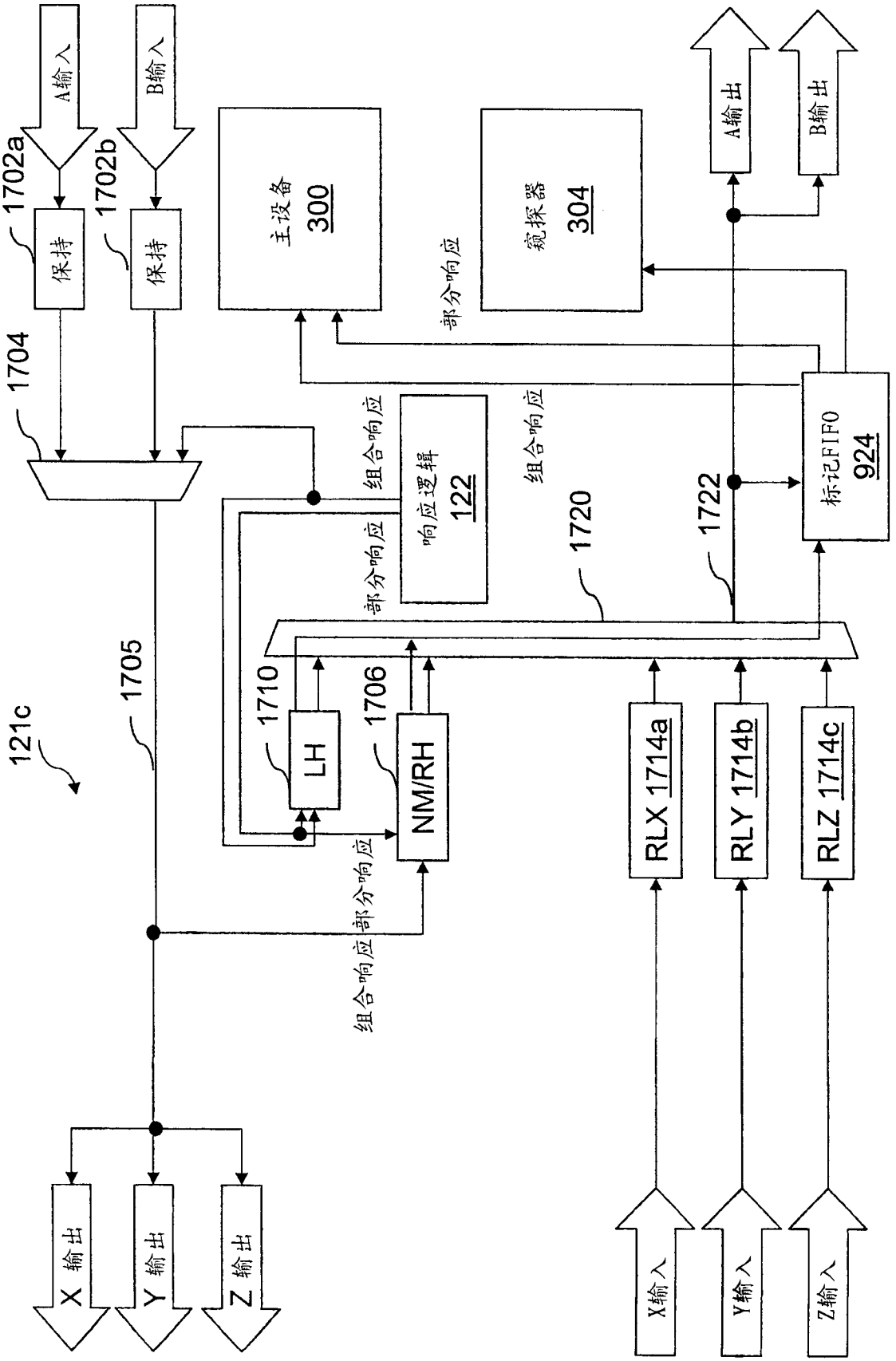

A输入
B输入
1702a
1702b
保持
保持
主设备
300
窥探器
304
A输出
B输出
部分响应
1704
组合响应
响应逻辑
122
组合响应
1720
1722
标记FIFO
924
部分响应
1705
1710
1706
121c
LH
NM/RH
RLX 1714a
RLY 1714b
RLZ 1714c
组合响应
部分响应
X 输出
Y 输出
Z 输出
X输入
Y输入
Z输入

图 16

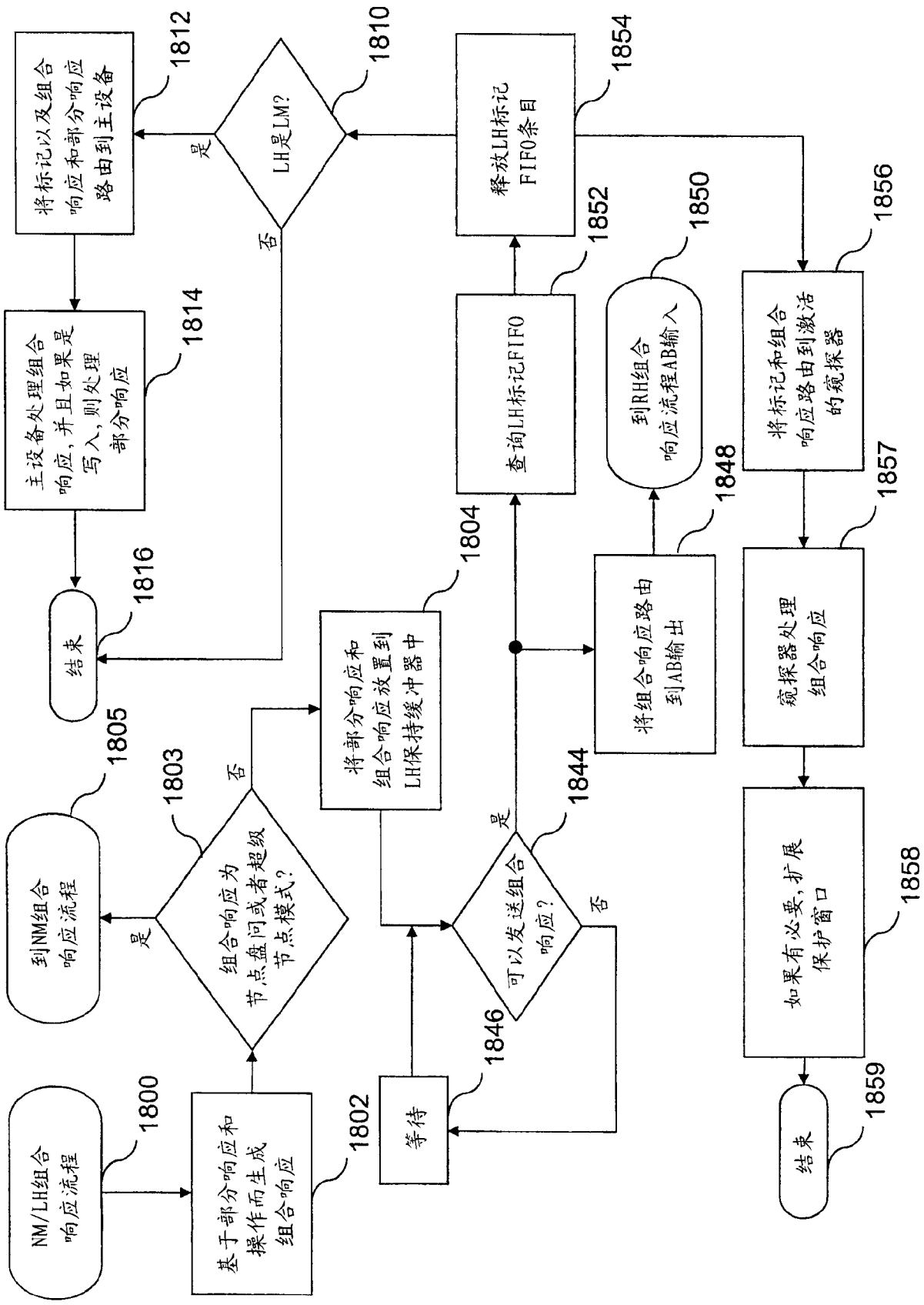

NM/LH组合响应流程
1800
基于部分响应和操作而生成组合响应
1802
组合响应为节点盘问或者超级节点模式?
1803
是
否
到NM组合响应流程
1805
将部分响应和组合响应放置到LH保持缓冲器中
1804
等待
1846
可以发送组合响应?
1844
是
否
将组合响应路由到AB输出
1848
到RH组合响应流程AB输入
1850
查询LH标记FIFO
1852
释放LH标记FIFO条目
1854
LH是LM?
1810
是
否
将标记以及组合响应和部分响应路由到主设备
1812
主设备处理组合响应,并且如果是写入,则处理部分响应
1814
结束
1816
窥探器处理组合响应
1857
将标记和组合响应路由到激活的窥探器
1856
如果有必要,扩展保护窗口
1858
结束
1859

图 17A

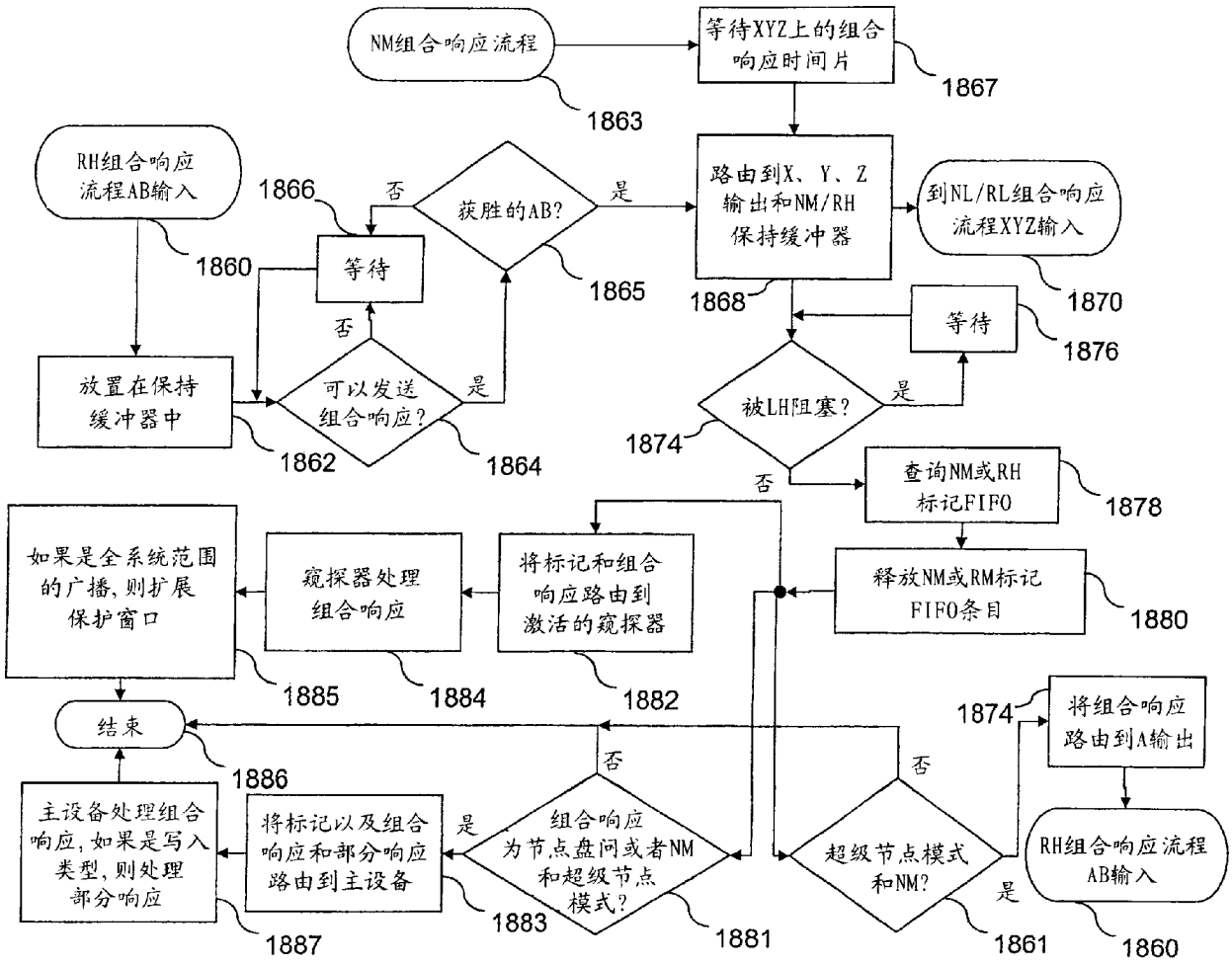

NM组合响应流程
1863
等待XYZ上的组合响应时间片
1867
RH组合响应流程AB输入
1860
1866
否
获胜的AB?
是
1865
路由到X、Y、Z输出和NM/RH保持缓冲器
1868
到NL/RL组合响应流程XYZ输入
1870
等待
否
等待
1876
放置在保持缓冲器中
1862
可以发送组合响应?
是
1864
被LH阻塞?
是
1874
否
查询NM或RH标记FIFO
1878
如果是全系统范围的广播,则扩展保护窗口
1885
窥探器处理组合响应
1884
将标记和组合响应路由到激活的窥探器
1882
释放NM或RM标记FIFO条目
1880
结束
1886
否
否
1874
将组合响应路由到A输出
主设备处理组合响应,如果是写入类型,则处理部分响应
1887
将标记以及组合响应和部分响应路由到主设备
是
组合响应为节点盘问或者NM和超级节点模式?
1883
1881
超级节点模式和NM?
是
1861
RH组合响应流程AB输入
1860

图 17B

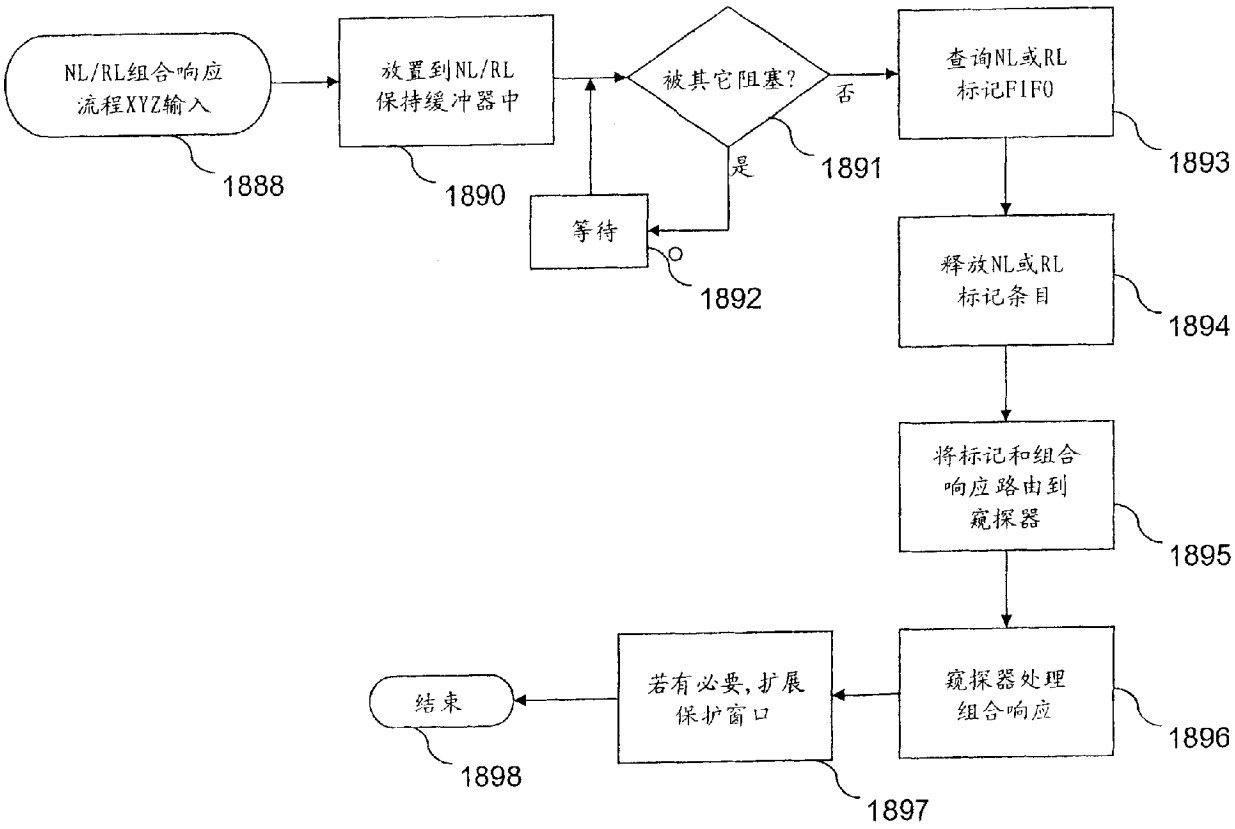

NL/RL组合响应
流程XYZ输入
1888
放置到NL/RL
保持缓冲器中
1890
被其它阻塞?
否
是
1891
查询NL或RL
标记FIFO
1893
等待
1892
释放NL或RL
标记条目
1894
将标记和组合
响应路由到
窥探器
1895
窥探器处理
组合响应
1896
若有必要,扩展
保护窗口
1897
结束
1898

图 17C