[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 13/10

[12] 发明专利申请公开说明书

[21] 申请号 02804290.5

[43] 公开日 2004 年 6 月 30 日

[11] 公开号 CN 1509440A

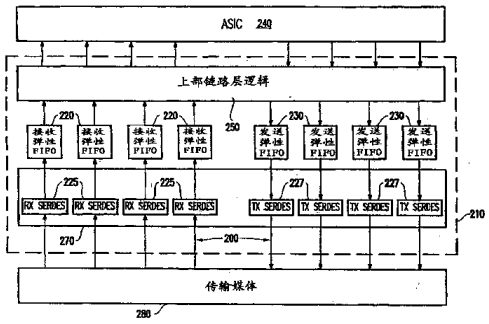
[22] 申请日 2002.2.15 [21] 申请号 02804290.5
[30] 优先权
[32] 2001. 3.23 [33] US [31] 09/816,979
[86] 国际申请 PCT/EP2002/001637 2002.2.15
[87] 国际公布 WO2002/077833 英 2002.10.3
[85] 进入国家阶段日期 2003.7.29
[71] 申请人 国际商业机器公司
地址 美国纽约
[72] 发明人 格里戈里·J·曼
[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 吴丽丽

权利要求书 3 页 说明书 9 页 附图 5 页

[54] 发明名称 用于校正并行/串行接口中的波动的缓冲网

[57] 摘要

一种用于输入/输出接口的弹性型先进先出(FIFO)缓冲器网络，使得高链路层时钟频率被赋予这些“并行－串行”高速链路接口的固定发送时钟频率。 该网络尤其可以应用于 InfiniBand 型硬件内的接口部件。



ISSN1008-4274

1. 一种用于在并行－串行体系结构中的传输媒体与处理器之间提供通信的核心，所述核心包括:

逻辑层;

至少一个串行通路，用于将所述逻辑层连接到所述传输媒体;以及

至少一个缓冲器，插入每个串行通路中，

其中每个缓冲器对所述传输媒体内的波动进行校正，并改变沿所述串行通路处理的信号的频率。

2. 根据权利要求 1 所述的核心，该核心进一步包括连接在所述传输媒体与每个缓冲器之间的串行化器/去串行化器。

3. 根据权利要求 1 所述的核心，其中所述缓冲器包括弹性先进先出（FIFO）缓冲器。

4. 根据权利要求 1 所述的核心，其中每个缓冲器位于所述逻辑层的外部。

5. 根据权利要求 1 所述的核心，其中所述缓冲器包括多个缓冲器，并且第一组所述缓冲器改变从所述逻辑层传送到所述传输媒体的信号的频率。

6. 根据权利要求 5 所述的核心，其中第二组所述缓冲器改变从所述传输媒体传送到所述逻辑层的信号的频率。

7. 根据权利要求 1 所述的核心，其中所述处理器是从包括主机通道适配器、目标通道适配器以及互连交换机的组中选择的单元。

8. 一种并行－串行体系结构网络，它包括传输媒体和通过权利要求 1 所述的核心连接到所述传输媒体的至少一个处理器，其中所述核心在所述传输媒体与所述处理器之间提供通信，所述核心包括:

逻辑层;

多个串行通路，用于将所述逻辑层连接到所述传输媒体；以及

所述串行通路内的多个接收缓冲器和发送缓冲器;

其中所述接收缓冲器和所述发送缓冲器改变沿所述串行通路处理的信号的频率。

9. 根据权利要求 8 所述的并行－串行体系结构网络，该网络进一步包括位于所述串行通路内的多个串行化器/去串行化器。

10. 根据权利要求 8 所述的并行－串行体系结构网络，其中所述接收缓冲器和所述发送缓冲器包括弹性先进先出（FIFO）缓冲器。

11. 根据权利要求 8 所述的并行－串行体系结构网络，其中所述接收缓冲器和所述发送缓冲器位于所述逻辑层的外部。

12. 根据权利要求 8 所述的并行－串行体系结构网络，其中所述发送缓冲器改变从所述层逻辑传送到所述传输媒体的信号的频率。

13. 根据权利要求 8 所述的并行－串行体系结构网络，其中所述接收缓冲器对从所述传输媒体传送到所述逻辑层的信号进行处理。

14. 根据权利要求 8 所述的并行－串行体系结构网络，其中所述处理器包括主机通道适配器、目标通道适配器以及互连交换机之一。

15. 根据权利要求 1 所述的核心，用于在字节带并行－串行 InfiniBand 体系结构中在传输媒体与处理器之间提供通信，所述核心包括:

逻辑层;

多个串行通路，用于将所述逻辑层连接到所述传输媒体；以及

所述串行通路内的多个接收缓冲器和发送缓冲器，

其中所述接收缓冲器对所述传输媒体上的波动进行校正，并改变沿所述串行通路处理的信号的频率。

16. 根据权利要求 15 所述的核心，该核心进一步包括多个位于所述串行通路上的串行化器/去串行化器。

17. 根据权利要求 15 所述的核心，其中所述接收缓冲器和所述发送缓冲器包括弹性先进先出（FIFO）缓冲器。

18. 根据权利要求 15 所述的核心，其中所述接收缓冲器和所述

发送缓冲器位于所述逻辑层的外部。

19. 根据权利要求 15 所述的核心，其中所述发送缓冲器改变从所述层逻辑传送到所述传输媒体的信号的频率。

20. 根据权利要求 15 所述的核心，其中所述接收缓冲器对从所述传输媒体传送到所述逻辑层的信号进行处理。

用于校正并行/串行接口中的波动的缓冲网

技术领域

本发明涉及输入/输出（I/0）数据传输设备，更具体地说，本发明涉及I/O数据传输通路中的先进先出（FIFO）缓冲设备

背景技术

InfiniBand（InfiniBand Trade Association, Portland, Oregon的注册商标）体系结构是对基于通道的、交换结构技术拟定的新通用I/O规范，整个硬件业和软件业均可以采用这种规范。图1a示出与InfiniBand网络100有关的网络和部件。基于InfiniBand的网络用于满足需要大量带宽的网络应用，例如通过因特网综合了语音、数据和视频的网络应用。InfiniBand体系结构正由包括许多硬件公司和软件公司的InfiniBand Trade Association开发。其鲁棒的分层设计可以使多个计算机系统与外围设备更容易一起作为一个高性能的、高可用性服务器工作。

作为以结构为中心的、基于消息的体系结构，InfiniBand最适于多种网络应用中的群集、输入/输出扩展以及本机附件。InfiniBand技术可以用于建立远程卡笼15，也可以连接到附加主机35、路由器40或磁盘阵列。InfiniBand的特征还在于，增强故障隔离、支持冗余以及内置故障切换（failover）能力，从而提供高网络可靠性和可用性。因为以高性能和高可靠性为特征，所以这些设备对包括服务器和存储区网络的大量网络基础设施部件提供解决方案。

在图1b中，以示例形式示出部分图1a所示网络内的InfiniBand部件的方框图。这些部件具有分别构成部分目标通道适配器（TCA）10、主机通道适配器（HCA）20、互连交换设备30以及路由器40的各输入/输出接口，它们分别具有包括InfiniBand技术链

路协议引擎（IBT－LPE）核心的专用集成电路（ASIC）核心接口，在 InfiniBand 技术（IBT）网络 100 中，InfiniBand 技术链路协议引擎核心通过链路 25 将 ASIC 连接在每个这些部件之间。IBT－LPE 核心支持位于较上部的物理层和下部链路层的所有 IBT 设备要求的大量功能。它还处理达到并且包括以每秒 2.5 千兆位工作的 4 宽带链路的全部 IBT 带宽要求。位于较上部的物理层的 IBT－LPE 核心（大规模集成电路设计）和 ASIC 的链路层核心遵守 InfiniBand Trade Association 在 IBTA 1.0 规范（2001）中设立的标准。利用基于通道的点到点连接，而不利用共享总线、负载以及存储结构，这种体系结构将 I/O 子系统与存储器去耦合。

TCA 10 对 InfiniBand 型数据存储部件和通信部件提供接口。通过利用合作、协同处理方法设计 InfiniBand 和本机 I/O 适配器，可以创建利用 InfiniBand 体系结构的性能优势的 InfiniBand 适配器。TCA 10 对 InfiniBand 结构提供高性能接口，并且，利用包括队列、共享存储模块以及门铃的非常简单的接口，主机通道与基于主机的 I/O 控制器进行通信。同时，TCA 和 I/O 控制器用作 InfiniBand I/O 通道深度适配器。TCA 以硬件方式实现在队列之间移动数据以及共享主机总线上的存储器和 InfiniBand 网络上的分组所需的全部机制。将具有最佳排队的基于硬件的数据移动和与基于主机的 I/O 控制器功能块并行工作的互连交换优先级仲裁方案组合在一起，可以使 InfiniBand 适配器的性能最大化。HCA 20 可以实现从主机总线连接到双 1X 或者 4X InfiniBand 网络。这样可以使现有服务器连接到 InfiniBand 网络并通过 InfiniBand 结构与其他节点通信。连接到 InfiniBand HCA 的主机总线集成了双 InfiniBand 接口适配器（物理层、链路层以及传输层）、主机总线接口、直接存储器目标访问（DMA）引擎以及管理支持。它实现了分层存储器结构，在该分层存储器结构中，将与连接有关的信息存储到直接安装在 HCA 上的通道设备上（on-device）存储器或通道设备外（off-device）存储器。其特征是，在两个方向上进行适配器流水线标题与数据处理。两个嵌

入式 InfiniBand 微处理器和各独立直接存储器访问（DMA）引擎使得可以同时对数据通路进行接收和发送处理。

互连交换机 30 可以是引入了 8 个 InfiniBand 端口和一个管理接口的 8 端口 4X 交换机。每个端口均可以连接到另一个交换机、TCA 10 或者 HCA 20，从而实现了具有多个服务器和外围设备在基于 InfiniBand 的高性能网络内一起工作的配置。互连交换机 30 集成了每个端口的物理层和链路层，并且执行滤波、映射、排队以及仲裁功能。它包括多点广播支持以及性能和差错计数器。管理接口连接到实现配置和控制功能的管理处理器。互连交换机 30 通常可以提供 64 千兆位的最大集合通道吞吐量，它集成了缓冲存储器，并且每个端口支持多达 4 个数据虚拟通路（VL）和一个管理 VL。

图 2 示出用于将 InfiniBand 传输媒体 280（图 1b 所示的链路 25）连接到专用集成电路（ASIC）240（例如，TCA 10、HCA 20、交换机 30、路由器 40 等，如图 1b 所示）的核心逻辑 210。利用以下披露的发明改进图 2 所示的核心逻辑 210。图 2 所示的核心逻辑 210 不一定是现有技术，并且，在提交本发明时，它通常可能不被本技术领域内的普通技术人员所知。尽管在图 2 中将核心逻辑 210 与 ASIC 240 分开示出，但是本技术领域内的普通技术人员明白，核心逻辑通常是 ASIC 的一部分。

接收和发送数据传输媒体时钟 280 可以以不同的频率运行（例如，接收通道为 250MHz +/-百万分之 100，而核心逻辑 210 发送数据通道可以以 250 MHz 运行）。此外，与 ASIC 240 时钟速度相比（例如，312MHz），核心逻辑 210 又可以以不同的频率工作。

为了适应所处理的各数据信号的不同速率，核心逻辑 210 包括串行化部分 270，串行化部分 270 包括串行化/去串行化单元 225、227。这种串行化/去串行化单元的结构和运行过程为本技术领域内的普通技术人员所知，因此，为了不使本发明的显著特征无谓的模糊不清，在此不对它们做详细说明。

InfiniBand 传输媒体 280 由形成链路 25 的大量串行传输通路构

成。接收串行化/去串行化单元 225 去串行化来自传输媒体 280 的信号，并充分进行变换以将频率降低到核心逻辑 210 可以接受的频率。例如，如果串行化/去串行化接收单元 225 运行以便一次去串行化 10 位，则出现 10 比 1 的降低，这样将传输媒体 280 上每秒 2.5 千兆位的速度降低到核心逻辑 210 可以接受的 250 MHz 频率。

核心逻辑 210 还包括频率校正单元 260。沿传输媒体 280 传播的信号的频率不可能始终以该线速出现，但是它可以少许高于或者低于要求的频率（例如，至多高于或者低于百万分之 100）。频率的这种不一致性会通过串行化/去串行化单元 225 传送。频率校正单元 261 包括 FIFO 缓冲器，FIFO 缓冲器 261 缓存串行化/去串行化单元 225 输出的信号，以将 250MHz 均匀频率的信号送到上部链路层逻辑 250。

上部链路层逻辑 250 包括附加 FIFO 缓冲器 251，附加 FIFO 缓冲器 251 将频率校正单元 260 输出的信号的频率变换为 ASIC 240 可以接受的频率。在信号从 ASIC 240 传输到传输媒体 280 期间，进行相反的处理，并且上部链路层逻辑 250 采用不同的 FIFO 缓冲器 253。同样，串行化单元 270 使用其他传输串行化/去串行化单元 227。请注意，频率校正单元 262 不需要对传送到传输媒体 280 的信号进行校正，因为 ASIC 240 通常产生不需要进行校正的信号。

图 2 所示的核心逻辑 210 的一个缺点是，上部链路层逻辑 250 和频率校正单元 260 所需的大量缓冲器 251、253、261。这些缓冲器消耗大量电路功率，并降低了利用核心逻辑 210 处理数据的操作速度。因此，为了降低这种功率消耗并提高处理速度，需要减少核心逻辑 210 内的缓冲器的数量。

发明内容

基于上述问题，推出本发明。本发明的目的是提供一种并行－串行体系结构网络，它包括传输媒体和至少一个通过核心与该传输媒体相连的处理器。该核心用于在传输媒体与处理器提供通信。

核心包括与处理器相连的逻辑层、将逻辑层连接到传输媒体的串行通路以及位于串行通路上的接收和发送缓冲器。接收缓冲器对传输媒体内的波动进行校正，并改变沿串行通路处理的各信号的频率。

本发明还可以在串行通路内包括串行化器/去串行化器。接收缓冲器和发送缓冲器优先是弹性先进先出（FIFO）缓冲器，并且接收缓冲器和发送缓冲器均在逻辑层的外部。发送缓冲器改变从该层逻辑传送到传输媒体的信号的频率，而接收缓冲器处理从传输媒体传送到逻辑层的信号。“处理器”可以是主机通道适配器、目标通道适配器或网络的互连交换机。

对于本发明，接收缓冲器执行以前由图 2 所示结构中的 FIFO 缓冲器 251 和 FIFO 缓冲器 261 执行的功能。因此，本发明减少了核心逻辑 210 内的缓冲器的数量。减少核心逻辑 210 内的缓冲器的数量就降低了功耗，提高了处理速度并且减小了核心逻辑 210 占据的芯片面积（例如，脚印）。

将频率校正处理和频率调整处理集成到输入接收弹性 FIFO 220 还可以使上部层逻辑 250 的时钟频率高于与其相连的外部部件的时钟频率。因此，与图 2 所示的结构相比，本发明将时钟域变换过程转移到下级逻辑。

附图的简要说明

根据以下参考附图详细说明本发明的优选实施例，可以更好地理解本发明的上述以及其他目的、方面和优点，附图包括:

图 1a 是其中优先采用了本发明的用于进行数据传输的典型 InfiniBand 网络的原理图;

图 1b 是具有接口部件的 InfiniBand 网络的一部分;

图 2 是用于在 ASIC 与传输媒体之间实现传输的核心的原理图;

图 3 是用于在 ASIC 与传输媒体之间实现传输的核心的原理图; 以及

图 4 是部分图 3 所示核心逻辑的更详细原理图。

优选实施例的详细说明

如上所述，需要减少核心逻辑 210 内的缓冲器的数量。通过组合缓冲器 251、261 的操作并从上部链路层逻辑 250 中去除缓冲器 251、253，图 3 所示的本发明第一实施例减少核心 210 内的缓冲器的数量。更具体地说，如图 3 所示，弹性缓冲器 220、230 位于上部链路层逻辑 250 与串行化部分 270 之间。从图 3 所示的结构中去除了频率校正部分 260（如图 2 所示）。

现在，接收弹性 FIFO 缓冲器 220 起频率校正部分 260 的作用，并对沿传输媒体 280 可能出现的频偏进行校正。然而，FIFO 缓冲器 220 还将信号的频率调整到 ASIC 240 要求的频率，这就是图 2 所示 FIFO 缓冲器 251 单独实现的功能。

因此，FIFO 缓冲器 220 执行以前由图 2 所示 FIFO 缓冲器 251 和 261 执行的功能，因此减少了核心逻辑 210 内的缓冲器的数量。减少核心逻辑 210 内的缓冲器的数量就降低了功耗，提高了处理速度并且减小了核心逻辑 210 占据的芯片面积。弹性发送 FIFO 缓冲器 230 所起的作用与图 2 所示发送 FIFO 253 所起的作用类似。

将频率校正处理和频率调整处理集成到输入接收弹性 FIFO 220 还可以使上部层逻辑 250 的时钟频率高于与其相连的外部部件的时钟频率。例如，上部层逻辑部分 250 可以具有比 250 MHz 高的速度，而缓冲器 220、230 以及串行化部分 270 可以以约 250 MHz 工作（与图 2 所示的网络相比，图 3 所示的网络将时钟域变换过程转移到下级逻辑）。

如上所述，InfiniBand 网络内的某个硬件具有以不同速度工作的部件，因为采用了不同标准。例如，以 250 MHz 工作的、InfiniBand 网络内的某些设备必须与诸如以 312 MHz 工作的、基于“光纤通道”的部件的非 InfiniBand 接口部件通信。本发明可以解决这些不同的速度差别。通过将用于进行时钟域变换的时钟补偿

FIFO 251 与由 I/O 部件的下级接收逻辑部分使用的、本发明的弹性 FIFO 220 中的频率校正 FIFO 251 集成在一起，本发明通过缩短数据通过该设备的等待时间提高网络性能。

现在，参考图 4，图 4 示出核心 210 的设计的更详细原理图。为了使传输媒体 280（通过并行－串行高速物理层）与上部层逻辑 250 之间具有不同的时钟速度，通过字节带（byte striped）串行发送通路 200、各自又通过串行化/去串行化（TX SERDES）转换器 227，发送数据。将用于对上部发送层逻辑 250 调步的逻辑控制器电路系统引入其内，以防止 FIFO 发生溢出。逻辑控制器检测弹性 FIFO 缓冲器 220、230 何时几乎要满，然后中断上部层逻辑 250 的时钟（暂停数据流），以在弹性 FIFO 220、230 几乎要满时，防止过多的数据流入这些弹性 FIFO 220、230。

本技术领域内的普通技术人员众所周知，这种弹性 FIFO 缓冲器 220、230 分别具有多个将数据连续输入到其上的存储器位置。弹性 FIFO 是本发明采用的 FIFO 的优选形式，因为它们可以容许不同数量的数据（例如，是可扩充的）。作为一种选择，可以采用常规的 FIFO（例如，非弹性），但是存在局限性，因为在任何瞬间，它们内只能含有固定数量的数据。以与输入数据时同样的连续顺序，从 FIFO 输出数据。

此外，众所周知，还对输入加以控制，以指示 FIFO 缓冲器闩锁当前输入，并将当前输入存储到下一个存储器位置，并且对输出加以控制，以指示 FIFO 缓冲器对输出展示下一个存储器位置。还存在设备 220、230 发出的、关于该设备内当前存在多少数据的指示。从该设备内删除数据的频率不必与将数据放入该设备的频率相关，这样允许 FIFO 转换信号频率。然而，用于控制该设备的逻辑必须控制它，以避免在该设备内没有数据时，指示该输出前进到下一个项目，并且避免在该设备充满数据时，指示该输入将数据存入下一个项目。

为了实现上述功能，弹性 FIFO 220、230 针对每个 FIFO 通路包括对于数据字节信号 211、FIFO 满指示 212、数据选通信号 213 以

及上部层时钟信号 214 的连接。此外，数据字节输出信号 216、数据取得选通取得信号 217 以及媒体时钟信号 218 用于数据信号传输控制。

FIFO 230 使用 data_byte_get_strobe 信号 217 被断言的 data_byte_out_clk 信号 218 的每个闩锁边缘，以释放 FIFO 中的项目，并将该项目内的数据送到该 FIFO 的输出端。该 FIFO 230 使用 data_byte_put_strobe 信号 213 被断言的 data_byte_in_clk 信号 214 的每个闩锁边缘以将项目送入该 FIFO 内。FIFO 利用 data_count 指出 FIFO 内当前有多少数据。在插入或者删除数据时，更新该值。上部层逻辑部分 250 利用 data_count 输出监测 FIFO 的状态。如果 FIFO 内的所有项目已经被使用，则上部层逻辑将重新断言 data_byte_put_strobe 信号 213，直到 data_count 值指出存在可用项目。在采用上述操作时，上部层逻辑部分 210 可以以较高频率工作，并实现时钟域变换。

如上所述，对于本发明，FIFO 缓冲器 220 执行以前由图 2 所示结构内的 FIFO 缓冲器 251 和 261 单独执行的功能，从而减少了核心逻辑 210 内的缓冲器的数量。减少核心逻辑 210 内的缓冲器的数量就降低了功耗，提高了处理速度并且减小了核心逻辑 210 占据的芯片面积。将频率校正处理和频率调整处理集成到输入接收弹性 FIFO 220 还可以使上部层逻辑 250 的时钟频率高于与其相连的外部部件的时钟频率（例如，上部层逻辑部分 250 可以具有高于 250 MHz 的速度，而缓冲器 220、230 以及串行化单元 270 可以以约 250 MHz 工作）。因此，与图 2 所示的结构相比，本发明将时钟域变换过程转移到下级逻辑。此外，尽管图 3 示出本发明的优选结构，但是如规定的运行模式所要求的，本发明可以单独用作数据输入或数据输出处理。

本发明还允许与弹性 FIFO 220、230 一起使用的时钟设备具有较低精度（并且成本较低）。更具体地说，上部链路层逻辑 250 内的设备要求时钟信号具有非常高的精度。通过从上部链路层逻辑 250 内去除缓冲器 220、230，本发明降低了核心逻辑 210 对高精度时钟信

号的要求。通过允许将较低精度的时钟信号送到 FIFO 缓冲器 220、230，本发明降低了核心逻辑 210 的成本，因为本发明允许将较低精度、较低成本的时钟信号产生设备用于缓冲器 220、230。相反，图 2 所示的 FIFO 缓冲器 251、253 则要求更高成本、更高精度的时钟信号产生设备。

因此，通过减少核心逻辑 210 内的 FIFO 缓冲器的数量，并且还通过去除上部链路层逻辑 250 内的缓冲器，本发明在许多方面得到节省。本发明实现具有较高处理速度、较小脚印并且比先前结构廉价的核心。

尽管根据优选实施例对本发明进行了说明，但是本技术领域内的熟练技术人员明白，利用所附权利要求所述的本发明实质范围内的变换例也可以实现本发明。

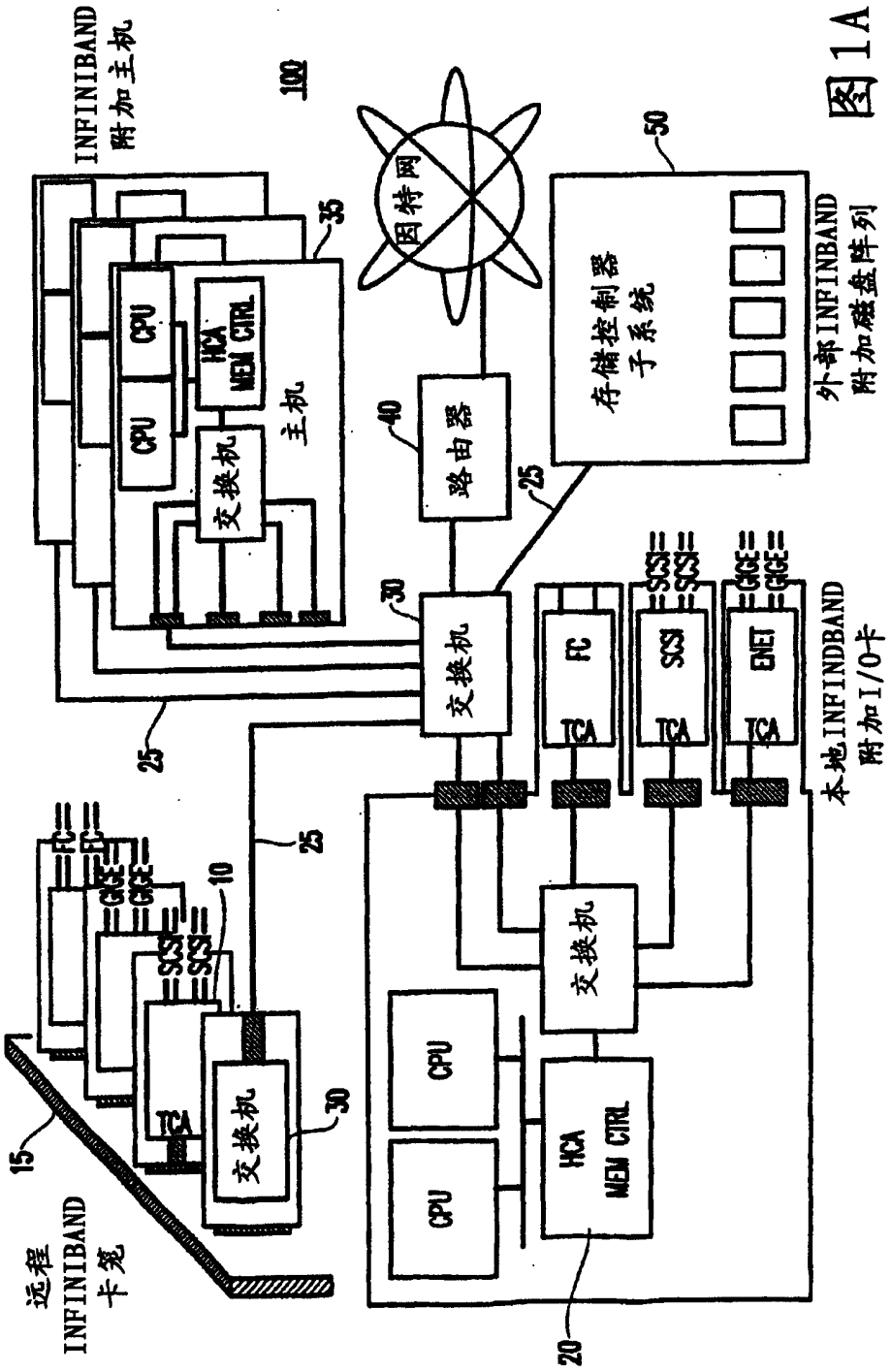

INFINIBAND
附加主机
100
35
CPU
CPU
HCA
MEM CTRL
交换机
主机
因特网
50
存储控制器
子系统
外部INFINBAND
附加磁盘阵列
40
路由器
25
30
交换机
FC
SCSI
ENET
本地INFINDBAND
附加I/O卡
10
15
远程
INFINIBAND
卡笼
CPU
CPU
HCA
MEM CTRL
20

图 1A

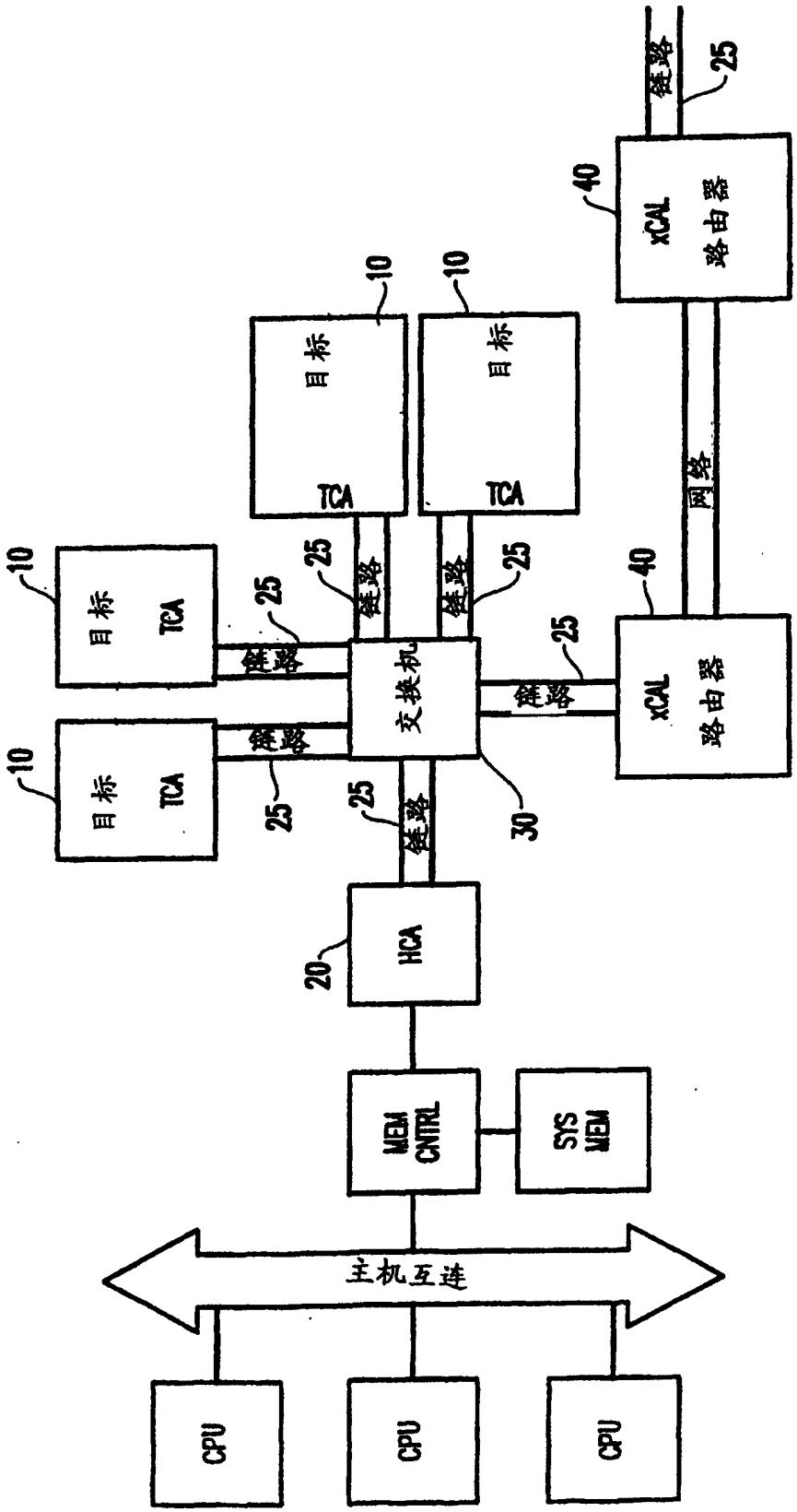

链路
25
40
xCAL
路由器
网络
10
目标
TCA
交换机
30
HCA
20
MEM
CNTRL
SYS
MEM
主机互连
CPU

图1B

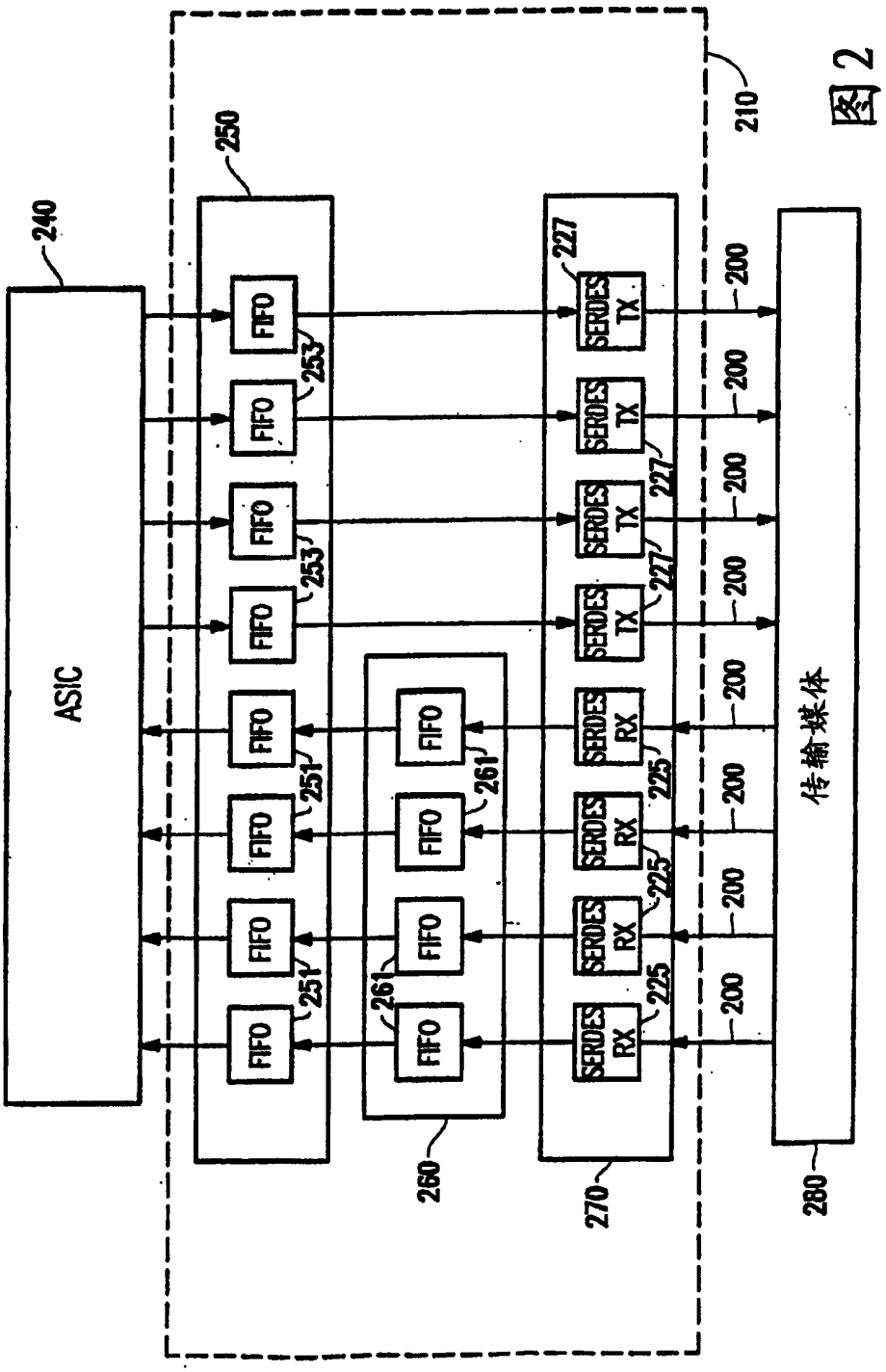

ASIC
240
250
210
FIFO
253
251
261
260
270
SERDES TX
SERDES RX
227
225
200
传输媒体
280

图 2

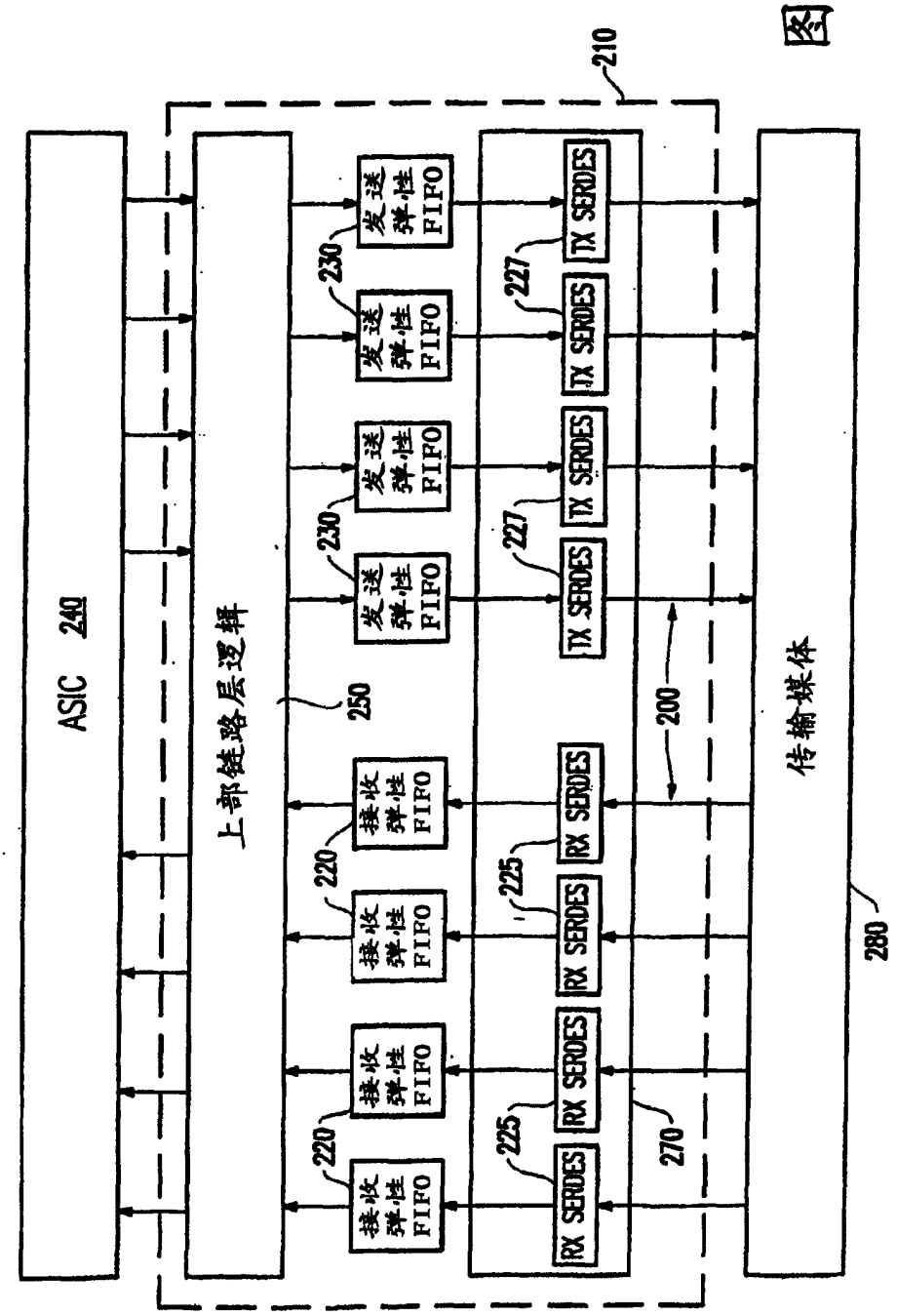
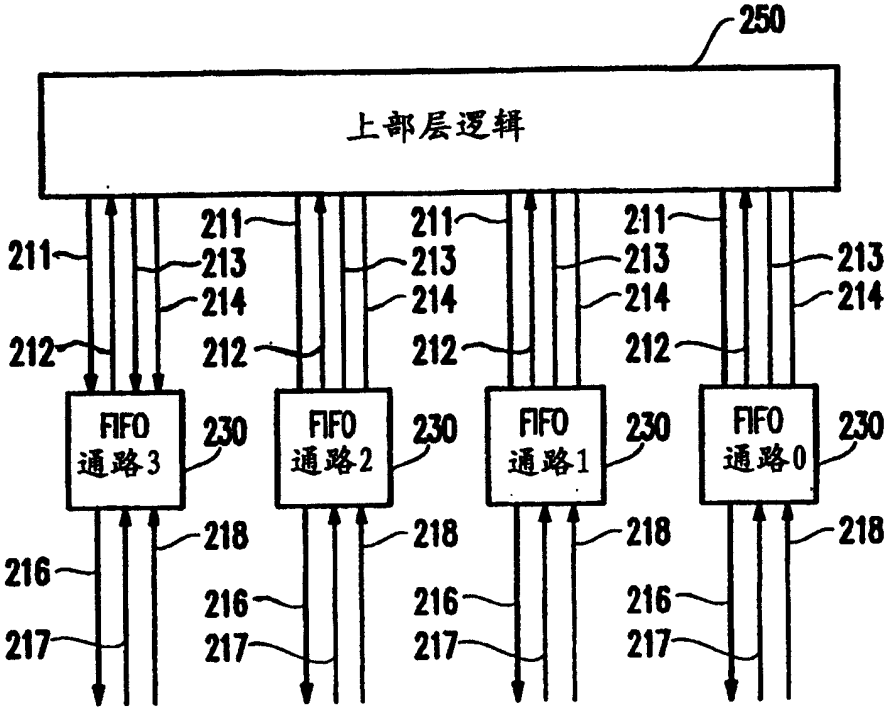

ASIC 240
上部链路层逻辑
250
210
发送弹性FIFO
230
TX SERDES
227
接收弹性FIFO
220
RX SERDES
225
200
270
传输媒体
280

图 3

图4


250
上部层逻辑
211
213
214
212
FIFO
通路3
FIFO
通路2
FIFO
通路1
FIFO
通路0
230
218
216
217