[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
H04L 29/02 (2006.01)
H04L 12/58 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200710093626.1

[43] 公开日 2008 年 1 月 9 日

[11] 公开号 CN 101102305A

[22] 申请日 2007.3.26
[21] 申请号 200710093626.1
[30] 优先权
[32] 2006. 3.31 [33] US [31] 60/788,396
[32] 2006. 5.16 [33] US [31] 11/435,075
[71] 申请人 美国博通公司
地址 美国加州尔湾市奥尔顿公园路 16215 号
[72] 发明人 范 勘

[74] 专利代理机构 深圳市顺天达专利商标代理有限公司
代理人 蔡晓红 李 琴

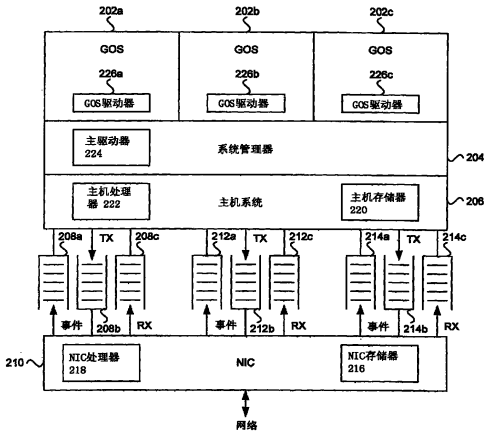
权利要求书 2 页 说明书 16 页 附图 8 页

[54] 发明名称

管理网络信息处理的系统和方法

[57] 摘要

本发明提供了一种操作系统虚拟认知网络接口卡(NIC)的方法和系统。 NIC 可为主机系统中多个同时运行的客户操作系统(GOS)中的每个提供直接 I/O 功能。 所述 NIC 包括针对每个 GOS 的 GOS 队列，其中每个 GOS 队列包括发送(TX)队列、接收(RX)队列和事件队列。 所述 NIC 可通过对应的 TX 队列和 RX 队列与 GOS 传输数据。 所述 NIC 可通过对应的事件队列通知 GOS 发生了事件，例如，下行链路、上行链路、数据包发送和数据包接收。 所述 NIC 还可支持 GOS 之间的单播、广播和/或多播通信。 如果地址对应主机系统中运行的 GOS 之一，所述 NIC 也可验证缓存的地址。



1、一种管理网络信息处理的方法，其特征在于，所述方法包括：通过网络接口卡内的多个客户操作系统队列中对应的至少一个队列，与主机系统内同时运行的多个客户操作系统中的每一个客户操作系统传输数据。

2、根据权利要求 1 所述的方法，其特征在于，所述多个客户操作系统队列中的每个队列包括发送队列、接收队列和事件队列中至少其一。

3、根据权利要求 2 所述的方法，其特征在于，所述方法进一步包括：通过所述对应的客户操作系统队列中的发送队列，从所述网络接口卡传输数据至所述多个同时运行的客户操作系统之一。

4、根据权利要求 2 所述的方法，其特征在于，所述方法进一步包括：通过所述对应的客户操作系统队列中的接收队列，从所述多个同时运行的客户操作系统之一传输数据至所述网络接口卡。

5、根据权利要求 2 所述的方法，其特征在于，所述方法进一步包括：通过所述对应的客户操作系统队列中的事件队列，将通信事件的发生从所述网络接口卡指示给所述多个同时运行的客户操作系统之一。

6、一种机器可读存储器，其特征在于，其内存储的计算机程序具有至少一个用于管理网络信息处理的代码段，所述至少一个代码段由机器执行后使得所述机器执行如下步骤：通过网络接口卡内的多个客户操作系统队列中对应的至少一个队列，与主机系统内同时运行的多个客户操作系统中的每一个客户操作系统传输数据。

7、根据权利要求 6 所述的机器可读存储器，其特征在于，所述多个客户操作系统队列中的每个队列包括发送队列、接收队列和事件队列中至少其一。

8、一种管理网络信息处理的系统，其特征在于，所述系统包括网络接口卡中的电路，允许所述网络接口卡与主机系统内同时运行的多个客户操作系统中的每一个客户操作系统之间，通过所述网络接口卡内的多个客户操作系统队列中对应的至少一个队列传输数据。

9、根据权利要求 8 所述的系统，其特征在于，所述多个客户操作系统队

列中的每个队列包括发送队列、接收队列和事件队列中至少其一。

10、根据权利要求 9 所述的系统，其特征在于，所述系统进一步包括有位于所述网络接口卡内的、用于通过所述对应的客户操作系统队列中的发送队列从所述网络接口卡传输数据至所述多个同时运行的客户操作系统之一的电路。

管理网络信息处理的系统和方法

技术领域

本发明涉及 TCP（传输控制协议）数据和相关的 TCP 信息的处理，更具体地说，涉及一种操作系统虚拟认知网络接口卡（virtualization-aware network interface card）系统和方法。

背景技术

在网络系统中，一个服务器可同时支持多个服务器操作或服务。例如，一个服务器可提供对商业应用程序的访问，并可同时用作电子邮件服务器、数据库服务器和/或交换服务器。服务器通过使用一个操作系统（OS）支持各种服务器操作。通过一个 OS，服务器操作使用服务器处理资源，例如，中央处理器（CPU）、存储器、网络接口卡（NIC）、外设声卡和/或图形卡。在许多情况中，服务器资源可能不被充分地使用，因为服务器操作的需求基于提供的服务和/或用户的需要而变化。将服务器的服务合并为一个操作可改善服务器的效率。然而，合并后操作的安全性也不如分离的操作的安全性高。例如，如果操作被合并，数据库服务器的崩溃或故障可导致电子邮件服务、交换服务和/或应用程序服务的丢失。

另一种改善服务器效率的方法是同时运行多个操作系统，每个操作系统支持不同的服务器操作。该多个操作系统可称作客户操作系统（GOS）。这种方法在服务器操作未被合并的情况下保持安全性级别，并同时优化服务器的处理资源的使用。多客户操作系统的使用又称为 OS 虚拟化，因为每个 GOS 都认为有充分访问服务器的硬件资源。在这点上，GOS 不知道还有其它的 GOS 在同一服务器上运行。为了实现 OS 虚拟化，需要有软件层来仲裁对服务器硬件资源的访问。例如，所述软件层可以是系统管理器（hypervisor）或虚拟机（VM）监视器。该系统管理器可允许多个 GOS 以分时（time-sharing）的方式访问硬

件资源。

NIC（网络接口卡）可一种被至少一个服务器操作或服务频繁使用的硬件。在这点上，所述系统管理器或 VM 监视器可允许创建 GOS 所使用的 NIC 的软件表示。所述 NIC 的软件表示可称作“虚拟 NIC”。然而，虚拟 NIC 在提供给 GOS 的 NIC 的功能或特性上受到限制。例如，虚拟 NIC 不能支持传输控制协议（TCP）卸载功能。从 CPU 卸载至少一部分 TCP 处理至 NIC 上的处理器可提高网络吞吐量。没有卸载功能，服务器的数据传输率在某些情况下可能受限制。另一个限制是，虚拟 NIC 仅可对多个 GOS 提供基础层 2（L2）网络功能。虚拟 NIC 在提供 GOS 和网络之间的数据通信方面也受到限制。在这点上，虚拟 NIC 不能支持其它高级特征，例如远程直接存储器访问（RDMA）和/或互联网小型计算机系统接口（ISCSI）。

除了虚拟 NIC 可提供的特性的限制之外，在管理多个 GOS 和服务器硬件资源之间的交互时，还要求系统管理器处理大量的工作。使用系统管理器层会引入更多的开销至基础 L2 网络操作，而在使用一个 OS 时，不会出现所述开销。例如，在从任意 GOS 传输数据时，需要系统管理器仲裁对 NIC 的访问。在 NIC 接收到数据时，需要系统管理器确定是哪个恰当的 GOS 发送该接收的数据的。此外，因为每个 GOS 和该系统管理器会使用存储器的不同部分，系统管理器还具有管理数据从存储器的一个部分传输至另一部分的附加责任。

随着在一个服务器内支持多个 GOS 的需求增加，需要有新的解决方案来降低系统管理器引入的开销，和/或通过一部分服务器硬件资源例如网络接口卡的虚拟化来支持高级特性，。

比较本发明后续将要结合附图介绍的系统，现有技术的其它局限性和弊端对于本领域的普通技术人员来说是显而易见的。

发明内容

本发明提供一种操作系统（OS）虚拟认知网络接口卡系统和/或方法，在后续部分给出充分的展示和/或结合至少一个附图进行了描述，并在权利要求中对技术方案进行了完整的记载。

根据本发明的一个方面，提供一种管理网络信息处理的方法，所述方法包括：通过网络接口卡内的多个客户操作系统（GOS）队列中对应的至少一个队列，与主机系统内同时运行的多个客户操作系统中的每一个客户操作系统传输数据。

优选地，所述多个客户操作系统队列中的每个队列包括发送（TX）队列、接收（RX）队列和事件队列中至少其一。

优选地，所述方法进一步包括：通过所述对应的客户操作系统队列中的发送队列，从所述网络接口卡传输数据至所述多个同时运行的客户操作系统之一。

优选地，所述方法进一步包括：通过所述对应的客户操作系统队列中的接收队列，从所述多个同时运行的客户操作系统之一传输数据至所述网络接口卡。

优选地，所述方法进一步包括：通过所述对应的客户操作系统队列中的事件队列，将通信事件的发生从所述网络接口卡指示给所述多个同时运行的客户操作系统之一。

优选地，所述通信事件是上行链路事件、下行链路事件、数据包发送事件和数据包接收事件之一。

优选地，所述方法进一步包括：通过所述网络接口卡，在所述多个同时运行的客户操作系统中至少两个客户操作系统之间传输数据。

优选地，所述方法进一步包括：验证缓存在所述网络接口卡中的客户操作系统媒体访问控制（MAC）地址。

根据本发明的一个方面，提供一种机器可读存储器，其内存储的计算机程序具有至少一个用于管理网络信息处理的代码段，所述至少一个代码段由机器执行后使得所述机器执行如下步骤：通过网络接口卡内的多个客户操作系统（GOS）队列中对应的至少一个队列，与主机系统内同时运行的多个客户操作系统中的每一个客户操作系统传输数据。

优选地，所述多个客户操作系统队列中的每个队列包括发送（TX）队列、

接收（RX）队列和事件队列中至少其一。

优选地，所述机器可读存储器进一步包括有用于通过所述对应的客户操作系统队列中的发送队列，从所述网络接口卡传输数据至所述多个同时运行的客户操作系统之一的代码。

优选地，所述机器可读存储器进一步包括有用于通过所述对应的客户操作系统队列中的接收队列，从所述多个同时运行的客户操作系统之一传输数据至所述网络接口卡的代码。

优选地，所述机器可读存储器进一步包括有用于通过所述对应的客户操作系统队列中的事件队列，将通信事件的发生从所述网络接口卡指示给所述多个同时运行的客户操作系统之一的代码。

优选地，所述通信事件是上行链路事件、下行链路事件、数据包发送事件和数据包接收事件之一。

优选地，所述机器可读存储器进一步包括有用于通过所述网络接口卡，在所述多个同时运行的客户操作系统中至少两个客户操作系统之间传输数据的代码。

优选地，所述机器可读存储器进一步包括有验证缓存在所述网络接口卡内的客户操作系统媒体访问控制（MAC）地址的编码。

根据本发明的一个方面，提供一种管理网络信息处理的系统，所述系统包括网络接口卡（NIC）中的电路，允许所述网络接口卡与主机系统内同时运行的多个客户操作系统（GOS）中的每一个客户操作系统之间，通过所述网络接口卡内的多个客户操作系统队列中对应的至少一个队列传输数据。

优选地，所述多个客户操作系统队列中的每个队列包括发送（TX）队列、接收（RX）队列和事件队列中至少其一。

优选地，所述系统进一步包括有位于所述网络接口卡内的、用于通过所述对应的客户操作系统队列中的发送队列从所述网络接口卡传输数据至所述多个同时运行的客户操作系统之一的电路。

优选地，所述系统进一步包括有位于所述网络接口卡内的、用于通过所述

对应的客户操作系统队列中的接收队列从所述多个同时运行的客户操作系统之一传输数据至所述网络接口卡的电路。

优选地，所述系统进一步包括有位于所述网络接口卡内的、用于通过所述对应的客户操作系统队列中的事件队列将通信事件的发生从所述网络接口卡指示给所述多个同时运行的客户操作系统之一的电路。

优选地，所述通信事件是上行链路事件、下行链路事件、数据包发送事件和数据包接收事件之一。

优选地，所述系统进一步包括有位于所述网络接口卡内的、用于通过所述网络接口卡在所述多个同时运行的客户操作系统中至少两个客户操作系统之间传输数据的电路。

优选地，所述系统进一步包括有位于所述网络接口卡内的、验证缓存在所述网络接口卡内的客户操作系统媒体访问控制（MAC）地址的电路。

本发明的各种优点、各个方面和创新特征，以及其中所示例的实施例的细节，将在以下的说明书和附图中进行详细介绍。

附图说明

下面将结合附图及实施例对本发明作进一步说明，附图中：

图 1 是本发明通信连接至支持多个客户操作系统（GOS）的主机系统的网络接口卡的模块图；

图 2A 是本发明操作系统（OS）虚拟认知 NIC 的一个实施例的模块图；

图 2B 是本发明 OS 虚拟认知 NIC 的另一实施例的模块图；

图 2C 是本发明通过 OS 虚拟认知 NIC 发送和接收数据包的流程图；

图 2D 是本发明通过 OS 虚拟认知 NIC 发送和接收数据包的过程中 GOS 和主驱动器的操作的流程图；

图 3 是本发明支持统计值采集（statistics collection）的 OS 虚拟认知 NIC 的一个实施例的模块图；

图 4A 是本发明支持主机系统中 GOS 之间通信的第二级（L2）交换的 OS 虚拟认知 NIC 的一个实施例的模块图；

图 4B 是本发明通过 OS 虚拟认知 NIC 进行单播、多播和/或广播的步骤流程图。

具体实施方式

本发明的各个实施例涉及一种操作系统（OS）虚拟认知网络接口卡（NIC）的方法和系统。所述系统包括 NIC，为主机系统中多个同时运行的客户操作系统（GOS）提供直接 I/O 功能。所述 NIC 包括有用于每个 GOS 的 GOS 队列，其中每个 GOS 队列包括发送（TX）队列、接收（RX）队列和事件队列。所述 NIC 可与 GOS 通过对应的 TX 队列和 RX 队列传输数据。所述 NIC 可通过对应的事件队列，通知 GOS 有事件发生，例如下行链路、上行链路、数据包发送和数据包接收。所述 NIC 也可支持 GOS 之间的单播、广播和/或多播通信。当有地址与主机系统中的 GOS 之一相对应时，所述 NIC 也可验证该被缓存的地址。

图 1 是本发明通信连接至支持多个客户操作系统（GOS）的主机系统的网络接口卡的模块图。参照图 1，展示了第一 GOS 102a、第二 GOS 102b、第三 GOS 102c、系统管理器 104、主机系统 106、发送（TX）队列 108a、接收（RX）队列 108b 和 NIC 110。NIC 110 包括 NIC 处理器 118 和 NIC 存储器 116。主机系统 106 包括主机处理器 122 和主机存储器 120。

主机系统 106 可包括恰当的逻辑、电路和/或编码，例如，可进行数据处理和/或网络操作。在某些例子中，主机系统 106 还包括有其它硬件资源，例如，图形卡和/或外设声卡。主机系统 106 可通过系统管理器 104 支持第一 GOS 102a、第二 GOS 102b 和第三 GOS 102c 的操作。主机系统 106 通过使用系统管理器 104 支持的 GOS 的数量不限于图 1 中的实施例所示。例如，主机系统 106 可支持两个或多个 GOS。

系统管理器 104 可用作实现主机系统 106 中硬件资源 OS 虚拟化和/或通信连接至主机系统 106 的硬件资源虚拟化的软件层，例如，NIC 110。系统管理器 104 也可实现 GOS 和主机系统 106 中的硬件资源和/或连接至主机系统 106 的硬件资源之间的数据传输。例如，系统管理器 204 可实现主机系统 106

所支持的 GOS 和 NIC 110 之间通过 TX 队列 108a 和/或 RX 队列 108b 传输数据包。

主处理器 122 可包括恰当的逻辑、电路和/或编码，可控制和/或管理与主机系统 106 相关的数据处理和/或网络操作。主机存储器 120 包括恰当的逻辑、电路和/或编码，可存储主机系统 106 所使用的数据。主机存储器 120 可被分割为多个存储区。例如，主机系统 106 所支持的每个 GOS 在主机存储器 120 中具有对应的存储区。此外，系统管理器 104 在主机存储器 120 中具有对应的存储区。因此，系统管理器 104 可通过控制数据从对应一个 GOS 的存储器 120 的一部分传输到对应另一个 GOS 的存储器 120 的另一部分，实现 GOS 之间的数据传输。

NIC 110 包括恰当的逻辑、电路和/或编码，可实现与网络的数据传输。例如，NIC 110 可进行基础级 2（L2）交换操作。TX 队列 108a 包括恰当的逻辑、电路和/或编码，可登记（posting）数据以通过 NIC 110 发送。RX 队列 108b 包括恰当的逻辑、电路和/或编码，可登记通过 NIC 110 接收到的数据以供主机系统 106 处理。因而，NIC 110 可登记 RX 队列 108b 中从网络接收的数据，并可获取 TX 队列 108a 中由主机系统 106 登记的数据以发送到该网络。例如，TX 队列 108a 和 RX 队列 108b 可集成在 NIC 110 中。NIC 处理器 118 包括恰当的逻辑、电路和/或编码，可控制和/或管理 NIC 110 中的数据处理和/或网络操作。NIC 存储器 116 包括恰当的逻辑、电路和/或编码，可存储 NIC 110 所使用的数据。

第一 GOS 102a、第二 GOS 102b 和第三 GOS 102 每个均对应一个操作系统，可运行或执行操作或服务，例如，应用程序、电子邮件服务器操作、数据库服务器操作和/或交换服务器操作。第一 GOS 102a 包括虚拟 NIC 112a，第二 GOS 102b 包括虚拟 NIC 112b，第三 GOS 102c 包括虚拟 NIC 112c。例如，虚拟 NIC 112a、虚拟 NIC 112b 和虚拟 NIC 112c 对应于 NIC 110 资源的软件表示。因而，NIC 110 资源包括 TX 队列 108a 和 RX 队列 108b。通过虚拟 NIC 112a、虚拟 NIC 112b 和虚拟 NIC 112c 的 NIC 110 资源的虚拟化，可使得系统管理器 104 提供 NIC 110 所提供的 L2 交换支持给第一 GOS 102a、第二 GOS 102b 和

第三 GOS 102。然而，在这个例子中，通过系统管理器 104 实现的 NIC 110 资源的虚拟化，可能不支持其它高级功能，例如，GOS 中的 TCP 卸载、iSCSI 和/或 RDMA。

操作中，当图 1A 中所示的 GOS 需要发送数据包至网络时，该数据包传输可至少部分地由系统管理器 104 控制。如果不止一个 GOS 需要发送数据包至网络，系统管理器 104 对访问 NIC 110 资源进行仲裁。在这点上，作为仲裁结果，系统管理器 104 可使用虚拟 NIC 将 NIC 110 传输资源的当前可用性通知给对应的 GOS。系统管理器 104 可依据仲裁操作的结果将数据包登入 TX 队列 108a 内，从而协调 GOS 的数据包传输。数据包传输所发生的仲裁和/或协调操作将增加系统管理器 104 的开销。

通过 NIC 110 从网络接收数据包时，系统管理器 104 确定与该数据包相关联的媒体访问控制（MAC）地址，以便将接收到的数据包传送给恰当的 GOS。在这点上，系统管理器 104 可从 RX 队列 108b 接收数据包，并对该数据包解多路复用，以传输至恰当的 GOS。在为接收的数据包确定 MAC 地址和恰当的 GOS 后，系统管理器 104 将接收的数据包从主机存储器 120 的系统管理器区内的缓存中传输到主机存储器 120 的对应恰当 GOS 的存储区内的缓存中。与接收数据包和传输数据包至恰当的 GOS 相关的操作也会增加系统管理器 104 的开销。

图 2A 是本发明操作系统（OS）虚拟认知 NIC 的一个实施例的模块图。参照图 2A，展示了第一 GOS 202a、第二 GOS 202b、第三 GOS 202c、系统管理器 204、主机系统 206、事件队列 208a、212a 和 214a、发送（TX）队列 208b、212b 和 214b、接收（RX）队列 208c、212c 和 214c，以及 NIC 210。NIC 210 包括 NIC 处理器 218 和 NIC 存储器 216。主机系统 206 包括主机处理器 222 和主机存储器 220。系统管理器 204 包括主驱动器 224。

主机系统 206 包括恰当的逻辑、电路和/或编码，可进行数据处理和/或网络操作。在某些例子中，主机系统 206 也可包括其它硬件资源，例如，图形卡和/或外设声卡。主机系统 206 通过系统管理器 204 支持第一 GOS 202a、第二 GOS 202b 和第三 GOS 202c 的操作。第一 GOS 202a、第二 GOS 202b 和第三

GOS 202 每个均对应于可运行或执行操作或服务的操作系统，例如，所述操作或服务可以是应用程序、电子邮件服务器操作、数据库服务器操作和/或交换服务器操作。主机系统 206 通过使用系统管理器 104 所支持的 GOS 的数量不限于图 2A 中描述的实施例。例如，主机系统 206 可支持两个或多个 GOS。

系统管理器 204 可运行为软件层，实现主机系统 206 内的硬件资源的虚拟化和/或通信连接至主机系统 206 的硬件资源的虚拟化，例如，NIC 210。系统管理器 204 也可实现 GOS 和主机系统 206 中硬件资源和/或连接至主机系统 206 的硬件资源之间的数据通信。例如，系统管理器 204 可通过事件队列 208a、212a 和 214a、TX 队列 208b、212b 和 214b 和/或 RX 队列 208c、212c 和 214c 实现主机系统 206 支持的 GOS 和 NIC 210 之间的通信。在这点上，第一 GOS 202a 和 NIC 210 之间的通信可通过事件队列 208a、TX 队列 208b 和 RX 队列 208c 发生。同样地，第二 GOS 202b 和 NIC 210 之间的通信可通过事件队列 212a、TX 队列 212b 和 RX 队列 212c 发生。第三 GOS 202c 和 NIC 210 之间的通信可通过事件队列 214a、TX 队列 214b 和 RX 队列 214c 发生。每组队列彼此之间单独和独立地运行。

系统管理器 204 包括主驱动器 224，其可协调 GOS 和队列之间的数据传输。主驱动器 224 可与 GOS 202a 中的 GOS 驱动器 226a、GOS 202b 中的 GOS 驱动器 226b 和/或 GOS 202c 中的 GOS 驱动器 226c 通信。每个 GOS 驱动器对应于一部分 GOS，通过主驱动器 224 进行 GOS 所执行的操作或服务与恰当的队列之间的数据传输。例如，来自第一 GOS 202a 中的操作或服务传输的数据包和/或数据包描述符可通过 GOS 驱动器 226a 传输至 TX 队列 208b。在另一个例子中，由 NIC 210 登入事件队列 208a 中以指示网络条件或报告数据发送或数据接收的数据，将被传输至由 GOS 驱动器 226a 登记的缓存中。在另一个例子中，由 NIC 210 从网络接收的、具有对应第一 GOS 202a 的 MAC 地址的数据包，可从 RX 队列 208c 传输至由 GOS 驱动器 226a 登记的缓存中。

主机处理器 222 包括恰当的逻辑、电路和/或编码，可控制和/或管理与主机系统 206 相关的数据处理和/或网络操作。主机存储器 220 包括恰当的逻辑、电路和/或编码，可存储主机系统 206 所使用的数据。主机存储器 220 可被分

割为多个存储区。例如，主机系统 206 所支持的每个 GOS 在主机存储器 220 中具有对应的存储区。此外，系统管理器 204 在主机存储器 220 中具有对应的存储区。因而，系统管理器 204 可通过控制数据从对应一个 GOS 的存储器 220 的存储区传输至对应另一个 GOS 的存储器 220 的另一存储区，来实现 GOS 之间的数据传输。

NIC 210 包括恰当的逻辑、电路和/或编码，可实现与网络传输数据。NIC 210 可实现基础 L2 交换、TCP 卸载、iSCSI 和/或 RDMA 操作。NIC 210 可称为 OS 虚拟认知 NIC，因为与每个 GOS 的通信通过独立的队列组完成。NIC 210 可确定所接收的数据包的 MAC 地址，并可将接收的数据包发送给与具有恰当 MAC 地址的 GOS 相对应的 RX 队列。同样地，NIC 210 可通过协调和/或仲裁 TX 队列中登记的数据包被发送的顺序，来实现从 GOS 到网络的数据包传输。在这点上，NIC 210 可实现直接输入/输出（I/O）或系统管理器旁路操作。

事件队列 208a、212a 和 214a 包括恰当的逻辑、电路和/或编码，可通过 NIC 210 登入数据以表示事件的发生。例如，NIC 210 可在事件队列中登入数据以表示下行链路或上行链路。链路的当前状态，无论是上行的还是下行的，都将登记给所有事件队列。

TX 队列 208b、212b 和 214b 包括恰当的逻辑、电路和/或编码，可通过 NIC 110 从第一 GOS 202a、第二 GOS 202b 和第三 GOS 202c 登记数据。RX 队列 208c、212c 和 214c 包括恰当的逻辑、电路和/或编码，可登入通过 NIC 110 接收到的数据以供第一 GOS 202a、第二 GOS 202b 和第三 GOS 202c 处理。TX 队列 208b、212b 和 214b 和/或 RX 队列 208c、212c 和 214c 可集成在 NIC 210 内。

NIC 处理器 218 包括恰当的逻辑、电路和/或编码，可控制和/或管理 NIC210 中的数据处理和/或网络操作。NIC 存储器 216 包括恰当的逻辑、电路和/或编码，可存储 NIC 210 所使用的数据。

图 2B 是本发明 OS 虚拟认知 NIC 的另一实施例的模块图。参照图 2B，所示的主机系统 206 可支持 N 个 GOS 和一个 NIC 210，该 NIC 210 可支持 N 组队列。主机系统 206 如图 2A 中描述，并可支持 GOS 202_1、…、GOS 202_N

的操作，其中 1≤N。每个 GOS 可用于提供单独的操作或服务。系统管理器 204 和主驱动器 224 可支持 N 个 GOS 和队列组 228_1、…、228_N 之间的数据通信。主机存储器 220 的一部分可与每个 GOS 202_1、…、GOS 202_N 以及系统管理器 204 相关联。图 2B 中展示的 GOS 驱动器 228_1、…、228_N 可用于在 GOS202_1、…、GOS 202_N 执行的操作或服务和对应的队列 228_1、…、228_N 之间传输数据。GOS 驱动器 228_1、…、228_N 和对应的队列 228_1、…、228_N 之间的数据传输可通过主驱动器 224 发生。在这点上，GOS 驱动器和主驱动器 224 如图 2A 中展示。

NIC 210 如图 2A 描述，并也可称作 OS 虚拟认知 NIC。NIC 210 可通过队列组 226_1、…、226_N 实现网络和 N 个 GOS 的每个之间的通信。例如，网络和 GOS 202_1 之间的通信可通过队列组 226_1 发生。在另一个例子中，网络和 GOS 202_N 之间的通信可通过队列组 226_N 发生。每组队列可包括事件队列、发送（TX）队列和接收（RX）队列。队列组 226_1，…，226_N 中的事件队列、TX 队列和 RX 队列如图 2A 所描述。

图 2C 是本发明通过 OS 虚拟认知 NIC 发送和接收数据包的流程图。参照图 2C，展示了流程图 230。起始步骤 232 后，步骤 234 中，当数据包准备好通过图 2A-2B 中的 OS 虚拟认知 NIC 210 从 GSO 传输至网络时，流程图 230 中的处理可进入步骤 236。步骤 236 中，GOS 中的 GOS 驱动器可通过系统管理器 204 中的主驱动器 224 发送将要登入对应 TX 队列中的数据包。步骤 238 中，将数据包从 TX 队列传输至 NIC 210 用于传输。步骤 240 中，NIC 210 可传输该数据包至与网络连接的终端和/或设备。在这点上，NIC 210 可在与发起数据包传输的 GOS 相对应的事件队列中登记一个标识，以此报告数据包已经被传输至网络。步骤 240 后，流程图 230 中的处理进入结束步骤 242。

回到步骤 234，当数据包将通过图 2A-2B 中的 OS 虚拟认知 NIC 210 从网络接收到时，流程图 230 可进入步骤 244。步骤 244 中，NIC 210 可基于为每个数据包确定的 MAC 地址对从网络接收的数据包进行解多路复用。步骤 246 中，NIC 210 可将该数据包登入对应的 RX 队列中，该 RX 队列与所确定的 MAC 地址对应的 GOS 相关联。此外，NIC 210 可在与所确定的 MAC 地址对

应的 GOS 相关联的事件队列中登记一个标识，以此报告已经从网络中接收到数据包。步骤 248 中，将数据包从 RX 队列传输至由对应的 GOS 中的 GOS 驱动器所登记的缓存中。在这点上，该传输可通过系统管理器 204 中的主驱动器 224 发生。步骤 248 后，流程图 230 中的处理进入结束步骤 242。

图 2D 是本发明通过 OS 虚拟认知 NIC 发送和接收数据包的过程中 GOS 和主驱动器的操作流程图。参照图 2D，展示了流程图 252。步骤 254 中，当数据包准备好通过图 2A-2B 中的 OS 虚拟认知 NIC 210 从 GOS 传输至网络时，流程图 230 中的处理可进入步骤 256。步骤 256 中，GOS 中的 GOS 驱动器发送数据包，通过系统管理器 204 中的主驱动器 224 将该数据包登入对应的 TX 队列中。步骤 258 中，将该数据包从 TX 队列传输至 NIC 210 以供传输。NIC 210 可传输数据包至与网络连接的设备和/或终端。在这点上，NIC 210 可在与发起数据包传输的 GOS 相对应的事件队列中登记一个标识，以报告数据包已经传输至网络。步骤 258 后，流程图 250 的处理进入结束步骤 260。

回到步骤 254，当对从网络接收的数据包使用一个中断时，流程图 250 的处理进入步骤 264。步骤 264 中，NIC 210 为收到的数据包确定 MAC 地址，且该数据包可登入对应的 RX 队列中。此外，NIC 210 可生成数据包到达标识并将该标识登记在对应的事件队列中。步骤 266 中，NIC 210 可生成中断信号并将其传送至系统管理器 204 内的主驱动器 224。步骤 268 中，主驱动器 224 通知与收到的数据包的 MAC 地址相对应的 GOS 内的 GOS 驱动器，数据包已经登记在对应的 RX 队列中。对应该合适 GOS 的主机存储器 220 中存储区内的缓存被登入数据包。步骤 270 后，流程图 250 的处理进入结束步骤 260。

回到步骤 262，当多信号中断（MSI）方法用于从网络接收的数据包时，流程图 250 的处理可进入步骤 272。步骤 272 中，NIC 210 确定收到的数据包的 MAC 地址，并且所述数据包被登入对应的 RX 队列中。此外，NIC 210 可产生数据包到达通知，并且可将所述通知登入对应的事件队列中。步骤 274 中，在 NIC 210 和主机系统 206 之间激活 MSI。在这点上，NIC 210 可产生多个中断信号，且该多个中断信号被传输至系统管理器 204 所使用的存储器位置内。NIC 210 可写所述存储器位置以表示特定的 GOS 接收到了数据包。步骤

276 中，在读取了包含有关于多个中断信号的信息的存储器位置后，主驱动器 224 可通知对应的 GOS 驱动器数据包已经到达。步骤 278 中，GOS 驱动器登记一个缓存用于存储 RX 队列中登入的数据包。该缓存可位于与合适的 GOS 相对应的主机存储器 220 内的一个存储区内。步骤 278 后，流程图 250 的处理可进入结束步骤 260。

图 3 是本发明支持统计值采集的 OS 虚拟认知 NIC 的模块图。参照图 3，展示了图 2A-2B 中描述的 NIC 210。在这点上，NIC 210 可包括有存储或存储器缓存，例如，存储器 302a、存储器 302b、存储器 302c 和存储器 304，其中 NIC 210 可存储与网络传输数据包相关的统计信息。例如，标记为存储器 302a、存储器 302b 和存储器 302c 的缓存可基于 NIC 存储器 216 实现，并可存储对应主机系统 206 支持的每个 GOS 的统计信息。例如，存储器 302a 可存储 NIC 210 产生的关于 GOS 202a 的数据包通信的统计信息。例如，存储器 302b 可存储 NIC 210 产生的关于 GOS 202b 的数据包通信的统计信息。例如，GOS 202c 和网络之间的通信有关的统计信息可存储在存储器 302c 内。在本发明的这个实施例中，每个 GOS 的统计信息可存储在单独的缓存中。在本发明的另一个实施例中，所有统计信息可存储在一个缓存中。

缓冲存储器 302a、302b 和 302c 可用于存储每个 GOS 的统计信息，例如，NIC 210 为每个 GOS 接收的正确数据包的数量、接收的数据包中字节的数量和/或正确地传递至每个 GOS 的数据包的数量。所述统计数据可称作“好”统计值（good statistics），并可由 NIC 210 用于通信操作。

缓冲存储器 304 可存储与主机系统 206 所支持的任意 GOS 相对应的关于数据包错误的统计信息。例如，缓存 304 可用于存储统计信息，例如，不符合循环冗余校验（CRC）的数据包和/或长度短于以太网通信规范的数据包。因为这些错误不能使 NIC 210 确定数据包对应的 GOS，NIC 210 可将这些统计值收集并存储在单个缓冲存储器 304 中。这些统计信息可称作“坏”统计信息（bad statistics），并可由 NIC 210 用于通信操作。

图 4A 是本发明支持主机系统中 GOS 之间通信的第二级（L2）交换的 OS 虚拟认知 NIC 的模块图。参照图 4A，展示了图 2A-2B 中描述的 NIC 210。如

图所示，NIC 210 包括 L2 交换机（switch）400。L2 交换机 400 包括恰当的逻辑、电路和/或编码，可使得 NIC 210 支持 GOS 和网络之间和/或 GOS 之间的数据包通信。L2 交换机 400 可支持单播、广播和/或多播操作。单播操作指的是到一个 MAC 地址的数据包传输。广播操作指的是到所有 MAC 地址的数据包传输。多播操作指的是到一组特定 MAC 地址的数据包传输。

例如，图 2A 中的 GOS 202a 可发送数据包至与网络连接的至少一个设备。在这个情况中，GOS 驱动器 226a 可传输数据包至对应于 GOS 202a 的 TX 队列 208b。L2 交换机 400 可接收来自 TX 队列 208b 的该数据包，并可确定 MAC 地址对应于网络上的某个或多个设备。然后 NIC 210 可传输该数据包至对应的 MAC 地址。

又例如，GOS 202a 可发送数据包至 GOS 202b 和/或 GOS 202c。在这个例子中，GOS 驱动器 226a 可传输数据包至对应 GOS 202a 的 TX 队列 208b。L2 交换机 400 从 TX 队列 208b 接收数据包，并可确定 MAC 地址对应于 GOS 202b 和/或 GOS 202c 的地址。L2 交换机 400 可传输该数据包至对应于 GOS 202b 和/或 GOS 202c 的 RX 队列 212c 和/或 RX 队列 214c。GOS 驱动器 226b 和/或 GOS 驱动器 226c 可被通知接收到数据包，并可在主机存储器 220 的恰当存储区内登记缓存。执行在 GOS 202b 和/或 GOS 202c 上的操作或服务可从该被登记的缓存中读取接收到的数据包。

图 4A 中的 NIC 210 还包括有地址验证器 402。地址验证器 402 包括恰当的逻辑、电路和/或编码，可验证由 GOS 驱动器所登记的、用以存储接收的数据包的缓存的地址。例如，在将 RX 队列中的数据包传输至被登记的缓存之前，地址验证器 402 可验证被登记的缓存是位于与接收的数据包相关的 GOS 对应的地址或存储器位置中。如果地址通过验证，则将该接收的数据包从 RX 队列传输至被登记的缓存中。如果地址未通过验证，则 GOS 驱动器需要登记一个新的缓存来接收来自 RX 队列的数据包。

图 4B 是本发明通过 OS 虚拟认知 NIC 进行单播、多播和/或广播的步骤流程图。参照图 4，展示了流程图 410。开始步骤 402 后，在步骤 404 中，图 2A-2B 中描述的主机系统 206 支持的 GOS 可产生数据包以供传输。GOS 驱动器可传

输该数据包至恰当的 TX 队列。图 4A 中的 L2 交换机 400 可从 TX 队列接收数据包，并确定目的地 MAC 地址。步骤 406 中，基于对应数据包目的地的 MAC 地址，L2 交换机 400 可确定数据包传输是单播、广播还是多播。如果该数据包的传输是多播或广播，流程 410 进入步骤 408。

步骤 408 中，L2 交换机 400 可传输数据包至网络上列为多播或广播传输的一部分的恰当的 MAC 地址。步骤 410 中，L2 交换机 400 还传输数据包至具有列为多播或广播传输的一部分的恰当的 MAC 地址的每个 GOS 的 RX 队列。列出的每个 GOS 的 GOS 驱动器可被通知已接收到数据包，并可在主机存储器 220 的恰当存储区中登记缓存。执行在列出的每个 GOS 上的操作或服务可从登记的缓存中读取接收的数据包。步骤 410 后，流程 410 进入结束步骤 418。

回到步骤 406，如果该数据包的传输是单播传输，流程 410 进入步骤 412。步骤 412 中，L2 交换机 400 可确定 MAC 地址是否对应于主机系统 206 支持的 GOS 或者对应于线缆或网络上的设备。如果将被传输的数据包的 MAC 地址对应于某个 GOS，流程 410 进入步骤 414。步骤 414 中，L2 交换机 400 传输该数据包至与具有恰当 MAC 地址的 GOS 相对应的 RX 队列。GOS 驱动器可被通知数据包已接收到，并在主机系统 220 的恰当存储区部分内登记一个缓存。GOS 上执行的操作或服务可从被登记的缓存中读取所接收到的数据包。步骤 414 后，流程 410 进入结束步骤 418。

回到步骤 412，如果将被传输的数据包的 MAC 地址对应网络上的设备，流程 410 进入步骤 416。步骤 416 中，L2 交换机 400 传输数据包至网络上恰当的 MAC 地址。步骤 416 后，流程 410 进入结束步骤 418。

本申请中描述的 OS 虚拟认知 NIC 可实现 OS 虚拟化，降低系统管理器层在 GOS 和网络之间和/或 GOS 之间传输数据包的开销要求。OS 虚拟认知 NIC 可支持多个 GOS。此外，OS 虚拟认知 NIC 可实现高级特性的虚拟化，例如，TCP 卸载功能、RDMA 和/或 iSCSI 接口。

本发明可以通过硬件、软件，或者软、硬件结合来实现。本发明可以在至少一个计算机系统中以集中方式实现，或者由分布在几个互连的计算机系统中

的不同部分以分散方式实现。任何可以实现所述方法的计算机系统或其它设备都是可适用的。常用软硬件的结合可以是安装有计算机程序的通用计算机系统，通过安装和执行所述程序控制计算机系统，使其按所述方法运行。在计算机系统中，利用处理器和存储单元来实现所述方法。

本发明还可以通过计算机程序产品进行实施，所述程序包含能够实现本发明方法的全部特征，当其安装到计算机系统中时，通过运行，可以实现本发明的方法。本文件中的计算机程序所指的是：可以采用任何程序语言、代码或符号编写的一组指令的任何表达式，该指令组使系统具有信息处理能力，以直接实现特定功能，或在进行下述一个或两个步骤之后实现特定功能：a)转换成其它语言、编码或符号；b)以不同的格式再现。

本发明是通过几个具体实施例进行说明的，本领域技术人员应当明白，在不脱离本发明范围的情况下，还可以对本发明进行各种变换及等同替代。另外，针对特定情形或具体情况，可以对本发明做各种修改，而不脱离本发明的范围。因此，本发明不局限于所公开的具体实施例，而应当包括落入本发明权利要求范围内的全部实施方式。

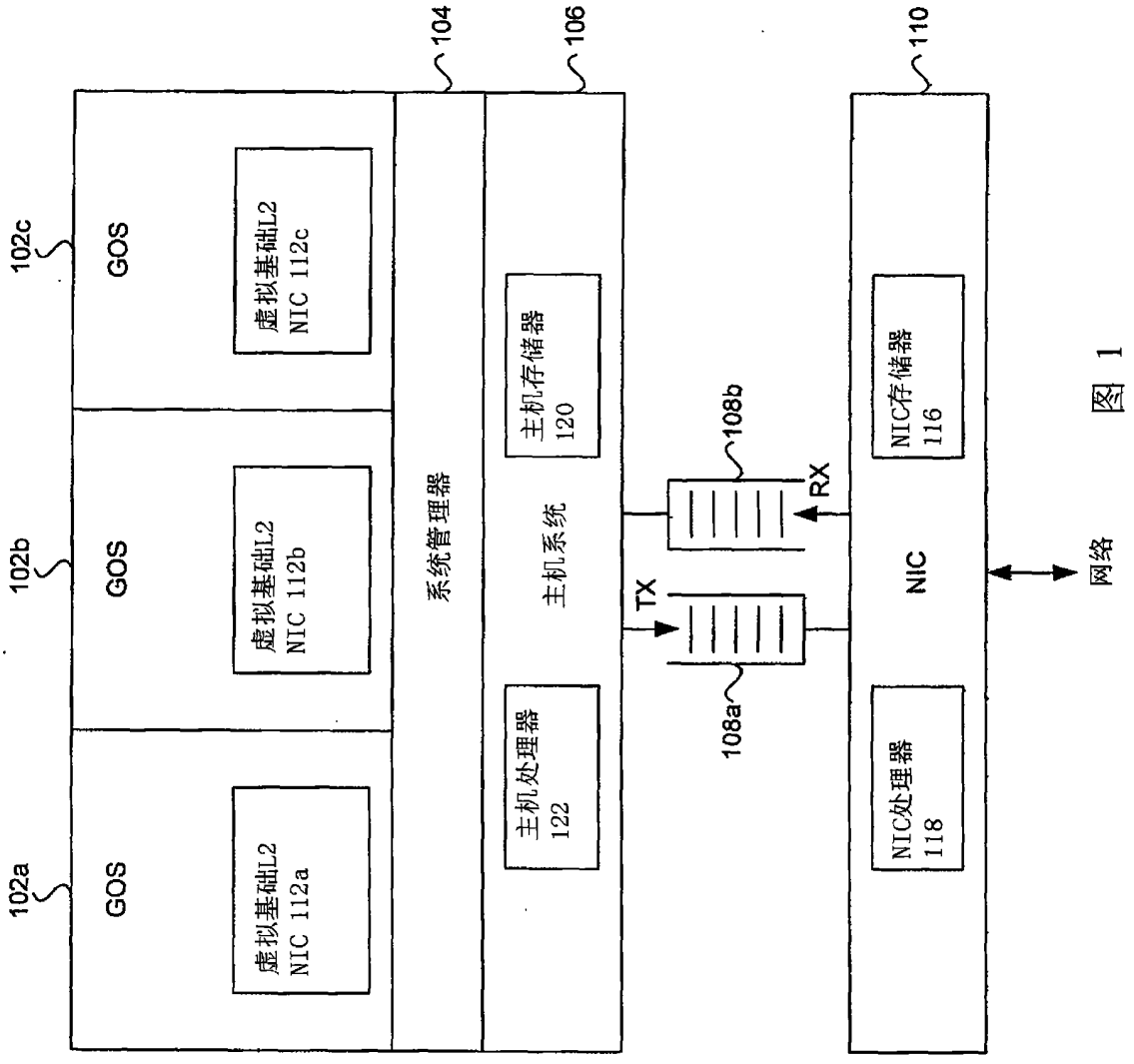

102a
GOS
虚拟基础L2
NIC 112a
102b
GOS
虚拟基础L2
NIC 112b
102c
GOS
虚拟基础L2
NIC 112c
104
系统管理器
106
主机系统
主机处理器
122
主机存储器
120
108a
TX
108b
RX
110
NIC
NIC处理器
118
NIC存储器
116
网络

图 1

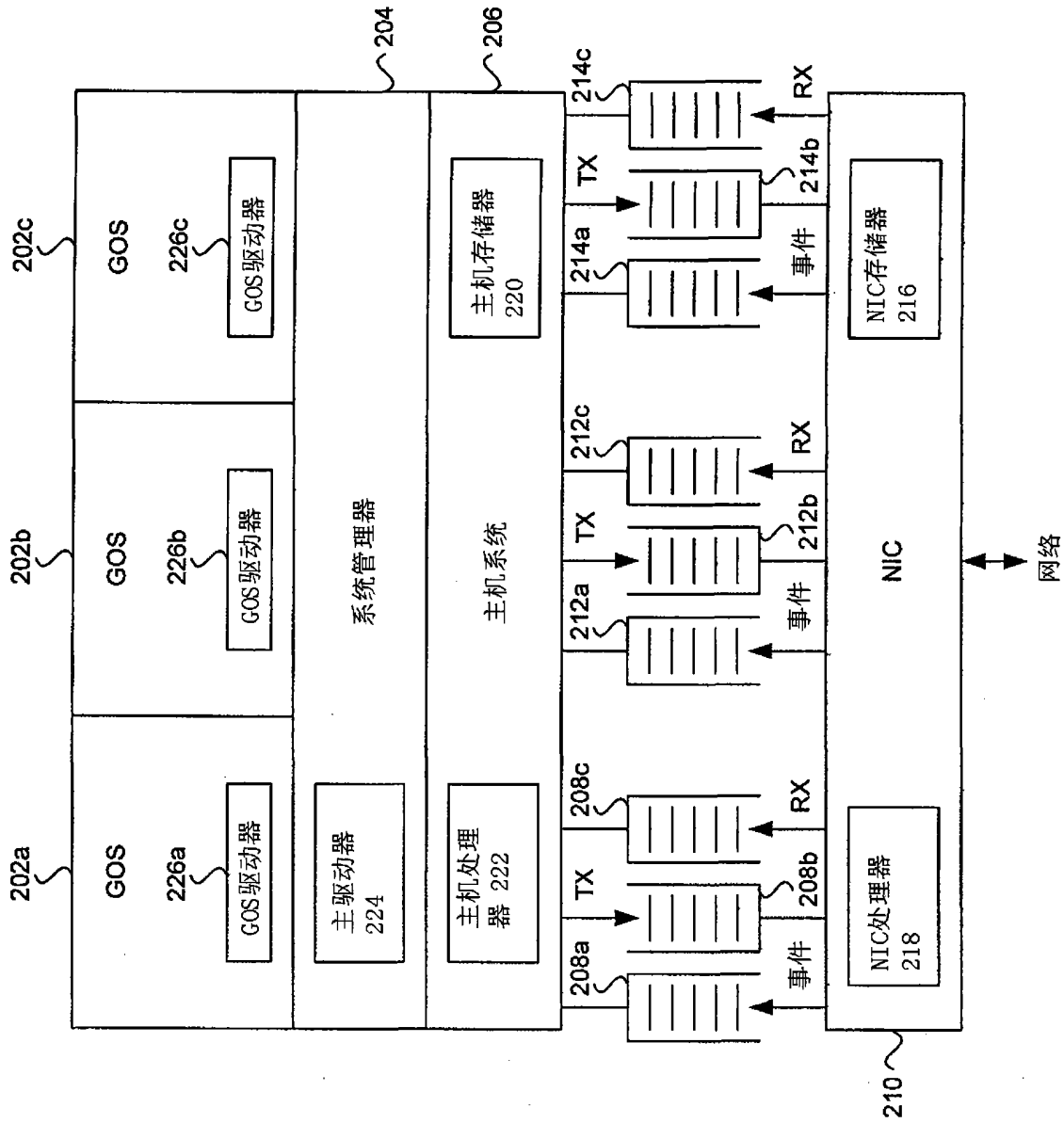

202a
GOS
226a
GOS驱动器
202b
GOS
226b
GOS驱动器
202c
GOS
226c
GOS驱动器
204
系统管理器
主驱动器
224
206
主机系统
主机处理器 222
主机存储器
220
208a
208b
208c
TX
RX
事件
212a
212b
212c
214a
214b
214c
NIC
NIC处理器
218
NIC存储器
216
210
网络

图 2A

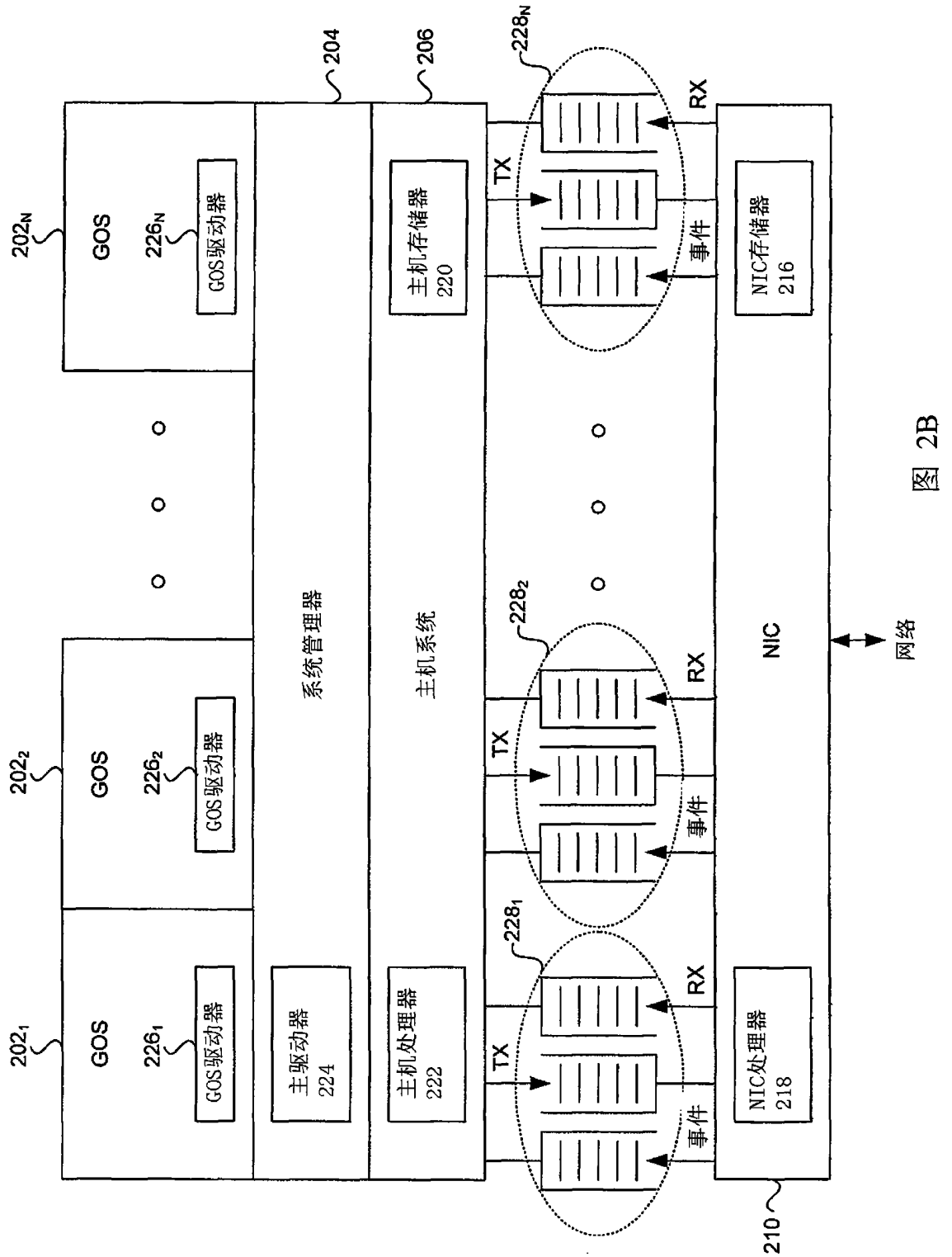

202_N
GOS
226_N
GOS驱动器
202_2
GOS
226_2
GOS驱动器
202_1
GOS
226_1
GOS驱动器
系统管理器
主驱动器
224
204
主机系统
主机处理器
222
主机存储器
220
206
228_N
228_2
228_1
TX
RX
事件
NIC
NIC处理器
218
NIC存储器
216
210
网络

图 2B

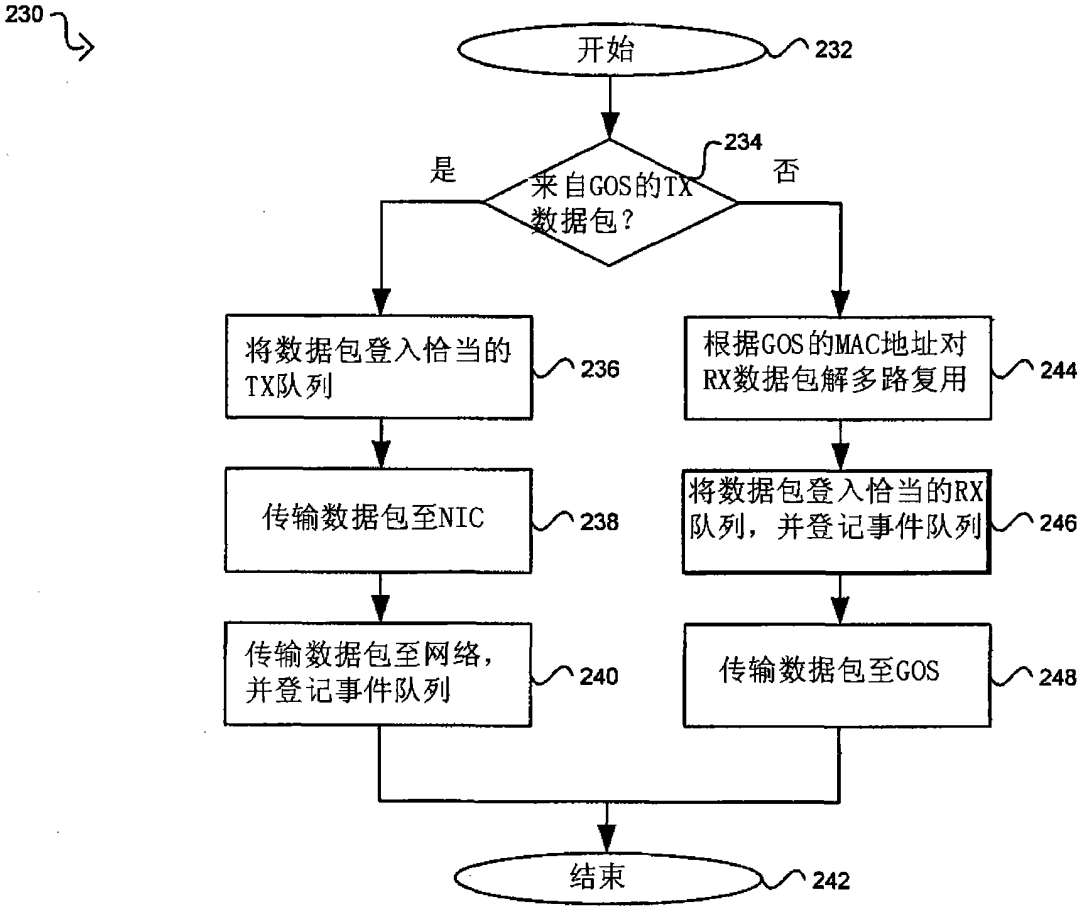

230
开始
232
234
来自GOS的TX数据包？
是
否
将数据包登入恰当的TX队列
236
传输数据包至NIC
238
传输数据包至网络，并登记事件队列
240
根据GOS的MAC地址对RX数据包解多路复用
244
将数据包登入恰当的RX队列，并登记事件队列
246
传输数据包至GOS
248
结束
242

图 2C

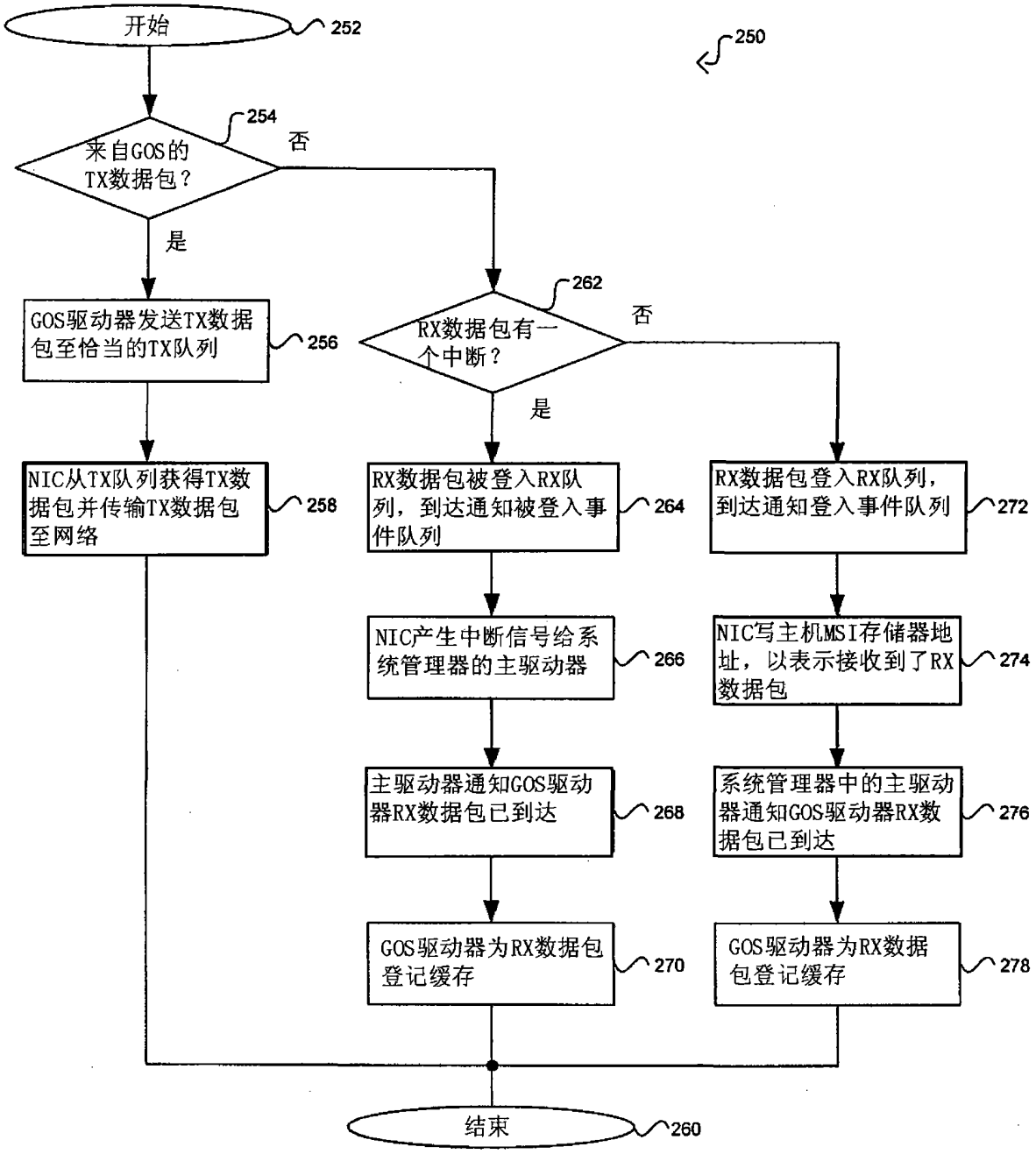

开始
252
250
254
来自GOS的TX数据包？
否
是
GOS驱动器发送TX数据包至恰当的TX队列
256
262
RX数据包有一个中断？
否
是
NIC从TX队列获得TX数据包并传输TX数据包至网络
258
RX数据包被登入RX队列，到达通知被登入事件队列
264
RX数据包登入RX队列，到达通知登入事件队列
272
NIC产生中断信号给系统管理器的主驱动器
266
NIC写主机MSI存储器地址，以表示接收到了RX数据包
274
主驱动器通知GOS驱动器RX数据包已到达
268
系统管理器中的主驱动器通知GOS驱动器RX数据包已到达
276
GOS驱动器为RX数据包登记缓存
270
GOS驱动器为RX数据包登记缓存
278
结束
260

图 2D

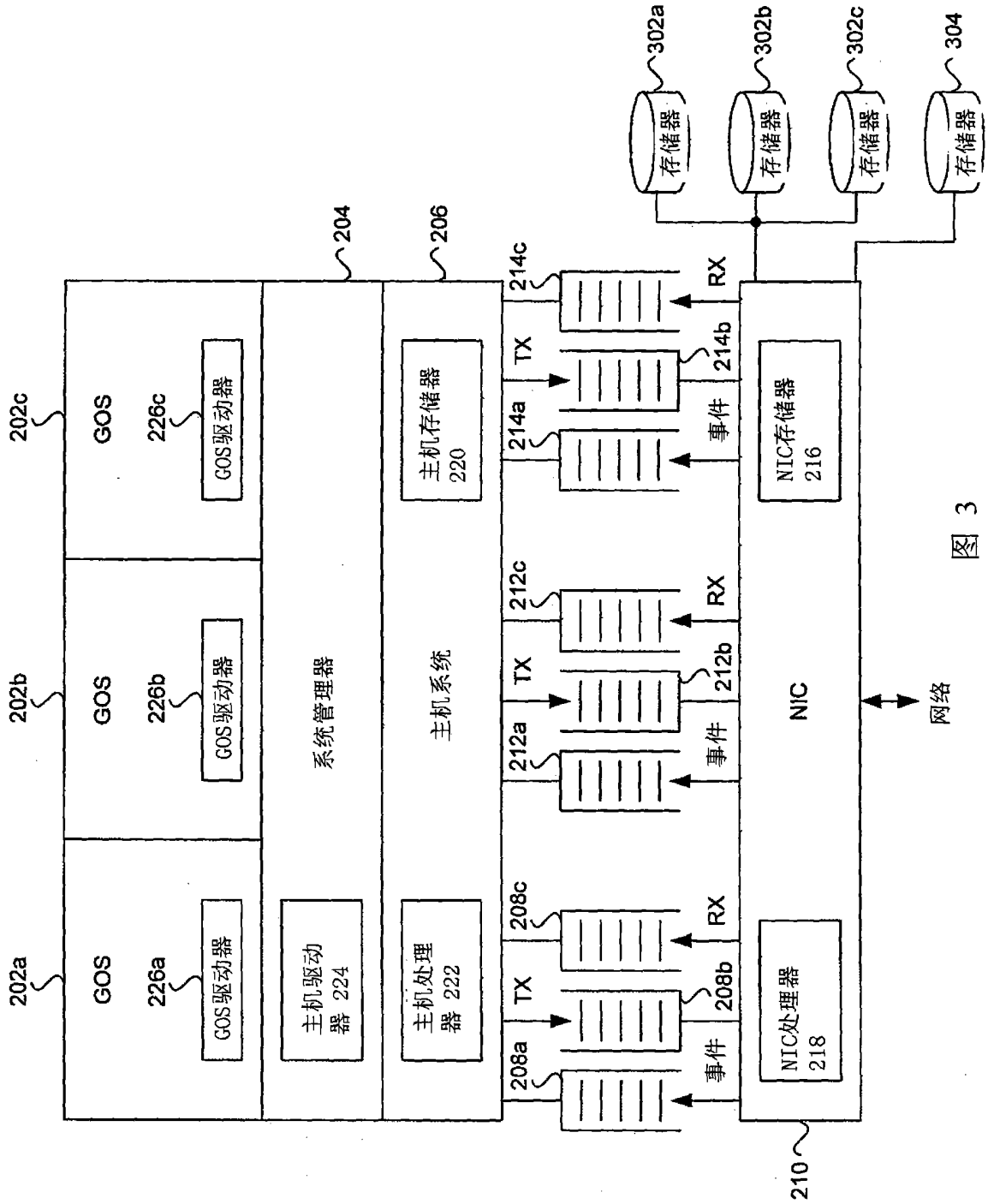

202a
202b
202c
GOS
GOS
GOS
226a
226b
226c
GOS驱动器
GOS驱动器
GOS驱动器
204
206
系统管理器
主机驱动器 224
主机处理器 222
主机系统
主机存储器 220
208a
208b
208c
212a
212b
212c
214a
214b
214c
TX
RX
事件
NIC
NIC处理器 218
NIC存储器 216
210
网络
302a
302b
302c
304
存储器

图 3

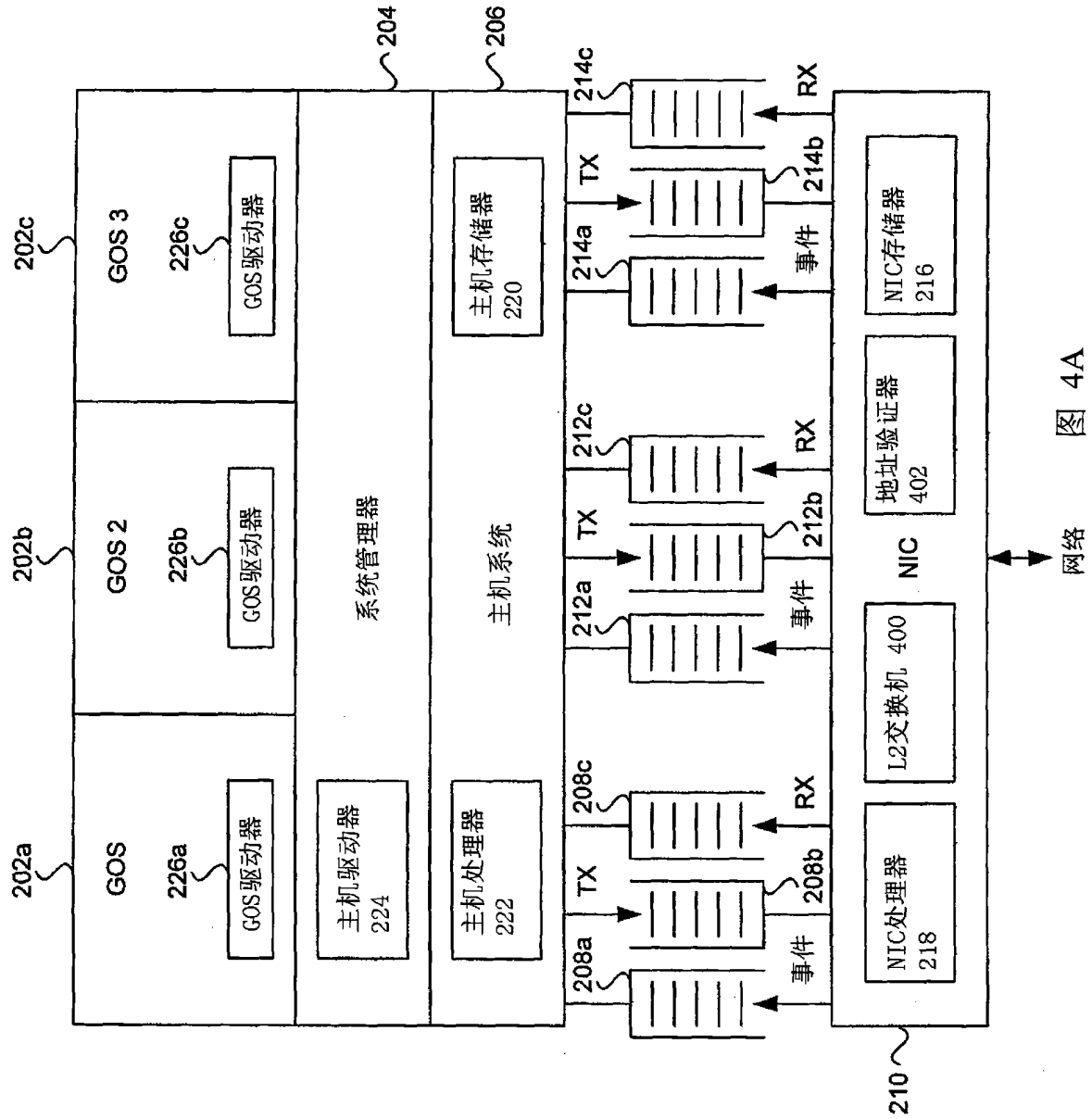

202a
GOS
226a
GOS驱动器
202b
GOS 2
226b
GOS驱动器
202c
GOS 3
226c
GOS驱动器
204
系统管理器
主机驱动器
224
206
主机系统
主机处理器
222
主机存储器
220
208a
TX
208b
事件
208c
RX
212a
TX
212b
事件
212c
RX
214a
TX
214b
事件
214c
RX
210
NIC
NIC处理器
218
L2交换机 400
地址验证器
402
NIC存储器
216
网络

图 4A

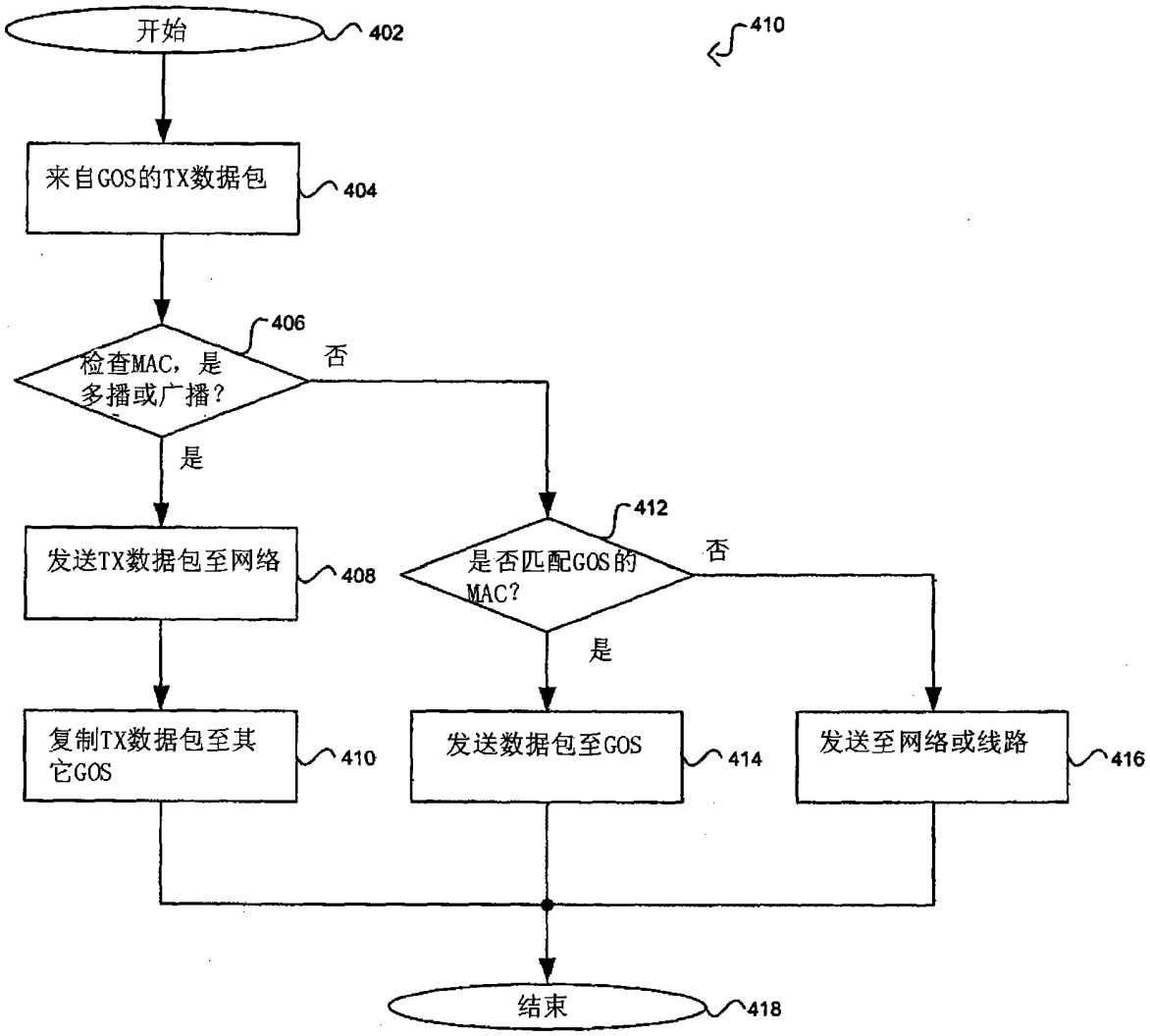

开始
402
410
来自GOS的TX数据包
404
406
检查MAC，是多播或广播？
否
是
412
发送TX数据包至网络
408
是否匹配GOS的MAC?
否
是
复制TX数据包至其它GOS
410
发送数据包至GOS
414
发送至网络或线路
416
结束
418

图 4B