(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110221769 B
(45)授权公告日 2024.12.03

(21)申请号 201910053414.3

(22)申请日 2019.01.21

(65)同一申请的已公布的文献号
申请公布号 CN 110221769 A

(43)申请公布日 2019.09.10

(30)优先权数据
62/637,340 2018.03.01 US
16/007,949 2018.06.13 US

(73)专利权人 三星电子株式会社
地址 韩国京畿道水原市

(72)发明人 颂蓬·保罗·奥莱瑞兹
弗雷德·沃利

(74)专利代理机构 北京铭硕知识产权代理有限公司 11286
专利代理师 姜长星 张川绪

(51)Int.Cl.
G06F 3/06(2006.01)

(56)对比文件
CN 107659518 A,2018.02.02
US 2014372678 A1,2014.12.18
US 6052738 A,2000.04.18

审查员 郑明月

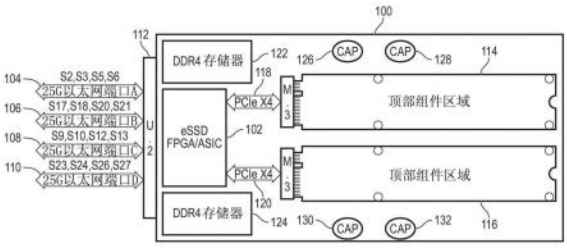
权利要求书2页 说明书13页 附图12页

(54)发明名称

存储系统、存储装置和控制存储装置的方法

(57)摘要

存储系统、存储装置和控制存储装置的方法。根据本发明的一些示例实施例,一种系统包括:至少一个主板;至少一个基板管理控制器(BMC);中平面;至少一个存储装置,其中,所述至少一个存储装置被配置为:基于通过中平面经由多个装置端口从所述至少一个主板或所述至少一个BMC接收的第一输入,在第一模式或第二模式下进行操作;当在第二模式下进行操作时,所述至少一个存储装置被配置为:基于经由所述多个装置端口从中平面接收的第二输入,以多个操作速度之中的第一速度进行操作。



CN 110221769 B

1.一种存储系统,包括:

至少一个主板;

至少一个基板管理控制器;

中平面;

至少一个存储装置,其中,所述至少一个存储装置被配置为:基于通过中平面经由多个装置端口从所述至少一个主板或所述至少一个基板管理控制器接收的第一输入,在第一模式或第二模式下进行操作;

当在第二模式下进行操作时,所述至少一个存储装置被配置为:基于经由所述多个装置端口从中平面接收的第二输入,从多个操作速度选择第一速度并以第一速度进行操作,其中,所述多个操作速度分别对应于多个网络速度,

其中,所述至少一个存储装置的第一模式和第二模式分别是非易失性存储器传输标准NVMe模式和通过网络的非易失性存储器传输标准NVMe-oF模式。

2.根据权利要求1所述的存储系统,其中,所述至少一个存储装置包括经由连接器和至少一条总线与至少一个现场可编程门阵列FPGA进行通信的多个固态驱动器SSD。

3.根据权利要求2所述的存储系统,其中,连接器是SSD连接器,所述至少一条总线是外围组件快速互连PCIe总线。

4.根据权利要求1所述的存储系统,其中,第一输入通过使用所述至少一个主板的机箱上的物理引脚或通过来自所述至少一个基板管理控制器的带内命令来控制。

5.根据权利要求1所述的存储系统,其中,第二输入通过使用两个通用输入/输出GPIO引脚来控制,其中,两个GPIO引脚由所述至少一个基板管理控制器或所述至少一个主板的本地中央处理器CPU或者所述至少一个存储装置的现场可编程门阵列FPGA内的一个或多个内部寄存器控制。

6.根据权利要求1所述的存储系统,其中,所述至少一个存储装置的所述多个操作速度是10G以上的两个或更多个速度。

7.根据权利要求1所述的存储系统,其中,所述多个装置端口经由U.2连接器连接到所述至少一个存储装置,所述至少一个存储装置被配置为在高可用性HA模式或非HA模式下进行操作。

8.根据权利要求7所述的存储系统,其中,所述至少一个存储装置经由U.2连接器支持两个串行附接的小型计算机系统接口端口和外围组件快速互连PCIe X4总线的多达四条PCIe X4通路,其中,所述两个串行附接的小型计算机系统接口端口用作网络附接的端口,其中,网络附接的端口包括以太网端口、网络通道端口和无限带宽端口中的至少一个。

9.根据权利要求8所述的存储系统,其中,当在第二模式下进行操作时,所述至少一个存储装置经由U.2连接器支持所述四条PCIe X4通路中的两条PCIe X4通路用于两个网络附接的端口的控制平面以及所述四条PCIe X4通路中的剩余两条PCIe X4通路作为额外的网络附接的端口。

10.根据权利要求1所述的存储系统,其中,所述至少一个存储装置包括第一存储装置和第二存储装置,其中,在第一时间,第一存储装置和/或第二存储装置以所述多个操作速度之中的第一速度或第二速度在第一模式或第二模式下进行操作。

11.一种控制存储装置的方法,包括:

在存储装置处接收第一输入,其中,第一输入在存储装置处通过中平面经由多个装置端口从至少一个主板或至少一个基板管理控制器被接收;

基于在存储装置处接收的第一输入,通过存储装置确定是在第一操作模式下还是在第二操作模式下进行操作;

当在第二操作模式下进行操作时,在存储装置处经由所述多个装置端口从中平面接收第二输入;

基于第二输入,通过存储装置从多个操作速度选择存储装置的操作速度,其中,所述多个操作速度分别对应于多个网络速度,

其中,存储装置的第一操作模式和第二操作模式分别是非易失性存储器传输标准NVMe模式和通过网络的非易失性存储器传输标准NVMe-oF模式。

12.根据权利要求11所述的方法,其中,第一输入通过使用所述至少一个主板的机箱上的物理引脚或通过来自所述至少一个基板管理控制器的带内命令来控制。

13.根据权利要求11所述的方法,其中,第二输入通过使用两个通用输入/输出GPIO引脚来控制,其中,两个GPIO引脚由所述至少一个基板管理控制器或所述至少一个主板的本地中央处理器CPU或者存储装置的现场可编程门阵列FPGA内的一个或多个内部寄存器控制。

14.根据权利要求11所述的方法,其中,存储装置的所述多个操作速度是10G以上的两个或更多个速度。

15.一种存储装置,包括:

多个固态驱动器,经由第一连接器和至少一条总线与至少一个现场可编程门阵列FPGA进行通信,其中,所述存储装置被配置为基于经由第二连接器接收的第一输入在第一模式或第二模式下进行操作,其中,

当在第二模式下进行操作时,所述存储装置被配置为:基于经由第二连接器接收的第二输入,以多个操作速度之中的第一速度进行操作,其中,所述多个操作速度分别对应于多个网络速度,

其中,所述存储装置的第一模式和第二模式分别是非易失性存储器传输标准NVMe模式和通过网络的非易失性存储器传输标准NVMe-oF模式。

16.根据权利要求15所述的存储装置,其中,通过中平面经由多个装置端口从交换机的主板或基板管理控制器BMC接收第一输入,经由所述多个装置端口从中平面接收第二输入,

其中,第一输入通过使用主板的机箱上的物理引脚或通过来自BMC的带内命令来控制,第二输入通过使用两个通用输入/输出GPIO引脚来控制,其中,两个GPIO引脚由BMC或主板的本地中央处理器CPU或者FPGA内的一个或多个内部寄存器控制。

17.根据权利要求15所述的存储装置,其中,第一连接器是固态驱动器连接器,所述至少一条总线是外围组件快速互连PCIe总线,第二连接器是U.2连接器,所述存储装置的所述多个操作速度是10G以上的两个或更多个速度。

18.根据权利要求15所述的存储装置,其中,在第一时间,所述多个固态驱动器中的第一固态驱动器和/或第二固态驱动器以所述多个操作速度之中的第一速度或第二速度在第一模式或第二模式下进行操作。

存储系统、存储装置和控制存储装置的方法

[0001] 本申请要求于2018年3月1日提交的第62/637,340号美国临时专利申请和于2018年6月13日提交的第16/007,949号美国专利申请的权益,所述申请的内容通过引用全部包含于此。

技术领域

[0002] 根据本发明的实施例的一个或多个方面涉及网络附接装置,更具体地讲,涉及用于支持多模式和/或多速度通过网络的非易失性存储器(NVM)传输标准(NVMe)(Non-volatile memory express over fabrics,NVMe-oF)装置的系统和方法。

背景技术

[0003] 在NVMe-oF配置(例如,使用以太网附接的NVMe固态驱动器(SSD)的配置)的情况下,改善以太网和SSD成本以及性能可正受到挑战。例如,在SSD性能可依赖于外围组件快速互连(PCIe)接口和NAND技术的同时,以太网速度已随着50G/100G技术的到来而增加。网络附接的SSD可提出用于支持擦除编码数据保护的额外的独特设计挑战,这是因为每个装置可提供点对点连通。

[0004] 因此,可期望能够支持NVMe和NVMe-oF协议二者并以不同以太网速度操作的存储装置。

[0005] 上面背景技术部分中公开的信息仅是为了增加对本发明的背景技术的理解,因此,它可能包含不构成现有技术的信息。

发明内容

[0006] 本公开的实施例的各方面指向一种系统,所述系统包括:至少一个主板;至少一个基板管理控制器(BMC);中平面;至少一个存储装置,其中,所述至少一个存储装置被配置为:基于通过中平面经由多个装置端口从所述至少一个主板或所述至少一个BMC接收的第一输入,在第一模式或第二模式下进行操作;当在第二模式下进行操作时,所述至少一个存储装置被配置为:基于经由所述多个装置端口从中平面接收的第二输入,以多个操作速度之中的第一速度进行操作。

[0007] 在一个实施例中,所述至少一个存储装置包括经由连接器和至少一条总线与至少一个现场可编程门阵列(FPGA)进行通信的多个固态驱动器(SSD)。在一个配置中,连接器是SSD连接器,所述至少一条总线是外围组件快速互连(PCIe)总线。

[0008] 在另一实施例中,所述至少一个存储装置的第一模式和第二模式分别是非易失性存储器传输标准(NVMe)模式和通过网络的NVMe(NVMe-oF)模式。在一个配置中,第一输入通过使用所述至少一个主板的机箱上的物理引脚或通过来自所述至少一个BMC的带内命令来控制。

[0009] 在一个实施例中,第二输入通过使用两个通用输入/输出(GPIO)引脚来控制,其中,两个GPIO引脚由所述至少一个BMC或所述至少一个主板的本地中央处理器(CPU)或者所

述至少一个存储装置的现场可编程门阵列(FPGA)内的一个或多个内部寄存器控制。在另一实施例中,所述至少一个存储装置的所述多个操作速度是10G以上的两个或更多个速度(例如,10G与100G之间,或者10G与100G以上的速度之间)。

[0010] 在一个方面,所述多个装置端口经由U.2连接器连接到所述至少一个存储装置,存储装置被配置为在高可用性(HA)模式或非HA模式下进行操作。在另一方面,所述至少一个存储装置经由U.2连接器支持两个串行附接的小型计算机系统接口(SAS)端口和外围组件快速互连(PCIe)X4总线的多达四条PCIe X4通路,其中,所述两个SAS端口用作网络附接的端口,其中,网络附接的端口包括以太网端口、网络通道端口和无限带宽端口。

[0011] 在一个配置中,当在第二模式下进行操作时,所述至少一个存储装置经由U.2连接器支持所述四条PCIe X4通路中的两条PCIe X4通路用于两个网络附接的端口的控制平面以及所述四条PCIe X4通路中的剩余两条PCIe X4通路作为额外的网络附接的端口。在另一配置中,所述至少一个存储装置包括第一存储装置和第二存储装置,其中,在第一时间,第一存储装置和/或第二存储装置以所述多个速度之中的第一速度或第二速度在第一模式或第二模式下进行操作。

[0012] 根据本发明的实施例,提供一种方法,所述方法包括:在存储装置处接收第一输入,其中,第一输入在存储装置处通过中平面经由多个装置端口从至少一个主板或基板管理控制器(BMC)被接收;基于在存储装置处接收的第一输入,通过存储装置确定是在第一操作模式下还是在第二操作模式下进行操作;当在第二模式下进行操作时,在存储装置处经由所述多个装置端口从中平面接收第二输入;基于第二输入,通过存储装置从多个操作速度选择存储装置的操作速度。

[0013] 在一个实施例中,存储装置的第一模式和第二模式是非易失性存储器传输标准(NVMe)模式和通过网络的NVMe(NVMe-oF)模式。在一个实施例中,第一输入通过使用所述至少一个主板的机箱上的物理引脚或通过来自BMC的带内命令来控制。

[0014] 在一个方面,第二输入通过使用两个通用输入/输出(GPIO)引脚来控制,其中,两个GPIO引脚由BMC或所述至少一个主板的本地中央处理器(CPU)或者存储装置的现场可编程门阵列(FPGA)内的一个或多个内部寄存器控制。在另一方面,其中,存储装置的所述多个操作速度是10G以上的两个或更多个速度。

[0015] 根据本发明的另一实施例,提供一种存储装置,包括:多个固态驱动器(SSD),经由第一连接器和至少一条总线与至少一个现场可编程门阵列(FPGA)进行通信,其中,所述存储装置被配置为基于经由第二连接器接收的第一输入在第一模式或第二模式下进行操作,其中,当在第二模式下进行操作时,所述存储装置被配置为:基于经由第二连接器接收的第二输入,以多个操作速度之中的第一速度进行操作。

[0016] 在一个方面,通过中平面经由多个装置端口从交换机的主板或基板管理控制器(BMC)接收第一输入,经由所述多个装置端口从中平面接收第二输入,其中,第一输入通过使用主板的机箱上的物理引脚或通过来自BMC的带内命令来控制,第二输入通过使用两个通用输入/输出(GPIO)引脚来控制,其中,两个GPIO引脚由BMC或主板的本地中央处理器(CPU)或者所述至少一个存储装置的现场可编程门阵列FPGA内的一个或多个内部寄存器控制。

[0017] 在另一方面,第一连接器是SSD连接器,所述至少一条总线是外围组件快速互连

(PCIe)总线,第二连接器是U.2连接器,所述存储装置的第一模式和第二模式分别是非易失性存储器传输标准(NVMe)模式和通过网络的NVMe(NVMe-oF)模式,所述存储装置的所述多个操作速度是10G以上的两个或更多个速度。

[0018] 在一个配置中,在第一时间,所述多个SSD中的第一SSD和/或第二SSD以所述多个速度之中的第一速度或第二速度在第一模式或第二模式下进行操作。

附图说明

[0019] 参照说明书、权利要求书和附图,将了解和理解本发明的一些示例实施例的这些和其他特征,其中,在附图中:

[0020] 图1A示出根据本发明的一些示例实施例的NVMe-oF装置的配置;

[0021] 图1B示出根据本发明的一些示例实施例的示例SSD的框图;

[0022] 图2示出根据本发明的一些示例实施例的中平面(mid-plane)或复杂可编程逻辑器件(CPLD)的框图;

[0023] 图3示出说明根据本发明的一些示例实施例的根据示例多模式NVMe-oF装置的配置的U.2连接器的示例使用的示例表;

[0024] 图4A和图4B示出根据本发明的一些示例实施例的以非高可用性(非HA)模式操作的示例NVMe-oF装置的框图;

[0025] 图4C和图4D示出根据本发明的一些示例实施例的以高可用性(HA)模式操作的示例NVMe-oF装置的框图;

[0026] 图5A示出根据本发明的一些示例实施例的示例现场可编程门阵列(FPGA)装置的框图;

[0027] 图5B示出根据本发明的一些示例实施例的示例FPGA装置的另一框图;

[0028] 图6示出根据本发明的一些示例实施例的示例FPGA装置的又一框图;

[0029] 图7示出根据本发明的一些示例实施例的示例交换机的框图;

[0030] 图8示出根据本发明的一些示例实施例的用于操作存储装置的方法的流程图。

具体实施方式

[0031] 下面结合附图阐述的具体实施方式意图作为根据本发明提供的用于支持多模式和/或多速度NVMe-oF装置的系统和方法的一些示例实施例的实施方式,并不意图表示可构造或利用本发明的仅有形式。本说明书结合示出的实施例阐述本发明的特征。然而,将理解,相同或等同的功能和结构可通过也意图包含在本发明的范围内的不同实施例来实现。如这里别处所示,相同的元件标号意图指示相同的元件或特征。

[0032] NVMe是定义用于通过PCIe总线与非易失性存储器子系统(例如,SSD)进行通信的主机软件的寄存器级接口的标准。NVMe可以是小型计算机系统接口(SCSI)标准的替代,其中,小型计算机系统接口(SCSI)标准用于连接主机和外围目标存储装置或系统并在主机和外围目标存储装置或系统之间传送数据。PCIe连接的NVMe SSD允许应用与存储器直接对话。

[0033] NVMe的物理连接基于PCIe总线。典型以太网SSD(eSSD)具有U.2连接器以通过PCIe总线经由中平面与系统相连接。U.2连接器(SFF-8639)是用于连接SSD和计算机的计算机接

口。U.2连接器支持一个串行高级技术附件(SATA)端口、两个串行附接的SCSI(SAS)端口或PCIe SSD中的并行I/O的多达四个通路(X4)。U.2连接器针对NVMe被标准化并支持PCIe 3.0 X4,提供典型SATA SSD的五倍速度。

[0034] NVMe-oF是对通过PCIe之外的各种网络(或互连)实现操作的NVMe标准的扩展。这里,术语“网络(fabric)”表示网络拓扑,其中,在该网络拓扑中,网络节点能通过各种互连协议、端口和交换机彼此传递数据。例如,与以太网附接(Ethernet-attached)的SSD可以直接附接到网络,在这种情况下,网络是以太网。

[0035] NVMe-oF实现对PCIe的代替传送的使用,其中,代替传送延伸NVMe主机装置和NVMe存储驱动器或子系统能够连接的距离。因此,NVMe-oF是被设计为使用基于非易失性存储器传输标准消息的命令以通过网络(诸如,以太网、光纤通道(FC)或无限宽带)在主机计算机和目标固态存储装置(例如,eSSD或NVMe-oF装置)或系统之间传送数据的技术规范。当被配置为支持NVMe-oF标准,系统能够支持各种网络,所述各种网络不仅包括以太网还包括光纤通道、无限宽带和其他网络架构。例如,eSSD可直接附接到网络,在这种情况下,网络是以太网。eSSD可表示可支持NVMe-oF协议的SSD。为了示出的方便,下面的示例和实施例可示出以太网附接的NVMe-oF装置。然而,应注意,在不脱离本公开的范围的情况下,可使用任何其他类型的NVMe-oF装置。

[0036] 如上所述,NVMe的物理连接基于PCIe总线。最近,随着PCIe 4.0的出现,带宽失配误差(bandwidth mismatch)可高于PCIe 3.0。单个25G以太网不具有足够的带宽来从SSD通过X4PCIe 4.0(多达8GB/s)跟上后端。50G或100G以太网提供从SSD的X4PCIe 4.0的更好的匹配。因此,本发明的实施例可包括能够支持NVMe和NVMe-oF协议二者的存储装置(例如,eSSD),并且当在NVMe-oF模式下时,存储装置在对其硬件进行任何改变的情况下能够以从10G到100G的以太网速度进行操作。

[0037] 此外,本发明的实施例可使任何平台系统能够支持来自不同供应商的不同类型的NVMe-oF装置。通过使用通用构造块(诸如,交换机板、中平面和以太网SSD),生态系统提供商能够比现有系统更快地进入市场,并提供各种价格/性能产品以满足各种消费者。一些示例实施例还可实现能够支持具有标准U.2连接器的NVMe-oF装置(诸如,PM1725a或PM1735)和新出现的诸如M.3SSD的通用系统平台。在一种配置中,M.3SSD还可被称为基于NGSFF的SSD。

[0038] 本系统的不同实施例提供可支持NVMe和NVMe-oF装置二者的通用构造块和单个平台,并且当在NVMe-oF模式下时,该系统能够在不改变它们的硬件的情况下以从10G到100G的以太网速度进行操作。可支持NVMe和NVMe-oF装置二者的通用构造块可包括交换机板、中平面和eSSD。本系统可通过添加更多类似的装置和/或机箱(chassis)来线性扩展。本系统还可提供能够支持具有标准U.2连接器的NVMe-oF装置(诸如,PM1725a或PM1735)和基于M.3或NGSFF的SSD的通用系统平台。本系统的不同配置还可与技术进步(诸如,50G和100G以太网以及PCIe 4.0)兼容。尤其,本申请的不同配置提供能支持NVMe和NVMe-oF协议二者以及各种类型的网络附接的SSD(例如,eSSD)的系统。此外,本申请的不同配置可提供关于机箱和装置(例如,eSSD)如何调整装置的信息以理解它被部署在哪儿以及作为它在数据集合的特定地方的部署的结果它应该采用哪种性能。

[0039] 这里公开的网络附接的SSD(eSSD)是可用在与NVMe和NVMe-oF标准兼容的多个系

统中的单个通用装置。以太网SSD可使用U.2连接器以经由中平面与主机装置相连接。U.2连接器可支持25Gbps(100G-KR4)以太网。这里公开的eSSD可以是可用在与NVMe和NVMe-oF标准兼容的多个系统中的装置。照此,网络附接的SSD(eSSD)还可被称为多模式NVMe-oF装置。多模式NVMe-oF装置还可通过从已知位置(例如,中平面或主板的机箱类型引脚E6)检测产品信息来支持NVMe或NVMe-oF标准。例如,U.2连接器上的限定信号(例如,模式信号)可指示驱动器是在NVMe模式还是在NVMe-oF模式下进行操作。如果限定信号出现在NVMe机箱中,则U.2连接器的X4通路PCIe将由PCIe引擎驱动。在这种情况下,装置将禁用网络附接的端口(例如,以太网端口、网络通道端口和无限带宽端口中的至少一个),而全部的NVMe协议和功能被支持或启用。如果出现在NVMe-oF机箱中,则网络附接的端口将仅使用未使用的SAS引脚。

[0040] 图1A示出NVMe-oF装置100的配置。图1A的NVMe-oF装置100包括eSSD FPGA/ASIC 102、第一M.3格式SSD 114、第二M.3格式SSD 116、两个双数据速率第四代(DDR4)存储器(DDR4存储器122和144)和四个电容器CAP(电容器126、128、130和132)。NVMe-oF装置100还连接到U.2连接器112,并且,经由U.2连接器112,NVMe-oF装置100可支持两个SAS端口(SAS0和SAS1)和PCIe X4总线的多达四个PCIe X4通路。两个SAS端口(SAS0和SAS1)可被NVMe-oF装置100用作以太网端口。NVMe-oF装置100能够在对硬件不进行任何改变的情况下以从10G到100G的不同以太网速度进行操作。

[0041] 图1B示出根据一个实施例的示例M.3格式SSD 134的框图。M.3格式SSD 134可以是图1A的SSD 114或SSD 116。M.3格式SSD 134的新形状因子1(NF1)可容纳两行的NAND闪存封装,从而最大化模块化SSD的容量。M.3格式SSD 134可经由PCIe X4总线136连接到FPGA。FPGA可以是图1A的FPGA 102,PCIe X4总线136可以是图1A的PCIe X4总线118或120中的一个。M.3格式SSD 134还可使用用于连接到FPGA(例如,图1A中的FPGA 102)的M.2连接器的未使用或保留的(例如,高速)引脚,来容纳至少一个可选的以太网端口138(以太网端口A和/或以太网端口B)。至少一个以太网端口138可使M.3格式SSD 134能够在NVMe-oF模式下进行操作。

[0042] 在一些实施例中,M.3格式SSD 134可符合2.5英寸硬盘驱动器形状因子(或小形状因子(SFF))标准。在其他实施例中,M.3格式SSD 134可符合标准PCIe卡形状因子(例如,全高度、全长度(FH-FL)卡轮廓,或全高度、半长度(FH-HL)卡轮廓)。M.3格式SSD 134还可包括控制器、缓冲存储器和闪存。控制器可执行存储在例如控制器中(或与控制器分开的)的只读存储器或缓冲存储器中的软件、NVMe命令和/或固件。在图1B中,eSSD ASIC 140是包含闪存转换层(FTL)和闪存控制器的增强的SSD控制器。

[0043] 回到图1A,NVMe-oF装置100可被配置为在NVMe模式或NVMe-oF模式下进行操作。在NVMe-oF模式下,NVMe-oF装置100可将两条PCIe通路(PCIe 1和PCIe 2)配置为以太网端口106(以太网端口B)和以太网端口110(以太网端口D)。NVMe-oF装置100还可将两个SAS端口(SAS0和SAS1)配置为以太网端口104(以太网端口A)和以太网端口108(以太网端口C)。第一25G以太网端口104可连接在U.2连接器112的引脚S2、S3、S5和S6,第二25G以太网端口106可连接在U.2连接器112的引脚S17、S18、S20和S21,第三25G以太网端口108可连接在U.2连接器112的引脚S9、S10、S12和S13,第四25G以太网端口110可连接在U.2连接器112的引脚S23、S24、S26和S27。当在NVMe-oF模式下时,NVMe-oF装置100能够以从10G到100G的不同以太网

速度进行操作。

[0044] 在图1A中,当NVMe-oF装置100被配置在NVMe-oF模式下时,FPGA 102可提供四个25Gbps以太网端口(以太网端口104、106、108和110)和两个M.3格式SSD(SSD 114和116)之间的接口。四个25G以太网端口(以太网端口104、106、108和110)可根据NVMe-oF装置100的操作的模式通过中平面连接到主板。在一个配置中,主板可包括一个或多个开关元件、一个或多个存储器元件和一个或多个I/O元件等。FPGA 102经由第一PCIe X4总线118与第一M.3格式SSD 114相连接,FPGA 102经由第二PCIe X4总线120与第二M.3格式SSD 116相连接。第一PCIe X4总线118和第二PCIe X4总线120可通过它们各自的M.3连接器连接到第一M.3格式SSD 114和第二M.3格式SSD 116。在这种情况下,附接到SSD 114和SSD 116的M.2连接器的未使用或保留的(例如,高速)引脚可用于以太网连接。在这种模式下,FPGA 102可用作NVMe-oF目标。在eSSD ASIC/FPGA102上实现的NVMe-oF目标提供所有网络和存储协议处理,这免去了目标装置中对于主板上的基于x86的中央处理器(CPU)的需要。对于基于NVMe-oF的系统,不再需要主板上的基于x86的CPU在发起者(例如,主机软件)和目标装置(即,NVMe-oF装置)之间移动数据,因为目标装置通过它自己能够移动数据。

[0045] 在一个配置中,NVMe-oF装置100可被配置在NVMe模式下。在NVMe模式下,NVMe-oF装置100可(在单端口模式下)使用所有的四个PCIe X4通路,以通过PCIe X4总线携带PCIe信号。PCIe X4总线连接到中平面,PCIe X4总线在数据和控制信号之间被共享。在一个方面中,在给定的时间,NVMe-oF装置100的第一M.3格式SSD 114和第二M.3格式SSD 116都可在NVMe模式或NVMe-oF模式下进行操作。在另一方面中,在给定的时间,第一M.3格式SSD 114可在NVMe模式下进行操作,第二M.3格式SSD 116可在NVMe-oF模式下进行操作。在又一方面中,在给定的时间,第一M.3格式SSD 114可在NVMe-oF模式下进行操作,第二M.3格式SSD 116可在NVMe模式下进行操作。

[0046] NVMe-oF装置100的操作模式可自配置或使用物理引脚(例如,主板的机箱上的存在引脚(引脚E6))或通过来自主板的基板管理控制器(BMC)的带内命令进行外部设置。通过以太网得到的可管理性信息被称为“带内”信息,而通过PCIe总线得到的可管理性信息被称为“带外”信息。当被配置为NVMe-oF装置时,多模式NVMe-oF装置100(机箱类型引脚E6=低)可被配置在单端口NVMe-oF模式或双端口NVMe-oF模式下。在单端口NVMe-oF模式下,U.2连接器112的引脚E25可为高;在双端口NVMe-oF模式下,U.2连接器112的引脚E25可为低。

[0047] 在一个配置中,如图2所示,NVMe-oF装置100的不同操作速度可通过使用位于中平面或CPLD 204上的两个额外的通用输入/输出(GPIO)(ESpeed[1:0])引脚202来实现,其中,ESpeed[1:0]引脚202可由NVMe-oF装置100的FPGA 102内的内部寄存器或主板的本地CPU或BMS控制,并可经由连接在U.2连接器112的引脚E23和E24的I2C总线来访问。在前述选项中,ESpeed引脚202与CPLD 204内的U.2I2C引脚进行复用,并可在重置Reset#207已被断言(逻辑高至逻辑低)之后被锁存。在一个配置中,对于复用器MUX的输入由主板的本地CPU或BMC或CPLD 204驱动。然而,在一些配置中,MUX的位置在CPLD 204内。MUX的数量等于机箱支持的NVMe-oF装置的最大数量。通过这种方式,每个装置都可独自和独立地被重置。这种方法可支持热插拔事件(热添加和热移除)。如果存在一个MUX被所有设备共享,则当新装置被热插入时,它可由于重置周期而影响机箱中的现有装置,这可能是不可取的。

[0048] 如果CPLD 204由主板的BMC或本地CPU控制,则来自BMC或本地CPU的控制信号可在

位于CPLD 204的CPLD/BMC GPIO引脚206接收。下面的表1示出NVMe-oF装置100的不同操作速度期间的ESpeed引脚的状态。

[0049]

以太网速度	ESpeed 引脚 1	ESpeed 引脚 0
10G	逻辑高	逻辑高
25G	逻辑高	逻辑低
50G	逻辑低	逻辑高
100G	逻辑低	逻辑低

[0050]

[0051] 表1

[0052] 此外,在上电或重置期间,使用系统管理总线SMBus 208选择以太网速度。例如,SMBus 208可传输来自BMC/本地CPU的SMBus时钟(SMBus CLK)和SMBus数据(SMBusData)等。在一个配置中,通过上电或重置信号207选择复用器,并且,当重置207为低时,复用器可对FPGA 102选择Espeed[0:1]202,当重置207为高时,复用器可将每个槽之间的SMBus 208连接到交换机板上的BMC。

[0053] 在一个配置中,多模式NVMe-oF装置100可被配置为单端口NVMe模式、双端口NVMe模式、单端口NVMe-oF模式和双端口NVMe-oF模式。

[0054] 图3示出根据多模式NVMe-oF装置100的配置的U.2连接器112的示例使用。当被配置为NVMe装置(机箱类型引脚E6=低)时,多模式NVMe-oF装置100可被配置为单端口NVMe模式或双端口NVMe模式。在单端口NVMe模式下,U.2连接器112的双端口EN#引脚E25可被断言为高。在单端口NVMe模式下,U.2连接器112的PCIe通路0-3用于携带PCIe信号。在双端口NVMe模式下,U.2连接器112的双端口EN#引脚E25可被断言为低。在双端口NVMe模式下,PCIe通路0-3被分成2×2的通路;PCIe通路0和1用于第一端口(端口A),PCIe通路2和3用于第二端口(端口B)。

[0055] 在一个配置中,当多模式NVMe-oF装置100处于NVMe-oF模式时,NVMe-oF装置100可具有对于第一以太网端口(端口A)和第二以太网端口(端口B)的控制平面的两个X1PCIe通路(PCIe 0和PCIe 3)。这样的配置可使两个PCIe通路(PCIe 1和PCIe2)可用于额外的25G以太网端口(以太网端口B和D)。在一些配置中,当NVMe-oF装置100在NVMe-oF模式下以10G(单端口或双端口)或25G(单端口或双端口)速度进行操作时,SAS端口0用于以太网端口A(第一端口),而SAS端口1不被使用。PCIe通路0和3被用作第一(端口A)和第二(端口B)以太网附接的NVMe-oF控制器的控制平面。PCIe通路1用于以太网端口B(第二端口),而PCIe通路2不被使用。在一些配置中,当NVMe-oF装置100在NVMe-oF模式下以50G(单端口或双端口)或100G(仅单端口)速度进行操作时,SAS端口0用于以太网端口A(第一端口),SAS端口1用作以太网端口C(第三端口)。PCIe通路0和3被用作第一(端口A)和第二(端口B)以太网附接的NVMe-oF控制器的控制平面,PCIe通路1用于以太网端口B(第二端口),而PCIe通路2用于以太网端口D(第四端口)。

[0056] 在一些配置中,如果产品信息存储在机箱中,则U.2连接器112上的PCIe总线的两条通路(单端口模式下)或四条通路(双端口模式下)由PCIe引擎驱动。在这种情况下,多模式NVMe-oF装置100可禁用一个或多个以太网引擎,并且NVMe协议和功能被支持或启用。如

果产品信息存储在NVMe-oF机箱中,则以太网附接的NVMe-oF控制器根据多模式NVMe-oF装置的设计仅使用PCIe通路1和2和/或SAS引脚。

[0057] 在一些情况下,用于自配置的产品信息以重要产品数据(VPD)存储在机箱中。在启动期间,多模式NVMe-oF装置100可从机箱得到VPD,并基于VPD配置自己。在一些配置中,多模式NVMe-oF装置100可在不脱离本公开的范围的情况下以各种方式进行配置。例如,多模式NVMe-oF装置100可通过由与多模式NVMe-oF装置100连接的主板的BMC经由PCIe总线发出的控制命令来配置。本系统提供具有最小硬件改变而可在非高可用性(非HA)模式(即,单路径输入/输出(I/O))或HA模式下支持各种类型的NVMe和NVMe-oF装置的平台。

[0058] 图4A和图4B示出根据一个实施例的在非HA模式下操作的示例NVMe-oF装置400的框图。NVMe-oF装置400可以是图1A中的NVMe-oF装置100。在这个示例中,NVMe-oF装置400可支持单路径I/O,并能够在没有任何硬件改变的情况下以从10G到100G的不同以太网速度进行操作。NVMe-oF装置400可包括FPGA/ASCI 402、第一M.3格式SSD 414、第二M.3格式SSD 416和两个DDR4存储器(422和424)。FPGA 402经由第一PCIe X4总线418与第一M.3格式SSD 414相连接,FPGA 402经由第二PCIe X4总线420与第二M.3格式SSD 416相连接。NVMe-oF装置400还连接到U.2连接器412,并且经由U.2连接器412,NVMe-oF装置400可支持两个SAS端口(SAS0和SAS1)以及PCIe X4总线的多达4条PCIe X4通路。两个SAS端口(SAS0和SAS1)可用作以太网端口。NVMe-oF装置400可将两条PCIe通路(PCIe 1和PCIe 2)配置为以太网端口406(以太网端口B)和以太网端口410(以太网端口D)。NVMe-oF装置400还可将两个SAS端口(SAS0和SAS1)配置为以太网端口404(以太网端口A)和以太网端口408(以太网端口C)。第一25G以太网端口404可连接在U.2连接器412的引脚S2、S3、S5和S6,第二25G以太网端口406可连接在U.2连接器412的引脚S17、S18、S20和S21,第三25G以太网端口408可连接在U.2连接器412的引脚S9、S10、S12和S13,第四25G以太网端口410可连接在U.2连接器112的引脚S23、S24、S26和S27。

[0059] 在一个配置中,以太网端口404、406、408和410可通过中平面426经由若干高速莫莱克斯连接器428连接到主板401,其中,高速莫莱克斯连接器428可共同地携带全部的以太网端口404、406、408和410以及其他非高速控制信号(诸如,SMBus、重置和时钟等)。若干高速莫莱克斯连接器428可包括PCIe时钟分配和12V电力B到B。中平面426还可包括12V电力B到B。12V电力B到B二者可将电力供应给主板401和中平面426中的元件。具体地,中平面426的12V电力B到B可将电力供应给中平面426中的VPD电可擦除可编程只读存储器(EEPROM)。主板401中的12V热插拔控制器可控制中平面426中的12V电力B到B。由12V电力B到B二者供应的电力可由双冗余电源单元(PSU)、电力分配单元和12V电力输入生成并分配。主板401可包括连接到管理物理层(MGMT PHY)的本地CPU 434(即,CPU Broadwell-DE 434)、BMC432、连接到2×3四通道小型可插拔(QSFP)28-100G端口的以太网交换机控制器436、PCIe交换机438和两个DDR4存储器(DDR4存储器440和442)。在一些配置中,主板401可通过中平面426经由以太网端口404、406、408和410将各种信号推送到NVMe-oF装置400,并对于在NVMe-oF装置400执行各种服务。例如,主板401可通过以太网端口404、406、408和410从NVMe-oF装置400接收装置专用信息,其中,装置专用信息包括但不限于NVMe-oF装置400的健康状态信息、场可替换单元(FRU)信息和传感器信息。主板401还可通过以太网端口404、406、408和410执行各种服务,其中,各种服务包括但不限于对于BMC(例如,BMC 432)或本地主机CPU

(例如,CPU 434)的发现服务和针对新eSSD固件的用于执行固件升级的下载服务。

[0060] 图4C和图4D示出根据一个实施例的在HA模式下操作的示例NVMe-oF装置400的框图。在这个示例中,NVMe-oF装置400可支持多路径I/O,并能够在没有任何硬件改变的情况下以从10G到100G的不同以太网速度进行操作。在HA模式下,NVMe-oF装置400可经由中平面426连接到两个主板401和403,以在双端口配置下(在HA模式下)支持多路径I/O。中平面426是能支持HA模式和非HA模式二者的共用中平面。根据系统配置,信号完整性可能需要被测试,以确保共用中平面426可支持两种配置。如果信号完整性不够,则系统能具有包括用于HA模式的第一中平面和用于非HA模式的第二中平面的两种版本的中平面。U.2连接器412的E25引脚可用于实现双端口配置。NVMe-oF装置400可使用物理引脚(例如,主板(主板401或403)的机箱上的存在引脚(引脚E6))或者通过来自主板(主板401或403)的BMC(BMC432或444)的带内命令(未示出)来对它的操作模式进行自配置。BMC(BMC432,444)二者可通过3×SMBus与中平面426中的CPLD(CPLD 430)结合。CPLD(CPLD 430)可基于3.3V辅助供电设备(aux)激活BMC标识(ID)。

[0061] 在一个配置中,当在HA模式下操作时,NVMe-oF装置400的以太网端口404(以太网端口A)和以太网端口406(以太网端口B)可通过中平面426经由高速莫莱克斯连接器428连接到主板401,其中,高速莫莱克斯连接器428可共同地携带以太网端口404和406以及其他非高速控制信号(诸如,SMBus、重置和时钟等)。此外,在HA模式下,NVMe-oF装置400的以太网端口408(以太网端口C)和以太网端口410(以太网端口D)可通过中平面426经由若干高速莫莱克斯连接器456连接到第二主板403(未示出),其中,高速莫莱克斯连接器456可共同地携带以太网端口408和410以及其他非高速控制信号(诸如,SMBus、重置和时钟等)。第二主板403可包括本地CPU 446(即,CPU Broadwell-DE 446)、BMC444、以太网交换机控制器448、PCIe交换机450和两个DDR4存储器(452和454)。

[0062] 因为多模式NVMe-oF装置400可在NVMe模式和NVMe-oF模式二者下进行操作,所以由于同一装置可在NVMe模式和NVMe-oF模式下使用,因此可降低用于开发和部署装置的成本。出于同样的原因,多模式NVMe-oF装置400可具有更快的上市时间。多模式NVMe-oF装置可用于各种产品和机箱。PCIe总线的两个通路通过控制平面被预留给标准特征。CPU、BMC和其他装置可在没有额外开销的情况下使用PCIe总线的两个通路作为用于与机箱内的每个NVMe-oF装置进行通信的控制平面。NVMe中平面可未修改的使用,并且由于位于中平面上的额外的GPIO引脚(例如,图2的ESpeed[1:0]202)而不需要对于NVMe-oF装置400的新的连接器。此外,可期望任何平台系统支持来自不同供应商的不同类型的NVMe-oF装置。通过使用诸如交换机板、中平面及以太网SSD的通用构造块,本申请的不同实施例可使现有的生态系统提供商能够快的多地进入市场并提供各种价格/性能的产品,以满足不同的消费者。还可期望拥有能够支持具有标准U.2的NVMe-oF装置(诸如,PM1725a或PM1735)和新出现的诸如基于M.3或NGSFF的SSD的通用系统平台。

[0063] 图5A-5B是示出根据一个实施例的示例FPGA装置500的框图。FPGA 500可以是图1A中的FPGA 102。作为edge FPGA和SSD控制器顶级块或FPGA顶部,FPGA 500可包括时钟和重置(Clk/Reset)、调试工具、激光发光二极管(Laser emitting diode,LED)、温度传感器内部集成电路(Inter-Integrated Circuit,I2C)、通用异步接收器/发送器(Universal Asynchronous Receiver/Transmitter,UART)和联合测试行为组织(Joint Test Action

Group,JTAG)。FPGA 500可提供U.2连接器502和多个闪存驱动器(例如,图1A的两个M.3SSD 114、116)之间的接口。FPGA 500还可连接到多个DDR4存储器(例如,图1A的两个DDR4存储器(DDR4存储器122和124))。例如,FPGA 500可经由可连接到PCIe X4总线(例如,图1A的第一PCIe X4总线118)的端口518与第一SSD SSD1(例如,图1A的SSD 114)相连接,FPGA 500可经由可连接到另一PCIe X4总线(例如,图1A的第二PCIe X4总线120)的端口522与第二SSD SSD2(例如,图1A的SSD 116)相连接。FPGA 500还可经由端口516与第一DDR4存储器DDR4 1(例如,图1A的DDR4存储器122)连接,并经由端口520与第二DDR4存储器DDR4 2(例如,图1A的DDR4存储器124)连接。

[0064] 连接到FPGA 500的U.2连接器502可支持两个SAS端口(SAS0和SAS1)和PCIe X4总线的多达四条PCIe X4通路。两个SAS端口(SAS0和SAS1)可用作以太网端口。在NVMe-oF模式下,两条PCIe通路(PCIe 1和PCIe 2)可被配置为以太网端口510(以太网端口1)和以太网端口512(以太网端口3)。在NVMe-oF模式下,两个SAS端口(SAS0和SAS1)可被配置为以太网端口506(以太网端口0)和以太网端口508(以太网端口2)。以太网端口506(以太网端口0)可连接在U.2连接器502的引脚S2、S3、S5和S6,以太网端口510(以太网端口1)可连接在U.2连接器502的引脚S9、S10、S12和S13,以太网端口508(以太网端口2)可连接在U.2连接器502的引脚S23、S24、S26和S27,以太网端口512(以太网端口3)可连接在U.2连接器502的引脚S17、S18、S20和S21。

[0065] 在一个配置中,在NVMe-oF模式下,U.2连接器502的两个X1PCIe通道(PCIe 0和PCIe 3)504和514可被配置为针对用于第一(端口0)和第二(端口1)以太网NVMe-oF控制器的控制平面。控制平面PCIe通路(PCIe 0)504可连接在U.2连接器502的引脚E10、E11、E13和E14,控制平面PCIe通路514(PCIe 3)可连接在U.2连接器502的引脚E17、E18、E20和E21。在一些配置中,如图5A所示,在NVMe-oF模式下,当FPGA 500以50G(单端口或双端口)速度进行操作时,SAS端口0用于以太网端口0(以太网端口506),SAS端口1用作以太网端口2(以太网端口508)。PCIe通路504(PCIe 0)和PCIe通路514(PCIe 3)用作控制平面,PCIe通路1(以太网端口510)用于以太网端口1,PCIe通路2(以太网端口512)用于以太网端口3。

[0066] 在另一配置中,如图5B所示,在NVMe-oF模式下,当FPGA 500以100G(单端口)速度进行操作时,SAS端口0用于以太网端口0(以太网端口506),SAS端口1用作以太网端口2(以太网端口508)。PCIe通路504(PCIe 0)用作控制平面,PCIe通路1(端口510)用于以太网端口1,PCIe通路2(端口512)用于以太网端口3。控制平面PCIe通路514(PCIe 3)未被使用。

[0067] 在一些配置中,端口504-510可根据附接到FPGA 500的SSD的操作的模式通过中平面(例如,图4A和图4B的中平面426)连接到主板(例如,图4A的主板401或图4B的主板403)。在一些配置中,FPGA 500可用作NVMe-oF目标,并免去了在目标装置中对于主板上的基于x86的CPU的需求。对于基于NVMe-oF的系统,为了在发起者(例如,主机软件)和目标装置(即,NVMe-oF装置)之间移动数据,不再需要主板上的基于x86的CPU,因为目标装置通过它自己能够移动数据。此外,连接到FPGA 500的多个闪存驱动器(例如,图1A的两个M.3格式SSD(SSD 114和116))的不同操作速度可通过使用位于中平面或CPLD(例如,图2的CPLD 204)上的两个额外的GPIO(ESpeed[1:0])引脚(例如,图2的引脚202)来实现,其中,GPIO(ESpeed[1:0])引脚可由FPGA 500内的一个或多个内部寄存器控制。

[0068] 在一个实施例中,当以太网数据包在FPGA 500被接收时,FPGA 500内的逻辑块(诸

如,处理器子系统1、处理器子系统2、写入缓冲器(WDB)、读取缓冲器(RDB)、DDR控制器、DDR4控制器、以太网1物理层(Ethernet 1-physical layer,Eth 1-PHY)、以太网2物理层(Ethernet 2-physical layer,Eth 2-PHY)、PCIe 1物理层(PCIe 1-physical layer,PCIe 1-PHY)、PCIe 2物理层(PCIe 2-physical layer,PCIe 2-PHY)、以太网媒体访问控制(Eth-MAC)、可扩展认证协议包(extensible authentication protocol packet,EAPP)、密码认证协议包(password authentication protocol packet,PAPP)、快速远程直接存储器访问网络接口卡(express remote direct memory access network interface card,eRNIC)、NVM-oF前端(NVMF FE)、NVM-oF后端(NVMF BE)和PCIe控制器)可参与剥离那个以太网数据包中的头文件的各种层。处理器子系统可以是片上独立的冗余磁盘阵列(RoC)。例如,在以太网数据包内,可存在TCP或UDP数据包,在TCP或UDP数据包内可存在远程直接存储器访问(RDMA)数据包,在RDMA数据包内可存在NVMe-oF数据包,在NVMe-oF数据包内可存在NVMe命令。FPGA 500内的逻辑块可表示各个层,其中,在各个层中,以太网数据包被剥离出来,以暴露下一层协议。经由端口518和522连接到FPGA 500的闪存驱动器或SSD可消费NVMe命令,就好像它已经经由U.2连接器502连接到以太网端口506、508、510和512另一端的主机或主板。

[0069] 图6是示出根据一个实施例的示例FPGA装置600的框图。FPGA装置600可包括FPGA 601。FPGA 601可提供U.2连接器606和多个闪存驱动器2Gbit SPI NOR(例如,闪存驱动器603A和603B)之间的接口。例如,FPGA 601可经由可连接到PCIe X4总线的端口607与闪存驱动器603A相连接,FPGA 601可经由可连接到另一PCIe X4总线的端口608与闪存驱动器603B相连接。FPGA 601还可连接到多个DDR4存储器(例如,DDR4存储器602A-602J)。FPGA 601还连接到时钟电路CLK Gen 614。

[0070] U.2连接器606可通过两个复用器610和612经由两条PCIe X4总线604和605连接到FPGA 601。PCIe X4总线604可用于通过复用器610经由U.2连接器606通过中平面将信号或数据包发送到主板,PCIe X4总线605可用于通过复用器612经由U.2连接器606通过中平面从主板接收信号或数据包。在一些配置中,连接到FPGA 601的多个闪存驱动器(例如,闪存驱动器603A和603B)的不同操作速度可通过使用位于中平面或CPLD(例如,图2的204)上的两个额外的GPIO(ESpeed[1:0])引脚(例如,图2的引脚202)来实现,其中,GPIO(ESpeed[1:0])引脚可由FPGA 601内的一个或多个内部寄存器控制。表1示出连接到FPGA 601的闪存驱动器603A和603B的不同操作速度期间的ESpeed引脚的状态。

[0071] 图7是示出根据一个实施例的包括两个主板的示例交换机的框图。交换机700包括用于经由中平面761在双端口配置中(在HA模式下)支持多I/O的两个主板701A和701B。主板701A包括以太网交换机控制器704A和PCIe交换机705A,主板701B包括以太网交换机控制器704B和PCIe交换机705B。主板701A和701B中的每个可包括其他组件和模块(例如,如图4A所示的示例主板401中的本地CPU(CPU 706A和706B(即,CPU Skylake-D 706A和CPU Skylake-D 706B))、BMC(BMC 707A和707B)和可选中继器等。

[0072] 一些eSSD可插入到交换机700的装置端口中。例如,每个eSSD使用U.2连接器连接到交换机700。每个eSSD可连接到主板701A和主板701B二者。在本示例中,插入到交换机700中的eSSD被配置为需要经由PCIe总线和以太网端口通过中平面761与交换机700连接的NVMe-oF装置。

[0073] 图8示出根据本发明的实施例的用于操作存储装置的方法的流程图。存储装置可

以是图4A的NVMe-oF装置400。

[0074] 在801,存储装置接收第一输入。存储装置可通过中平面经由多个装置端口从主板或BMC接收第一输入。例如,NVMe-oF装置400可通过图4A的中平面426经由装置端口404-406从主板401或BMC432接收第一输入。

[0075] 在一个配置中,第一输入可通过使用主板的机箱上的物理引脚或通过来自BMC的带内命令来控制。例如,在NVMe-oF装置400接收到的第一输入可通过使用图4A的主板401的机箱上的物理引脚(E6)来控制。

[0076] 在802,基于第一输入,存储装置确定是在第一操作模式下还是在第二操作模式下进行操作。在一个配置中,存储装置的第一操作模式和第二操作模式可以是NVMe模式和NVMe-oF模式。例如,基于第一输入,NVMe-oF装置400确定是在NVMe模式还是在NVMe-oF模式下进行操作。

[0077] 在802,基于第一输入,如果存储装置确定在第一操作模式下进行操作,则在803,存储装置根据第一操作模式执行存储操作。例如,如果基于第一输入,NVMe-oF装置400确定在NVMe模式下进行操作,则NVMe-oF装置400根据NVMe协议执行存储操作。

[0078] 然而,如果在802,基于第一输入,存储装置确定在第二操作模式下进行操作,则在804,存储装置根据第二操作模式执行存储操作。例如,如果基于第一输入,NVMe-oF装置400确定在NVMe-oF模式下进行操作,则NVMe-oF装置400根据NVMe-oF协议执行存储操作。

[0079] 在805,当在第二操作模式下进行操作时,存储装置接收第二输入。存储装置可经由多个装置端口从中平面接收第二输入。例如,当在NVMe-oF操作模式下进行操作时,NVMe-oF装置可经由图4A的多个装置端口404-406从中平面426接收第二输入。

[0080] 在一个配置中,第二输入通过使用位于中平面上的两个GPIO引脚来控制,其中,两个GPIO引脚由BMC或主板的本地CPU或者存储装置的FPGA内的一个或多个内部寄存器控制。例如,第二输入可通过使用位于中平面426上的两个GPIO引脚(例如,图2的(ESpeed[1:0])引脚202)来控制,其中,两个GPIO引脚由图4A的BMC432或主板401的本地CPU 434或者存储装置400的FPGA 402内的一个或多个内部寄存器控制。

[0081] 在806,基于第二输入,存储装置从用于存储装置的多个操作速度选择操作速度。例如,基于第二输入,NVMe-oF装置400从用于NVMe-oF装置400的多个操作速度选择操作速度。例如,当在NVMe-oF模式下进行操作时,基于第二输入,NVMe-oF装置400选择100G的操作速度。在一个配置中,存储装置的多个操作速度可以是10G与100G之间的任意数字的操作速度。

[0082] 在807,当在第二模式下进行操作时,存储装置以选择的操作速度进行操作。例如,当在NVMe-oF模式下进行操作时,NVMe-oF装置400以100G的操作速度进行操作。

[0083] 将理解,尽管可在这里使用术语“第一”、“第二”、“第三”等描述各种元件、组件、区域、层和/或部分,但是这些元件、组件、区域、层和/或部分不应该受这些术语限制。这些术语只是用于将一个元件、组件、区域、层或部分与另一个区域、层或部分进行区分。因此,在不脱离本发明构思的精神和范围的情况下,这里讨论的第一元件、组件、区域、层或部分可被称为第二元件、组件、区域、层或部分。

[0084] 空间相对术语(诸如,“在…以下”、“在…下面”、“低于”、“在…之下”、“在…之上”、“上面的”等)可在这里使用以便于说明书描述在附图中示出的一个元件或特征与另外的一

个或多个元件或一个或多个特征的关系。将理解,这样的空间相对术语意图包含除了在附图中描述的方向之外的装置在使用或操作中的不同方向。例如,如果翻转在附图中的装置,则被描述为在其他元件或特征“下面”或“以下”或“之下”的元件可被定向为在其他元件或特征的“上面”。因此,示例术语“在…下面”和“在…之下”可包含上面和下面的两个方向。装置可被另外定向(旋转90度或朝向其他方向),并且这里使用的与空间相关的描述符应被相应地解释。此外,还将理解,当层被描述为在两个层“之间”时,该层可以是两个层之间的唯一层,或者一个或多个中间层也可存在。

[0085] 这里使用的术语仅是用于描述特定的实施例的目的,而不意图限制本发明构思。如这里所使用的,术语“大体上”和“大约”以及类似术语用作近似的术语而不用作程度的术语,并且意图说明将被本领域技术人员理解的测量或计算的值的固有偏差。

[0086] 如这里所使用的,除非上下文明确地另有指示,否则单数形式也意图包括复数形式。还将理解,当在本说明书中使用术语“包括”和/或“包含”时,表明描述的特征、整体、步骤、操作、元件和/或组件的存在,但不排除一个或多个其他特征、整体、步骤、操作、元件、组件和/或它们的组的存在或添加。如这里所使用的,术语“和/或”包括一个或多个关联的列出项的任何和全部组合。当诸如“中的至少一个”的表述出现在一列元素之后时,其修饰整列元素,而不是修饰列表的单个元素。此外,当描述本发明构思的实施例时的“可”的使用表示“本发明的一个或多个实施例”。此外,术语“示例性”意图指示示例或例证。如这里所使用的,术语“使用”可被认为与术语“利用”同义。

[0087] 将理解,当元件或层被称为“在……上”、“连接到”、“结合到”或“邻近于”另一个元件或层时,它能直接在所述另一个元件或层上、连接到、结合到或邻近于所述另一个元件或层,或者可存在一个或多个中间元件或层。相反,当元件被称为“直接在……上”、“直接连接到”、“直接结合到”或“直接邻近于”另一个元件或层时,不存在中间元件或层。

[0088] 这里叙述的任何数字范围意图包括包含在叙述的范围内的相同数字精度的全部子范围。例如,范围“1.0至10.0”意图包括叙述的最小值1.0和叙述的最大值10.0之间(并包含叙述的最小值1.0和叙述的最大值10.0)的全部子范围(诸如,例如,2.4至7.6),也就是说,全部子范围具有等于或大于1.0的最小值以及等于或小于10.0的最大值。这里叙述的任何最大数字限制意图包括包含在其中的任何更低数字限制,本说明书中叙述的任何最小数字限制意图包括包含在其中的任何更高数字限制。

[0089] 虽然这里已经具体描述和示出了用于支持多模式和/或多速度NVMe-oF装置的系统和方法的示例性实施例,但是各种修改和变化对本领域技术人员来说是清楚的。因此,将理解,根据本发明的原理构造的用于支持多模式和/或多速度NVMe-oF装置的系统和方法可被实现为在此具体描述的方式之外的方式。本发明还被限定在上面的权利要求书和其等同物中。

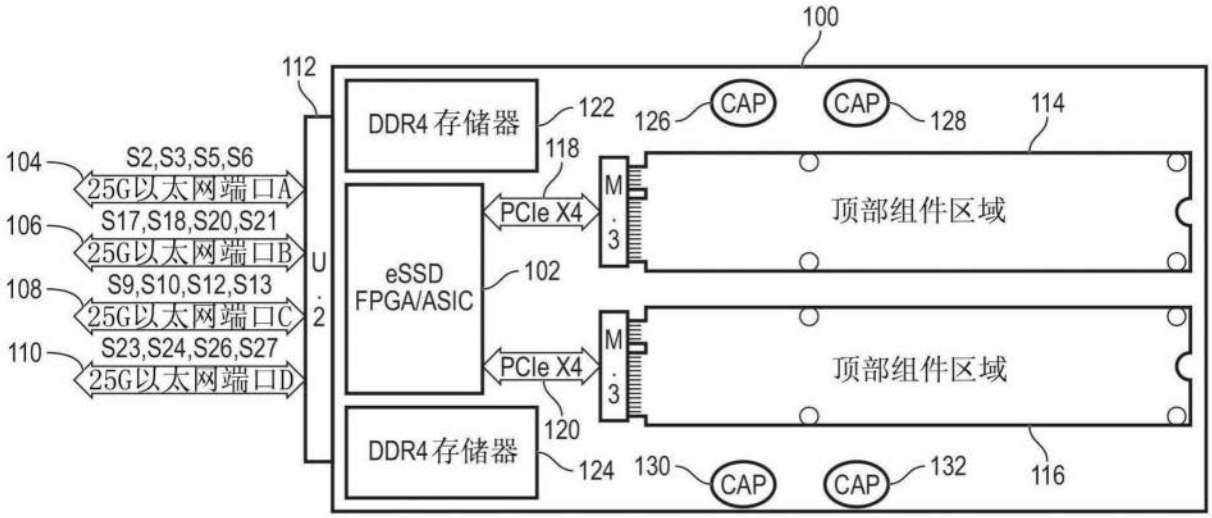

100
112
DDR4 存储器
122
126
CAP
CAP
128
114
104
S2,S3,S5,S6
25G以太网端口A
106
S17,S18,S20,S21
25G以太网端口B
108
S9,S10,S12,S13
25G以太网端口C
110
S23,S24,S26,S27
25G以太网端口D
U.2
118
PCIe X4
M.3
顶部组件区域
eSSD
FPGA/ASIC
102
PCIe X4
M.3
顶部组件区域
120
DDR4 存储器
124
130
CAP
CAP
132
116

图1A

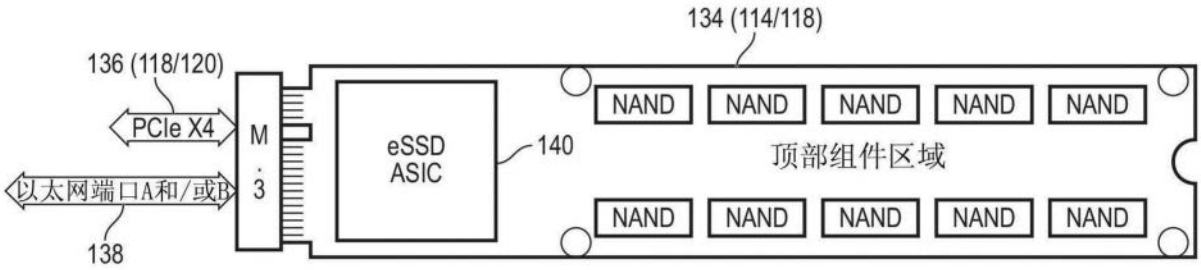

134 (114/118)
136 (118/120)
PCIe X4
以太网端口A和/或B
M.3
138
eSSD
ASIC
140
NAND
NAND
NAND
NAND
NAND
顶部组件区域
NAND
NAND
NAND
NAND
NAND

图1B

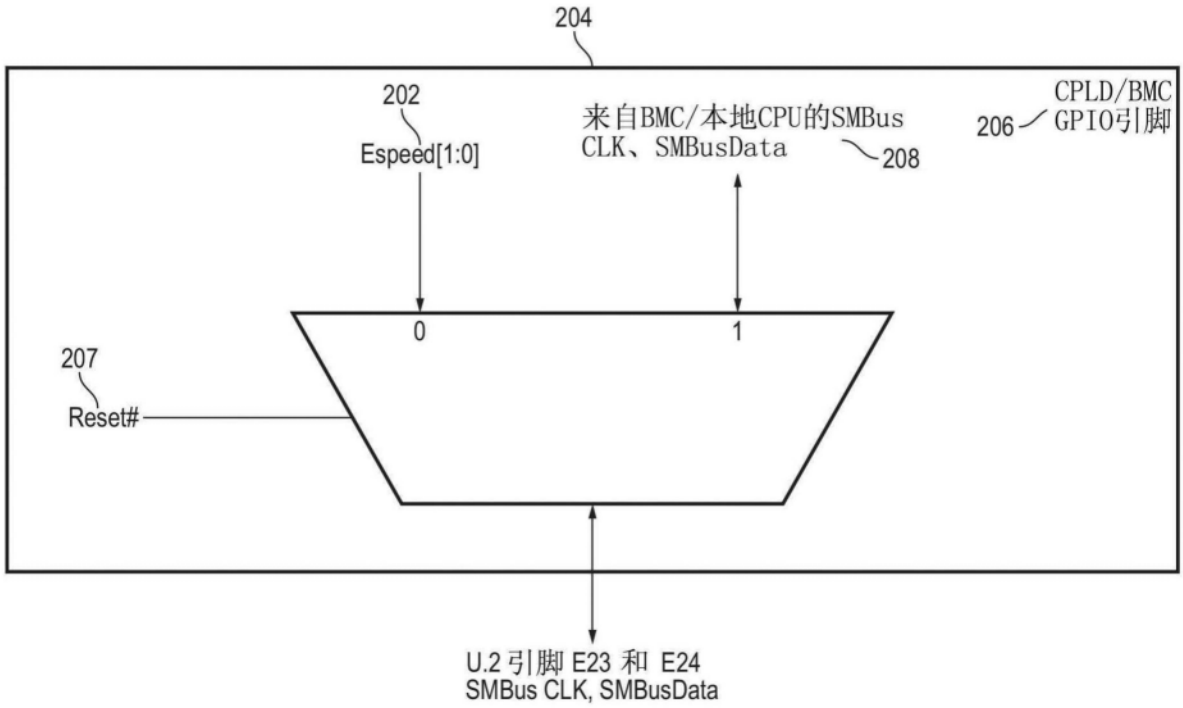
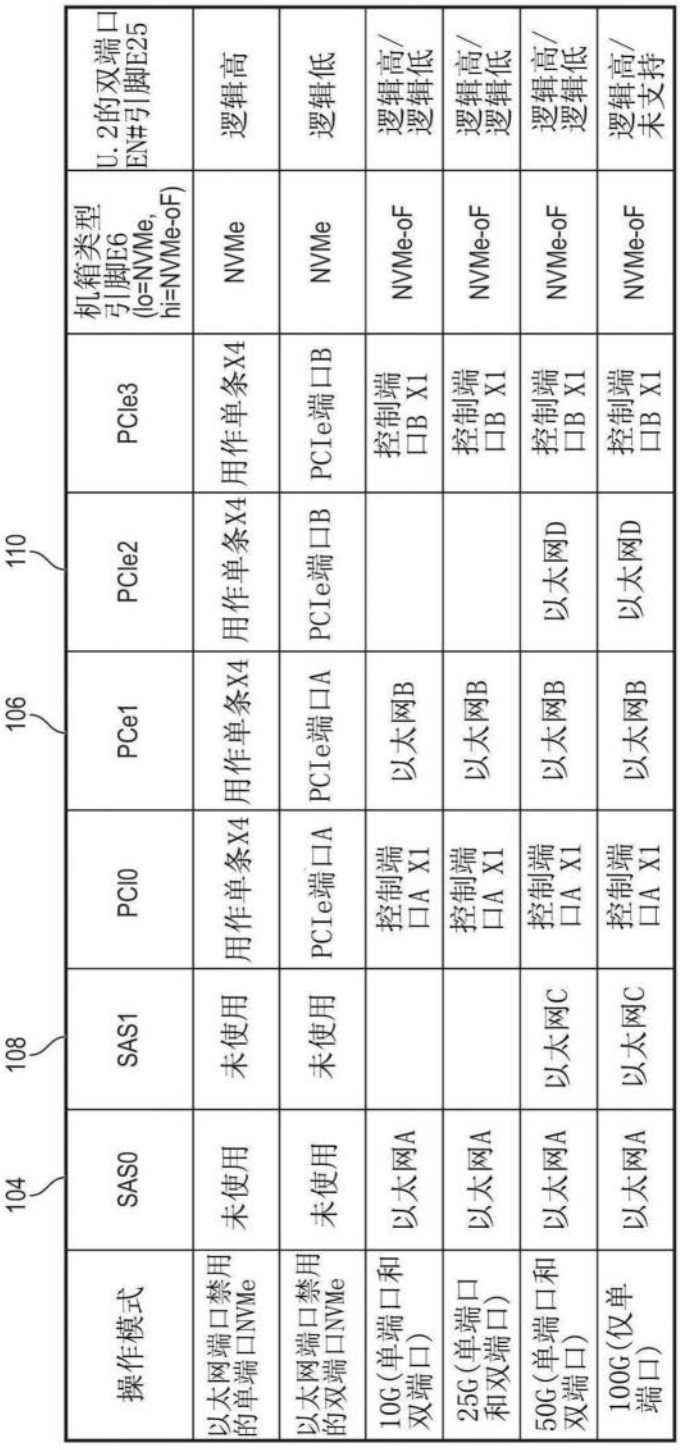

204
202
Espeed[1:0]
来自BMC/本地CPU的SMBus
CLK、SMBusData
208
CPLD/BMC
206
GPIO引脚
0
1
207
Reset#
U.2 引脚 E23 和 E24
SMBus CLK, SMBusData

图2



操作模式	SAS0	SAS1	PCI0	PCe1	PCIe2	PCIe3	机箱类型引脚E6 (lo=NVMe, hi=NVMe-oF)	U.2的双端口EN#引脚E25
以太网端口禁用的单端口NVMe	未使用	未使用	用作单条X4	用作单条X4	用作单条X4	用作单条X4	NVMe	逻辑高
以太网端口禁用的双端口NVMe	未使用	未使用	PCIe端口A	PCIe端口A	PCIe端口B	PCIe端口B	NVMe	逻辑低
10G(单端口和双端口)	以太网A		控制端口A X1	以太网B		控制端口B X1	NVMe-oF	逻辑高/逻辑低
25G(单端口和双端口)	以太网A		控制端口A X1	以太网B		控制端口B X1	NVMe-oF	逻辑高/逻辑低
50G(单端口和双端口)	以太网A	以太网C	控制端口A X1	以太网B	以太网D	控制端口B X1	NVMe-oF	逻辑高/逻辑低
100G(仅单端口)	以太网A	以太网C	控制端口A X1	以太网B	以太网D	控制端口B X1	NVMe-oF	逻辑高/未支持

图3

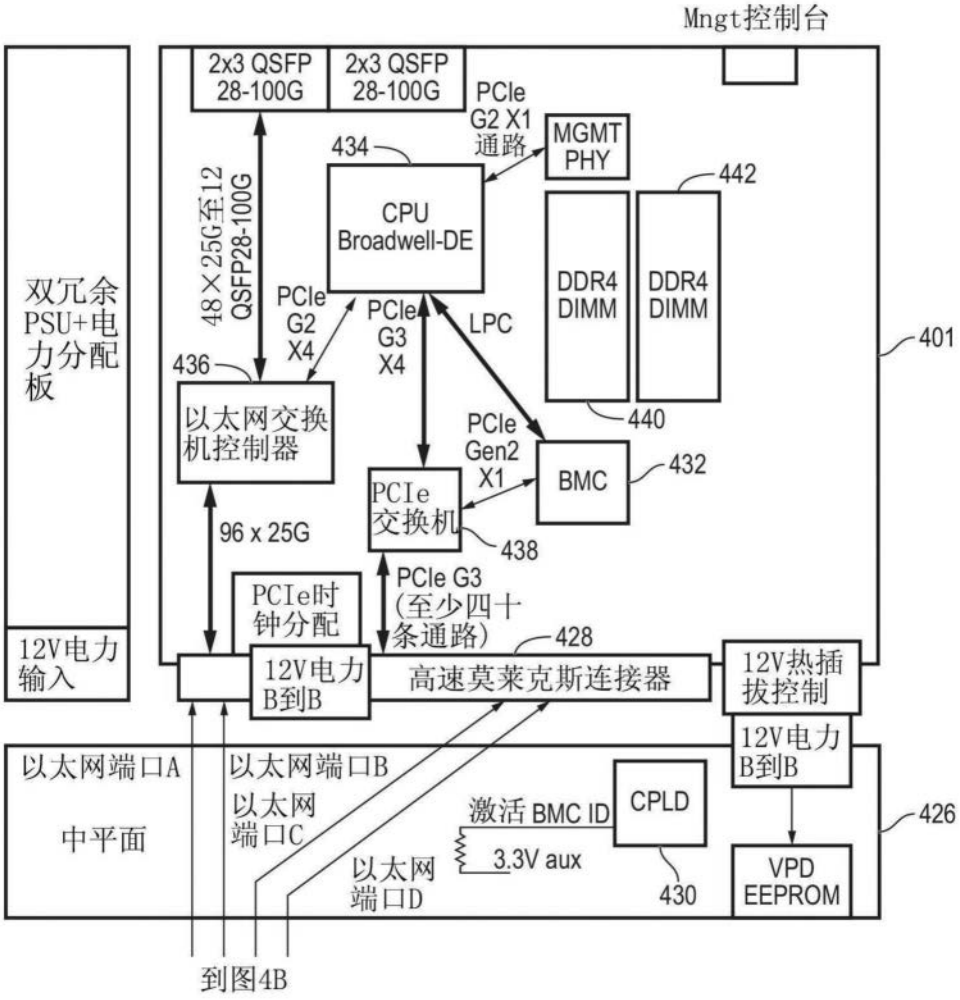

Mngt控制台
2x3 QSFP 28-100G
2x3 QSFP 28-100G
PCIe G2 X1 通路
MGMT PHY
434
CPU Broadwell-DE
442
DDR4 DIMM
DDR4 DIMM
48×25G至12 QSFP28-100G
PCIe G2 X4
PCIe G3 X4
LPC
401
双冗余 PSU+电力分配板
436
以太网交换机控制器
PCIe Gen2 X1
BMC
432
440
PCIe 交换机
438
96 x 25G
PCIe G3 (至少四十条通路)
PCIe时钟分配
428
12V电力输入
12V电力 B到B
高速莫莱克斯连接器
12V热插拔控制
12V电力 B到B
以太网端口A
以太网端口B
以太网端口C
以太网端口D
中平面
激活 BMC ID
3.3V aux
CPLD
430
VPD EEPROM
426
到图4B

图4A

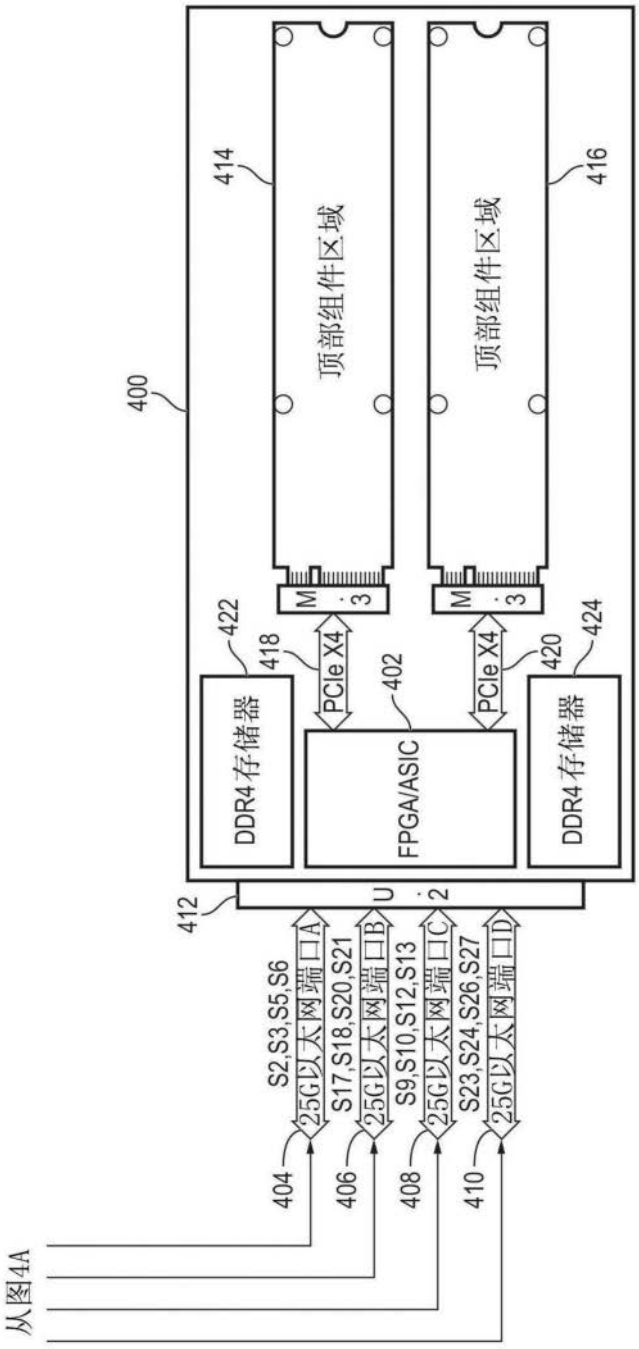

400
414
顶部组件区域
416
顶部组件区域
M.3
M.3
422
418
PCIe X4
402
PCIe X4
420
424
DDR4存储器
FPGA/ASIC
DDR4存储器
412
U.2
S2,S3,S5,S6
25G以太网端口A
S17,S18,S20,S21
25G以太网端口B
S9,S10,S12,S13
25G以太网端口C
S23,S24,S26,S27
25G以太网端口D
404
406
408
410
从图4A

图4B

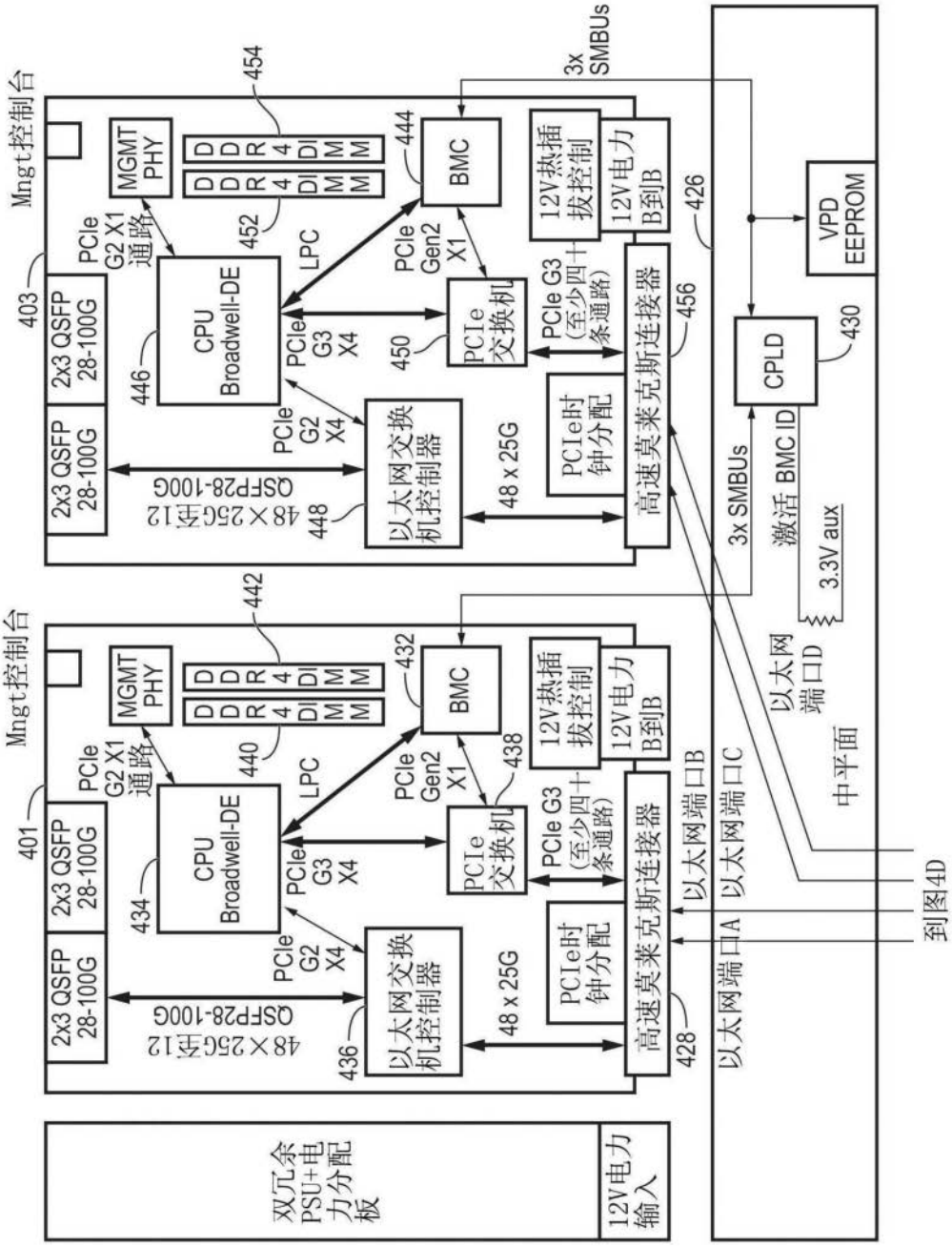

Mngt控制台
403
2x3 QSFP 28-100G
2x3 QSFP 28-100G
48×25G至12 QSFP28-100G
PCIe G2 X1 通路
MGMT PHY
DDR4 DIMM
454
452
CPU Broadwell-DE
446
LPC
PCIe G3 X4
PCIe G2 X4
448
以太网交换机控制器
PCIe Gen2 X1
444
BMC
PCIe 交换机
450
PCIe G3 (至少四十条通路)
12V热插拔控制
12V电力B到B
426
456
高速莫莱克斯连接器
48 x 25G
PCIe时钟分配
3x SMBUs
VPD EEPROM
CPLD
430
激活 BMC ID
3.3V aux
以太网端口D
中平面
以太网端口C
以太网端口A
到图4D
428
Mngt控制台
401
442
MGMT PHY
DDR4 DIMM
440
434
CPU Broadwell-DE
LPC
PCIe G3 X4
PCIe G2 X4
436
以太网交换机控制器
PCIe Gen2 X1
432
BMC
438
PCIe 交换机
PCIe G3 (至少四十条通路)
12V热插拔控制
12V电力B到B
高速莫莱克斯连接器
48 x 25G
PCIe时钟分配
以太网端口B
双冗余PSU+电力分配板
12V电力输入

图4C

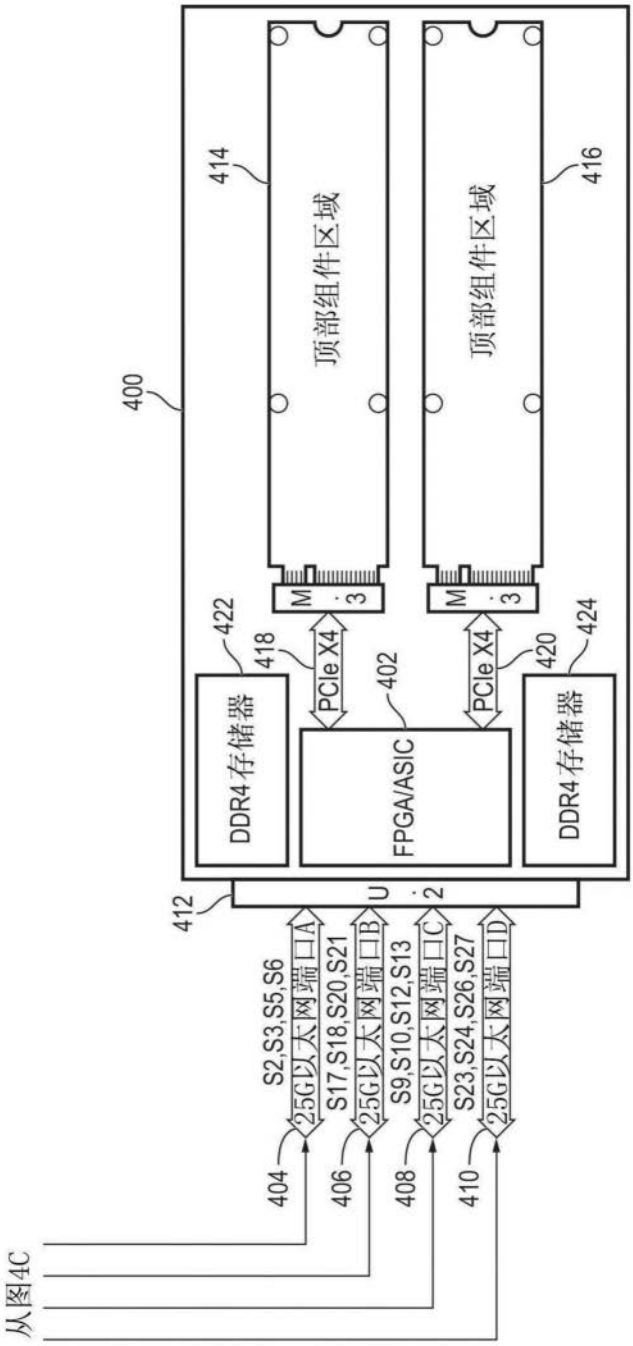

400
414
顶部组件区域
416
顶部组件区域
M.3
M.3
422
418
PCIe X4
402
PCIe X4
420
424
DDR4存储器
FPGA/ASIC
DDR4存储器
412
U.2
S2,S3,S5,S6
25G以太网端口A
S17,S18,S20,S21
25G以太网端口B
S9,S10,S12,S13
25G以太网端口C
S23,S24,S26,S27
25G以太网端口D
404
406
408
410
从图4C

图4D

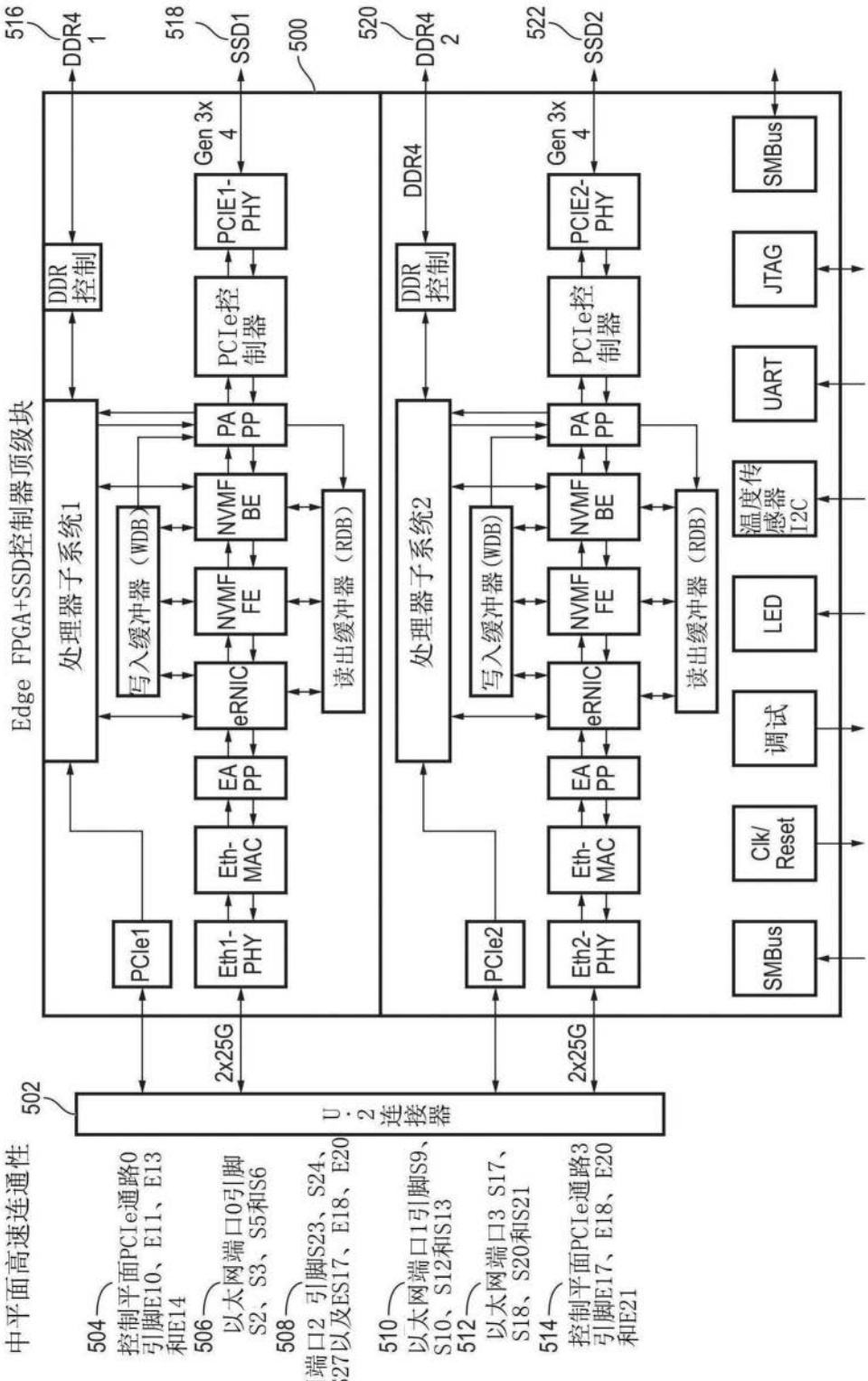

Edge FPGA+SSD控制器顶级块
处理器子系统1
处理器子系统2
DDR4 1
SSD1
DDR4 2
SSD2
500
516
518
520
522
DDR控制
PCIe控制器
PCIE1-PHY
PCIE2-PHY
Gen 3x 4
PA PP
NVMF BE
NVMF FE
eRNIC
EA PP
Eth-MAC
Eth1-PHY
Eth2-PHY
写入缓冲器（WDB）
读出缓冲器（RDB）
PCIe1
PCIe2
2x25G
502
U.2连接器
JTAG
UART
温度传感器I2C
LED
调试
Clk/Reset
SMBus
中平面高速连通性
504 控制平面PCIe通路0引脚E10、E11、E13和E14
506 以太网端口0引脚S2、S3、S5和S6
508 以太网端口2 引脚S23、S24、S26、S27以及ES17、E18、E20和E21
510 以太网端口1引脚S9、S10、S12和S13
512 以太网端口3 S17、S18、S20和S21
514 控制平面PCIe通路3引脚E17、E18、E20和E21

图5A

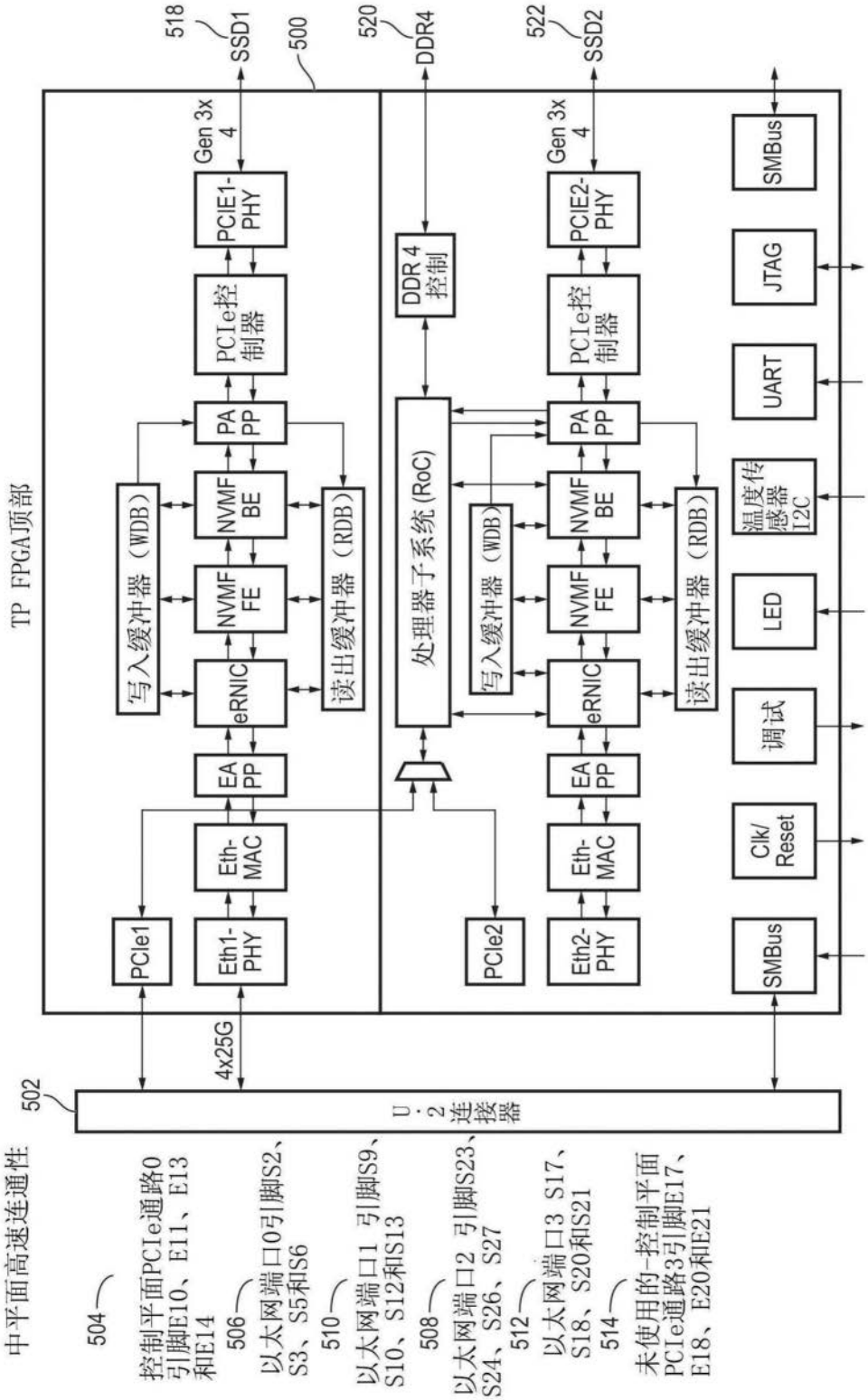

TP FPGA顶部
518
SSD1
500
520
DDR4
522
SSD2
Gen 3x 4
PCIE1-PHY
PCIe控制器
PA PP
NVMF BE
NVMF FE
eRNIC
EA PP
Eth-MAC
Eth1-PHY
写入缓冲器（WDB）
读出缓冲器（RDB）
PCIe1
4x25G
DDR 4 控制
处理器子系统(RoC)
PCIe2
Eth2-PHY
PCIE2-PHY
SMBus
JTAG
UART
温度传感器I2C
LED
调试
Clk/Reset
502
U.2连接器
中平面高速连通性
504
控制平面PCIe通路0引脚E10、E11、E13和E14
506
以太网端口0引脚S2、S3、S5和S6
510
以太网端口1 引脚S9、S10、S12和S13
508
以太网端口2 引脚S23、S24、S26、S27
512
以太网端口3 S17、S18、S20和S21
514
未使用的-控制平面PCIe通路3引脚E17、E18、E20和E21

图5B

以太网速度	ESpeed引脚1	Espeed引脚0
10G	逻辑高	逻辑高
25	逻辑高	逻辑低
50G	逻辑低	逻辑高
100G	逻辑低	逻辑低

表1

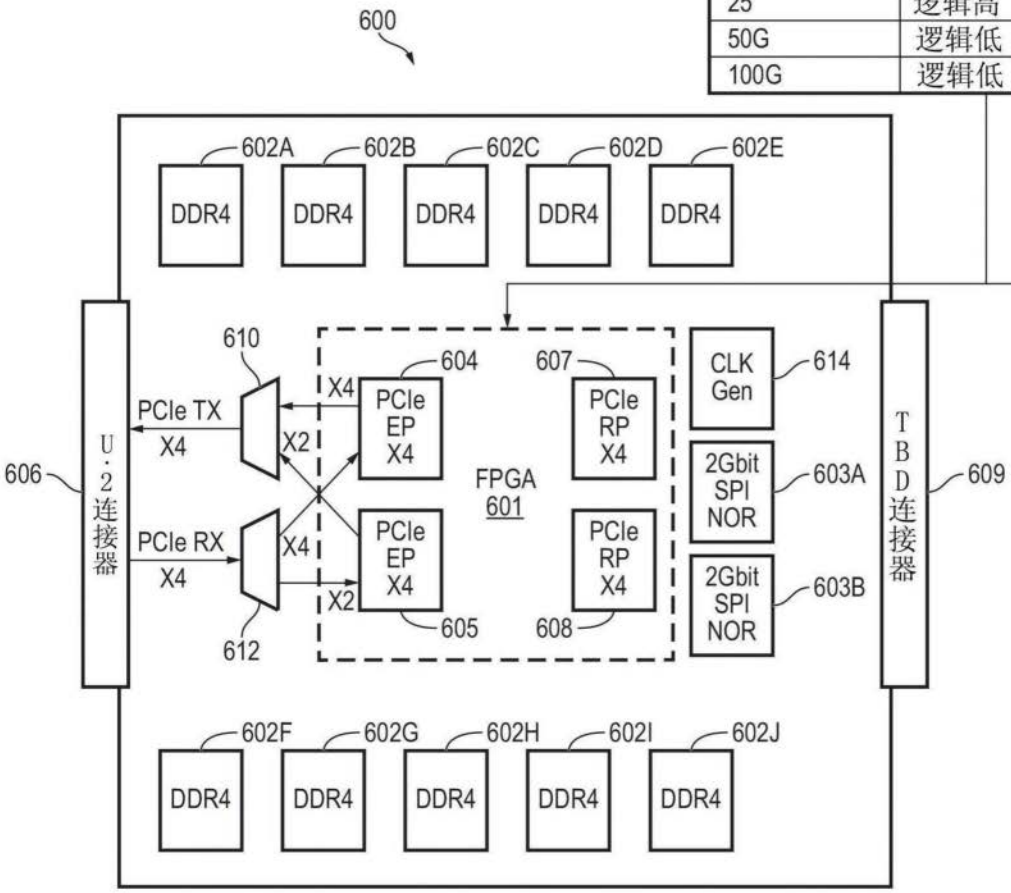


图6

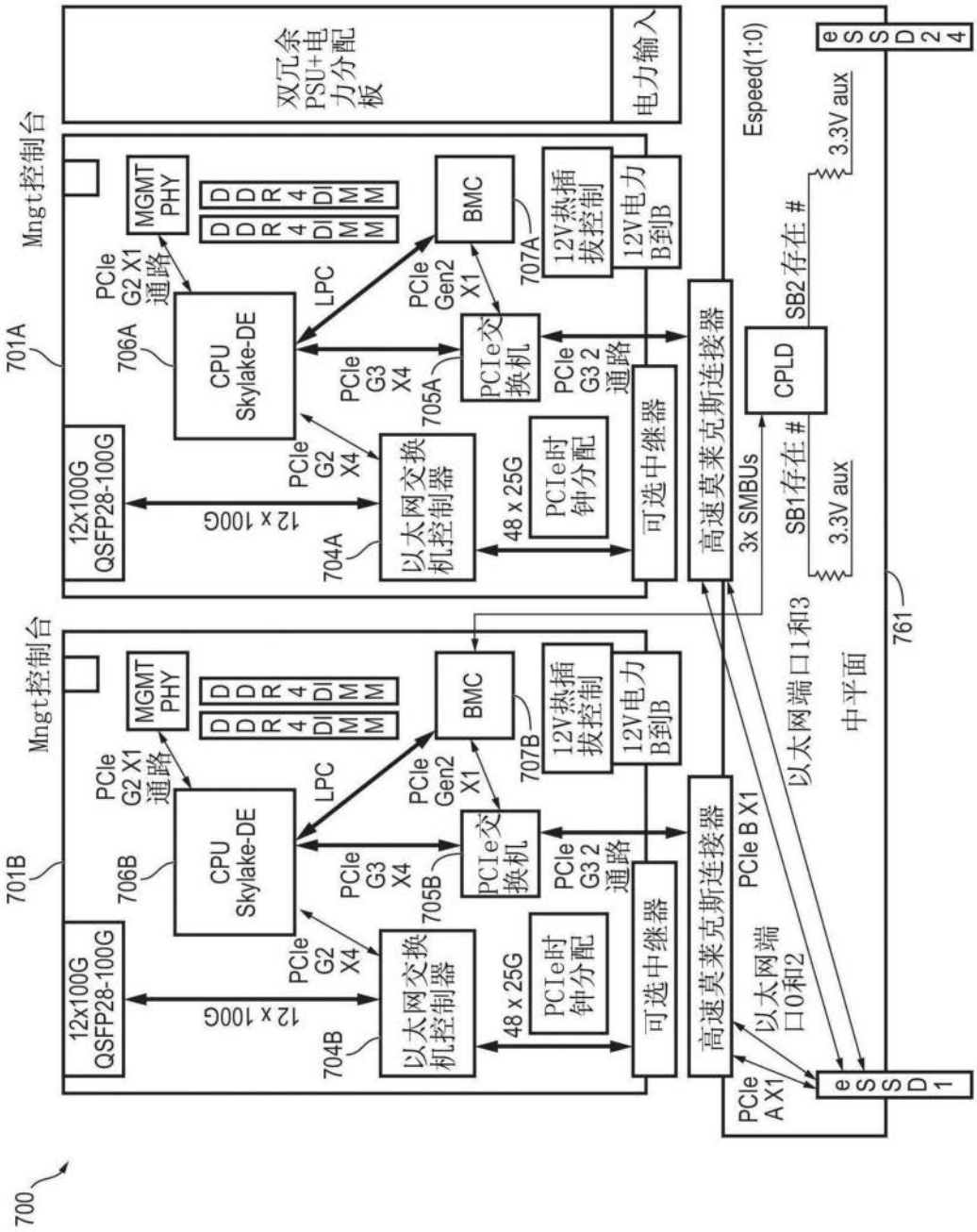

700
Mngt控制台
701A
701B
双冗余PSU+电力分配板
电力输入
MGMT PHY
DDR4 DIMM
PCIe G2 X1 通路
CPU Skylake-DE
706A
706B
LPC
BMC
707A
707B
12V热插拔控制
12V电力B到B
PCIe Gen2 X1
PCIe交换机
705A
705B
PCIe G3 X4
PCIe G3 2 通路
可选中继器
高速莫莱克斯连接器
PCIe G2 X4
以太网交换机控制器
704A
704B
12x100G QSFP28-100G
12 x 100G
48 x 25G
PCIe时钟分配
3x SMBUs
CPLD
SB1存在 #
SB2 存在 #
3.3V aux
Espeed(1:0)
761
中平面
以太网端口1和3
以太网端口0和2
PCIe B X1
PCIe A X1
eSSD24
eSSD1

图7

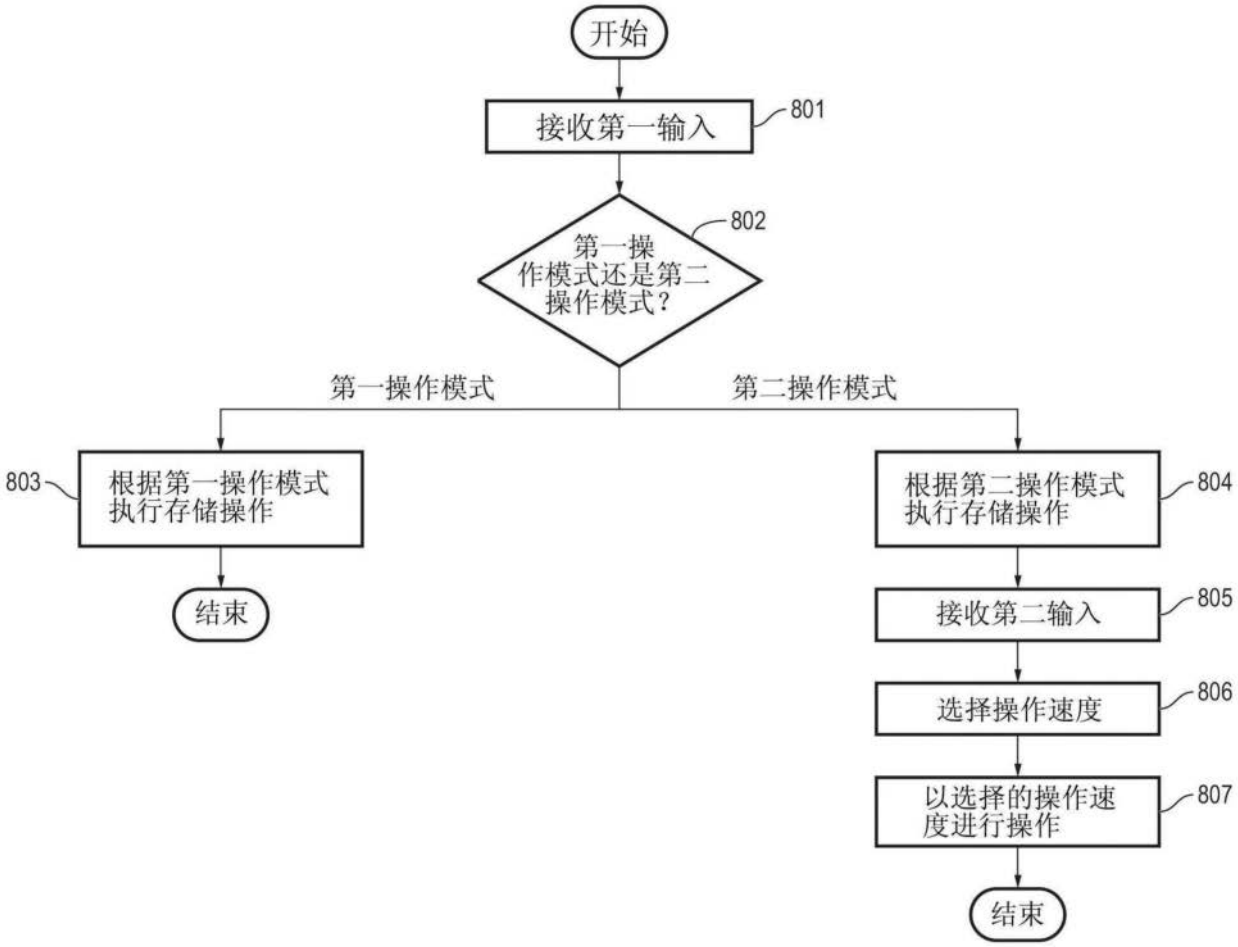

开始
接收第一输入
801
第一操作模式还是第二操作模式?
802
第一操作模式
第二操作模式
803
根据第一操作模式执行存储操作
结束
根据第二操作模式执行存储操作
804
接收第二输入
805
选择操作速度
806
以选择的操作速度进行操作
807
结束

图8