(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105589830 A
(43)申请公布日 2016.05.18

(21)申请号 201511001143.5

(22)申请日 2015.12.28

(71)申请人 浪潮(北京)电子信息产业有限公司
地址 100085 北京市海淀区上地信息路2号2-1号C栋1层

(72)发明人 王磊

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 罗满

(51)Int.Cl.
G06F 15/16(2006.01)

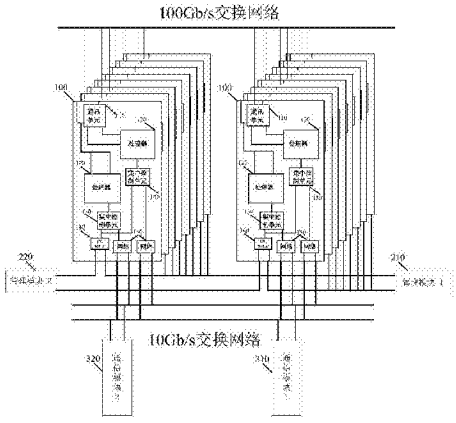
权利要求书1页 说明书5页 附图3页

(54)发明名称

一种刀片式服务器架构

(57)摘要

本发明公开了一种刀片式服务器架构，包括N个计算模块，管理模块组，通信模块组；每个计算模块包括：互为冗余的运算子系统组；每个运算子系统包括：与交换网络相连的通讯单元，与通讯单元相连的处理器，与处理器相连的集中控制单元，与集中控制单元相连的I2C管理总线切换器，与集中控制单元相连的网络传输单元；管理模块组中的每个管理模块与每个运算子系统中的I2C管理总线切换器相连；通信模块组中的每个通信模块与每个运算子系统中的网络传输单元相连，可见，在本实施例的服务器构架，满足了对大规模集中、高带宽数据应用的需求，同时通过通讯单元满足高性能计算领域对高宽带、低延时计算网络的需求。



CN 105589830 A

1.一种刀片式服务器架构,其特征在于,包括:

N个计算模块,管理模块组,通信模块组;其中,N为正整数;

每个计算模块包括:互为冗余的运算子系统组;其中,每个运算子系统包括:与交换网络相连的通讯单元,与所述通讯单元相连的处理器,与所述处理器相连的集中控制单元,与所述集中控制单元相连的I2C管理总线切换器,与所述集中控制单元相连的网络传输单元;其中,所述通讯模块支持至少100Gb的网络传输宽带;

所述管理模块组中的每个管理模块与每个运算子系统中的I2C管理总线切换器相连;所述通信模块组中的每个通信模块与每个运算子系统中的网络传输单元相连。

2.根据权利要求1所述的刀片式服务器架构,其特征在于,每个运算子系统包括:

通过PCIe3.0x2链路与处理器相连的互为冗余的存储单元组,用于存储系统的临时数据。

3.根据权利要求2所述的刀片式服务器架构,其特征在于,每个运算子系统中的处理器通过DMI链路与集中控制单元相连。

4.根据权利要求3所述的刀片式服务器架构,其特征在于,每个运算子系统中的集中控制单元通过PCIe2.0x8链路与网络传输单元相连。

5.根据权利要求4所述的刀片式服务器架构,其特征在于,每个运算子系统包括:

与集中控制单元和I2C管理总线切换器均相连的管理单元。

6.根据权利要求5所述的刀片式服务器架构,其特征在于,所述管理单元为ASP2400系列芯片组。

7.根据权利要求1所述的刀片式服务器架构,其特征在于,每个运算子系统中的处理器支持72个物理计算核心,6个内存扩展通道,提供384GB的内部数据存储空间,支持2个100Gb的高速扩展接口。

8.根据权利要求7所述的刀片式服务器架构,其特征在于,通过每颗处理器支持的2个100Gb高速扩展接口,所有运算子系统中的处理器能组建成任意拓扑形态的计算互联传输网络。

9.根据权利要求8所述的刀片式服务器架构,其特征在于,所述任意拓扑形态的计算互联传输网络,包括:

星型计算互联传输网络、环形计算互联传输网络、树型计算互联传输网络、簇星型计算互联传输网络或者网状计算互联传输网络。

10.根据权利要求9所述的刀片式服务器架构,其特征在于,每个运算子系统包括:

通过BCM54610与管理单元相连的交换单元。

一种刀片式服务器架构

技术领域

[0001] 本发明涉及计算机领域，更具体地说，涉及一种刀片式服务器架构。

背景技术

[0002] 随着大数据和高速网络的发展，数据容量正在呈几何级别的速度增长，传统的以太网的传输带宽能力和传统计算机的体系架构，已经无法满足大数据、HPC等这种对大规模集中、高带宽数据应用的需求。

[0003] 因此，如何提供一种服务器构架，以满足现有技术中大规模集中、高带宽数据应用的需求，是本领域技术人员需要解决的问题。

发明内容

[0004] 本发明的目的在于提供一种刀片式服务器架构，以满足现有技术中大规模集中、高带宽数据应用的需求。

[0005] 为实现上述目的，本发明实施例提供了如下技术方案：

[0006] 一种刀片式服务器架构，包括：

[0007] N个计算模块，管理模块组，通信模块组；其中，N为正整数；

[0008] 每个计算模块包括：互为冗余的运算子系统组；其中，每个运算子系统包括：与交换网络相连的通讯单元，与所述通讯单元相连的处理器，与所述处理器相连的集中控制单元，与所述集中控制单元相连的I2C管理总线切换器，与所述集中控制单元相连的网络传输单元；其中，所述通讯模块支持至少100Gb的网络传输宽带；

[0009] 所述管理模块组中的每个管理模块与每个运算子系统中的I2C管理总线切换器相连；所述通信模块组中的每个通信模块与每个运算子系统中的网络传输单元相连。

[0010] 其中，每个运算子系统包括：

[0011] 通过PCIe3.0x2链路与处理器相连的互为冗余的存储单元组，用于存储系统的临时数据。

[0012] 其中，每个运算子系统中的处理器通过DMI链路与集中控制单元相连。

[0013] 其中，每个运算子系统中的集中控制单元通过PCIe2.0x8链路与网络传输单元相连。

[0014] 其中，每个运算子系统包括：

[0015] 与集中控制单元和I2C管理总线切换器均相连的管理单元。

[0016] 其中，所述管理单元为ASP2400系列芯片组。

[0017] 其中，每个运算子系统中的处理器支持72个物理计算核心，6个内存扩展通道，提供384GB的内部数据存储空间，支持2个100Gb的高速扩展接口。

[0018] 其中，通过每颗处理器支持的2个100Gb高速扩展接口，所有运算子系统中的处理器能组建成任意拓扑形态的计算互联传输网络。

[0019] 其中，所述任意拓扑形态的计算互联传输网络，包括：

[0020] 星型计算互联传输网络、环形计算互联传输网络、树型计算互联传输网络、簇星型计算互联传输网络或者网状计算互联传输网络。

[0021] 其中，每个运算子系统包括：

[0022] 通过BCM54610与管理单元相连的交换单元。

[0023] 通过以上方案可知，本发明实施例提供的一种刀片式服务器架构，包括：N个计算模块，管理模块组，通信模块组；其中，N为正整数；每个计算模块包括：互为冗余的运算子系统组；其中，每个运算子系统包括：与交换网络相连的通讯单元，与所述通讯单元相连的处理器，与所述处理器相连的集中控制单元，与所述集中控制单元相连的I2C管理总线切换器，与所述集中控制单元相连的网络传输单元；其中，所述通讯模块支持至少100Gb的网络传输宽带；所述管理模块组中的每个管理模块与每个运算子系统中的I2C管理总线切换器相连；所述通信模块组中的每个通信模块与每个运算子系统中的网络传输单元相连，可见，在本实施例提供的服务器构架，满足了对大规模集中、高带宽数据应用的需求，同时通过通讯单元满足高性能计算领域对高宽带、低延时计算网络的需求。

附图说明

[0024] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0025] 图1为本发明实施例公开的一种刀片式服务器架构结构示意图；

[0026] 图2为本发明实施例公开的一种异构型计算子系统原理框图；

[0027] 图3为本发明实施例公开的一种异构计算单元模型示意图。

具体实施方式

[0028] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0029] 本发明实施例公开了一种刀片式服务器架构，以满足现有技术中大规模集中、高带宽数据应用的需求。

[0030] 参见图1，本发明实施例提供的一种刀片式服务器架构，包括：

[0031] N个计算模块100，管理模块组200，通信模块组300；其中，N为正整数；

[0032] 具体的，本实施例中的计算模块100的个数可以根据实际情况而设定，计算模块100的个数越多，其内置的处理器越多，运算能力就越强。参见图1，在本实施例中的计算模块100的个数设定为16个，即在12U的物理空间内，可以支持16个计算模块100，16个独立的计算模块100从物理形态上分为左右2组，每组各8个计算模块。

[0033] 每个计算模块100包括：互为冗余的运算子系统组；其中，每个运算子系统包括：与交换网络相连的通讯单元110，与所述通讯单元110相连的处理器120，与所述处理器120相连的集中控制单元130，与所述集中控制单元130相连的I2C管理总线切换器140，与所述集

中控制单130元相连的网络传输单元150；其中，所述通讯模块110支持至少100Gb的网络传输宽带；

[0034] 具体的，在本实施例中互为冗余的运算子系统设定为两个，即在每个计算模块中有两个互为冗余的运算子系统，每个运算子系统包括：通讯单元110、处理器120、集中控制单元130、I2C管理总线切换器140和网络传输单元150，并在在本实施例中，两个运算子系统中的通讯单元110和I2C管理总线切换器140可以共用，因此，在图1中只有一个通讯单元110和I2C管理总线切换器140。

[0035] 其中，每个运算子系统中的处理器支持72个物理计算核心，6个内存扩展通道，提供384GB的内部数据存储空间，支持2个100Gb的高速扩展接口，并且通过每颗处理器支持的2个100Gb高速扩展接口，所有运算子系统中的处理器能组建成任意拓扑形态的计算互联传输网络。

[0036] 其中，所述任意拓扑形态的计算互联传输网络，包括：

[0037] 星型计算互联传输网络、环形计算互联传输网络、树型计算互联传输网络、簇星型计算互联传输网络或者网状计算互联传输网络。

[0038] 具体的，本实施例中的有16个计算模块100，每个计算模块中有两个运算子系统，每个运算子系统中有一个处理器，因此，处理器的总数为36个。并且在本实施例中的每颗独立的处理器支持72个物理计算核心、6个内存扩展通道，可以提供384GB的内部数据存储空间，同时每颗处理器最大支持2*100Gb的高速扩展接口，通过每颗处理器支持的2*100Gb高速扩展接口，所有运算子系统中的处理器可以组建成任意拓扑形态的计算互联传输网络，可以是星型网络、环形网络、树型网、簇星型网或网状网等根据客户需求的任意形态，网络基于目前业界最领先的100Gb高速网络，可以满足高性能运算领域对高带宽、低延迟计算网络的需求。

[0039] 所述管理模块组200中的每个管理模块与每个运算子系统中的I2C管理总线切换器140相连；所述通信模块组300中的每个通信模块与每个运算子系统中的网络传输单元150相连。

[0040] 具体的，由于本实施例中的每个计算模块100中包括两个运算子系统，因此本实施例中的管理模块组200的个数为两个，即图1中的管理模块210和管理模块220，同理通讯模块组300的个数也为两个，即通信模块310和通信模块320。

[0041] 参见图2，基于上述技术方案，本实施例提供一种异构型计算子系统原理框图，可见，在本实施例中，每个运算子系统不仅包括：通讯单元110，与所述通讯单元110相连的处理器120，与所述处理器120相连的集中控制单元130，与所述集中控制单元130相连的I2C管理总线切换器140，与所述集中控制单130元相连的网络传输单元150；

[0042] 还包括：通过PCIe3.0x2链路与处理器120相连的互为冗余的存储单元组160，用于存储系统的临时数据。

[0043] 具体的，本实施例中的每颗独立的处理器120支持36Lane的PCIe 3.0扩展。本实施例中每个运算子系统包括两个存储单元，这两个存储单元分别通过1组PCIe 3.0x2的链路与处理器120连接，用来存储系统中处理的临时数据，同时支持软RAID0和1，可以提供冗余的数据保护。其中，每个运算子系统中的处理器120通过DMI链路与集中控制单元相连，每个运算子系统中的集中控制单元130通过PCIe2.0x8链路与网络传输单元150相连。

[0044] 还包括:与集中控制单元130和I2C管理总线切换器140均相连的管理单元170,所述管理单元170为ASP2400系列芯片组。

[0045] 具体的,在本实施例中的每个处理器通过DMI的链路连接集中控制单元,集中控制单元主要负责系统中低速设备的控制,主要通过PCIe 2.0x8的链路连接网络传输单元,通过PCIe 2.0x1和LPC链路与管理单元连接。管理单元采集单元用ASP2400系列芯片组,用来负责控制模块上所有器件温度、电压的监控。

[0046] 还包括:通过BCM54610与管理单元170相连的交换单元180。

[0047] 具体的,在本实施例中每个运算子单元中的网络传输单元会与架构中的交换模块进行连接,实现每个子计算系统与外界设备的通信。

[0048] 其中BCM54610是一个物理的PHY芯片,需要将管理单元170自身的网络信号转换成需要的Serdes类型的信号并与通讯模块组300中相应的通讯模块进行连接。

[0049] 由图2中的异构型计算子系统的原理框图可知,在每个计算模块中设计支持2个独立的运算子系统,在每个独立的系统中会由通讯单元110、处理器120、集中控制单元130、I2C管理总线切换器140、网络传输单元150、存储单元组160、管理单元170和交换单元180组成,其中每个系统采用1颗浪潮最新研制的基于异构计算的处理器,每颗独立的处理器支持72个物理计算核心、6个内存扩展通道、支持36Lane的PCIe 3.0扩展。同时每个处理器可以支持2个100Gb的高速信号,通过通讯单元110与外部提供高速的数据通信通道。在每个计算模块中共有2个通讯单元110,每个通讯单元100对应的1个处理器130,每个通讯单元100上含有2个QSFP28的接口。

[0050] I2C管理总线切换器140是一个四路输入、两路选择输出的模块,接收两个管理单元170的2组I2C链路输入,2组输出后与系统中的管理模块组200中相应的管理模块进行连接。交换单元180其中4个端口通过BCM54610与管理单元140连接,BCM54610是一个物理的PHY芯片,需要将管理单元自身的网络信号转换成我们需要的Serdes类型的信号与图1中的通讯模块进行连接。

[0051] 具体的,在图2中的BIOS Flash410为每个计算子节点上的一个Flash芯片,主要启动对系统的基本输入和输出控制,即系统底层硬件层面的配置及管理。前控模板420中为显示连接的鼠标、键盘是否工作的指示灯,从而能及时了解各个器件是否正常工作。

[0052] 参见图3,为本实施例提供的异构计算单元模型示意图,图中中部位置A处,有2个处理器,每个处理器支持72个物理计算核心、6个内存扩展通道和2个存储磁盘介质;图中前部位置B处设计2组风扇模块用来给2颗处理器提供散热和2块100Gb高速信号传输模块,此模块采用上下双层的布局设置、分别通过Cable与每颗处理器进行连接;图中后部设计2个I/O应用模块,分别通过1组PCIex16的信号与每颗处理器连接,用来最为不同IO应用的扩展支持,可以支持以太网、FC网络或Infiniband网络的支持。位置C为扩展卡,系统前端支持2块半高半长模块,支持Infiniband EDR,支持100Gb卡,每卡2*100Gb端口;位置D为存储模块,支持4个M.2存储(Socket2接口),且单运算单元2个M.2,位置E为内存模组,12个DDR4DIMM,单处理器支持6个DIMM。

[0053] 本发明实施例提供的一种刀片式服务器架构,包括:N个计算模块,管理模块组,通信模块组;其中,N为正整数;每个计算模块包括:互为冗余的运算子系统组;其中,每个运算子系统包括:与交换网络相连的通讯单元,与所述通讯单元相连的处理器,与所述处理器相

连的集中控制单元，与所述集中控制单元相连的I2C管理总线切换器，与所述集中控制单元相连的网络传输单元；所述管理模块组中的每个管理模块与每个运算子系统中的I2C管理总线切换器相连；所述通信模块组中的每个通信模块与每个运算子系统中的网络传输单元相连，可见，在本实施例提供的服务器构架，满足了对大规模集中、高带宽数据应用的需求，同时通过通讯单元满足高性能计算领域对高宽带、低延时计算网络的需求。

[0054] 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。

[0055] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

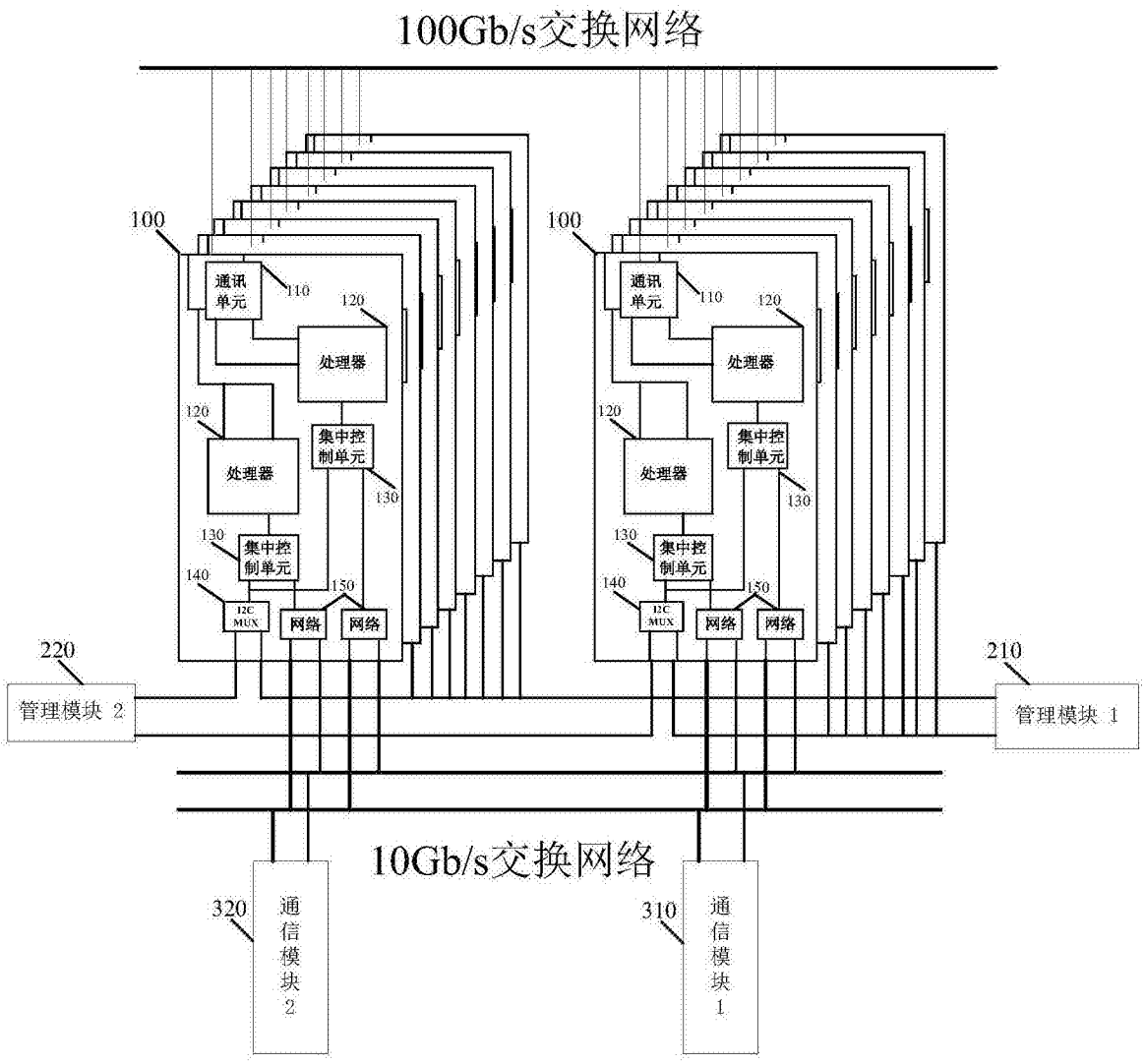

100Gb/s交换网络
100
通讯单元
110
120
处理器
120
处理器
集中控制单元
130
130
集中控制单元
140
I2C MUX
150
网络
网络
100
通讯单元
110
120
处理器
120
处理器
集中控制单元
130
130
集中控制单元
140
I2C MUX
150
网络
网络
220
管理模块 2
210
管理模块 1
10Gb/s交换网络
320
通信模块2
310
通信模块1

图1

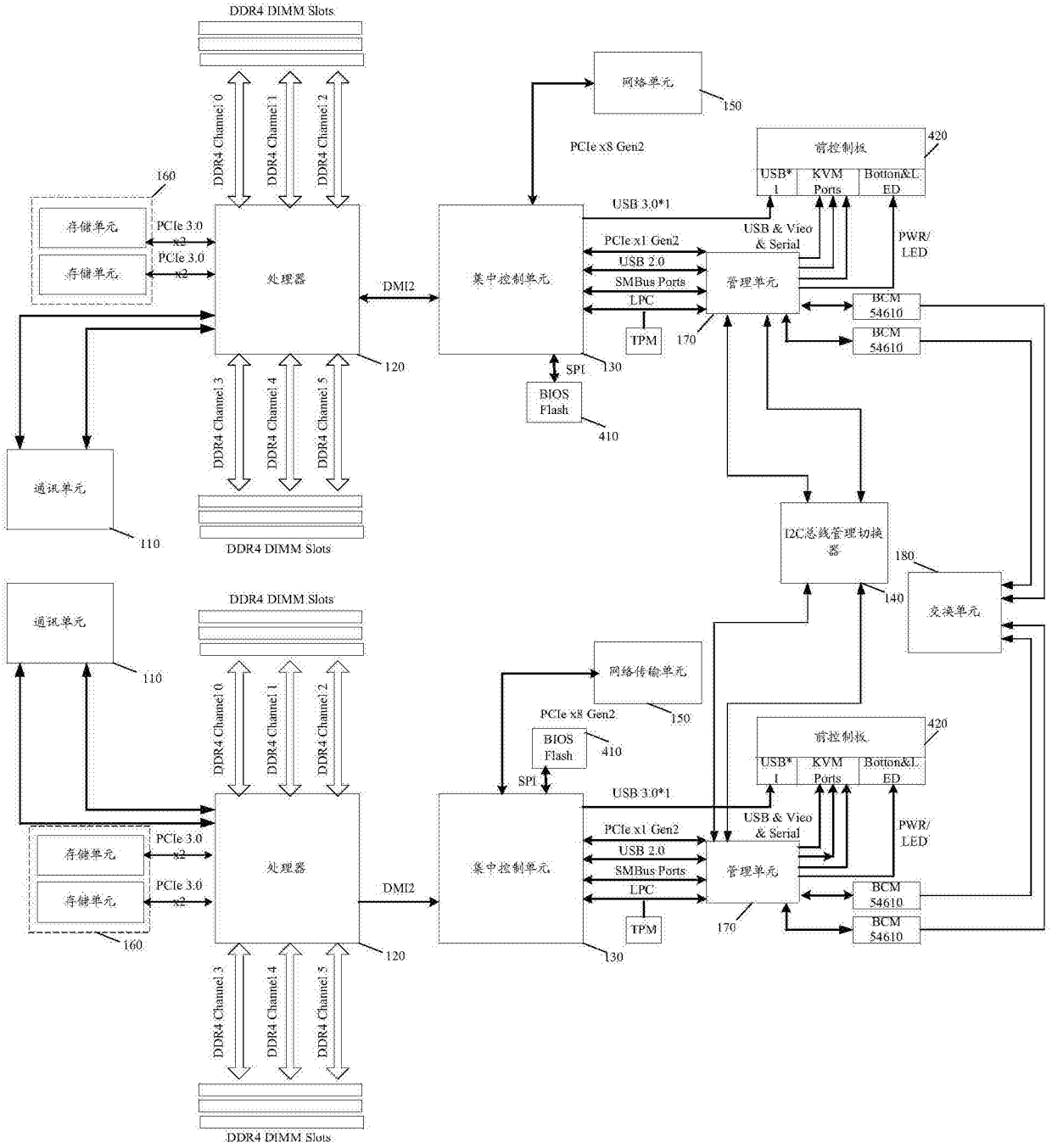

DDR4 DIMM Slots
DDR4 Channel 0
DDR4 Channel 1
DDR4 Channel 2
DDR4 Channel 3
DDR4 Channel 4
DDR4 Channel 5
DDR4 DIMM Slots
160
存储单元
存储单元
PCIe 3.0 x2
PCIe 3.0 x2
处理器
DMI2
120
通讯单元
110
集中控制单元
网络单元
150
PCIe x8 Gen2
前控制板
420
USB* 1
KVM Ports
Botton&LED
USB 3.0*1
PCIe x1 Gen2
USB 2.0
SMBus Ports
LPC
USB & Vieo & Serial
PWR/LED
管理单元
TPM
170
BCM 54610
BCM 54610
SPI
130
BIOS Flash
410
I2C总线管理切换器
180
140
交换单元
网络传输单元
150
PCIe x8 Gen2
BIOS Flash
410
SPI
USB 3.0*1
PCIe x1 Gen2
USB 2.0
SMBus Ports
LPC
USB & Vieo & Serial
PWR/LED
管理单元
TPM
170
BCM 54610
BCM 54610
130
120
160
110
DDR4 DIMM Slots

图2

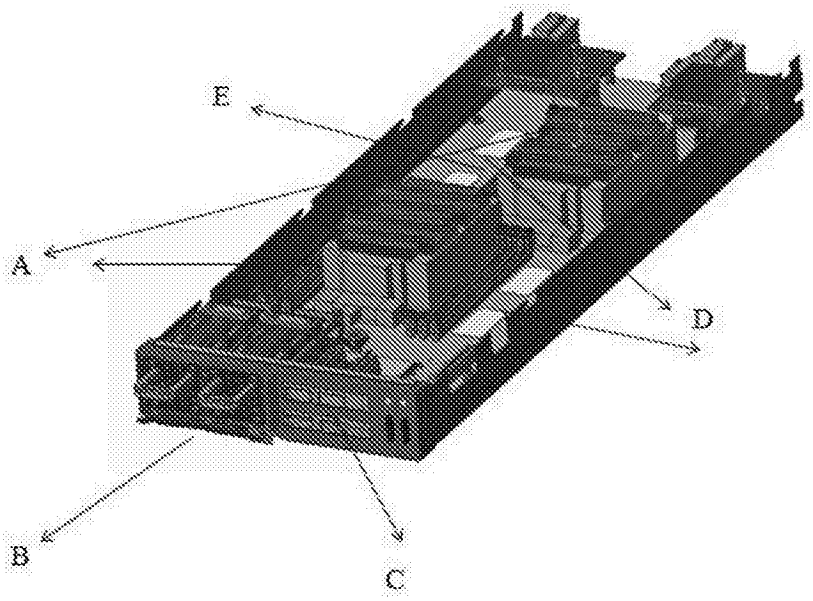


图3