(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 102929363 A
(43)申请公布日 2013.02.13

(21)申请号 201210411662.9

(22)申请日 2012.10.25

(71)申请人 浪潮电子信息产业股份有限公司
地址 250014 山东省济南市高新区舜雅路 1036 号

(72)发明人 王磊 王守昊

(51)Int.Cl.
G06F 1/18(2006.01)
G06F 1/16(2006.01)

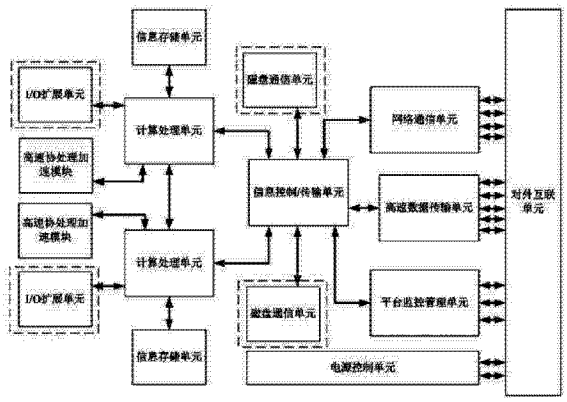
权利要求书 2 页 说明书 4 页 附图 2 页

(54)发明名称

一种高密度刀片服务器的设计方法

(57)摘要

本发明提供一种高密度刀片服务器的设计方法，该方法区别于传统的以处理器为中心的刀片服务器体系结构，是一种即能支持高速协处理器，又同时具备支持本地数据存储及丰富 I/O 扩展功能的高密度刀片服务器体系架构设计，在新的体系结构中，打破了原有刀片服务器体系架构只以通用处理器单元为中心并且 I/O 扩展性差的限制，通过创新设计基于通用处理器的高速转换模块作为通用处理单元与高速协处理器及 I/O 扩展单元之间的转换桥梁，可以在统一的体系架构中不但实现支持高速协处理器，以大幅度提升单个刀片服务器自身的信息及运算处理能力，避免原有服务器架构中通用处理器浮点运算能力低的问题。



CN 102929363 A

1. 一种高密度刀片服务器的设计方法， 其特征在于该系统包括：计算处理单元、信息存储单元、高速转换单元、磁盘通信单元、信息控制 / 传输单元、网络通信单元、高速数据传输单元、平台监控管理单元、电源控制单元和对外互联单元其中：

计算处理单元，主要采用业界通用的处理器设备，负责平台基本数据的运算、控制信息的分析及处理、控制命令的接收及发布，计算处理器单元间通过传输速率高达 8.0GT/s 高速的 QPI 传输链路实现两个计算处里单元间信息的共享、通讯及处理；

信息存储单元，直接与计算处理器单元通信，每个计算处理单元都具备独立的 4 个信息存书单元，每个信息存储单元最大能可设计支持 3 个信息扩展模块，每个信息扩展模块能支持业界通用的容量为 8GB、16GB 以及 32GB 的内存储模组，信息存储单元作为计算处理单元的信息及数据存储仓库；

高速转换单元，在整个体系架构中起着重要的作用，高速转换单元作为通用处理单元与高速协处理器及 I/O 扩展单元之间的转换桥梁，直接与每个通用计算处里单元连接，为计算处里单元提供双向 32GB/s 的通讯带宽，通过高速转换单元平台支持高速协处理器，计算单元将计算数据和任务通过高速转换单元传送到高速协处理器，利用高速协处理器计算核心多、计算频率高的特点处理复杂及大数据量得数据，最终运算完成得数据高速协处理器又通过高速转换单元传往通用处理器，同时通过高速转换模块，系统还支持通用的基于 PCIe 传输信道的 I/O 扩展设备，不但支持像 Infiniband、万兆以太网这样偏重高速运算的应用需求，还满足 SATA、SAS 及 FC 这种偏重数据存储应用的需求；

磁盘通信单元，提供本地数据的存储，主要存放通常计算处里单元访问频率不高的数据，磁盘通信单元为本地数据的储存提供 6Gb/s 的传输链路，磁盘通信单元采用基于灵活配置的设计方案，根据系统的需求进行独立的安装与拆卸，当使用高速协处理器时，系统面向高性能的数据传输应用，数据直接通过高速数据传输单元与外界系统交互，因此系统将不提供对磁盘通信单元的支持；当使用 I/O 扩展单元时，系统将提供对磁盘通讯单元的支持，用于低速数据的存储；

信息控制 / 传输单元，作为整个平台的传输控制中枢，负责计算处理单元与磁盘通信单元、网络通信单元、高速数据传输单元、平台监控管理单元间的通信；

网络通信单元与高速数据传输单元，作为整个平台与外部系统的通信桥梁，负责将平台的数据传送的外界平台以及接收外界平台发送的数据及运算任务；其中网络通信单元采用基于以太网作为通讯通路，提供两条 1Gb/s 的传输链路；高速数据传输单元采用基于 Infiniband 作为高速通讯通路，具备高带宽和低传输延时的特点，提供 56Gb/s 的高速传输链路，当系统面向高性能的数据传输应用时，为高速协处理单元和通用计算单元提供一条与外界高速数据传输的通路；

平台监控管理单元，负责对信息处理单元、高速信息交换单元、IO 扩展模块等系统中各模块状态的监控和配置管理；

高运算性的实现步骤如下： 在支持通用处理单元的基础上，设计高速转换模块通过 PCIe3.0 总线直接与每个通用计算处里单元连接，为计算处里单元提供双向 32GB/s 的通讯带宽，使平台支持高速的协处理器，通过扩展支持协处理器平台可以提供每秒 1.2 万亿次的浮点运算能力；

高可扩展的实现步骤如下：高速转换模块将原有的高速协处理单元换成基于 PCI-E 总

线的 IO 扩展单元，为信息处理单元方便的进行 IO 方面的扩展，包括 HCA 卡、SAS RAID 卡、万兆光纤网卡、图形处理卡，从而提高系统整体的可扩展性。

一种高密度刀片服务器的设计方法

技术领域

[0001] 本发明涉及计算机通信领域，具体涉及一种支持高速协处理器、本地存储及 IO 扩展的高密度刀片服务器系统的设计方法。

背景技术

[0002] 高性能计算技术的发展是伴随着计算机技术的发展而发展的，也就是说，从计算机技术诞生之日起，人们就在为追求更高计算能力的计算机系统而努力。在过去几十年间，可以说是高性能计算机体系结构和通信技术不断创新的年代，出现了包括 MPP、SMP、集群等各种各样的体系结构及网络互联技术。尤其是最近几年，集群技术发展迅速，已经成为构建超级计算机系统的主流架构之一。

[0003] GPU 计算是指把图形处理器 (GPU) 用作协处理器来为 CPU 加速，从而为通用科学和工程计算服务。一颗 CPU 包含四到八个 CPU 核心，而一颗 GPU 却包含数百个尺寸更小的核心。它们在应用程序中共同处理数据。 正是这种大规模并行架构让 GPU 能够拥有极高的计算性能。GPU 计算是指把图形处理器 (GPU) 用作协处理器来为 CPU 加速，从而为通用科学和工程计算服务。GPU 通过承担部分运算量繁重且耗时的代码，从而为那些在 CPU 上运行的应用程序加速。应用程序的剩余部分仍然交由 CPU 处理。 从用户的角度来看，应用程序之所以能更快速地运行是因为使用了 GPU 的大规模并行处理能力来提升性能。 这种方式就叫做「 异构」或「 混合型」计算。

[0004] 目前刀片服务器产品由于计算节点密度大、集成管理、交换等应用的特点成为搭建高性能集群的首先硬件平台。然而，在日益增长的高性能商业计算应用领域中，对系统的计算能力和扩展性提出了更高的要求，但现有刀片服务器体系结构受制于空间限制只以通用处理器单元为中心并且 I/O 扩展性差，在现有刀片服务器系统体系结构的基础上，提出一种支持高速协处理器、本地存储及 IO 扩展的高密度刀片服务器架构设计。

发明内容

[0005] 本发明的目的是提供一种高密度刀片服务器的设计方法。

[0006] 本发明的目的是按以下方式实现的，该系统包括 :计算处理单元、信息存储单元、高速转换单元、磁盘通信单元、信息控制 / 传输单元、网络通信单元、高速数据传输单元、平台监控管理单元、电源控制单元和对外互联单元其中 :

计算处理单元，主要采用业界通用的处理器设备，负责平台基本数据的运算、控制信息的分析及处理、控制命令的接收及发布，计算处理器单元间通过传输速率高达 8.0GT/s 高速的 QPI 传输链路实现两个计算处里单元间信息的共享、通讯及处理 ;

信息存储单元，直接与计算处理器单元通信，每个计算处理单元都具备独立的 4 个信息存书单元，每个信息存储单元最大能可设计支持 3 个信息扩展模块，每个信息扩展模块能支持业界通用的容量为 8GB、16GB 以及 32GB 的内存储模组，信息存储单元作为计算处理单元的信息及数据存储仓库 ;

高速转换单元，在整个体系架构中起着重要的作用，高速转换单元作为通用处理单元与高速协处理器及 I/O 扩展单元之间的转换桥梁，直接与每个通用计算处里单元连接，为计算处里单元提供双向 32GB/s 的通讯带宽，通过高速转换单元平台支持高速协处理器，计算单元将计算数据和任务通过高速转换单元传送到高速协处理器，利用高速协处理器计算核心多、计算频率高的特点处理复杂及大数据量得数据，最终运算完成得数据高速协处理器又通过高速转换单元传往通用处理器，同时通过高速转换模块，系统还支持通用的基于 PCIe 传输信道的 I/O 扩展设备，不但支持像无限宽带 Infiniband、万兆以太网这样偏重高速运算的应用需求，还满足 SATA、SAS 及 FC 这种偏重数据存储应用的需求；

磁盘通信单元，提供本地数据的存储，主要存放通常计算处里单元访问频率不高的数据，磁盘通信单元为本地数据的储存提供 6Gb/s 的传输链路，磁盘通信单元采用基于灵活配置的设计方案，根据系统的需求进行独立的安装与拆卸，当使用高速协处理器时，系统面向高性能的数据传输应用，数据直接通过高速数据传输单元与外界系统交互，因此系统将不提供对磁盘通信单元的支持；当使用 I/O 扩展单元时，系统将提供对磁盘通讯单元的支持，用于低速数据的存储；

信息控制 / 传输单元，作为整个平台的传输控制中枢，负责计算处理单元与磁盘通信单元、网络通信单元、高速数据传输单元、平台监控管理单元间的通信；

网络通信单元与高速数据传输单元，作为整个平台与外部系统的通信桥梁，负责将平台的数据传送的外界平台以及接收外界平台发送的数据及运算任务；其中网络通信单元采用基于以太网作为通讯通路，提供两条 1Gb/s 的传输链路；高速数据传输单元采用基于无限宽带 Infiniband 作为高速通讯通路，具备高带宽和低传输延时的特点，提供 56Gb/s 的高速传输链路，当系统面向高性能的数据传输应用时，为高速协处理单元和通用计算单元提供一条与外界高速数据传输的通路；

平台监控管理单元，负责对信息处理单元、高速信息交换单元、IO 扩展模块等系统中各模块状态的监控和配置管理；

高运算性的实现步骤如下： 在支持通用处理单元的基础上，设计高速转换模块通过 PCIe3.0 总线直接与每个通用计算处里单元连接，为计算处里单元提供双向 32GB/s 的通讯带宽，使平台支持高速的协处理器，通过扩展支持协处理器平台可以提供每秒 1.2 万亿次的浮点运算能力；

高可扩展的实现步骤如下：高速转换模块将原有的高速协处理单元换成基于 PCI-E 总线的 IO 扩展单元，为信息处理单元方便的进行 IO 方面的扩展，包括 HCA 卡、SAS RAID 卡、万兆光纤网卡、图形处理卡，从而提高系统整体的可扩展性。

[0007] 本发明的有益效果是：本发明区别于传统的以处理器为中心的刀片服务器体系结构，提出一种即能支持高速协处理器，又同时具备支持本地数据存储及丰富 I/O 扩展功能的高密度刀片服务器体系架构设计，在新的体系结构中，打破了原有刀片服务器体系架构只以通用处理器单元为中心并且 I/O 扩展性差的限制，通过创新设计基于通用处理器的高速转换模块作为通用处理单元与高速协处理器及 I/O 扩展单元之间的转换桥梁，可以在统一的体系架构中不但实现支持高速协处理器，以大幅度提升单个刀片服务器自身的信息及运算处理能力，避免原有服务器架构中通用处理器浮点运算能力低的问题。

[0008] 在整个体系架构中还单独设计了高速数据传输单元为配合高速协处理器的使用。

同时系统还可以支持通用的基于PCIe传输信道的I/O扩展设备，使新的刀片服务器体系架构具备很强的扩展性，不但可以支持像无限宽带 Infiniband、万兆以太网这样偏重高速运算的应用需求，还可以满足SATA、SAS及FC这种偏重数据存储应用的需求，完全突破了原有刀片服务器架构受空间限制的难题，提出全新支持高速新处理器、本次存储及I/O扩展的告密度刀片服务器架构体系，通过创新设计基于通用处理器的高速转换模块作为通用处理单元与高速协处理器及I/O扩展单元之间的转换桥梁，可以在统一的体系结构中实现对高速协处理器与I/O扩展功能的支持，使整个刀片服务器平台的运算性能、扩展性和适应性同时得到了大幅度的提升，使其更适用于复杂的高性能运算及商业应用领域，因而具有非常广阔的发展前景。

附图说明

[0009] 图1是支持高速协处理器、磁盘存储及I/O扩展的高密度刀片服务器结构原理图；

图2是高速转换模块与计算单元通讯原理图。

具体实施方式

[0010] 下面参照附图，对本发明的内容以具体实例来描述实现这一体系结构的过程。

[0011] 正如发明内容中所描述的，本发明体系结构主要包括：计算处理单元、信息存储单元、高速转换单元、磁盘通信单元、信息控制/传输单元、网络通信单元、高速数据传输单元、平台监控管理单元、电源控制单元和对外互联单元。

[0012] 计算处理单元基于通用的计算机体系架构，主要采用基于Intel Xeon处理器及相应芯片组构建处理单元计算平台；

高速转换单元，采用可热插拔的模块板卡设计，高速转换单元作为通用处理单元与高速协处理器及I/O扩展单元之间的转换桥梁，通过PCIe3.0总线直接与每个通用计算处里单元连接，为计算处里单元提供双向32GB/s的通讯带宽。

[0013] 协处理器模块，基于Nvidia Tesla和Intel MIC协助里模块，单个高速协处理模块可以提供每秒6000亿次的浮点运算能力，同时通过I/O扩展模块可以为信息处理单元方便的进行IO放面的扩展，例如HCA卡、SAS RAID卡、万兆光纤网卡、图形处理卡。

[0014] 网络通信单元基于目前通用的以太网传输技术设计，采用业界通用的以太网交换解决方案，提供两条千兆1Gb/s的传输通路通过对外互联单元与外界通讯，在处理单元与通讯单元进行数据通信时，负责数据包的转换，把以太网包格式转换成处理单元本地协议可识别的包格式。

[0015] 高速数据传输单元采用可热插拔的模块板卡设计，基于Mellanox公司的ConnectX3芯片作为主传输交换芯片设计方案，为高速协处理单元和通用计算单元提供一条56Gb/s的高速传输链路与外界通讯。

[0016] 平台监控管理单元采用基于标准的计算机管理总线设计，可以对信息处理单元、高速信息交换单元、通讯转换单元的状况进行监督，并可对上述单元进行配置和基于预定策略的管理。

[0017] 本发明的即能支持高速协处理器，又同时具备支持本地数据存储及丰富I/O扩展

功能的高密度刀片服务器体系架构设计方法，可以在统一的体系结构中实现对高速协处理器与 I/O 扩展功能的支持，使整个刀片服务器平台的运算性能、扩展性和适应性同时得到了大幅度的提升。本系统可以在一个统一的系统体系架构内实现基于运算密集型应用与商用 I/O 扩展应用的互换。

[0018] 与传统的刀片服务器体系结构相比，这种新型的体系结构打破了原有体系架构只以通用处理器单元为中心并且 I/O 扩展性差的限制，具有高运算性能、高可扩展，以及基于模块化部件灵活配置等特性。

[0019] 其中，高运算性的实现方式描述如下： 在支持通用处理单元的基础上，设计高速转换模块通过 PCIe3.0 总线直接与每个通用计算处里单元连接，为计算处里单元提供双向 32GB/s 的通讯带宽，使平台可以支持高速的协处理器，通过扩展支持协处理器平台可以提供每秒 1.2 万亿次的浮点运算能力。

[0020] 高可扩展的实现方式描述如下：在这种新型体系结构中，可以高速转换模块将原有的高速协处理单元换成基于 PCI-E 总线的 IO 扩展单元，可以为信息处理单元方便的进行 IO 放面的扩展，例如 HCA 卡、SAS RAID 卡、万兆光纤网卡、图形处理卡，从而提高系统整体的可扩展性。

[0021] 除说明书所述的技术特征外，均为本专业技术人员的已知技术。

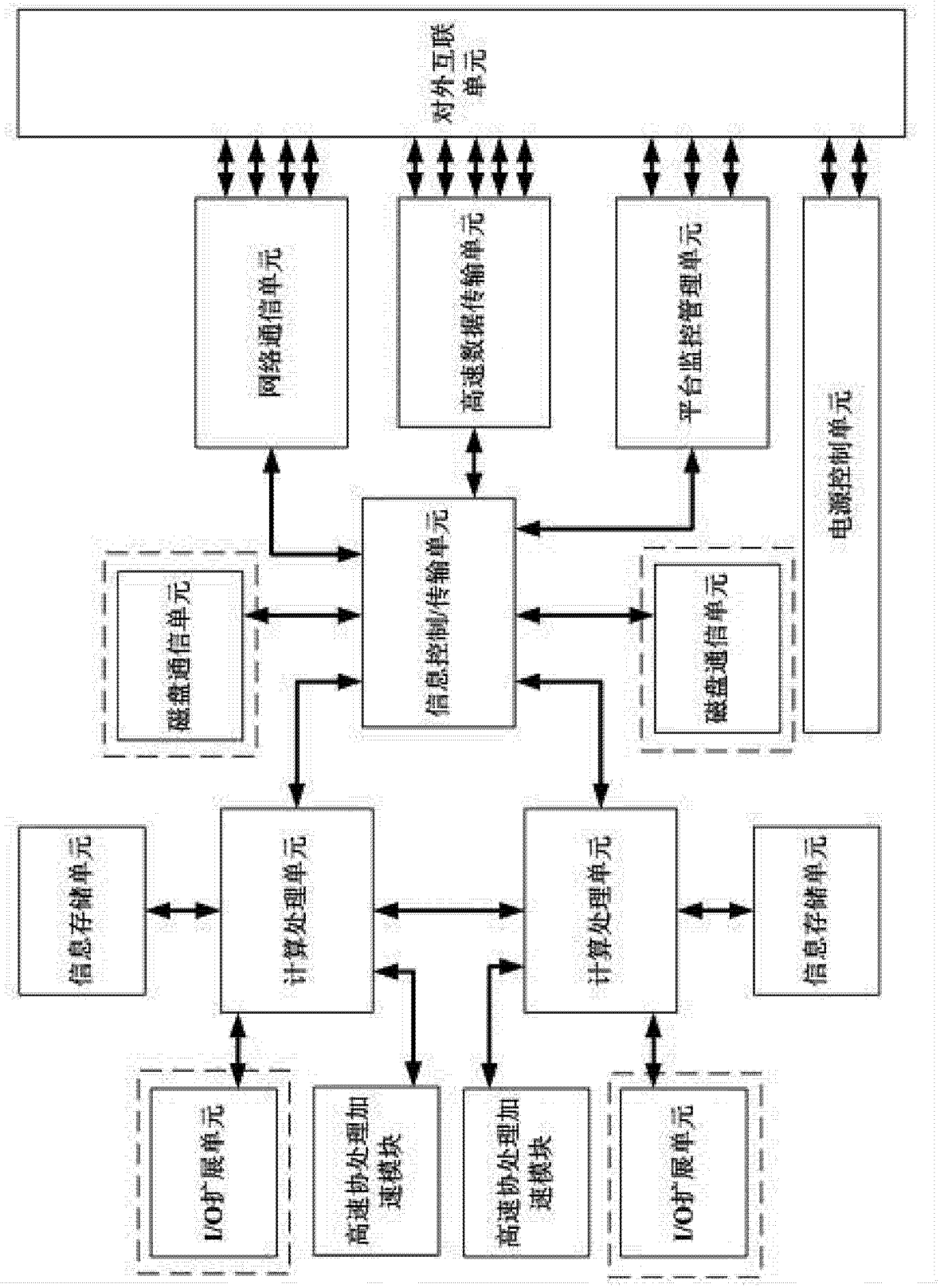

对外互联单元
网络通信单元
高速数据传输单元
平台监控管理单元
电源控制单元
磁盘通信单元
信息控制/传输单元
磁盘通信单元
信息存储单元
计算处理单元
计算处理单元
信息存储单元
I/O扩展单元
高速协处理加速模块
高速协处理加速模块
I/O扩展单元

图 1

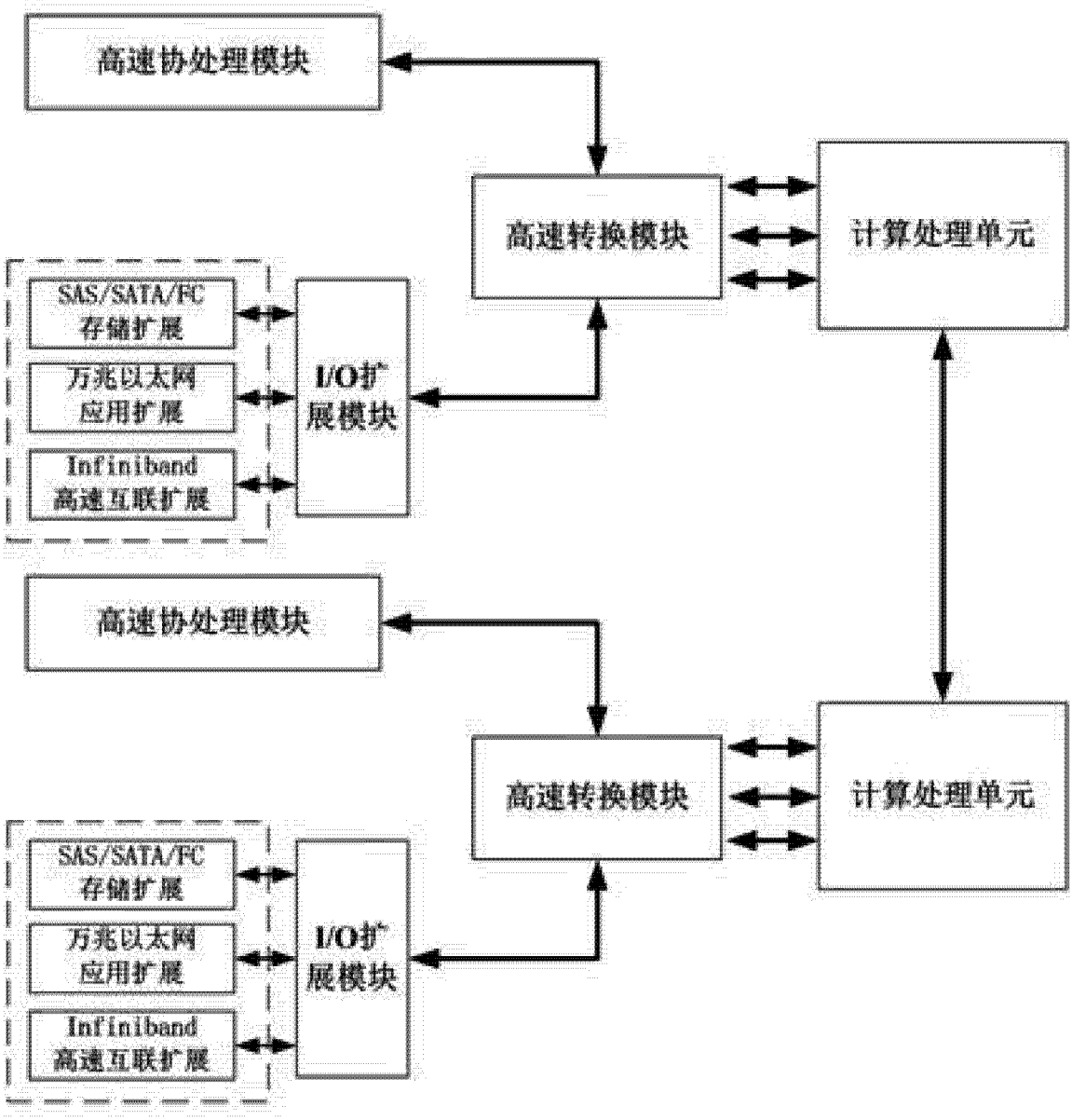

高速协处理模块
高速转换模块
计算处理单元
SAS/SATA/FC
存储扩展
万兆以太网
应用扩展
Infiniband
高速互联扩展
I/O扩
展模块
高速协处理模块
高速转换模块
计算处理单元
SAS/SATA/FC
存储扩展
万兆以太网
应用扩展
Infiniband
高速互联扩展
I/O扩
展模块

图 2