(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112166414 B
(45)授权公告日 2024.10.15

(21)申请号 201980031994.4
(22)申请日 2019.05.15
(65)同一申请的已公布的文献号
申请公布号 CN 112166414 A
(43)申请公布日 2021.01.01
(30)优先权数据
18172497.2 2018.05.15 EP
(85)PCT国际申请进入国家阶段日
2020.11.12
(86)PCT国际申请的申请数据
PCT/EP2019/062483 2019.05.15
(87)PCT国际申请的公布数据
WO2019/219747 EN 2019.11.21

(73)专利权人 派泰克集群能力中心有限公司
地址 德国慕尼黑
(72)发明人 B·福罗维特 T·利珀特
(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 马景辉

(51)Int.Cl.
G06F 9/50(2006.01)
G06F 1/324(2006.01)
G06F 1/329(2006.01)
G06F 1/3296(2006.01)
(56)对比文件
US 2011167229 A1,2011.07.07
审查员 谢丹

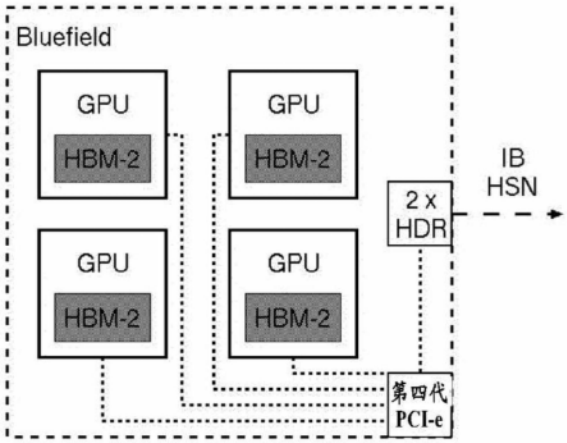
权利要求书1页 说明书5页 附图2页

(54)发明名称

用于高效并行计算的装置和方法

(57)摘要

本发明提供了一种用于在并行计算系统中操作的计算单元,该计算单元包括多个处理元件和用于将计算单元连接到该计算系统的其它组件的接口,其中,每个处理元件具有标称最大处理速率NPR并且每个处理元件包括相应的存储器单元使得可以以预定的最大数据速率MBW从该存储器单元传输数据,并且接口提供最大数据传输速率CBW,其中为了提供用于计算单元的预定的峰值计算性能PP,该峰值计算性能PP可由数量n个处理元件在标称最大处理速率下操作而使得每秒进行PP=n×NPR次操作而获得,该计算单元包括整数倍f乘以n个处理元件,其中,f大于1并且每个处理元件被限制为以NPR/f的处理速率进行操作。



CN 112166414 B

1.一种用于在并行计算系统中操作的计算单元,所述计算单元包括:

多个处理元件,其中,每个处理元件具有标称最大处理速率NPR;每个处理元件包括相应的随机存取存储器单元,使得能够以预定的最大数据速率MBW从所述存储器单元传输数据;以及

用于将所述计算单元连接到所述计算系统的其它组件的接口,所述接口提供最大数据传输速率CBW;

其中,为了提供用于所述计算单元的预定的峰值计算性能PP,该峰值计算性能PP能由数量n个处理元件在标称最大处理速率下操作而使得每秒进行PP = n × NPR次操作而获得,并且为了增大比率R = MBW/PP,所述计算单元包括整数倍f乘以n个处理元件,其中, f大于1并且每个处理元件被限制为以NPR/f的处理速率进行操作。

2.根据权利要求1所述的计算单元,其中,所述单元展现出的存储器带宽是具有n个处理元件的假设计算单元的存储器带宽的f倍。

3.根据权利要求1或权利要求2所述的计算单元,其中,所述计算单元的所述处理元件是图形处理单元。

4.根据任一前述权利要求所述的计算单元,其中,所述处理元件通过接口单元被连接在一起,每个所述处理元件通过多个即S个串行数据通道被连接到所述接口单元。

5.根据任一前述权利要求所述的计算单元,其中,所述计算单元是包括被布置为控制所述处理元件的多核处理器的计算机卡。

6.根据任一前述权利要求所述的计算单元,其中,所述最大数据传输速率在处理元件到处理元件的通信速率的30%内。

7.一种使计算单元进行操作的方法,所述计算单元包括多核处理器和多个图形处理元件GPU,每个GPU具有每秒PPG次操作的标称峰值性能,并且每个GPU具有相应的随机存取存储器单元,使得能够以预定的最大数据速率MBW从所述存储器单元传输数据,所述方法包括使所述GPU以所述GPU的标称峰值性能速率的分数1/f进行操作,其中f大于1,以及其中所述计算单元提供每秒PP次操作的预定峰值计算性能并且所述计算单元具有n乘以f个GPU使得PP等于PPG的n倍,以及其中增大比率R = MBW/PP。

用于高效并行计算的装置和方法

技术领域

[0001] 本发明涉及并行处理系统,并且具体地涉及在性能/能耗方面具有改进效率的并行处理系统。

背景技术

[0002] 在典型的并行处理系统中,各包括一个或多个处理元件的多个计算节点通过高速网络连接。计算节点的处理元件各具有内部存储器。处理元件在其计算节点内连接。该计算中节点的连接性可以用高速网络、分离的计算中节点高速网络或公共存储器(如例如在对称多处理SMP系统中)的技术来实现。这种布置图示在图1中。

[0003] 图1示出了多个计算节点CN的布置,其中每个计算节点包括各具有相应存储器MEM的多个处理元件PE。计算节点经由高速网络HSN彼此连接,其中每个计算节点包括用于连接到高速网络的网络接口控制器NIC。单独的处理元件被连接在一起,并且还连接到网络接口控制器。

[0004] 处理元件具有峰值性能PP,该峰值性能PP是该处理元件每秒可以执行的(浮点)运算的数量的上限,其被测量为每秒浮点运算次数或简称“flops”(虽然引用的是浮点运算,但该运算可以等同地是整数运算)。计算节点的峰值性能PPCN是其处理元件的峰值性能的总和。一般而言,给定应用A只可以实现峰值性能的分数η_A,其中$0<\eta<1$,η被称为持续效率。这样的原因在于,数据传输速率即处理元件的存储器到其计算寄存器之间的存储器带宽(MBW)是有限的,因此对于给定应用将把峰值性能的开发降低至η_A。对于输入/输出数据传输速率即处理元件到另一处理元件非计算节点的通信带宽(CBW),可以提出类似的论点,潜在地进一步降低对峰值性能的开发。

[0005] 根据经验,高性能计算的从业者认为,针对数据密集型应用中的大多数,比率R=MBW/PP为1字节/浮点运算是实现使η接近1的必要要求。取决于给定应用A耗尽峰值性能所需的数据速率,处理元件的实际存储器带宽确定了可以针对应用A实现的η_A。

[0006] 当前的高端CPU遭受低至0.05至0.1字节/浮点运算的R,该数字在最近十年里随着处理元件的计算核心的数量增加而不断减小。当前的高端GPU仅实现低于0.15字节/浮点运算的R,其主要是为了满足图形应用以及就在最近的深度学习应用的数据要求而设计的。不利的结果是,大多数数据密集型应用将先验地实现在当前CPU上低于5%至10%的η_A和在当前GPU上大约15%的η_A,而与由正在被执行的算法或在所需处理元件方面的并行性而引起的任何进一步降低无关。对于处理元件的给定R,应用越数据密集,效率η_A变得越低。

[0007] 这个问题已被其他人认识到,正如例如由Al Wegner在2011年的《Electronic Design》上发表的标题为“The Memory Wall is Ending Multicore Scaling(存储器墙正在终结多核扩展)”的文章中描述的,该文章可从http://www.electronicdesign.com/analog/memory-wall-ending-multicore-scaling获得。

[0008] 可以针对给定处理元件的通信带宽进行类似考虑,该通信带宽描述了到另一处理元件的计算中节点和非计算节点的数据传输速率。这里重要的是通信带宽对代码可扩展性

的影响。

[0009] 就通信带宽计算中节点而言,可以区分以下三种情况:其中处理元件经由高速网络连接的计算节点、其中处理元件被连接到再次连接到高速网络的计算节点上的分离网络的计算节点、以及通过公共存储器交换数据计算中节点的计算节点。

[0010] 关于通信非计算节点,高性能计算的从业者认为比率r＝CBW/MBW>0.1至0.2对于实现众多应用的可扩展性是适当的。显然,通信带宽越接近于存储器带宽,可扩展性的条件越好。

[0011] 理论上可能的通信带宽是由从处理元件到高速网络可用的串行通道数量确定的(这对于CPU和GPU均适用)。该数量被受当前芯片技术约束的串行器-解串器实现方式所限制。

[0012] 重要的是,计算节点的网络接口控制器NIC被适当地确定尺寸,以维持数据流往返于计算节点的处理元件。

[0013] US 2005/0166073 A1描述了使用系统处理器的可变工作频率以便最大化系统存储器带宽。

[0014] US 2011/0167229 A1描述了一种计算系统,该计算系统包括多个计算设备,该计算设备各被连接到与存储器相对的存储设备(诸如硬盘驱动器)。该系统的目的是使检索所存储数据的数据速率与处理速度相匹配。该文档中的建议是使用具有较高数据传输速率的存储单元(即作为硬盘驱动器的替代或补充的固态驱动器)来与以较低时钟速率下操作的特定低功率处理器相结合。

[0015] US 3025/0095620描述了一种用于估计计算系统中的工作负载的可扩展性的技术。该系统具有单个多核处理器。

发明内容

[0016] 本发明提供了一种用于在并行计算系统中操作的计算单元,该计算单元包括多个处理元件和用于将计算单元连接到计算系统的其它组件的接口,其中,每个处理元件具有标称最大处理速率NPR并且每个处理元件包括相应的存储器单元,使得可以以预定的最大数据速率MBW从存储器单元传输数据,并且接口提供最大数据传输速率CBW,其中为了提供用于计算单元的预定峰值计算性能PP,该峰值计算性能PP可由数量n个处理元件在标称最大处理速率下操作而使得每秒进行PP＝n×NPR次操作而获得,该计算单元包括整数倍f乘以n个处理元件,其中,f大于1并且每个处理元件被限制为以NPR/f的处理速率进行操作。

[0017] 在另一方面,本发明提供了一种操作计算单元的方法,该计算单元包括多核处理器和多个图形处理元件GPU,每个GPU具有每秒PPG次操作的标称峰值性能,该方法包括使GPU以其标称峰值性能速率的分数(fraction)1/f进行操作,其中计算单元提供每秒PP次操作的预定峰值计算性能并且计算单元具有n乘以f个GPU使得PP等于PPG的n倍。

[0018] 本发明涉及这样的事实,即处理元件的时钟频率v按因数f减小可以使处理元件的能耗按因数f或更多来减小。该过程被称为“降频”。

[0019] 以下近似公式适用于处理元件的器件功耗:$P \propto CV^2v$,其中,C是电容,V是电压而P是功耗。这意味着P与v成线性比例,与V成二次比例。

[0020] 关于作为示例的GPU的时钟频率,近年来,已经发表了关于功率建模的大量文章,

这些文章除了其它内容以外,力求将功耗分配给处理元件的各个部分。利用最新的NVIDIA GPU,可以改变流式多处理器(SM)的频率。这是由硬件越来越多地进行动态设计和自主控制,以便最佳地使用可用功率预算。据文献记载,存储器子系统的频率不可以改变,并且在当前一代中被自主计时。这使得其性能被存储器带宽限制的应用有可能通过稍微降低SM的频率来改进能量平衡。以这种方式,可以预期有约10%的效果。

[0021] 降频机器的性能经常好于预期。在正常桌面使用环境下,很少要求完全的处理元件性能。即使当系统繁忙时,通常也花费大量时间等待来自存储器或其它设备的数据。

[0022] 该事实原则上允许在频率v下操作的安装在计算节点上的处理元件能够被在频率v/f下操作的多个(f个)处理元件取代,而无需改变计算节点的累积计算能力PPCN。另外,计算节点的能耗被保持或潜在地减少。实际上,将选择f=2或f=3。

[0023] 本发明的关键方面在于,对于如现代CPU和GPU那样的处理元件,可以降低计算频率f,而同时没有减小处理元件的存储器带宽。结果,在该修改中,比率R按因数f增大。要注意的是,在不调整存储器速度的情况下不可能提高计算核心的工作频率。

[0024] 其次,将计算节点的处理元件的数量按因数f增加使计算节点上的可用串行通道的总数量按因数f增加。因此,输入/输出操作非计算节点的比率r也被按因数f提高。

[0025] 这些改进使每个计算节点的并发性按因数f增加。这要求针对各种高度可扩展的应用调谐算法方法,但原则上这并没有问题。

[0026] 虽然能耗预计会保持恒定,但处理元件的数量增加乍一看可能会增加投资成本。然而,这些成本中的大量成本将归因于存储器,其可以在每个计算节点的存储器总量保持恒定的同时针对每个处理元件按因数f降低。此外,在较低频率下使用高端处理元件可能会允许开发无法在峰值频率下操作的成本低得多的收益部门。

[0027] 作为第二种措施,可以执行降低处理元件的工作电压V,并且将导致能耗进一步降低。因功耗与电压成二次比例,对电压的依赖性可能很大。该“降电伏(undervolting)”可以与降频一起使用或者分开使用,并且是本发明为改进处理元件的计算部分的能耗的策略的另一要素。

[0028] 本发明提供了提高了并行处理系统在性能和能耗方面的效率的装置。引入技术修改,其降低处理元件的工作频率并相应地增加处理元件的数量以在增加的应用性能下实现整个系统的相同峰值性能。这些修改影响了影响整体效率的两个系统参数;存储器用于将按处理节点的峰值性能划分的数据带宽以及处理节点的数据带宽注册到按处理节点的峰值性能划分的并行系统的高速网络中。这允许在节点的能耗恒定或甚至更低的情况下节点的并行性能够增加。以这种方式,可以将系统调谐成最佳的应用性能。可以针对任何所期望的措施选择最佳值,例如,某个应用组合的平均应用性能或针对某个应用的最佳性能。由于在保持存储器和输入/输出性能的同时所使用的处理元件将在其处理单元的计算核心的较低工作频率下操作,因此总投资成本也预期保持类似。

[0029] 所提出的本发明允许根据所选择的期望标准(例如,针对某个应用组合的平均最大功率或针对某个应用的最大功率)选择因数f,该因数f确定了处理元件频率的降低和计算节点上的处理元件数量的对应增加。实际上,这两种修改也可以被独立应用,这取决于系统关键参数的影响,诸如能耗和投资成本以及最佳性能,尤其是关于相对于可扩展性的架构和应用的交互。

附图说明

[0030] 现在将只以示例的方式参考附图来描述本发明的优选实施例,在附图中:

[0031] 图1是常规并行处理系统的简化示意图;

[0032] 图2是包括两个图形处理单元GPU和每秒25万亿次浮点运算的峰值性能速率的计算节点的示意图;以及

[0033] 图3是包括多达图2的布置两倍的图形处理单元但峰值性能速率相同的计算节点的示意图。

具体实施方式

[0034] 本发明可以利用现有技术来实现。举例而言,它可能是加速在模块化超级计算系统内的提升模块(booster module)中的应用性能的方法,该超级计算系统目标在于到2021年达到峰值百亿亿级性能,如WO 2012/049247 A1和后续申请EP 16192430.3和EP 18152903.3中所述,其出于所有目的通过引用结合于此。本发明的目标是相比于任何其它架构设计,针对数据密集型计算按因数f提高计算中节点应用性能,并且另外,增加通信带宽以便与存储器带宽同步,以更好地扩展具有大通信要求非计算节点的许多应用。

[0035] 实现方式由一组使用Mellanox BlueField(BF)多核片上系统技术的计算节点给出。BlueField卡可以包括多个图形处理器单元GPU、第四代PCIe(PCIe gen 4)开关和一个或多个高数据速率HDR开关。每个BlueField卡可以配备多达四个GPU。BF卡各包括两个Mellanox主机通道适配器HCA,因此在非计算节点可以实现高达两倍的HDR性能。

[0036] AMD Radeon Vega 20 GPU被认为是该处理元件的具体示例,其预计将于2018年中期正式交付。Vega-20 GPU可以通过16个第四代PCIe通道连接到BF计算节点上的PCI-e接口。GPU预计配备有32GB的HBM-2存储器,其被分割成各8GB的四个存储器库。16GB的HBM-2也是可能的,其被同样组织成各4GB的四个存储器库。因此,存储器速率对于这两种配置可以是相同的。

[0037] 在预期存储器带宽为每秒1.28TB且预期峰值性能为每秒12.5万亿次浮点运算(双精度)的情况下,R=0.1。尽管这与从业者的1字节/浮点运算的规则相差为因数10之远,但它仍然是可用的最优比率R之一。

[0038] 通信带宽由16个第四代PCIe通道限制,这些通道在每个通道和方向能够各有2GB。在r=64GB/1.28TB=0.05的情况下,对于数据密集型应用肯定会遇到严重的可扩展性问题。R和r的任何改进在这方面将是有帮助的。

[0039] 这由图2和图3示意性地图示。

[0040] 让标准配置包括每个BF-CN的两个GPU作为处理元件,其在峰值频率v下操作从而实现峰值性能。图2中描绘了初始配置。就计算节点而言,给出或预计以下的系统参数:

[0041] • 每个计算节点的GPU的数量:2

[0042] • f:1

[0043] • 每个计算节点的能耗:2×150W=300W

[0044] • 每个计算节点的存储器:64GB

[0045] • 每个计算节点的存储器带宽:每秒2.56TB

[0046] • 每个计算节点的峰值性能:每秒25万亿次浮点运算dp

[0047] • 每个计算节点的R:0.1

[0048] • 每个计算节点的第四代PCIe通道:32

[0049] • 每个计算节点的通信速度双向:128GB/s(1/2用于从处理元件到处理元件,1/2用于到NIC)

[0050] • 可能的2×Mellanox HDR:每秒100GB双向

[0051] • 每个计算节点的r:0.05

[0052] • 与通信不平衡的NIC

[0053] 图3中示出的改进配置包括每个BF计算节点的四个GPU作为处理元件,其在峰值频率v的一半下操作(其中f＝2),从而提供相同的计算标称节点峰值性能值。在这种情况下,处理元件将进行操作以达到标准配置的峰值性能的一半。至于改进的计算节点,给出或预计以下的系统参数:

[0054] • 每个计算节点的GPU的数量:4

[0055] • f:2

[0056] • 所预计的每个计算节点的能耗:4×75W＝300W

[0057] • 每个计算节点的存储器:每个GPU 64GB@16GB或128GB@32GB

[0058] • 每个计算节点的存储器带宽:每秒5.12TB

[0059] • 每个计算节点的峰值性能:每秒25万亿次浮点运算dp

[0060] • 每个计算节点的R:0.2

[0061] • 每个计算节点的第四代PCIe通道:64

[0062] • 每个计算节点的通信速度双向:256GB/s(1/2用于从处理元件到处理元件,1/2用于到NIC)

[0063] • 可能的2×Mellanox HDR:每秒100GB双向

[0064] • 每个计算节点的r:0.05

[0065] • 与通信平衡的NIC

[0066] 可以在降频的基础上添加降电伏,以进一步降低能耗。与施加全电压的情况相比,处于降电伏的处理元件的稳定性可能受到的影响较小。

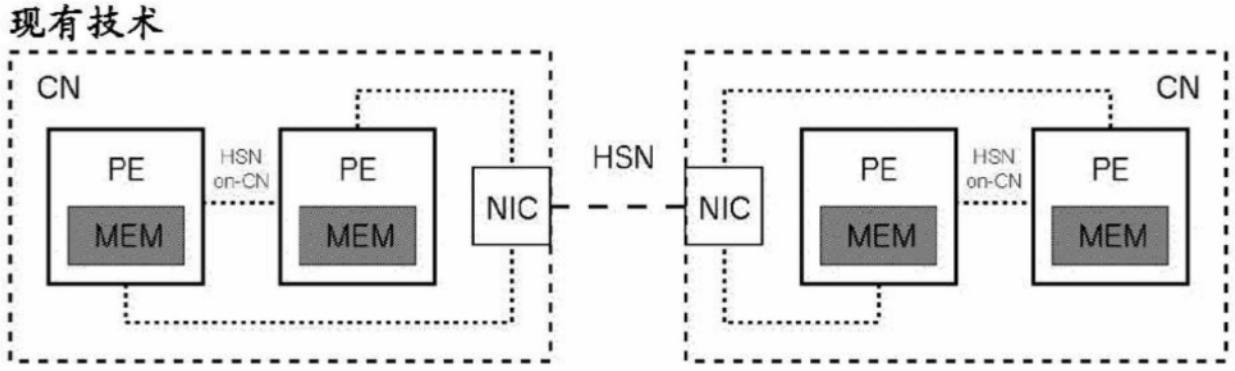

现有技术
CN
PE
MEM
HSN
on-CN
PE
MEM
NIC
HSN
NIC
CN
PE
MEM
HSN
on-CN
PE
MEM

图1

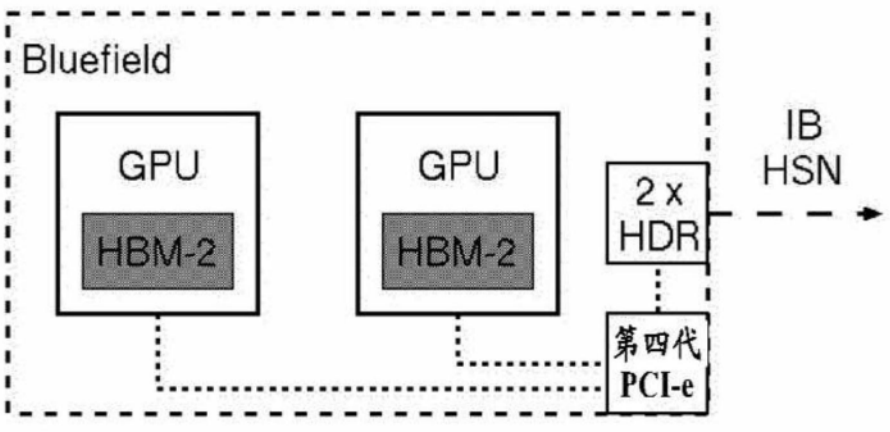

Bluefield
GPU
HBM-2
GPU
HBM-2
2 x
HDR
IB
HSN
第四代
PCI-e

图2

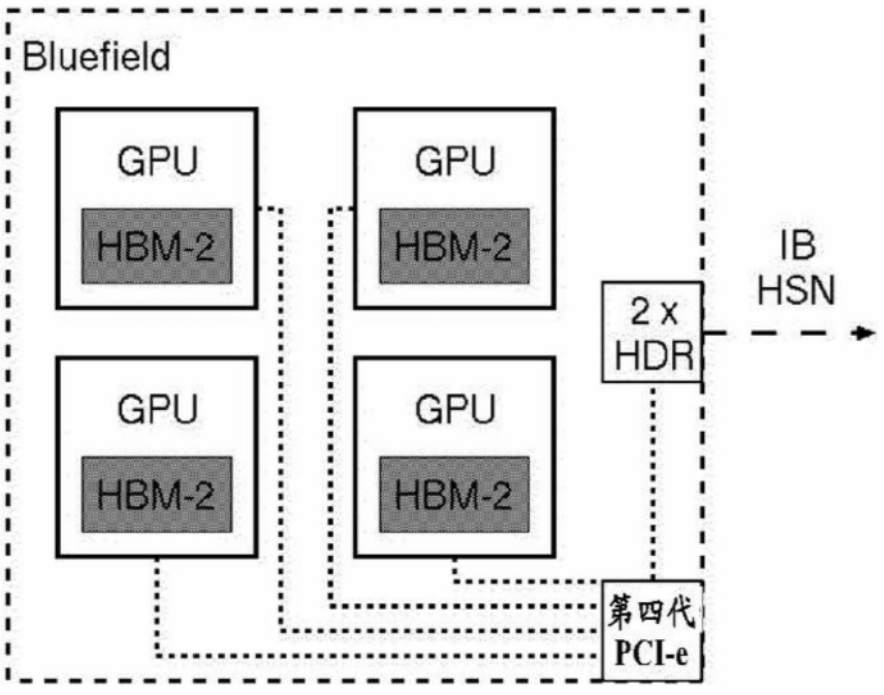

Bluefield
GPU
HBM-2
GPU
HBM-2
GPU
HBM-2
GPU
HBM-2
2 x
HDR
IB
HSN
第四代
PCI-e

图3