(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103229146 B
(45)授权公告日 2018.12.11

(21)申请号 201180056850.8
(22)申请日 2011.10.13
(65)同一申请的已公布的文献号
申请公布号 CN 103229146 A
(43)申请公布日 2013.07.31
(30)优先权数据
10187436.0 2010.10.13 EP
(85)PCT国际申请进入国家阶段日
2013.05.27
(86)PCT国际申请的申请数据
PCT/EP2011/067888 2011.10.13
(87)PCT国际申请的公布数据
WO2012/049247 EN 2012.04.19

(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 托马斯·利珀特
(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 李镇江

(51)Int.Cl.
*G06F 9/50*(2006.01)
(56)对比文件
US 2005/0097300 A1,2005.05.05,
US 2004/0257370 A1,2004.12.23,
US 2009/0213127 A1,2009.08.27,
审查员 武守秋

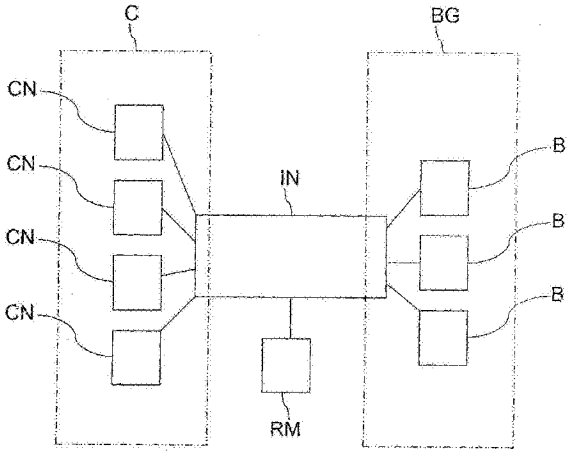
权利要求书2页 说明书9页 附图8页

(54)发明名称

用于处理计算任务的计算机集群布置及其操作方法

(57)摘要

本发明涉及一种计算机集群布置和一种用于所介绍计算机集群布置的操作方法。计算机集群布置包含计算节点CN,其将具体计算任务动态外包给增强器B。因此,增强器B到计算节点CN的分配技术得到介绍。该分配在运行时间动态发生。本发明找到在高性能集群技术中的应用。




CN 103229146 B

1.一种用于处理计算任务的计算机集群布置,所述计算机集群布置包含:

-多个计算节点(CN),每个计算节点联接通信基础设施(IN),至少两个节点被布置成共同计算所述计算任务的至少第一部分;特征在于所述计算机集群布置还包括:

-多个增强器(B),所述多个增强器(B)中的至少一个增强器(B)被布置成在被分配给计算节点之后计算所述计算任务的至少第二部分,每个增强器(B)与所述通信基础设施(IN)联接并且其中所述增强器具有包括比计算节点的处理器设计的算数逻辑单元更大规模的算数逻辑单元以及比计算节点的处理器设计的控制结构更简单的控制结构的处理器设计;和

-资源管理器(RM),被布置成执行将所述至少一个增强器(B)分配给计算节点(CN)用于所述计算任务的所述第二部分的计算,所述分配根据预定的分配度量来完成,

其中所述多个计算节点以及多个增强器被布置使得在所述计算任务的处理期间,能够提供计算节点以及增强器的分配使得至少

(i)所述多个计算节点的一个或者多个计算节点被布置以与所述多个增强器的一个或者多个增强器通信,

(ii)增强器中的一个或者多个增强器是能够被所述多个计算节点的多于一个的计算节点所共享的,以及

(iii)增强器中的每个增强器是能够分配给计算节点中的每个计算节点的,

并且其中所述资源管理器被布置以在开始处理所述计算任务时通过使用预定分配的度量来执行分配并且以在所述处理期间更新所述分配度量并且因此通过以下步骤在运行时间期间执行将增强器动态分配到计算节点:

(i)在处理的开始时通过使用预定的分配度量来初始化分配,

(ii)更新所述分配度量,并且

(iii)在处理计算任务期间通过使用更新的分配度量来执行重新分配。

2.根据权利要求1所述的计算机集群布置,其中所述预定的分配度量是根据一组度量规范技术中的至少一个形成的,所述一组度量规范技术包含:时序逻辑、分配矩阵、分配表、概率函数和成本函数。

3.根据权利要求1-2中之一所述的计算机集群布置,其中所述预定的分配度量是根据一组分配参数中的至少一个来指定的,所述一组分配参数包括:资源信息、成本信息、复杂性信息、规模可变性信息、计算日志记录、编译程序信息、优先级信息和时间戳。

4.根据权利要求1-2中之一所述的计算机集群布置,其中至少一个增强器(B)至所述多个计算节点(CN)之一的所述分配触发一组信号中的至少一个,所述一组信号包含:远程过程调用、参数移交以及数据传输。

5.根据权利要求1-2中之一所述的计算机集群布置,其中每个计算节点(CN)和每个增强器(B)分别经由联接单元(IU1;IU2)与所述通信基础设施(IN)联接。

6.根据权利要求5所述的计算机集群布置,其中所述联接单元(IU1;IU2)包含一组组件中的至少一个,所述一组组件包含:虚拟接口、接管、插件、网络控制器和网络设备。

7.根据权利要求1-2中之一所述的计算机集群布置,其中所述通信基础设施(IN)包含一组组件中的至少一个,所述一组组件包含:总线、通信链路、切换单元、路由器和高速网络。

8.根据权利要求1-2中之一所述的计算机集群布置，其中每个计算节点(CN)包含一组组件中的至少一个，所述一组组件包含：多核处理器、集群、计算机、工作站和多功能处理器。

9.根据权利要求1-2中之一所述的计算机集群布置，其中所述至少一个增强器(B)包含一组组件中的至少一个，所述一组组件包含：多核处理器、标量处理器、协处理器、图形处理单元、多核处理器的集群和单片处理器。

10.根据权利要求1-2中之一所述的计算机集群布置，其中所述资源管理器(RM)被布置成在所述计算任务的至少一部分的计算期间更新所述预定的分配度量。

11.一种操作用于处理计算任务的计算机集群布置的方法，尤其是根据权利要求1-10中至少一个所述的计算机集群，所述方法包含：

-通过所述多个计算节点(CN)中的至少两个计算所述计算任务的至少第一部分，每个计算节点(CN)与通信基础设施(IN)联接；

-通过至少一个增强器(B)计算所述计算任务的至少第二部分，每个增强器(B)与所述通信基础设施(IN)联接；并且

-通过资源管理器(RM)来将所述至少一个增强器(B)分配给所述多个计算节点(CN)之一，以计算所述计算任务的第二部分，所述分配根据预定的分配度量来完成，

其中在所述计算任务的处理期间，能够提供计算节点以及增强器的分配使得至少

(i)所述多个计算节点的一个或者多个计算节点与所述多个增强器的一个或者多个增强器通信，

(ii)增强器中的一个或者多个增强器是能够被所述多个计算节点的多于一个的计算节点所共享的，以及

(iii)增强器中的每个增强器是能够分配给计算节点中的每个计算节点的，

并且其中所述资源管理器在开始处理所述计算任务时通过使用预定的度量来执行分配并且以在所述处理期间更新分配度量并且因此通过以下步骤在运行时间期间执行将增强器动态分配到计算节点：

(i)在处理的开始时通过使用预定的分配度量来初始化分配，

(ii)更新所述分配度量，以及

(iii)在处理计算任务期间通过使用更新的分配度量来执行重新分配。

# 用于处理计算任务的计算机集群布置及其操作方法

## 技术领域

[0001] 本发明涉及一种计算机集群布置。尤其是，它涉及一种具有改进的资源管理的计算机集群布置，改进的资源管理是关于用于处理规模可改变的计算任务以及复杂计算任务的计算节点的应用。本发明尤其指向一种用于处理计算任务的计算机集群布置和该计算机集群布置的一种操作方法。根据本发明的计算机集群布置使用加速功能性，其辅助计算节点完成给定的计算任务。此外，本发明还指向一种被配置成用于完成该方法的计算机程序产品以及一种用于储存该计算机程序产品的计算机可读介质。

## 背景技术

[0002] 在本技术领域已知的是，计算机集群布置包含包括至少一个处理器的计算节点以及与耦接计算节点紧密耦接的加速器，用于高资源要求的外包计算。加速器至计算节点的紧密耦接导致静态分配并且导致加速器的过度预订(subscription)或预订不足。这可以导致资源缺乏或可以导致资源过度供应。此外，加速器至计算节点的这种静态分配在加速器故障的情况下不提供容错。

[0003] JoséDuato、Rafael Mayo等的出版物“rCUDA:reducing the number of GPU-based accelerators in high performance clusters(远程统一计算设备架构:减少高性能群集中基于图形处理器(GPU)的加速器的数量)”，高性能计算和模拟(HPCS)国际会议，出版日期:2010年6月28日-2010年7月2日，在第224-231页上，描述了一种在高性能集群中能够远程GPU加速，因而允许减少安装在集群上的加速器的数量的框架。这可以导致能源、采购、维护及空间的节省。

[0004] 耶路撒冷希伯来大学的计算机科学系的Amnon Barak等的出版物“A package for open CL based heterogeneous computing on clusters with many GPU devices(一种用于对具有许多个GPU设备的集群进行基于开放运算语言(Open CL)的异构计算的包)”描述了一种用于对具有许多个GPU设备的集群运行OpenMP、C++、未修改的OpenCL的应用的包。此外，提供允许在一个托管节点(hosting-node)上应用以便透明地利用集群范围内设备的OpenCL规范的实施方式和应用编程接口(OpenMP API)的扩展。

[0005] 图1示出根据本技术领域当前水平的计算机集群布置。该计算机集群布置包含数个计算节点CN，其是互相连接的并且共同计算一个计算任务。每个计算节点CN都与加速器Acc紧密耦接。从图1能够看出，计算节点CN包含加速器单元ACC，其与例如中央处理单元CPU的微处理器一起虚拟集成在计算节点CN上。如上所述，根据计算任务，加速器Acc至计算节点CN的固定耦接会导致加速器Acc的过度预订或预订不足。此外，在加速器Acc之一故障的情况下不提供容错。在根据图1的已知的计算机集群布置中，计算节点CN通过基础设施互相通信，其中加速器Acc不直接交换信息，但是需要计算节点CN联接(interfacing)基础设施IN，用于数据交换。

## 发明内容

[0006] 因此，本发明的一个目的是提供一种计算机集群布置，其允许关于加速器和计算节点之间数据交换的通信灵活性以及计算节点至任何一个和每个加速器的直接存取。此外，本发明的一个目的是在运行时间提供一种加速器至计算节点的动态耦接。

[0007] 这些目的通过具有根据专利权利要求1所述的特征的计算机集群布置来解决。

[0008] 因此，提供一种用于处理计算任务的计算机集群布置，该计算机集群布置包含：

[0009] -多个计算节点，每个计算节点都联接通信基础设施，至少两个计算节点被布置成共同计算计算任务的至少第一部分；

[0010] -至少一个增强器，其被布置成计算计算任务的至少第二部分，每个增强器都联接通信基础设施；和

[0011] -资源管理器，其被布置成将至少一个增强器分配给多个计算节点中的至少一个计算节点，用于计算任务的第二部分的计算，该分配依据预定分配度量的一个功能而完成。

[0012] 在这个计算机集群布置中，通过独立的增强器提供加速功能。所描述计算机集群布置允许那些增强器至计算节点的弱耦接，其也可以被称为计算节点。因此，在此通过计算节点共享具有增强器形式的加速器是可行的。对增强器至计算节点的分配而言，可以提供具有资源管理器模块或资源管理器节点形式的资源管理器。资源管理器可以在开始处理计算任务时建立静态分配。可选地或附加地是，可以在运行时间建立动态分配，其意味着在处理计算任务期间。

[0013] 资源管理器被布置成将分配信息提供给计算节点，以便用于计算任务从至少一个计算节点到至少一个增强器的外包部分。资源管理器可以被实施为具体的硬件单元、虚拟单元或其任何组合。尤其是，可以通过下列任何一个来形成资源管理器：微处理器、硬件组件、虚拟化硬件组件或守护器。此外，部分资源管理器可以在系统上分布并且经由通信基础设施进行通信。

[0014] 增强器之间的通信通过网络协议完成。因此，增强器分配依据应用需要而被执行，其意味着依赖于处理具体计算任务。在增强器故障的情况下提供容错并且规模可变性被促进。当增强器独立于计算节点被提供时，通过支持渐增的系统开发使得规模可改变性成为可能。因此，计算节点的数量和所提供增强器的数量可以不同。因而，建立在提供硬件资源中的最大灵活性。此外，所有计算节点都共享相同的成长容量(growth capacity)。

[0015] 计算任务可以借助于算法、源代码、二进制代码进行定义并且还可以是它们的任何组合。计算任务可以例如是模拟，其通过计算机集群布置进行计算。此外，计算任务可以包含数个子问题，也被称为子任务，其全面描述整个计算任务。将计算任务分成数个部分是可能的，例如计算任务的至少第一部分和计算任务的至少第二部分。对于计算机集群布置而言，并行或顺序解决部分计算任务也是可能的。

[0016] 每个计算节点都联接通信基础设施，也被称为互相连接。类似地是，每个增强器都联接通信基础设施。因此，计算节点以及增强器借助于通信基础设施进行交互。因此，每个计算节点都通过通信基础设施与每个增强器通信，从计算节点到增强器交换数据时无需涉及另一个通信节点。因而，计算节点到增强器的动态分配得以建立，其中计算节点处理至少一部分计算任务并且不需要从一个计算节点到一个增强器传递信息。因此，将增强器直接耦接至通信基础设施，不需要如本技术领域当前水平通常实施的中间计算节点的是可能的。

[0017] 为了完成增强器和计算节点之间的分配，需要具体的规则集(set of rules)。因此，提供分配度量，其作为决定哪一个增强器与哪一个计算节点耦接的基础。分配度量可以通过资源管理器进行管理。管理分配度量指的是建立和更新命名至少一个增强器的规则，其被分配给至少一个其他被命名的计算节点。因此，在运行时间更新分配度量是可能的。这种分配规则可以依据负荷平衡而被产生，其探测计算机集群布置的工作负荷，尤其是增强器的工作负荷。此外，探测增强器的计算容量并且还探测计算任务要求及分配选定的增强器是可能的，其给计算节点提供所需容量。为了确定增强器到计算节点的初始分配，分配度量被预定，但可以在运行时间改变。因此，在开始处理计算任务时提供静态分配，而在运行时间提供动态分配。

[0018] 在本发明的一个实施例中，根据度量规范技术组中至少一个形成确定的分配度量，该组包含：时序逻辑、分配矩阵、分配表、概率函数和成本函数。因此，可以为分配增强器考虑时间依赖性。可以是这种情况，即，在增强器上定义时间顺序，这确保在另一个增强器故障的情况下一个具体增强器总是被分配给计算节点，以便解决至少一部分计算任务。因此，能够为增强器的分配考虑增强器之间的层次(hierarchy)。分配度量可以命名计算节点的身份并且还可以定义能够被分配的兼容性增强器的身份。概率函数可以例如描述在用于计算某计算任务的一个具体增强器故障的情况下，另一个增强器可以在一个具体概率下解决相同的计算任务。此外，成本函数可以应用于评估所需资源容量并且还评估被提供的增强器的计算容量。因此，某些需要的计算任务能够被转发至适当的增强器。

[0019] 计算历史，也被称为计算日志记录，也可以应用于动态分配。因此，能够通过在至少一个第一个增强器上的计算和记录响应时间、并且还通过在至少一个其他增强器上处理相同计算任务和记录响应时间来经验性地评估计算任务。因此，增强器的容量能够被记录、经验性地评估，并且因此依据所需容量和它们被提供容量而被分配给计算节点。具体计算任务可以包含优先级信息，其指示必须多么紧急地计算这个具体计算任务。也可以是具体计算节点提供优先级的情况，该优先级指示处理计算任务有多么紧急，或至少一部分计算任务与起源于其它计算节点的计算任务的其它部分被比较。因此，提供关于计算任务的单个部分的优先级信息以及参考计算节点的优先级信息是可能的。

[0020] 一旦增强器被分配给计算节点，增强器就处理计算任务的具体部分。这可以通过远程过程调用、参数移交或数据传输来完成。该部分计算任务的复杂性可以依据参数移交而被评估。在参数含有矩阵的情况下，参数移交的复杂性能够通过矩阵的维数被评估。

[0021] 为了联接通信基础设施，可以提供联接单元，其被布置在一个计算节点和通信基础设施之间。不同于第一个联接单元的其他联接单元可以被布置在增强器和通信基础设施之间。联接单元能够不同于计算节点并且也不同于增强器。联接单元仅仅提供网络功能，无需被布置成处理部分计算任务。联接单元仅仅提供关于计算任务的管理和通信问题的功能。例如可以提供关于参考计算任务的数据的路由选择和传输的功能。

[0022] 此外，也能够通过从至少一个增强器到至少一个计算节点外包至少一部分计算任务而反向执行加速。因此，控制和信息流关于上面所介绍的本发明的多个方面是反向的。

[0023] 根据本发明的一个方面，可以根据至少一组矩阵规范技术形成预定分配，该组包含：时序逻辑、分配矩阵、分配表、概率函数和成本函数。这可以提供预定分配度量可以在使用正式或半正式模型或数据类型的情况下被形成的优势。

[0024] 根据本发明的另一方面，预定分配度量依据一组分配参数中的至少一个而被指定，该组包含：资源信息、成本信息、复杂性信息、规模可变性信息、计算日志记录、编译程序信息、优先级信息和时间戳。这可以提供在考虑不同运行时间参数并且响应于具体计算任务特性的情况下在运行时间动态执行分配的优势。

[0025] 根据本发明的另一方面，至少一个增强器至多个计算节点之一的分配触发一组信号中至少一个，该组包含：远程过程调用、参数移交和数据传输。这可以提供至少一部分计算任务能够从一个计算节点被转发到至少一个增强器的优势。

[0026] 根据本发明的另一方面，每个计算节点和每个增强器都分别经由联接单元联接通信基础设施。这可以提供数据能够经由通信基础设施被通信而无需中间计算节点的优势。因此，不需要将增强器与计算节点直接耦接但达到动态分配。

[0027] 根据本发明的另一方面，联接单元包含至少一组组件，该组包含：虚拟接口、接管(stub)、插件、网络控制器和网络设备。这可以提供计算节点以及增强器也能够被虚拟连接至通信和基础设施的优势。此外，现有通信基础设施能够被容易地存取。

[0028] 根据本发明的另一方面，通信和基础设施包含一组组件中的至少一个，该组包含：总线、通信链路、切换单元、路由器和高速网络。这可以提供能够使用现有通信基础设施并且能够通过公共可用的网络设备产生新通信基础设施的优势。

[0029] 根据本发明的另一方面，每个计算节点都包含一组组件中的至少一个，该组包含：多核处理器、集群、计算机、工作站和多功能处理器。这可以提供计算节点规模可大大改变的优势。

[0030] 根据本发明的另一方面，至少一个增强器包含至少一组组件，该组包含：多核处理器、标量处理器、协处理器、图形处理单元、多核处理器的集群和单片处理器。这可以提供增强器被实施成高速处理具体问题的优势。

[0031] 当数个计算任务必须被同时处理时，计算节点通常应用包含大规模(extensive)控制单元的处理器。当与计算节点处理器比较时，在增强器中应用的处理器通常包含大规模算术逻辑单元和简单的控制结构。例如单指令多数据流(SIMD)，也称为单指令多数据计算机，可以找到在增强器中的应用。因此，在计算节点中应用的处理器与在增强器中应用的处理器相比较的不同之处在于它们的处理器设计。

[0032] 根据本发明的另一方面，资源管理器被布置成在至少一部分所述计算任务的计算期间更新所述预定分配度量。这可以提供增强器至计算节点的分配能够在运行时间被动态执行的优势。

[0033] 该目的也通过根据专利权利要求11所述的特征的用于操作计算机集群布置的一种方法来解决。

[0034] 相应地，提供一种计算机集群布置的操作方法来处理计算任务，该方法包含：

[0035] -通过多个计算节点中的至少两个计算计算任务的至少第一部分，每个计算节点都联接通信基础设施；

[0036] -通过至少一个增强器计算计算任务的至少第二部分，每个增强器都联接通信基础设施；和

[0037] -通过资源管理器将至少一个增强器分配给多个计算节点之一，用于计算任务的第二部分的计算，该分配依据预定分配度量而完成。

[0038] 此外，提供被配置用于完成所介绍方法的计算机程序以及用于储存该计算机程序产品的计算机可读介质。

## 附图说明

[0039] 现在将参考附图、仅仅通过例示来描述本发明：

[0040] 图1示出根据本技术领域当前水平的计算机集群布置。

[0041] 图2示出根据本发明一个方面的计算机集群布置的示意性例示。

[0042] 图3示出根据本发明另一方面的计算机集群布置的示意性例示。

[0043] 图4示出根据本发明一个方面的计算机集群布置的操作方法的示意性例示。

[0044] 图5示出根据本发明另一方面的计算机集群布置的操作方法的示意性例示。

[0045] 图6示出根据本发明另一方面的计算机集群布置的控制流的示意性例示。

[0046] 图7示出根据本发明另一方面的控制流实施计算机集群布置的反向加速的示意性例示。

[0047] 图8示出根据本发明另一方面的计算机集群布置的控制流的示意性例示。

[0048] 图9示出根据本发明一个方面的计算机集群布置的网络拓朴的示意性例示。

[0049] 如果不另外指示，则在下文中相同概念将用相同参考标记表示。

## 具体实施方式

[0050] 图2示出包含集群C以及增强器组BG的计算机集群布置。在本实施例中该集群包含四个计算节点，也被称为CN，以及三个增强器，也被称为B。通过诸如互相连接的通信基础设施IN建立增强器至计算节点的灵活耦接。能够例如通过使用无限带宽而实施这类通信基础设施IN。因此，每个增强器B能够被任何一个计算节点CN共享。此外，能够完成针对集群级别的虚拟化。每个增强器、或至少一部分增强器能够被虚拟化并且虚拟化地可用于计算节点。

[0051] 在本实施例中，通过至少一个计算节点CN处理计算任务并且至少一部分计算任务可以被转发至至少一个增强器B。增强器B被布置成计算具体问题并且提供具体处理能力。因此，问题能够从计算节点CN之一至增强器B被外包，通过增强器进行计算，并且结果可以被传送回至计算节点。增强器ESB到计算节点CN的分配能够通过也被称为RM的资源管理器完成。资源管理器初始化第一分配并且进一步建立增强器B到计算节点CN的动态分配。

[0052] 为了增强器和通信节点之间的通信，能够提供也被称为API的应用编程接口。增强器B可以通过各自的API功能调用被计算节点透明地控制。API提取并且加强增强器的实际本机编程模型。此外，API可以在增强器故障的情况下提供用于容错的手段。涉及API调用的通信协议可以在通信层的顶层被分层。在下文中，提供根据本发明一个方面的一套API调用的简短说明，其中参数“加速器”可以指定编址增强器：

[0053] -aanInit(加速器)

[0054] 使用前初始化增强器

[0055] -aanFinalize(加速器)

[0056] 使用后释放增强器上的记账信息

[0057] -aanMemAlloc(地址，大小，加速器)

[0058] 在所参考增强器上分配多个大小字节(size Bytes)的内存

[0059] 返回所分配设备内存的地址

[0060] -aanMemFree(地址,加速器)

[0061] 释放从所参考增强器上的地址开始的内存

[0062] -aanMemCpy(dst,src,大小,方向,加速器)

[0063] 拷贝从src到dst内存地址的多个大小字节

[0064] 拷贝操作的方向能够是下列方向:

[0065] (i)增强器到主机,

[0066] (ii)主机到增强器

[0067] -aanKernelCreate(文件_名,功能_名,内核,加速器)

[0068] 产生内核,该内核由用于在所参考增强器上执行的文件名(文件_名)和功能名(功能_名)进行定义

[0069] 将句柄(handle)返回至内核

[0070] -aanKernelSetArg(内核,指数,大小,对准,值)

[0071] 定义变元,用于通过变元列表、大小、对准要求(对准)和值的形式的其指数而内核执行

[0072] -aanKernelRun(内核,网格_维度,块_维度)

[0073] 在与之前调用ac内核产生()的内核相关的增强器上开始内核执行。线程数量通过每块的线程数量(块_维度)和网格内的块数量(网格_维度)来确定。

[0074] -aanKernelFree(内核)

[0075] 释放与内核相关的资源

[0076] 图3示出根据本发明一个方面的另一个集群布置。所描绘的计算机集群布置被布置成计算科学计算任务,尤其是在高性能集群技术的背景中。科学高性能集群应用代码的产品组合(portfolio)的特性的深入分析揭示出:具有百亿亿级(Exascale)需要的许多代码一方面包括十分适合百亿亿级规模的代码块,而另一方面这种代码块太复杂以至于不是这样规模可改变的。在下文中,规模可大大改变和非常复杂之间的明显差别在代码块级别上形成,并且我们介绍概念百亿亿级规模代码块(ECB)和复杂代码块(CCB)。

[0077] 显然,不存在纯粹的规模可大大改变的代码,也不存在绝对复杂的代码。每个代码都具有规模可大大改变和规模改变不大的复杂元件。事实上,两个极端之间存在连续性。有趣地是,代码的许多规模改变不大的元素不需要高度的规模可变性,而是需要大的本地内存。同样明显地是,全部-全部的通信元素在较小并行性的情况下具有高度优势。

[0078] 对于这种问题,其中在内存相对量(即内存相对量的自由处理度,即ECB对CCB的自由处理度)、执行时间和被交换的数据方面,ECB和CCB之间的恰当平衡是给定的,它建议自身借助于具体架构解决方案来适应这个状况。该方案由传统集群计算机手段与百亿亿级增强器一起组成,百亿亿级增强器具有紧密连接的增强器并通过集群的网络与集群连接。这个二元手段具有使纯粹百亿亿级系统的可预料的狭窄应用领域大大加宽的潜力。

[0079] 粗粒度架构模型形成,其中应用代码的规模可大大改变的部分或ECB在并行的多核架构上执行,其是动态存取的,而CCB在维度合适的传统集群系统上执行,传统集群系统包括连接性和精确的动态资源分配系统。

[0080] 为了保证适应性和可靠性,以百亿亿级计算的集群需要虚拟化元素。虽然本地加

速器原则上考虑对整个系统上的简单视角，并且尤其是能够利用非常高的本地宽带时，但是它们绝对是静态硬件元件，十分适合粗放式(farming)或主从式并行化。因此，在虚拟化软件层中包括它们将是困难的。另外，如果加速器故障则将没有容错，并且不容许过度预订或预订不足。

[0081] 集群的计算节点CN通过例如Mellanox无限带宽的标准集群互相连接而内部耦接。这个网络扩展至还包括增强器(ESB)。在图中我们已经绘制了三个这种增强器。每个ESB都由通过具体的快速低延迟网络连接的许多个多核加速器组成。

[0082] CN与ESB的这个连接非常灵活。计算节点之间共享加速器容量成为可能。针对集群级别的虚拟化不受模型约束，并且全部的ESB并行性都能够被开发。ESB至CN的分配经由动态资源管理器RM进行。在开始时间，静态分配在运行时间能够成为动态的。所有CN-ESB通信经由集群网络协议进行。内部AC通信将需要新的解决方案。ESB分配能够遵循应用需要，并且当所有计算节点共享相同的成长容量时，在加速器故障的情况下保证容错。

[0083] 由于可以应用增强器Intel的多核处理器Knight's Corner(KC)的计算元件。因此KC-芯片将由多于50个核组成并且期望提供超过每芯片万亿次浮点运算/每秒的DP计算容量。使用10.000个元件将达到10千万亿次浮点计算/每秒的总性能。KC的前身，Knight的Ferry处理器(KF)将用于产生基于PCIe的试验性系统的项目中以便研究集群-增强器(CN-ESB)概念。

[0084] 由于KF的计算速度超过目前商品处理器大约10倍，因此内部ESB通信系统必须具有相应的维度。ESB的通信系统需要至少每卡兆兆位/每秒(双工)。通信系统EXTOLL可以用作总线系统的实施方式，总线系统提供每卡1.44兆兆位/每秒的通信率。它实现了提供每卡6个链路的3d拓朴。关于它的简单性，这个拓朴呈现为适用于基于多核加速器的增强器。即使具有为切入路由(cut-through routing)保留的两个方向，EXTOLL也能够使PCI Express性能饱和，只要数据率被关注。当基于专用集成电路(ASIC)实现时，延迟能够达到0.3μs。目前，EXTOLL借助于现场可编程门阵列(FPGA)实现。

[0085] 图4示出用于例示根据本发明的一种计算机集群布置的操作方法的一个方面的流程图。在第一步骤100中，通过多个计算节点CN中的至少两个计算计算任务的至少第一部分，每个计算节点CN都联接通信基础设施IN。此外，在步骤101中，通过至少一个增强器B执行计算任务的至少第二部分的计算，每个增强器B都联接通信基础设施IN。另外，在步骤102中，通过资源管理器RM执行将至少一个增强器B分配给多个计算节点CN之一，用于执行计算任务的第二部分。如图4中右箭头指示，控制流可以指回步骤100。在步骤102中，将至少一个增强器B分配给多个计算节点CN中的至少一个后，该分配能够被通信至计算节点CN，其在进一步的外包步骤中使用被传输的分配。因此，在步骤101中依据分配步骤102的一个执行计算计算任务的至少第二部分。

[0086] 图5示出例示根据本发明一个方面的一种计算机集群布置的操作方法的流程图。在本实施例中，在步骤202中将至少一个增强器B分配给多个计算节点CN之一后，执行计算计算任务的至少第二部分的步骤201。因此，选定具体增强器B并且基于在步骤202中建立的分配，增强器B计算计算任务的至少第二部分是可能的。这在计算任务的至少第二部分被转发至资源管理器RM的情况下可以是优势，资源管理器RM将增强器B分配给计算任务的第二部分。然后，资源管理器RM能够将计算任务的第二部分传输至增强器B，无需计算节点CN直

接联系增强器B。

[0087] 参考图4和5，本领域技术人员理解，任何步骤都能够反复地、以不同顺序执行并且可以包含进一步的子步骤。例如，步骤102可以在步骤101前执行，其导致计算任务的第一部分的计算、一个增强器到一个计算节点的分配以及最终计算任务的第二部分的计算。步骤102可以包含子步骤，诸如将所计算的计算任务的至少第二部分返回至计算节点CN。因此，增强器B将计算结果返回至计算节点CN。计算节点CN可以使用返回值来计算进一步的计算任务的，并且可以再次将计算任务的至少另一部分转发至至少一个增强器B。

[0088] 图6示出根据本发明一个方面的计算机集群布置的控制流的框图。在本实施例中，计算节点CN接收计算任务并且请求增强器B外包至少一部分接收到的计算任务。因此，资源管理器RM被存取，其将该部分计算任务转发至选定的增强器B。增强器B计算该部分计算任务并且返回结果，该结果通过最右边的箭头指示。根据本实施例的另一方面，返回值能够被传递回计算节点CN。

[0089] 图7示出根据本发明一个方面的实施计算机集群布置的反向加速的控制流的框图。在本实施例中，通过将至少一个计算节点CN分配给至少一个增强器B执行通过至少一个增强器B计算的计算任务的计算加速。因此，控制和信息流关于图6所示实施例是反向的。因此，能够通过从增强器B到至少一个计算节点CN外包计算任务而使任务的计算加速。

[0090] 图8示出根据本发明另一方面的计算机集群布置的控制流的框图。在本实施例中，资源管理器RM不将计算任务的至少一个部分传递至增强器B，但计算节点CN请求地址或进一步的增强器B的身份，其被布置成计算具体的计算任务的至少一个部分。资源管理器RM将所需地址返回至计算节点CN。计算节点CN现在能够借助于通信基础设施IN直接存取增强器B。在本实施例中，通信基础设施IN经由联接单元被存取。计算节点CN通过联接单元IU1对通信基础设施IN进行存取，而增强器B通过联接单元IU2联接通信基础设施IN。

[0091] 此外，资源管理器RM被布置成评估增强器B的资源容量并且依据每个增强器B的被评估的资源容量执行分配，这意味着增强器B的选定。为了这样做，资源管理器RM可以存取分配度量，其可以储存在数据库DB或任何类型的数据源内。资源管理器RM被布置成更新分配度量，其能够在使用数据库管理系统的情况下被执行。数据库DB能够被实施为任何类型的储存器。它可以例如被实施为表、寄存器或高速缓存。

[0092] 图9示出根据本发明一个方面的计算机集群布置的网络拓朴的示意性例示。

[0093] 在一个实施例中，计算节点共享公共的第一通信基础设施，例如具有中央切换单元S的星型拓朴。提供另一个第二通信基础设施用于计算节点CN与增强器节点BN的通信。提供第三通信基础设施用于在增强器节点BN之间通信。因此，用于在增强器节点BN之间通信的高速网络接口能够设有具体的BN-BN通信接口。能够将BN-BN通信基础设施实施为3d拓朴。

[0094] 在另一个实施例中，提供两个通信基础设施，一个用于在计算节点CN之间通信，而另一个通信基础设施用于在增强器节点BN之间通信。两个通信基础设施都能够通过从第一个网络至第二个网络或从第二个网络至第一个网络的至少一个通信链路进行耦接。因此，一个选定的计算节点CN或一个选定的增强器节点BN分别与其它网络连接。在本图9中，在使用切换单元S的情况下，一个增强器节点BN与计算节点CN的通信基础设施连接。

[0095] 在另一个实施例中，增强器组BG本身可以连接至计算节点CN的通信基础设施或中

间通信基础设施。

[0096] 通信基础设施一般可以在它们的拓朴、带宽、通信协议、吞吐量和消息交换方面的其它特性当中存在差异。增强器B例如可以包含1至10.000个增强器节点BN，但不限制在这个范围内。资源管理器RM一般可以管理增强器节点BN的多个部分，并且因此能够给增强器节点BN的总数量分区，并且由所述数量的增强器节点BN动态形成增强器B。可以通过开关、路由器或任何网络设备实施切换单元S。

[0097] 本领域技术人员理解计算机集群布置的组件的其它布置。例如数据库DB可以通过计算机集群布置的其它组件、各自的节点进行存取。示出的计算节点CN以及示出的增强器组BG可以分别是许多其它计算节点CN之一以及许多增强器组BG之一，其对资源管理RM和/或通信基础设施IN进行存取。此外，还能够通过从至少一个增强器B到至少一个计算节点外包至少一部分计算任务而反向执行加速。

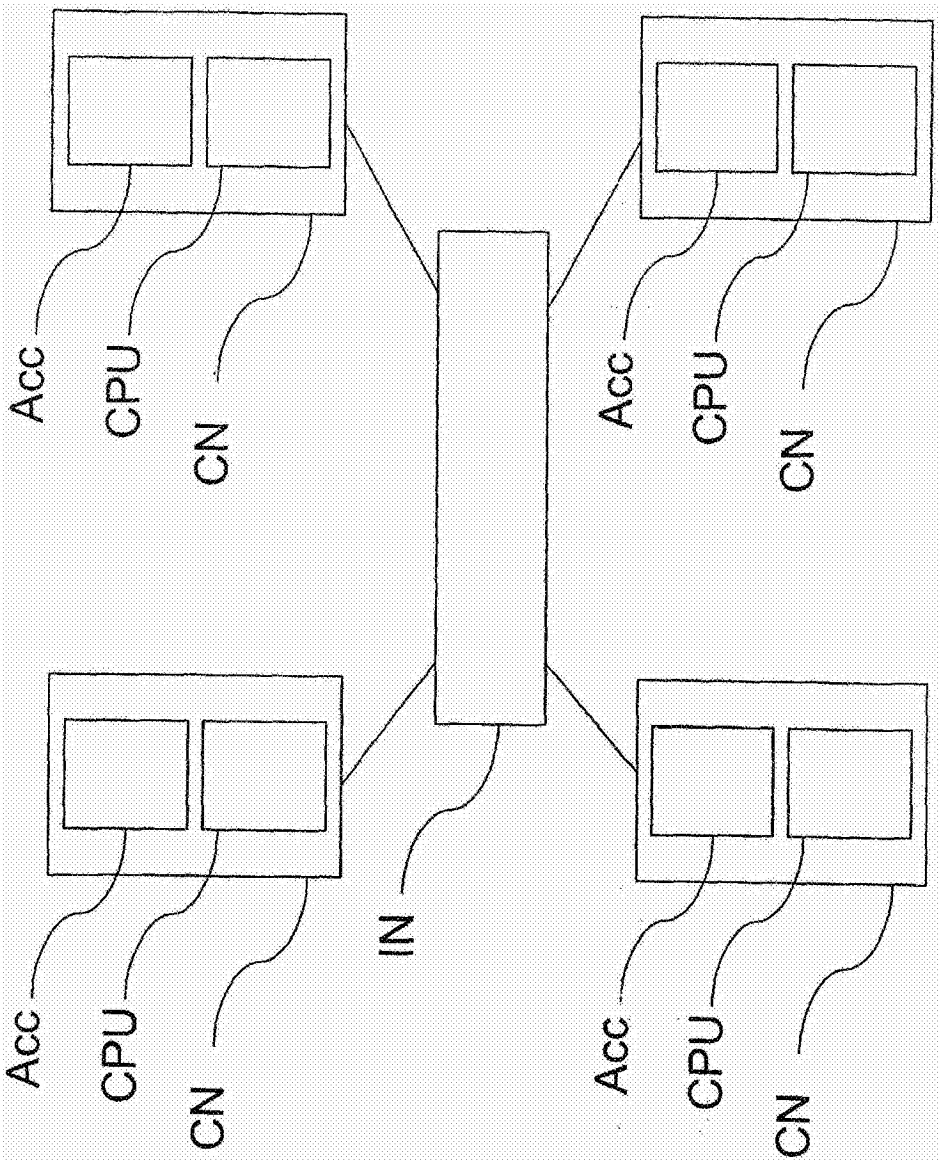



图1

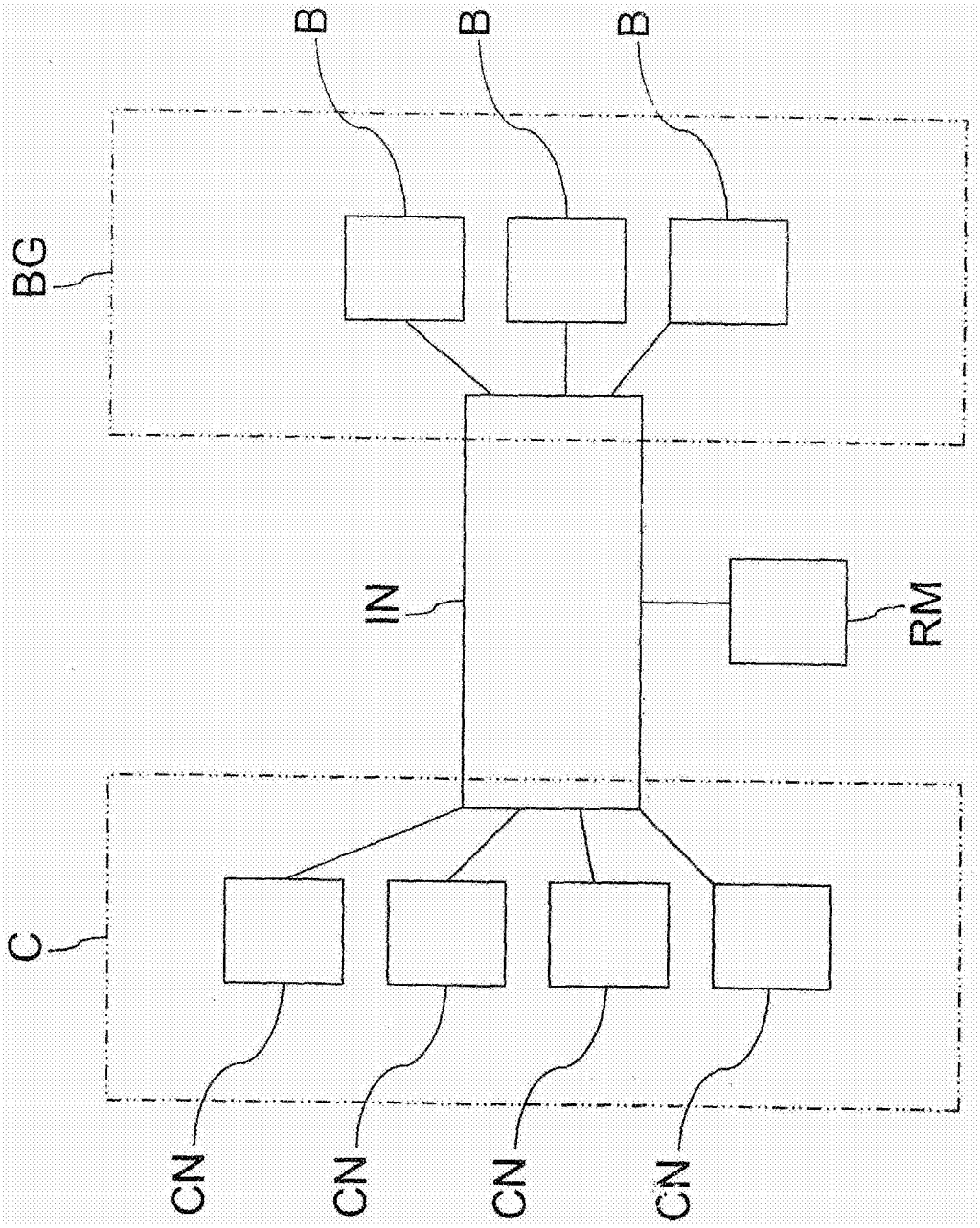



图2

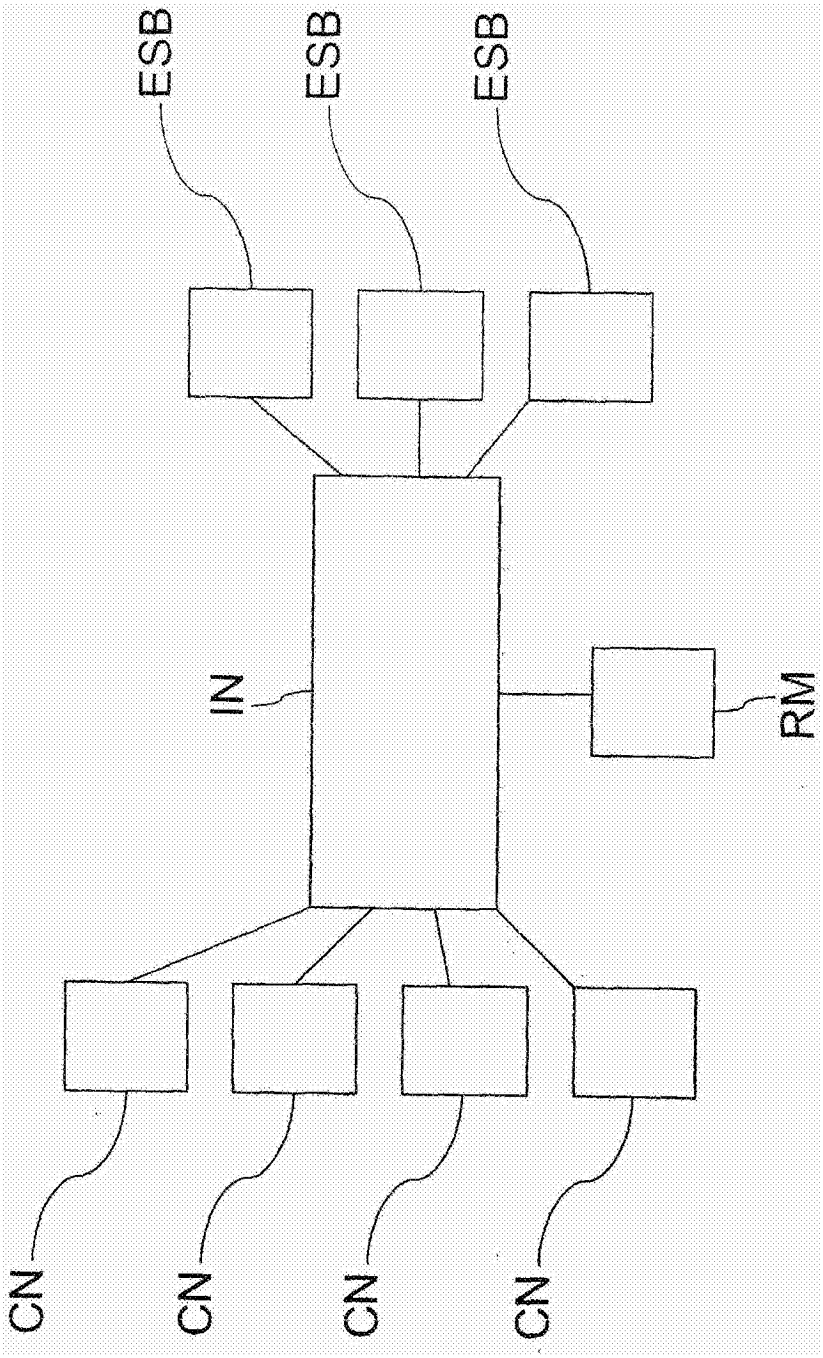



图3

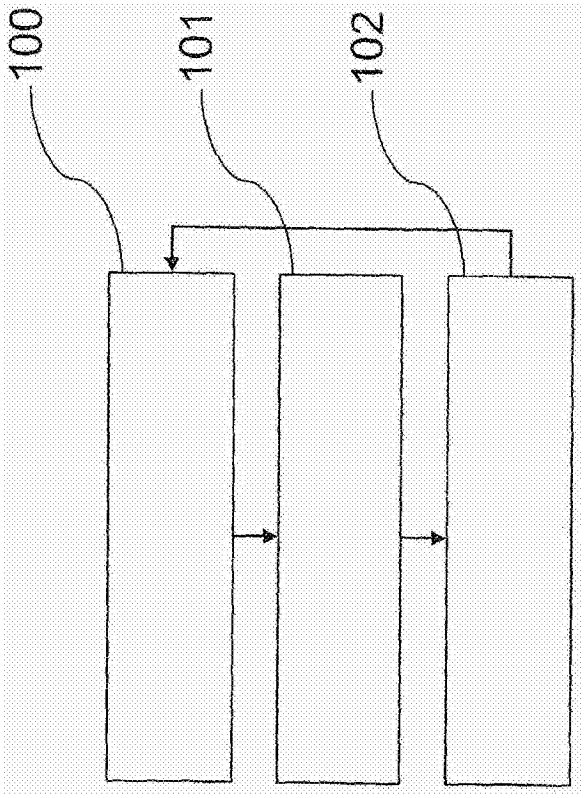



图4

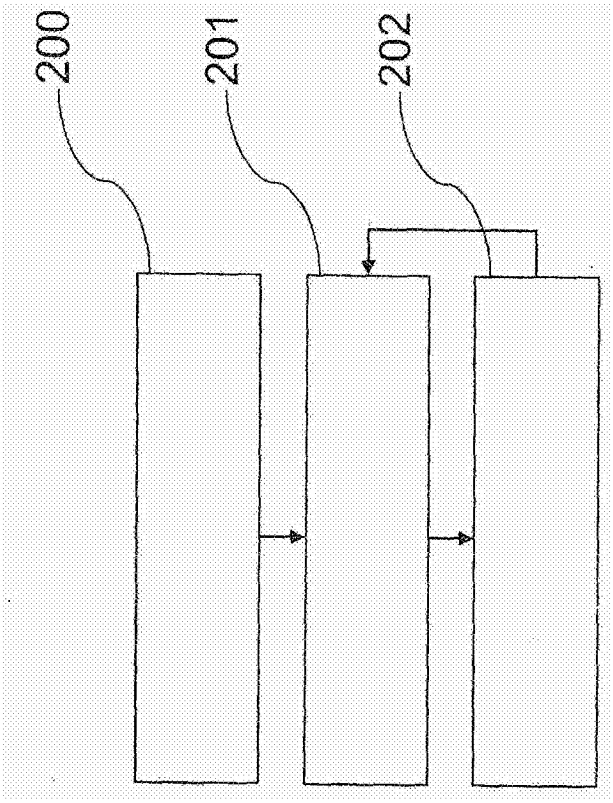



图5

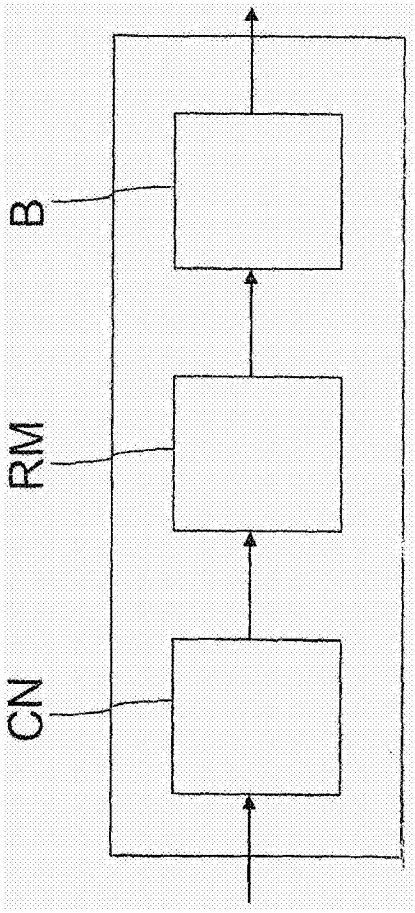



图6

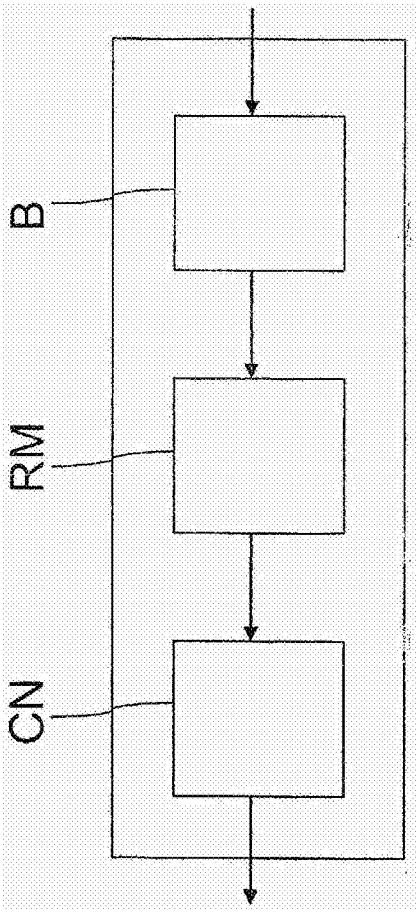



图7

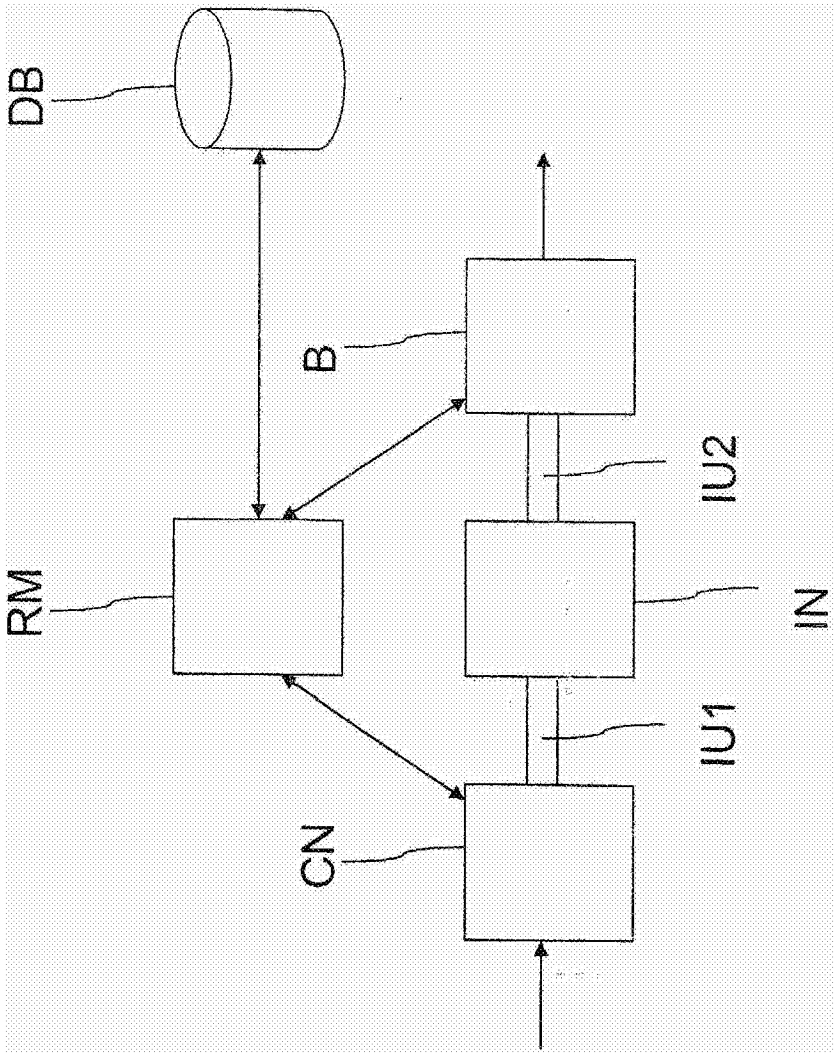



图8

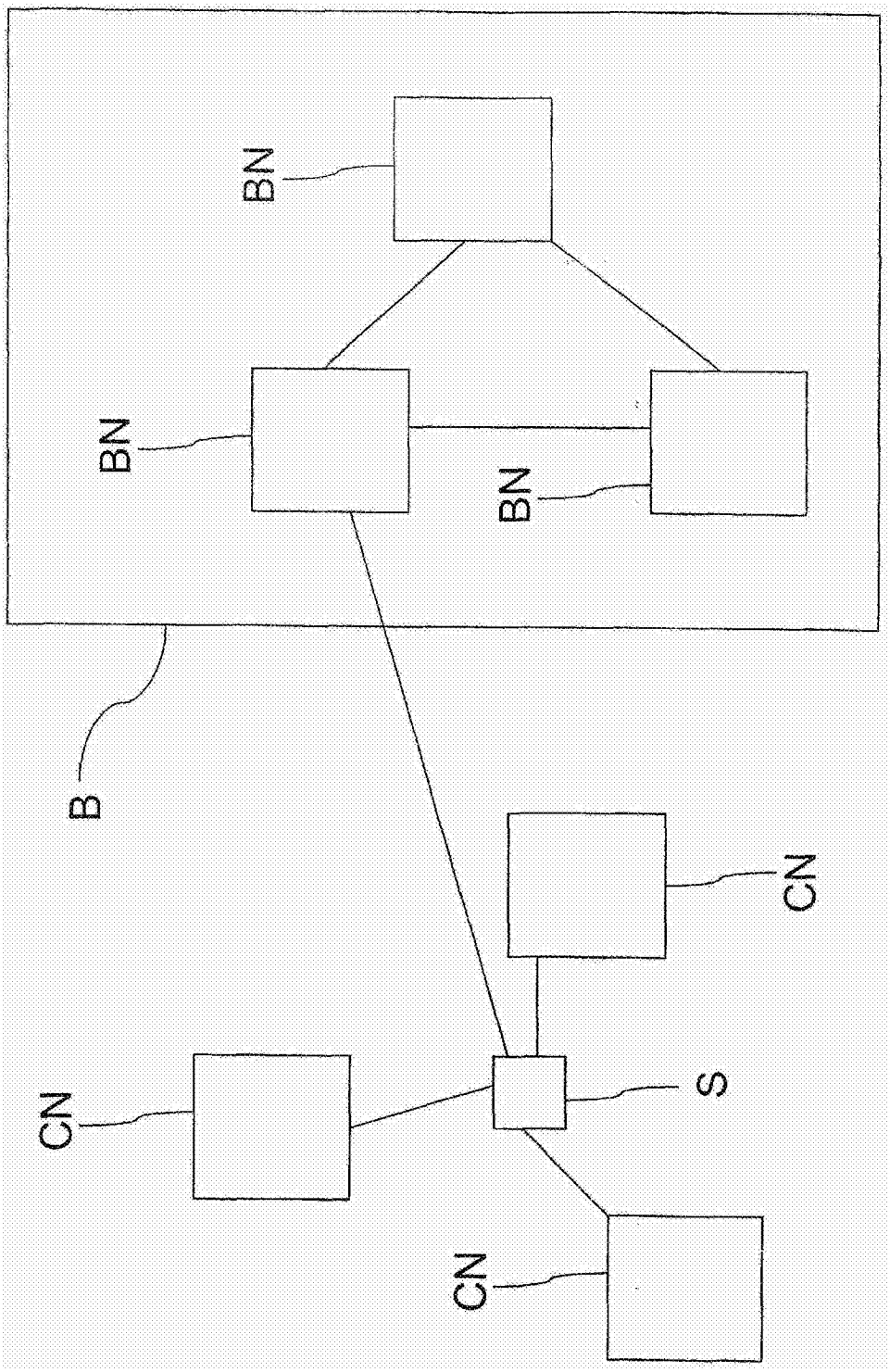



图9