(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 105408864 A
(43)申请公布日 2016.03.16

(21)申请号 201480042796.5
(22)申请日 2014.07.03
(30)优先权数据
13306091.3 2013.07.29 EP
(85)PCT国际申请进入国家阶段日
2016.01.28
(86)PCT国际申请的申请数据
PCT/EP2014/001841 2014.07.03
(87)PCT国际申请的公布数据
WO2015/014431 EN 2015.02.05
(71)申请人 阿尔卡特朗讯
地址 法国布洛涅-比扬古
(72)发明人 F·弗兰克 I·比得尼
(74)专利代理机构 北京市金杜律师事务所 11256
代理人 王茂华 张凡
(51)Int.Cl.
G06F 9/50(2006.01)
H04L 12/24(2006.01)
G06F 15/177(2006.01)
H04L 29/08(2006.01)

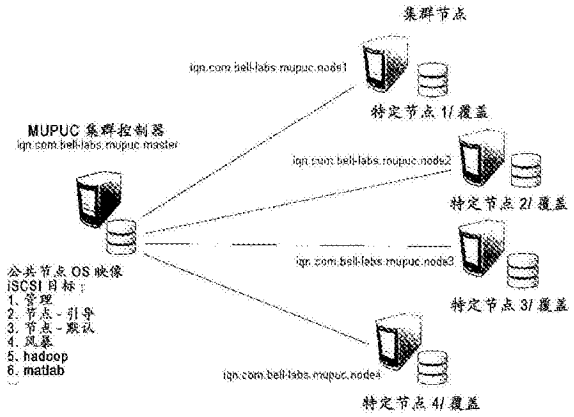
权利要求书2页 说明书12页 附图1页

(54)发明名称

数据处理

(57)摘要

公开了一种数据处理网络、集群控制器、数据处理节点、方法和计算机程序产品。数据处理网络包括：集群控制器，可操作为存储多个配置；以及数据处理节点的集群，所述集群控制器和所述数据处理节点的集群可操作为进行合作以使所述多个配置中的一个配置可用作对所述集群中的每个数据处理节点分配的只读配置，每个数据处理节点可操作为使用所述配置来进行引导，所述集群控制器和数据处理节点的集群进一步可操作为进行合作以对所述集群中的每个数据处理节点分配读/写存储区域，以用于访问在该数据处理节点在引导之后的操作期间要利用的数据。该方法提供了集群的灵活建立，其可以动态地加载特定配置并且自动地对不确定数目的活动节点操作。同时，该方法允许可用机器的分离，以能够并行运行不同的HPC服务。



CN 105408864 A

1. 一种数据处理网络，包括：

集群控制器，可操作为存储多个配置；以及

数据处理节点的集群，所述集群控制器和所述数据处理节点的集群可操作为进行合作以使所述多个配置中的一个配置可用作对所述集群中的每个数据处理节点分配的只读配置，每个数据处理节点可操作为使用所述配置来引导，所述集群控制器和数据处理节点的集群进一步可操作为进行合作以对所述集群中的每个数据处理节点分配读/写存储区域，以用于访问在该数据处理节点在引导之后的操作期间要利用的数据。

2. 根据权利要求 1 所述的数据处理网络，其中所述集群控制器和所述数据处理节点的集群可操作为进行合作，以通过从所述集群控制器向所述集群中的每个数据处理节点传输所述配置中的至少一部分来使得所述分配的只读配置可用。

3. 根据权利要求 1 或 2 所述的数据处理网络，其中每个配置包括盘映像。

4. 根据任何一项前述权利要求所述的数据处理网络，其中每个配置包括组合的操作系统和至少一个应用的盘映像。

5. 根据权利要求 2 或 3 所述的数据处理网络，其中所述集群中的每个数据处理节点可操作为将所述盘映像安装在所述集群控制器上作为本地盘和引导盘中的至少一个。

6. 根据任何一项前述权利要求所述的数据处理网络，其中所述读/写存储区域被安装作为所述盘映像的根目录上的所述文件系统覆盖。

7. 根据权利要求 6 所述的数据处理网络，其中所述文件系统覆盖被分配比所述盘映像更高的优先级，以使得文件系统覆盖修改文件能够优先于对应的盘映像文件而被访问。

8. 根据任何一项前述权利要求所述的数据处理网络，其中所述读/写存储区域位于每个数据处理节点处，并且每个数据处理节点可操作为当被指令执行去激活和重新配置中的一项时，将所述读/写存储区域的内容转移到集中式存储装置。

9. 根据任何一项前述权利要求所述的数据处理网络，其中所述集群控制器可操作为指令数据处理节点的所述集群利用所述配置进行重新引导。

10. 一种数据处理网络的方法，包括：

在集群控制器处存储多个配置；以及

提供数据处理节点的集群；

使得所述多个配置中的一个配置可用作对所述集群中的每个数据处理节点分配的只读配置并且使用所述配置来进行引导；以及

对所述集群中的每个数据处理节点分配读/写存储区域，以用于访问在该数据处理节点在引导之后的操作期间要利用的数据。

11. 一种用于数据处理网络的集群控制器，包括：

存储装置，可操作为存储多个配置；以及

合作逻辑，可操作为与数据处理节点的集群进行合作，以使所述多个配置中的一个配置可用作对所述集群中的每个数据处理节点分配的只读配置，并且可操作为对所述集群中的每个数据处理节点分配读/写存储区域，以用于访问在该数据处理节点在引导之后的操作期间要利用的数据。

12. 一种集群控制器的方法，包括：

存储多个配置；

与数据处理节点的集群进行合作，以使所述多个配置中的一个配置可用作对所述集群中的每个数据处理节点分配的只读配置；以及

对所述集群中的每个数据处理节点分配读/写存储区域，以用于访问在该数据处理节点在引导之后的操作期间要利用的数据。

13. 一种用于数据处理网络的数据处理节点，包括：

合作逻辑，可操作为与集群控制器进行合作，以使所述集群控制器所存储的多个配置中的一个配置可用作分配的只读配置；以及

引导逻辑，可操作为使用所述配置进行引导，所述合作逻辑进一步可操作为进行合作以分配读/写存储区域，以用于访问在引导之后的操作期间要利用的数据。

14. 一种数据处理节点的方法，包括：

与集群控制器进行合作以使所述集群控制器所存储的多个配置中的一个配置可用作分配的只读配置；以及

使用所述配置进行引导。

15. 一种计算机程序产品，可操作为当在计算机上执行时，执行根据权利要求 10、12 或 14 中的任何一项所述的方法步骤。

数据处理

技术领域

[0001] 本发明涉及数据处理网络、集群控制器、数据处理节点、方法和计算机程序产品。

背景技术

[0002] 高性能计算 (HPC) 是集群计算内的专业领域，其中数据处理节点的集群的基础设施可能对其运行的软件的性能有很大影响。这意味着 HPC 应用常常非常专用于其配置中以实现多数的底层计算硬件。HPC 集群的建立通常由三个整体装载阶段构成 :操作系统装载，这对于集群中的所有节点来说通常是公共的 ;软件应用层，这是特定于应用领域 (例如 Hadoop、实时流传送框架、Matlab、科学专用代码) ;和节点必须计算的数据的特定配置和集合。

[0003] 虽然提供 HPC 集群可以在适当配置时提供显著的性能数据处理优势，但是也可能发生不期望的后果。因此，期望提供一种改进的布置。

发明内容

[0004] 根据第一方面，提供了一种数据处理网络，包括 :集群控制器，可操作为存储多个配置 ;以及数据处理节点的集群，集群控制器和数据处理节点的集群可操作为进行合作以使多个配置中的一个可用作对集群中的每个数据处理节点分配的只读配置，每个数据处理节点可操作为使用该配置来引导集群控制器，并且数据处理节点的集群进一步可操作为进行合作以对集群中的每个数据处理节点分配读 / 写存储区域，以用于访问在该数据处理节点在引导之后的操作期间要利用的数据。

[0005] 第一方面认识到，诸如例如 HPC 应用的应用通常在其配置中专用于实现多数底层计算硬件，并且这进而使其不适合布置在可重新配置的云状环境中。这压低了成本效益以及传统 HPC 布置的实现容易度。具体地，在操作系统、应用软件以及特定配置和数据集之间的分离使得实现动态可重新配置的 HPC 系统，即按需 HPC 作为服务 HPCaaS、硬性和挑战性技术任务。实际上，具有许多不同的目的意味着集群应当具有运行许多不同种类的软件的能力，并且科学 / 网格计算软件通常需要对所讨论的软件唯一的非常特定的设置以便以最高性能运行。为了促进该在配置中的灵活性，使对集群的维护保持为最小并且提供一种用于在多模式集群在一步操作中需要一系列的技术不便时从一个配置切换为另一个的简便方式，该技术不便应当被解决以保持系统一致和可操作。此外，目前没有有效的方式来在没有虚拟化或专用软件装载的情况下提供专用机器。然而，第一方面还认识到，虚拟化在硬件和软件之间引入了重新定向层，对性能产生了不期望和不可预测的影响。对该方法的替代是非虚拟化专用软件装载，这避免了这些缺陷。然而，该方法引入了集群的更静态的配置，其中该软件必须通过多引导解决方案被装载在每个单个机器上，这消耗盘资源或者不允许特定配置。换言之，当前存在两个主要方式来进行 HPC 平台的布置，第一个是布置专用于一个 HPC 任务的平台 (例如 Hadoop、风暴等)。其优点是，可以充分利用 (通常是非常昂贵的) 运转平台的基础设施，但是其缺点是，如果没有该具体类型的工作可用于执行，则硬件将保

持不被利用。第二个是在顶部布置具有虚拟化层的通用基础设施。这允许系统完全用于任何数目的任务。然而，这还意味着，HPC 软件在虚拟化环境内运行——就如同该类型的应用通常遭受高性能损失。这些解决方案在如下布置情形下都没有多大的意义，在该布置情形中，HPC 任务在不同的软件系统之间变化，诸如例如研究环境或者在基础设施作为服务市场中。不期望在完全硬件利用和最优灵活度之间进行选择。

[0006] 因此，可以提供数据处理网络。数据处理网络可以包括集群控制器。集群控制器可以存储多于一个的配置。数据处理网络还可以包括数据处理节点的集群。集群控制器和集群可以进行协作或一起起作用，以使得配置中的一个可用于每个数据处理节点或可由每个数据处理节点访问。所分配的配置可以被提供为只读配置。然后，每个数据处理节点可以使用所分配的配置来引导。集群控制器和数据处理节点的集群还可以一起起作用，以提供用于每个数据处理节点的读和 / 或写存储区域。所分配的读 / 写存储区域可以用于访问在数据处理节点在引导之后的操作期间使用的数据。该方法提供了集群的灵活建立，其可以动态地加载特定配置并且自动地在不确定的数目的活动节点上操作。同时，该方法允许可用机器的分离，以便于能够同时运行不同的 HPC 服务。这通过提供具有正确的顺序和精心策划的特征的组合而成为可能，允许对具有提供对每个节点的特定访问的能力的相同物理硬件驱动的共享。这将可重新配置的云布置框架的很多优点代入 HPC 情形，而不影响性能和可配置性。

[0007] 在一个实施例中，集群控制器和数据处理节点的集群可操作为进行合作，以使得多个配置中的同一个可用作对集群中的每个数据处理节点分配的只读配置。因此，集群控制器可以对集群中的每个数据处理节点提供相同的配置。这确保了集群内的每个节点通过相同的配置进行引导。

[0008] 在一个实施例中，集群控制器和数据处理节点的集群可操作为进行合作，以通过将配置的至少一部分从集群控制器传输到集群中的每个数据处理节点来使得所分配的只读配置可用。因此，可以将配置中的至少一些从集群控制器传输到集群中的每个数据处理节点。这使得单个集群控制器能够配置许多数据处理节点。

[0009] 在一个实施例中，每个配置包括盘映像 (disk image)。提供盘映像是用于确保每个数据处理节点以相同方式被配置的方便方式。

[0010] 在一个实施例中，每个配置包括组合的操作系统和至少一个应用的盘映像。因此，操作系统和应用二者可以通过盘映像来提供。这使得每个数据处理节点能够按需要容易地通过不同的操作系统和应用来重新配置。

[0011] 在一个实施例中，集群中的每个数据处理节点可操作为将盘映像安装 (mount) 在集群控制器上作为本地盘。因此，由集群控制器提供的盘映像可以被安装在每个数据处理节点上作为本地盘。

[0012] 在一个实施例中，集群中的每个数据处理节点可操作为将盘映像安装在集群控制器上作为引导盘。再次，这是特别方便的，因为数据处理节点可以被配置为使用该盘来进行引导。

[0013] 在一个实施例中，读 / 写存储区域位于集群控制器、数据处理节点和集中式存储中的至少一个处。因此，读 / 写存储区域可以位于网络内的可访问位置处。

[0014] 在一个实施例中，读 / 写存储区域被安装为文件系统覆盖 (overlay) 和联合安装。

[0015] 在一个实施例中，读 / 写存储区域被安装为盘映像的根目录上的文件系统覆盖。因此，读 / 写存储区域可以与盘映像组合。这使得盘映像的内容能够针对每个数据处理节点以受控的方式被有效地修改（而盘映像本身实际上不被修改 - 仅覆盖），以适应该数据处理节点的具体需要。

[0016] 在一个实施例中，文件系统覆盖被分配比盘映像更高的优先级，以使得文件系统覆盖修改文件能够优先于对应的盘映像文件被访问。因此，覆盖可以被配置为具有比映像更大的优先级，以便于覆盖内的文件优先于盘映像内的那些被呈现。

[0017] 在一个实施例中，读 / 写存储区域位于每个数据处理节点处，并且每个数据处理节点可操作为当被指令执行去激活和重新配置中的一个时，将读 / 写存储区域的内容转移到集中式存储装置。因此，存储区域可以由每个数据处理节点来提供。每个数据处理节点可以将该存储区域的内容转移到集中式存储装置，以便于在数据处理节点被去激活或重新配置时保持覆盖的内容。

[0018] 在一个实施例中，集群控制器可操作为指令数据处理节点的集群利用该配置进行重新引导。

[0019] 在一个实施例中，集群控制器可操作为指令数据处理节点的集群中的不同组来利用不同的配置进行重新引导。因此，完整的集群可以被分成不同的组或子组，这些中的每一个可以由集群控制器指令以利用不同的配置进行引导。这使得数据处理网络能够被配置为多于一个的 HPC 布置，以便于适应不同用户的需要。应当理解，上述和下述特征中的每一个可以由这样的组来使用。

[0020] 在一个实施例中，集群控制器可操作为响应于对这样的改变的请求来改变集群中的数据处理节点的数目。因此，当需要更多或更少的资源时，那么集群控制器可以请求集群内的数据处理节点的数目的改变。

[0021] 在一个实施例中，集群控制器可操作为指令数据处理节点的集群来执行通电和断电中的一个。

[0022] 在一个实施例中，每个数据处理节点可操作为被分配唯一标识符。

[0023] 在一个实施例中，唯一标识符是基于与每个数据处理节点相关联的媒体接入控制地址来确定的。

[0024] 根据第二方面，提供了一种数据处理网络的方法，包括 ：在集群控制器处存储多个配置 ；以及提供数据处理节点的集群 ；使得多个配置中的一个可用作对集群中的每个数据处理节点分配的只读配置并且使用该配置来进行引导 ；以及对集群中的每个数据处理节点分配读 / 写存储区域，以用于访问在该数据处理节点在引导之后的操作期间要利用的数据。

[0025] 在一个实施例中，使得的步骤包括 ：使得多个配置中的同一个可用作对集群中的每个数据处理节点分配的只读配置。

[0026] 在一个实施例中，使得的步骤包括 ：使得通过将配置的至少一部分从集群控制器传输到集群中的每个数据处理节点来使得所分配的只读配置可用。

[0027] 在一个实施例中，每个配置包括盘映像。

[0028] 在一个实施例中，每个配置包括组合的操作系统和至少一个应用的盘映像。

[0029] 在一个实施例中，该方法包括将盘映像安装在集群控制器上作为集群中的每个数

据处理节点的本地盘。

[0030] 在一个实施例中，该方法包括将盘映像安装在集群控制器上作为集群中的每个数据处理节点的引导盘。

[0031] 在一个实施例中，分配的步骤包括：将读/写存储区域定位在集群控制器、数据处理节点和集中式存储装置中的至少一个处。

[0032] 在一个实施例中，分配的步骤包括：安装读/写存储区域作为文件系统覆盖和联合安装中的至少一个。

[0033] 在一个实施例中，分配的步骤包括：安装读/写存储区域作为盘映像的根目录上的文件系统覆盖。

[0034] 在一个实施例中，分配的步骤包括：对文件系统覆盖分配比盘映像更高的优先级，以使得文件系统覆盖修改文件能够优先于对应的盘映像文件被访问。

[0035] 在一个实施例中，分配的步骤包括：将读/写存储区域定位在每个数据处理节点处，该方法包括下述步骤：当被指令执行去激活和重新配置中的一个时，将读/写存储区域的内容转移到集中式存储装置。

[0036] 在一个实施例中，该方法包括：指令数据处理节点的集群利用该配置进行重新引导。

[0037] 在一个实施例中，该方法包括：指令数据处理节点的集群的不同组来利用不同的配置进行重新引导。应当理解，上述和下述特征中的每一个可以由这样的组来使用。

[0038] 在一个实施例中，该方法包括：响应于对这样的改变的请求来改变集群中的数据处理节点的数目。

[0039] 在一个实施例中，该方法包括：指令数据处理节点的集群来执行通电和断电中的一个。

[0040] 在一个实施例中，该方法包括：对每个数据处理节点分配唯一标识符。

[0041] 在一个实施例中，该方法包括：基于与每个数据处理节点相关联的媒体接入控制地址来确定唯一标识符。

[0042] 根据第三方面，提供了一种用于数据处理网络的集群控制器，包括：存储装置，可操作为存储多个配置；以及合作逻辑，可操作为与数据处理节点的集群进行合作，以使多个配置中的一个可用作对集群中的每个数据处理节点分配的只读配置，并且可操作为对集群中的每个数据处理节点分配读/写存储区域，以用于访问在该数据处理节点在引导之后的操作期间要利用的数据。

[0043] 在一个实施例中，合作逻辑可操作为进行合作，以使得多个配置中的同一个可用作对集群中的每个数据处理节点分配的只读配置。

[0044] 在一个实施例中，合作逻辑可操作为进行合作，以通过将配置的至少一部分从集群控制器传输到集群中的每个数据处理节点来使得所分配的只读配置可用。

[0045] 在一个实施例中，每个配置包括盘映像。

[0046] 在一个实施例中，每个配置包括组合的操作系统和至少一个应用的盘映像。

[0047] 在一个实施例中，读/写存储区域位于集群控制器、数据处理节点和集中式存储装置中的至少一个处。

[0048] 在一个实施例中，集群控制器包括指令逻辑，可操作为指令数据处理节点的集群

利用该配置进行重新引导。

[0049] 在一个实施例中，集群控制器包括指令逻辑，可操作为指令数据处理节点的集群的不同组来利用不同的配置进行重新引导。应当理解，上述和下述特征中的每一个可以由这样的组来使用。

[0050] 在一个实施例中，集群控制器包括指令逻辑，可操作为响应于对这样的改变的请求来改变集群中的数据处理节点的数目。

[0051] 在一个实施例中，集群控制器包括指令逻辑，可操作为指令数据处理节点的集群来执行通电和断电中的一个。

[0052] 在一个实施例中，集群控制器包括指令逻辑，可操作为对每个数据处理节点分配唯一标识符。

[0053] 在一个实施例中，唯一标识符是基于与每个数据处理节点相关联的媒体接入控制地址来确定的。

[0054] 根据第四方面，提供了一种集群控制器方法，包括：存储多个配置；以及与数据处理节点的集群进行合作，以使多个配置中的一个可用作对集群中的每个数据处理节点分配的只读配置；以及对集群中的每个数据处理节点分配读/写存储区域，以用于访问在该数据处理节点在引导之后的操作期间要利用的数据。

[0055] 在一个实施例中，合作步骤包括：使得多个配置中的同一个可用作对集群中的每个数据处理节点分配的只读配置。

[0056] 在一个实施例中，合作步骤包括：通过将配置的至少一部分从集群控制器传输到集群中的每个数据处理节点来使得所分配的只读配置可用。

[0057] 在一个实施例中，每个配置包括盘映像。

[0058] 在一个实施例中，每个配置包括组合的操作系统和至少一个应用的盘映像。

[0059] 在一个实施例中，该方法包括：将读/写存储区域定位在集群控制器、数据处理节点和集中式存储装置中的至少一个处。

[0060] 在一个实施例中，该方法包括：指令数据处理节点的集群利用该配置进行重新引导。

[0061] 在一个实施例中，该方法包括：指令数据处理节点的集群的不同组来利用不同配置进行重新引导。应当理解，上述和下述特征中的每一个可以由这样的组来使用。

[0062] 在一个实施例中，该方法包括：响应于对这样的改变的请求来改变集群中的数据处理节点的数目。

[0063] 在一个实施例中，该方法包括：指令数据处理节点的集群来执行通电和断电中的一个。

[0064] 在一个实施例中，该方法包括：对每个数据处理节点分配唯一标识符。

[0065] 在一个实施例中，基于与每个数据处理节点相关联的媒体接入控制地址来确定唯一标识符。

[0066] 根据第五方面，提供了一种用于数据处理网络的数据处理节点，包括：合作逻辑，可操作为与集群控制器进行合作以使集群控制器存储的多个配置中的一个可用作分配的只读配置；以及引导逻辑，可操作为使用该配置进行引导，合作逻辑进一步可操作为进行合作以分配读/写存储区域，以用于访问在引导之后的操作期间要利用的数据。

[0067] 在一个实施例中，合作逻辑可操作为进行合作，以使得多个配置中的同一个可用作分配的只读配置。

[0068] 在一个实施例中，合作逻辑可操作为进行合作，以通过从集群控制器接收配置的至少一部分来使得分配的只读配置可用。

[0069] 在一个实施例中，每个配置包括盘映像。

[0070] 在一个实施例中，每个配置包括组合的操作系统和至少一个应用的盘映像。

[0071] 在一个实施例中，合作逻辑可操作为安装盘映像作为本地盘。

[0072] 在一个实施例中，合作逻辑可操作为进行合作以安装盘映像作为引导盘。

[0073] 在一个实施例中，读 / 写存储区域位于集群控制器、数据处理节点和集中式存储装置中的至少一个处。

[0074] 在一个实施例中，合作逻辑可操作为安装读 / 写存储区域作为文件系统覆盖和联合安装中的至少一个。

[0075] 在一个实施例中，合作逻辑可操作为安装读 / 写存储区域作为在盘映像的根目录上的文件系统覆盖。

[0076] 在一个实施例中，合作逻辑可操作为对文件系统覆盖分配比盘映像更高的优先级，以使得文件系统覆盖修改文件能够优先于对应的盘映像文件被访问。

[0077] 在一个实施例中，读 / 写存储区域位于数据处理节点处，并且合作逻辑可操作为，当被指令执行去激活和重新配置中的一个时，将读 / 写存储区域的内容转移到集中式存储装置。

[0078] 在一个实施例中，引导逻辑可操作为响应于来自集群控制器的指令来利用该配置重新引导。

[0079] 在一个实施例中，引导逻辑可操作为响应于来自集群控制器的指令来执行通电和断电中的一个。

[0080] 在一个实施例中，合作逻辑可操作为分配唯一标识符。

[0081] 在一个实施例中，唯一标识符是基于与数据处理节点相关联的媒体接入控制地址来确定的。

[0082] 根据第六方面，提供了一种数据处理节点方法，包括：与集群控制器进行合作以使集群控制器存储的多个配置中的一个可用作分配的只读配置；使用该配置进行引导；以及分配读 / 写存储区域，以用于访问在引导之后的操作期间要利用的数据。

[0083] 在一个实施例中，合作步骤包括：使得多个配置中的同一个可用作分配的只读配置。

[0084] 在一个实施例中，合作步骤包括：通过从集群控制器接收配置的至少一部分来使得分配的只读配置可用。

[0085] 在一个实施例中，每个配置包括盘映像。

[0086] 在一个实施例中，每个配置包括组合的操作系统和至少一个应用的盘映像。

[0087] 在一个实施例中，合作步骤包括：安装盘映像作为本地盘。

[0088] 在一个实施例中，合作步骤包括：安装盘映像作为引导盘。

[0089] 在一个实施例中，读 / 写存储区域位于集群控制器、数据处理节点和集中式存储装置中的至少一个处。

[0090] 在一个实施例中，合作步骤包括：安装读/写存储区域作为文件系统覆盖和联合安装中的至少一个。

[0091] 在一个实施例中，合作步骤包括：安装读/写存储区域作为盘映像的根目录上的文件系统覆盖。

[0092] 在一个实施例中，合作步骤包括：对文件系统覆盖分配比盘映像更高的优先级，以使得文件系统覆盖修改文件能够优先于对应的盘映像文件被访问。

[0093] 在一个实施例中，读/写存储区域位于数据处理节点处，并且其中合作的步骤包括：当被指令执行去激活和重新配置中的一个时，将读/写存储区域的内容转移到集中式存储装置。

[0094] 在一个实施例中，引导的步骤包括响应于来自集群控制器的指令来用该配置重新引导。

[0095] 在一个实施例中，该方法包括响应于来自集群控制器的指令来执行通电和断电中的一个。

[0096] 在一个实施例中，该方法包括分配分配唯一标识符。

[0097] 在一个实施例中，唯一标识符是基于与数据处理节点相关联的媒体接入控制地址来确定的。

[0098] 根据第七方面，提供了一种计算机程序产品，可操作为当在计算机上执行时，执行第二、第四或第六方面的方法步骤。

[0099] 在所附独立和从属权利要求中阐述了其他具体和优选方面。从属权利要求的特征可以与独立权利要求的特征适当地并且在除了权利要求中明确阐述的那些之外的组合中被组合。

[0100] 当装置特征被描述为可操作为提供功能时，应当理解，这包括装置特征，该装置特征提供了该功能或者适配或配置为提供该功能。

附图说明

[0101] 现在将参考附图来描述本发明的实施例，在附图中：

[0102] 图 1 图示了根据一个实施例的 MUPUC（多用途集群）HPC 集群

具体实施方式

[0103] 概述

[0104] 在更详细地讨论实施例之前，首先将提供概述。实施例提供了一种布置，其中提供了集中式集群控制器，该集中式集群控制器能够分配、去分配和/或重新配置集群内的数据处理节点。具体地，集群控制器存储多个不同的配置，其中的任何一个可以被提供给集群内或集群内的组内的每个数据处理节点。这些配置中的每一个可以包括例如特定的操作系统和/或一个或多个应用。响应于来自用户的请求而将这些配置中的一个分配给集群或组内的每个数据处理节点，以提供多个数据处理节点，其中的每一个具有该配置。一旦数据处理节点已经被分配给集群并且集群控制器已经配置了每个数据处理节点，数据处理节点然后就可以使用该配置来进行引导。这使得多个数据处理节点中的每一个能够以相同的配置被引导，如用户所请求的。通过使配置只读，由数据处理节点进行的任何改变不影响由集群

控制器所提供的配置。为了使得能够在数据处理节点的操作期间创建和存储数据，提供数据可以被读取和 / 或写入的读 / 写存储区域。这提供了灵活的布置。

[0105] 网络概述

[0106] 实施例提供了一种布置，该布置寻求使得能够获得具有完全可重配置的高性能计算 (HPC) 解决方案的优点，而不施加由于虚拟化环境而产生的限制。具体地，实施例提供了 HaaS(HPC 作为服务)，其将集群控制器引入集群系统的建立。为了与基于云的布置比较，集群控制器是用作用于集群中的剩余节点的一种管理程序的特殊节点。然而，当 HPC 软件直接在集群节点的硬件 (“集群节点”是集群中没有作为集群控制器进行操作的节点) 上运行时，没有引入虚拟化层，带来了在性能方面的所有优点。集群控制器仅用作编排器 (orchestrator)，并且没有必要参与由集群节点进行的计算。

[0107] 集群控制器

[0108] 集群控制器使用对允许其管理它们的低级功能的集群节点的接口 (例如智能平台管理接口 (IPMI) 或一些定制软件)，诸如它们必须打开或关闭还是重新引导等。此外，集群控制器保持系统映像的存储库，每个系统映像都保持集群用户能够请求的配置中的一个。每个系统映像可以被认为是可与虚拟机映像比较的 HaaS——然而，重要的区别是在节点上一次仅一个映像可以是活动的。当用户请求特定配置时，集群控制器使用其管理工具来指令集群中的任何数目的节点在该配置下通电。

[0109] 使系统映像驻留在集群控制器上提供了从基于云的系统已知的装载一次布置很多的功能。然而，该布置本身具有明显的缺点 :第一，当映像驻留在集群控制器上时，该节点必须通过网络安装的文件系统来对其进行访问。这样的文件系统的性能通常远低于本地盘文件系统的性能。第二，使多个节点共享单个映像在节点需要将数据写入文件系统时将产生问题。要从云系统获得的期望分离质量因此丢失。

[0110] 因此，实施例将映像存储库的网络引导过程与定制分区方案组合，允许用户透明地存储、修改和删除在每个节点本地的数据。

[0111] 因此，实施例提供了 :唯一的可共享系统映像装载点 ;系统映像与持久性数据的明确分离 ;特定引导节点的自动识别 ;能够在运行时间动态地读取和专门重新配置每个节点的参数化引导加载器 ;以及简单的配置切换器。

[0112] 系统映像

[0113] 唯一系统映像的功能允许在集群中的所有节点之间建立共享的操作系统映像。该特征具有许多优点。例如，这使得能够接通必要数目的节点，而不必将系统装载在每个节点上。即使在具有几百个节点的大的集群中，或者当新的机器被添加到集群时，仅必要的操作配置该机器来访问存储在集群控制器上的系统映像，并且该系统将能够引导到功能环境中。这里的技术难题是如何维护系统处于一致状态。具体地，如果许多节点使用同一共享系统分区，则文件系统将由于不同客户端的许多运行时写入访问而变得不一致。为了避免这样的问题，包含和共享系统装载的盘分区被呈现为在执行中保持不可变状态的只读盘。

[0114] 可写入存储装置

[0115] 因此，实施例还向每个节点提供可写入存储装置，在没有对操作系统软件的任何修改的情况下，集群可以持续运行。以下述方式安装盘的可写入部分 :该方式同时提供对可读分区的直接访问，而不对传统文件系统路径进行任何修改。

[0116] 在实施例中，HaaS 系统上的数据遵从成为两类的严格划分：应用映像和数据集配置覆盖（或简称为覆盖）。覆盖是给予用户的可变沙箱（sandbox）以存储他们的数据集、附加程序或任何他们需要以运行他们的应用的东西。每个覆盖完全独立于其他覆盖，并且独立于任何应用映像。因此，任何覆盖可以与系统上的应用映像中的任何一个一起使用，并且这取决于覆盖的所有者以保持其处于有用或正常状态。

[0117] 实际上，名字（moniker）覆盖被有意地选择，因为 HaaS 覆盖正是在节点已经引导的应用映像的顶部透明地层叠的文件系统。因此，覆盖不与数据库世界中的实际数据集相混淆，也不能与从例如虚拟机已知的操作系统快照相混淆。当已经创建和激活覆盖时（当节点已经通过覆盖被引导时），HaaS 用户进行的一切被直接存储到该具体覆盖中的盘。因此，不需要请求系统的“快照”以保存工作，这自动地发生，并且保存的一切将在下一次激活覆盖时可用。

[0118] 因为覆盖被层叠在整个文件系统的顶部，所以用户可以甚至更新应用、装载新的应用或从应用映像中删除数据，而不实际影响应用映像本身。在该意义上，作为个人沙箱的覆盖的想法是准确的类比。

[0119] 当覆盖表示层叠在由集群控制器服务的应用映像的顶部上的节点的盘上的数据时，覆盖将包含每个节点上的不同数据。这样，可以创建覆盖，其中节点 1 被配置为用作用于分布式计算系统的主节点，并且节点 4 是节点 1 的回退。然后，该配置在接下来激活覆盖时将持续，并且节点中的每一个将处于与在配置完成时相同的状态。

[0120] 唯一标识符

[0121] 实施例提供了用于每个集群节点的、在引导时间时可识别的唯一标识。应当理解，重要的是，在引导时明确区分每个单个机器，否则共享包含系统映像的相同物理盘的结果将导致不可使用的集群，其中所有的或一些节点具有相同的因特网协议（IP）地址和名称。因此，为了使系统可利用，针对每个节点自动地获得唯一标识符，使得其可在引导时直接可用。在一个实施例（如下所述）中，特定脚本获得引导机器的 MAC 地址，并且从特定的预填充文件进行读取，在标准引导加载器配置文件通过系统建立被读取之前修改运行中标准引导加载器配置文件。

[0122] 集群控制器通过定制工具来设置要在引导时加载的期望的系统映像和数据集覆盖，该定制工具通过在运行时动态地用正确的命令参数来改变节点的引导配置来允许设置期望值。

[0123] 实施例提供了多功能集群，该多功能集群向用户提供以简单的一步动作从一个配置切换为另一配置的能力，这提供了容易可重新配置的系统。而且，集群中的节点可以被分成不同的组，并且每个组可以被单独配置，由此同时有效提供多个不同集群。

[0124] 示例性实现

[0125] 图 1 图示了根据一个实施例的 MUPUC（多用途集群）HPC 集群。使用因特网小型计算机系统接口 iSCSI 和覆盖文件系统获得在统映像和应用 / 节点配置和数据之间的适当水平的分离。

[0126] 应当理解，iSCSI 是用于链接数据存储设施的基于 IP 的存储联网标准。通过在 IP 网络上承载 SCSI 命令，iSCSI 用于促进通过内联网的数据传输并且用于管理甚至通过长距离的存储。该协议允许客户端（称为发起者）将 SCSI 命令（CDB）发送到远程服务器上的

SCSI 存储设备（目标）。这是存储区域网络 (SAN) 协议，允许组织将存储整合成数据中心存储阵列，而向主机（诸如数据库和 web 服务器）提供本地附连盘的假象。不同于需要专用电缆的传统光纤信道，iSCSI 可以使用现有网络基础设施通过长距离运行。

[0127] 覆盖文件系统 (OFS) 是用于 Linux 的文件系统服务，其实现用于其他文件系统的联合安装。其允许称为分支的分离的文件系统目录和文件被透明地覆盖，形成单个一致的文件系统。在合并的分支内具有相同路径的目录的内容将在新的虚拟文件内的单个合并的目录中被看成一起。在安装分支时，指定一个分支相对其他的优先级。因此，当两个分支包含具有相同名称的文件时，一个得到高于另一个的优先级。不同的分支可以是只读和读写文件系统二者，使得对虚拟合并副本的写入被导引到特定实际的文件系统。这允许文件系统呈现为可写入的，但是实际上不允许写入改变文件系统，也称为副本上写入。该特征一方面在 MUPUC 中高度使用以将由以上示出的 iSCSI 功能管理的系统映像与持久性数据存储在物理上分离，并且另一方面在需要时允许用于每个节点的更细颗粒度的配置。

[0128] 系统设置

[0129] MUPUC 目前由一个集群控制器节点和四个工作者节点组成，但是可以理解，可以提供任何数目的工作者节点。集群控制器维护其系统映像的集合作为 iSCSI 目标，并且工作者节点能够从这些目标执行网络引导。集群控制器上的所有 iSCSI 目标被表示为只读盘，这使得其中的每一个是用于一个系统映像的不变容器。然而，当工作者节点使系统在线时，它们需要某个读写存储装置（用于日志文件、应用数据、系统文件等）。这是通过使用覆盖文件系统来实现的，其中本地盘存储被透明地安装在工作者工作者节点上的只读网络文件系统的顶部。

[0130] 工作者节点用户看到的文件系统因此是从集群控制器服务的只读 iSCSI 盘映像和保持用户已经对映像进行的所有修改的持久性覆盖的组合。该覆盖通常从本地附连的盘提供服务，并且因此对于在其上进行修改的节点是本地的。

[0131] 引导目标管理器是使得能够对 MUPUC 系统进行简单重新配置的集群控制器的特殊 MUPUC 工具。通过操纵工作者节点的引导参数，可以指令其中的每一个关闭给定系统映像并且将给定覆盖安装在其顶部。

[0132] 这些引导参数被传递给在工作者节点上加载的操作系统内核，并且在启动期间由工作者节点来检测。另一 MUPUC 工具 - 启动管理器 - 使得确保这些参数在集群用户可以达到节点之前生效。

[0133] 在 MUPUC 系统中发出的内核参数的示例：

[0134] frfranck@node2:- ～ $cat/proc/cmdline

[0135] BOOT_IMAGE = /vmlinuz-3.2.0-23-generic root = /dev/sdb1ro textonly

[0136] nomodeset

[0137] ip = cluster

[0138] iscsi__target_name = iqn.2012-o8.com.bell-labs.mupuc:node-default

[0139] overlay =

[0140] 由于通过系统图像布置的单个配置文件而导致这全都是可能的，其将在节点上的联网硬件映射成特定网络配置。该配置进而允许 MUPUC 系统在每个节点上布置所请求的配置。

[0141] 在MUPUC系统上使用的配置文件的示例：

[0142]

```
root@controller:~/scripts# more /mnt/administrativia/etc/mupuc/cluster_ethers
# Ethernet MAC      # IP address/subnet mask      # Server IP    # node hostname
# cluster nodes 1-4
e4:1f:13:80:c4:f9    172.16.200.101/255.255.255.0  172.16.200.1   node1
e4:1f:13:80:c9:c7    172.16.200.102/255.255.255.0  172.16.200.1   node2
e4:1f:13:80:cb:95    172.16.200.103/255.255.255.0  172.16.200.1   node3
e4:1f:13:80:cb:dd    172.16.200.104/255.255.255.0  172.16.200.1   node4
```

[0143] 用与要添加到集群的节点相对应的数据填充该文件表示必须手动进行的仅一步动作，其余全部被动态地和自动地处理。

[0144] 实施例使得能够在IaaS情形中布置真正的高性能计算，这为寻求布置HPC系统的用户提供了最小管理开销和最大灵活度的组合。

[0145] 此外，该系统的动态可重新配置的性质允许节点在它们没有使用时被完全断电，并且按需要被自动通电，向IaaS操作者提供节电的优点。

[0146] 该方法有许多优点，包括：

[0147] • 可重新配置性--MUPUC集群可以用于所有实验方式，并且可以非常容易地从一个配置切换为另一个。这意味着集群硬件被利用得远优于其本来仅在其上安装软件系统的情况。

[0148] •性能-因为MUPUC节点没有运行虚拟化层，在集群上运行的软件和硬件之间不存在迂回。这提供了虚拟化解决方案的显著性能优势。所享受的明确优势取决于在集群上运行的应用，但是能够将HPC软件优化为底层硬件可以容易地提供相对于非优化版本的数量级的性能改进。

[0149] • 可分离性-装载在一个系统映像中的软件不会干扰另一映像中的软件。这对MUPUC用户给予了以他们喜欢的任何方式进行系统实验的自由，而不必担心影响其他用户的应用。这从维护的角度来看是期望的，因为其保持运行MUPUC的管理开销低，而不牺牲灵活度。这对于IaaS提供商来说是重要的，因为其使不同客户的数据保持分离。

[0150] 实施例提供了带来很多优势的成本有效并且容易的解决方案。在大数据解决方案和云计算时代，实施例可以由许多公司、生产环境、研究实验室并且一般地在需要对HPC平台的ad-hoc访问的任何环境中被采用。

[0151] 本领域技术人员将容易地认识到，各种上述方法的步骤可以通过编程计算机来执行。在本文中，一些实施例还意在涵盖程序存储设备，例如数字数据存储介质，其为机器或计算机可读的并且编码机器可执行或计算机可执行的指令程序，其中所述指令执行所述上述方法的一些或所有步骤。该程序存储设备可以是例如数字存储器、诸如磁盘和磁带的磁存储介质、硬盘驱动器或光学可读数字数据存储介质。实施例还旨在于涵盖编程为执行上述方法的所述步骤的计算机。

[0152] 在附图中示出的各种元件的功能，包括标记为“处理器”或“逻辑”的任何功能块，可以通过使用专用硬件以及能够与适当软件相关联地执行软件的硬件来提供。当由处理器提供时，该功能可以通过单个专用处理器、单个共享处理器或者通过一些可以被共享的

多个独立的处理器来提供。此外，术语“处理器”或“控制器”或“逻辑”的明确使用不应当被解释为专指能够执行软件的硬件，并且可以暗示地包括但不限于，数字信号处理器 (DSP) 硬件、网络处理器、专用集成电路 (ASIC)、现场可编程门阵列 (FPGA)、用于存储软件的只读存储器 (ROM)、随机存取存储器 (RAM) 和非易失性存储装置。还可以包括其他常规和 / 或定制的硬件。类似地，附图中所示的任何切换器仅仅是概念性的。其功能可以通过程序逻辑的操作、通过专用逻辑、通过程序控制和专用逻辑的交互或甚至手动地来执行，具体技术可由实现者选择，如从该上下文中更具体理解的。

[0153] 本领域技术人员应当理解，本文的任何框图表示实现本发明的原理的说明性电路的原理图。类似地，应当理解，任何流程图、流程示图、状态转换图、伪代码等表示可以在计算机可读介质中实质性表示并且由计算机或处理器执行的各种过程，不论这样的计算机或处理器是否被明确示出。

[0154] 说明书和附图仅说明本发明的原理。因此，应当理解，本领域技术人员将能够设计各种布置，其虽然本文中没有被明确描述或示出，但是实现本发明的原理并且被包括在其精神和范围内。此外，本文列举的所有示例都主要旨在于明确仅用于教学目的，以帮助读者理解本发明的原理和发明人为促进本领域所贡献的概念，并且应当被解释为不限于这样的具体引用的示例和条件。而且，本文中记载本发明的原理、方面和实施例及其特定示例的所有陈述意在包含其等价物。

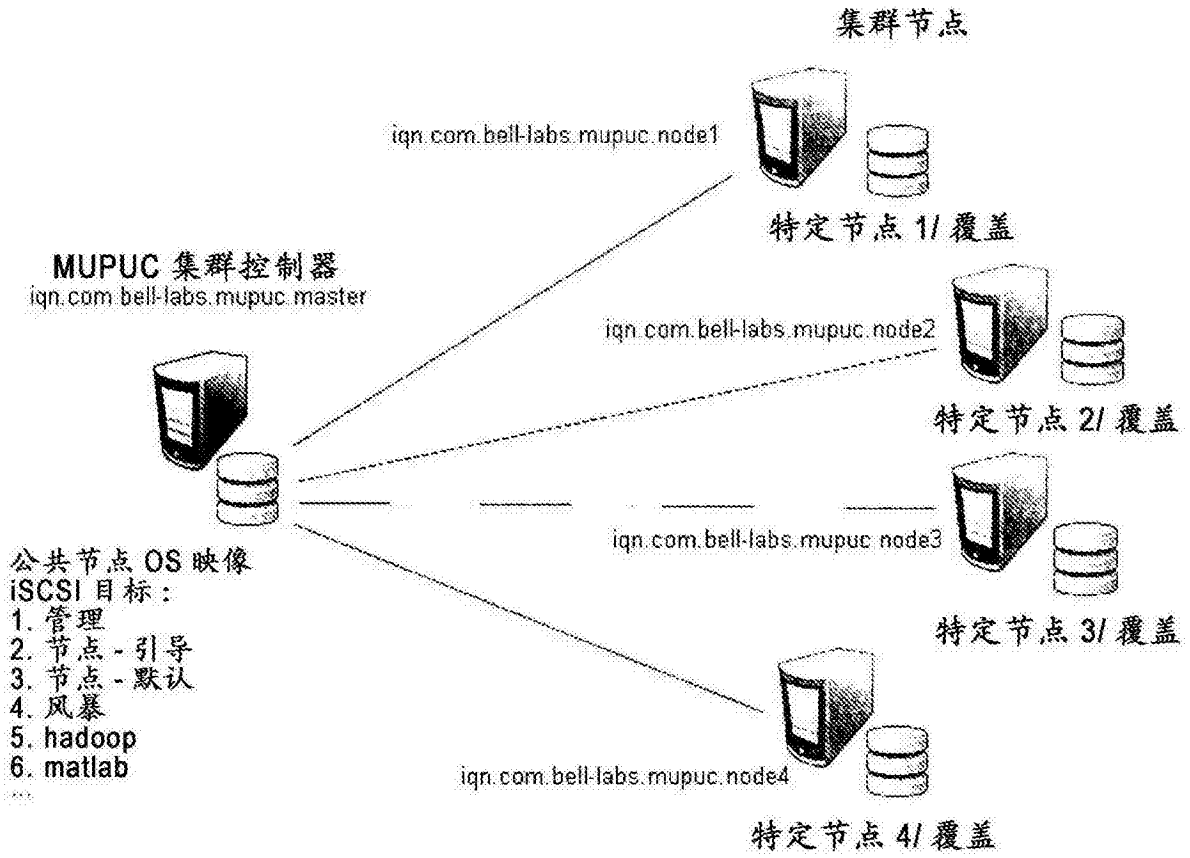

集群节点
iqn.com.bell-labs.mupuc.node1
特定节点 1/ 覆盖
MUPUC 集群控制器
iqn.com.bell-labs.mupuc.master
iqn.com.bell-labs.mupuc.node2
特定节点 2/ 覆盖
iqn.com.bell-labs.mupuc.node3
特定节点 3/ 覆盖
公共节点 OS 映像
iSCSI 目标：
1. 管理
2. 节点 - 引导
3. 节点 - 默认
4. 风暴
5. hadoop
6. matlab
...
iqn.com.bell-labs.mupuc.node4
特定节点 4/ 覆盖

图 1