

异构服务器系统及其使用方法

5 本申请要求于 2022 年 10 月 25 日提交中国专利局、申请号为 202211311808.2、申请名称为“异构服务器系统及其使用方法”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

10 本公开涉及人工智能的硬件计算设备，特别是涉及一种提供人工智能的复合服务的异构服务器系统。

背景技术

15 人工智能的模型的训练服务和推理服务通常需要不同的计算能力，因此通常设计不同的异构（例如，中央处理单元（CPU）+图形处理单元（GPU））服务器来满足不同的服务需求。例如，主流的训练服务通常使用 GPU 训练服务器或者基于 OAM（开放应用模型）的 UBB（通用基板，Universal Baseboard）基板为其提供算力，而推理服务通常使用单卡 GPU 机型服务器。

20 当前，为训练服务设计的异构服务器硬件（例如，CPU、GPU、内存的配比）对推理服务需求不匹配。如果利用训练服务器来进行推理服务，则无法充分发挥 CPU、GPU 算力，会造成算力的浪费。而且目前的训练服务器和推理服务器均无法灵活切换以实现训练、推理服务等灵活切换，从而无法跟进训练、推理需求的波峰波谷，充分调度 GPU 算力与服务需求匹配。因此，为了同时满足训练和推理服务需求，目前用户通常要同时购买训练服务器和推理服务器两者。然而，这容易造成两个服务器在各自服务需求波谷时段的算力的浪费。

25 因此，需要一种新型的异构服务器系统，其充分利用 GPU 算力来同时满足人工智能的至少两大类服务（例如训练与推理）的需求。

发明内容

30 本公开要解决的一个技术问题是提供一种异构服务器系统，其能够高效益地提供人工智能的至少两种服务。

35 根据本公开的第一个方面，提供了一种异构服务器系统，包括：第一计算节点，被配置为提供第一服务；第二计算节点，被配置为提供第二服务；以及计算资源节点，包括交换机和连接到所述交换机的计算处理单元。其中，所述计算处理单元用于执行所述第一服务或所述第二服务的至少部分计算任务。其中，所述交换机连接到所述第一计算节点和所述第二计算节点，并且能在第一状态与第二状态之间切换，其中在所述第一状态下所述交

交换机将所述计算处理单元连接到所述第一计算节点，而在所述第二状态下所述交换机将所述计算处理单元连接到所述第二计算节点。

5 可选地，所述交换机为 PCIe（外部设备互连扩展总线标准）交换机，所述计算处理单元经由 PCIe 线缆连接到所述交换机的下行端口，所述第一计算节点和所述第二计算节点经由 PCIe 线缆分别连接到所述交换机的第一端口和第二端口，并且在所述第一状态下所述第一端口被设为所述交换机的上行端口且所述第二端口被关闭，而在所述第二状态下所述第二端口被设为所述交换机的上行端口且所述第一端口被关闭。

可选地，所述计算资源节点还包括基板管理控制器，所述交换机在第一状态与第二状态之间的切换是通过所述基板管理控制器更改所述交换机的固件来实现的。

10 可选地，所述计算资源节点还包括基板管理控制器，所述交换机还包括内部处理器。所述交换机在第一状态与第二状态之间的切换是通过如下来实现的：所述交换机被配置为使所述基板管理控制器与所述内部处理器通信；所述内部处理器获取并保存所述下行端口的 PCIe 拓扑结构；在所述第一状态下，所述基板管理控制器将所述第一端口配置为所述交换机的上行端口且关闭所述第二端口，并且所述交换机将所述 PCIe 拓扑结构提供给所述第一计算节点；以及在所述第二状态下，所述基板管理控制器将所述第二端口配置为所述交换机的上行端口且关闭所述第一端口，并且所述交换机将所述 PCIe 拓扑结构提供给所述第二计算节点。

20 可选地，所述异构服务器系统包括多个第二计算节点，所述计算资源节点包括多个交换机和多个计算处理单元，其中，所述多个计算处理单元被分成多组，并且每组分别连接到所述多个交换机中的一个。其中，所述第一计算节点连接到所述多个交换机中的至少两个交换机。其中，所述多个第二计算节点中的每一个连接到所述多个交换机中的至少一个交换机，并且所述第二计算节点连接的交换机的数量是根据所述第二服务所需的计算处理单元数量、以及计算处理单元与交换机的连接架构而计算得出的。并且其中，所述第一计算节点连接的交换机的数量大于所述第二计算节点连接的交换机的数量。

25 可选地，所述计算处理单元为 GPU；并且/或者所述第一计算节点和所述第二计算节点分别包括 CPU 和内存；并且/或者所述第一服务为人工智能的训练服务；并且/或者所述第二服务为人工智能的推理服务。

30 可选地，所述计算资源节点还包括：第一接口，用于将所述第一计算节点连接到所述交换机的第一端口；以及/或者，第二接口，用于将所述第二计算节点连接到所述交换机的第二端口；以及/或者，存储器接口，用于将存储器连接到所述交换机的第三端口。

35 根据本公开的第二个方面，提供了一种使用根据本公开的第一个方面所述的异构服务器系统执行计算任务的方法，包括：确定交换机所连接的计算处理单元用于第一服务还是第二服务；在交换机所连接的计算处理单元用于第一服务的情况下，将所述交换机配置为第一状态；以及在交换机所连接的计算处理单元用于第二服务的情况下，将所述交换机配置为第二状态。

可选地，所述交换机为 PCIe 交换机，所述计算处理单元经由 PCIe 线缆连接到所述交换机的下行端口，所述第一计算节点和所述第二计算节点经由 PCIe 线缆分别连接到所述交换机的第一端口和第二端口，所述计算资源节点还包括基板管理控制器。将所述交换机配置为第一状态的步骤包括：通过所述基板管理控制器将所述交换机的固件设置为第一固
5 件，所述第一固件将所述第一端口设为所述交换机的上行端口从而连接到所述下行端口，且关闭所述第二端口；并且/或者，将所述交换机配置为第二状态的步骤包括：通过所述基板管理控制器将所述交换机的固件设置为第二固件，所述第二固件将所述第二端口设为所述交换机的上行端口从而连接到所述下行端口，且关闭所述第一端口。

可选地，所述交换机为 PCIe 交换机，所述计算处理单元经由 PCIe 线缆连接到所述交换机的下行端口，所述第一计算节点和所述第二计算节点经由 PCIe 线缆分别连接到所述交换机的第一端口和第二端口，所述计算资源节点还包括基板管理控制器，所述交换机还包括内部处理器。所述将所述交换机配置为第一状态的步骤包括：配置所述交换机以使所述基板管理控制器与所述内部处理器通信，通过所述内部处理器获取并保存所述下行端口的 PCIe 拓扑结构，以及通过所述基板管理控制器将所述第一端口配置为所述交换机的上行端口且关闭所述第二端口，并且通过所述交换机将所述 PCIe 拓扑结构提供给所述第一
10 计算节点；并且/或者，将所述交换机配置为第二状态的步骤包括：配置所述交换机以使所述基板管理控制器与所述内部处理器通信，通过所述内部处理器获取并保存所述下行端口的 PCIe 拓扑结构，以及通过所述基板管理控制器将所述第二端口配置为所述交换机的上行端口且关闭所述第一端口，并且通过所述交换机将所述 PCIe 拓扑结构提供给所述第二
15 计算节点。

可选地，所述异构服务器系统包括多个第二计算节点，所述计算资源节点包括多个交换机和多个计算处理单元，所述多个计算处理单元被分成多组，并且每组分别连接到所述多个交换机中的一个；所述第一计算节点连接到所述多个交换机中的至少两个交换机；所述多个第二计算节点中的每一个连接到所述多个交换机中的至少一个交换机，并且所述第二计算节点连接的交换机的数量是根据所述第二服务所需的计算处理单元数量、以及计算
20 处理单元与交换机的连接架构而计算得出的；所述第一计算节点连接的交换机的数量大于所述第二计算节点连接的交换机的数量。所述确定交换机所连接的计算处理单元用于第一服务还是第二服务的步骤包括：根据所述异构服务器系统要提供的第一服务和第二服务的数量，确定所述多个计算处理单元中分别用于第一服务和第二服务的计算处理单元的数量，
25 并且根据计算处理单元与交换机的连接架构、在第一计算节点、第二计算节点与交换机之间的连接架构、以及各服务所需的计算处理单元的数量，将所述多个计算处理单元分别分配给第一服务和第二服务。

根据本公开的第三个方面，提供了一种计算设备，包括：处理器；以及存储器，其上存储有可执行代码，当可执行代码被处理器执行时，使处理器执行如上述第二方面所述的方法。
30 35

根据本公开的第四个方面，提供了一种计算机程序产品，包括可执行代码，当所述可执行代码被电子设备的处理器执行时，使所述处理器执行如上述第二方面所述的方法。

根据本公开的第五个方面，提供了一种非暂时性机器可读存储介质，其上存储有可执行代码，当可执行代码被电子设备的处理器执行时，使处理器执行如上述第二方面所述的方法。

由此，本公开通过将至少两种服务节点与计算资源节点混合组网以融合为一个整体的物理机复合机型，可提供至少两种服务，并且可以利用交换机的灵活切换方案来提高计算资源节点的算力的利用率，有效提升总拥有成本（TCO，Total Cost of Ownership）受益。

附图说明

通过结合附图对本公开示例性实施方式进行更详细的描述，本公开的上述以及其它目的、特征和优势将变得更加明显，其中，在本公开示例性实施方式中，相同的参考标号通常代表相同部件。

图1示出了根据本公开一个实施例的异构服务器系统的示意性框图。

图2示出了根据本公开一个实施例的异构服务器系统的使用方法的示意性流程图。

图3示出了根据本公开一个实施例的异构服务器系统的一个具体示例的示意性框图。

图4示出了根据本公开一个实施例的计算设备的结构示意图。

具体实施方式

下面将参照附图更详细地描述本公开的优选实施方式。虽然附图中显示了本公开的优选实施方式，然而应该理解，可以以各种形式实现本公开而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了使本公开更加透彻和完整，并且能够将本公开的范围完整地传达给本领域的技术人员。

为了解决前述技术问题，本公开通过将至少两个提供不同服务的计算节点经由交换机连接到计算处理单元来实现计算处理单元的算力的灵活分配，从而高效地利用计算资源（即计算处理单元）来提供至少两个不同服务。

下面结合图1-2来描述本发明的基本构思。

图1示出了根据本公开实施例的异构服务器系统的一个基本构架的示意性框图。

如图1所示，异构服务器系统100包括第一计算节点110、第二计算节点120和计算资源节点130。计算资源节点130包括交换机140和计算处理单元150。第一计算节点110、第二计算节点120和计算处理单元150分别连接到交换机140的三个端口1-3。

图1中显示的交换机140中连接端口1和3的实线示意性地表示交换机140的第一状态，其将计算处理单元150连接到第一计算节点110。而图1中显示的连接端口2和3的虚线示意性地表示交换机140的第二状态，其将计算处理单元150连接到第二计算节点120。

如图1的大箭头所示，交换机140能在第一状态与第二状态之间切换，从而选择将计算处

理单元 150 连接到第一计算节点 110 还是第二计算节点 120。

本领域技术人员应该理解，图中显示的交换机 140 的结构只是对其功能的简单示意，并不代表本公开的交换机的物理结构；图中的所有连线表示的连接并不仅限于直接的物理连接，还可以包括经由中间接口的间接连接、或者无线连接等。

5 第一计算节点 110 和第二计算节点 120 被配置为分别提供第一服务和第二服务，例如人工智能的训练服务和推理服务。在本公开的一些实施方式中，第一计算节点 110 和第二计算节点 120 可以是通用计算机或服务器，其均可以包括 CPU 和内存以执行第一/第二服务的操作。而由于人工智能的服务通常要求更高的算力，因此，这些通用计算机或服务器需要额外的计算资源来满足其服务所需算力，也就是说，连接到计算资源节点 130 以利用
10 其中的计算处理单元 150 来执行第一或第二服务的至少部分计算任务。在另一些实施方式中，第一计算节点 110 和第二计算节点 120 也可以是专门设计的硬件架构，以利用计算资源节点 130 的算力来执行第一/第二服务的至少部分计算任务。

计算处理单元 150 可以为 GPU，但是本公开并不限于此，而是包括能为人工智能的各种服务提供所需算力的各种计算处理硬件，例如 ASIC 或 FPGA 等。

15 在本公开的一些实施方式中，交换机 140 可以为 PCIe 交换机 (switch)，并且经由 PCIe 线缆将第一计算节点 110、第二计算节点 120 和计算处理单元 150 分别连接到交换机 140 的端口 1-3。通过设置交换机 140，可以将端口 3 设为下行端口，并且根据需要将端口 1 和 2 中的一个设为上行端口而关闭另一个，由此实现交换机在两个状态之间的灵活切换。当然本发明不限于此，而是还可以通过例如网络接口控制器 (NIC, Network Interface Controller)
20 来实现各计算节点与计算处理单元之间的网络连接，例如通过 NIC 与 GPU 直接的远程直接内存访问 (RDMA, Remote Direct Memory Access) 技术，可将 GPU 算力灵活提供给第一或第二计算节点。与利用 NIC 的 RDMA 技术相比，利用 PCIe 交换机进行互连的解决方案的系统时延更小，并且软件复杂度更低。

本领域技术人员应该理解，尽管图 1 中仅示出了两个计算节点、一个交换机和计算处理单元，但是本发明不限于此。根据需要，系统 100 还可以包括另一种计算节点以提供另一种服务，和/或多个第一/第二计算节点，和/或多个交换机和计算处理单元。一个计算节点可以连接两个或更多个的交换机，一个交换机可以连接两个或更多个的计算处理单元，一个交换机也可以连接两个或更多个的计算节点。在实际应用中，可以根据系统需要提供的服务种类、各服务需求数量、各服务对计算资源的算力需求大小等条件来设计该系统包括的各部件的数量和连接架构。在一些情况下，第一服务所需的算力比第二服务要多，因此，可以将第一计算节点连接到更多的交换机，或者将第一计算节点连接到每一个交换机以便能利用全部算力来执行第一服务的计算任务。
30

下面结合图 2 来描述使用上述图 1 的异构服务器系统 100 实现本发明的基本构思的一种方法。

35 图 2 示出了根据本公开一个实施例的使用异构服务器系统执行计算任务的方法的示意

性流程图。

参见图 2，在步骤 S210，确定计算处理单元 150 用于第一服务还是第二服务。如果确定用于第一服务，则进行步骤 S220，将交换机 140 配置为第一状态，即如图 1 中的实线所示的连接关系。如果确定用于第二服务，则进行步骤 S230，将交换机 140 配置为第二状态，即如图 1 中的虚线所示的连接关系。通过这个方法，可以根据需要灵活调度计算处理单元 150 来执行哪个服务的计算任务。

该方法可以由该异构服务器系统自身（例如，系统中的各计算节点和计算资源节点的 CPU，或者系统中独立于这些节点之外的其它控制器）来实施，也可以由该异构服务器系统外部的控制装置来实施。

10 综合上面结合图 1-2 所述的，本发明将分别提供不同服务的各计算节点经由交换机来与所需的计算资源连接，且根据需要将计算资源灵活提供给各计算节点，从而可以以统一的复合物理机满足多种服务需求，并提高计算资源的利用率。

下面，为了更充分地理解本发明，将结合图 3 描述本公开的一个具体示例及其操作。

图 3 示出了根据本公开一个实施例的异构服务器系统的一个具体示例的示意性框图。
15 在本示例中，上述的第一计算节点为提供模型的训练服务的训练节点，第二计算节点为提供推理服务的推理节点，交换机为 PCIe 交换机，计算资源节点为 GPU 节点，计算处理单元为 GPU。

如图 3 所示，异构服务器系统 300 包括 1 个训练节点 310、4 个推理节点 320、4 个 PCIe 交换机 340 和 8 个 GPU 350。训练节点 310 连接到 4 个 PCIe 交换机 340，每个推理节点 320 连接到一个 PCIe 交换机 340，每个 PCIe 交换机 340 连接到 2 个 GPU 350。
20

本领域技术人员应该理解，本发明不限于图 3 所示的硬件数量及架构，而是可以根据训练和推理服务场景的硬件需求来配置 CPU 与 GPU 配比合理的复合系统。本公开的异构服务器系统也可以称为“异构服务器复合系统”。该复合系统通过将训练节点与 GPU 节点中所需数量的 GPU 经由交换机连接组网，将系统变成适合训练服务的复合物理机，而通过将推理节点与所需数量的 GPU 经由交换机连接组网，将系统变成适合推理服务的复合物理机，从而实现了将推理节点、训练节点与 GPU 节点组合成复合物理机以满足推理和训练需求的技术效果。
25

训练节点 310 和各推理节点 320 可以分别包括相同或不同数量的 CPU。在本示例中，可以设定训练节点 310 和各推理节点 320 分别包括 1 个 CPU，训练服务需要的 CPU 与 GPU 算力的配比为 1:8，推理服务需要的 CPU 与 GPU 算力的配比为 1:2，因此 GPU 节点 330 提供 8 个 GPU，并将其分成 4 组分别连到 4 个 PCIe 交换机，这样，每个 PCIe 交换机连接的 1 组 GPU 可分配给一个推理节点使用，所有 PCIe 交换机连接的全部 GPU 可同时分配给训练节点使用。
30

根据当前系统要提供的训练服务和推理服务的数量，可以将这 8 个 GPU 灵活分配给训练节点或推理节点。
35

例如，在一些实施方式中，在只有训练服务而无推理服务时可以将所有 GPU（即全部算力）分配给训练节点，而在训练途中来了优先级更高的推理服务请求时，可以将其中一个交换机切换到推理节点以满足推理计算任务需求而不需要暂停当前训练服务，由此系统可以同时运行训练服务和推理服务。在另一些实施方式中，可以按照优先级或其他方式管理训练和推理服务，最大利用效率地将所有 GPU 分配给训练和推理服务。例如，可以根据训练和推理服务需求的波峰和波谷时段，充分调度 GPU 算力与服务需求匹配。

5 确认了 GPU 到训练节点和/或推理节点的分配方案后，根据该方案配置好各个 PCIe 交换机的驱动，从而配置相应的 PCIe 交换机的上行端口连接到训练节点和推理节点中的一个，并关闭与另一个节点的 PCIe 通道，因此实现 GPU 的算力与各个计算节点的组合，以
10 满足相应服务的需求。

下面结合图 3 具体描述 PCIe 交换机 340 的配置操作。

如图 3 所示，每个 PCIe 交换机 340 有 6 个 PCIe 端口 PE1-PE6，分别通过 PCIe 线缆（图中标示的“PCIe X16/X8”表示 16 位/8 位的 PCIe 线）连接到各个 PCIe 设备，包括 GPU 350、接口 360、MCIO 380 以及插槽 390。当然，本发明的 PCIe 交换机并不限于此，而是可以
15 根据需要增加或减少 PCIe 端口，以及增加或减少所连接的各个 PCIe 设备。

在本示例中，训练节点 310 和推理节点 320 无法与 PCIe 交换机直连，而是由接口 360 进行转接，即通过线缆物理连接到接口 360 再经由接口 360 转接到 PCIe 交换机 340。在图 3 中，接口 360 位于 GPU 节点 330 中，但是本发明并不限制接口的位置，即，接口也可以独立于各节点或者安装于各计算节点中。

20 MCIO（Mini Cool edge I/O，小型冷边缘输入/输出接口）380 可以作为一种支持 PCIe 的存储器接口，用于将存储器（例如 SSD 或硬盘等）连接到 PCIe 交换机。当然本发明不限于此种存储器接口，而且服务所需的存储器也可以通过其它方式提供，并不限于如图所示地连接到交换机。

插槽 390 可以连接其它所需的 PCIe 设备，或者为将来需要连接的 PCIe 设备留有余地。

25 另外，图 3 示出了 GPU 节点 330 还包括基板管理控制器（BMC，Baseboard Management Controller）370，其连接到各个 PCIe 交换机 340。

如本领域技术人员所知的，PCIe 交换机采用的是树状连接结构，其只有一个上行端口，该上行端口连接到一个或多个下行端口。因此，根据本发明，将 GPU 350 的连接端口设为下行端口，通过将上行端口在 PE1 和 PE2 之间灵活切换来实现 GPU 算力在训练节点与推理节点间的灵活切换。
30

在一些实施方式中，PCIe 交换机 340 的上行端口的切换是通过 BMC 370 更改 PCIe 交换机 340 的固件来实现的。BMC 370 直接刷新交换机的固件，该固件把各个端口按需设置，从而实现所需的连接。

例如，系统 300 向 BMC 370 输出将训练节点连接到各个 GPU 的第一固件和将推理节点连接到各个 GPU 的第二固件。然后，BMC 370 根据服务需求所生成的 GPU 调度方案，
35

选择向各个 PCIe 交换机 340 加载第一固件或第二固件。第一和第二固件可以均将端口 PE3-PE6 设为交换机的下行端口。第一和第二固件的区别在于第一固件将端口 PE1 设为交换机的上行端口且关闭端口 PE2，而第二固件将端口 PE2 设为交换机的上行端口且关闭端口 PE1。

- 5 在另一些实施方式中，PCIe 交换机 340 的上行端口的切换是利用交换机的内部处理器 341 来实现的。

例如，将 PCIe 交换机 340 的模式配置为 ssw 模式（合成交换机模式，synthetic switch mode），并使能 secrouting 库。这里，secrouting 库是交换机的增强特性的一个库，支持交换机高级模式调测库。

- 10 然后，通过 secrouting 库接口，BMC 370 可以与内部处理器 341 通信，用于相关的配置和修改。

然后，内部处理器 341 获取 PCIe 交换机下层（即，各下行端口）的 PCIe 拓扑结构并将其保存在内部处理器的缓存中。

- 15 然后，BMC 370 通过 IIC（内部集成电路总线）带外通道进行端口 PE1 和 PE2 的配置，从而根据需要，将端口 PE1 和 PE2 之一设为上行端口，并关闭另一个。

然后，PCIe 交换机 340 将虚拟的 PCIe 树等资源同步给上行端口所连接的训练节点或推理节点，完成系统 PCIe 驱动资源配置。PCIe 树描述了该交换机的树状连接结构，其包括下行端口的 PCIe 拓扑结构。

- 20 如上所述，训练节点、推理节点与 GPU 节点根据服务需求通过 PCIe 线缆进行物理组网，并且通过 PCIe 交换机的配置实现 GPU 算力在训练节点与推理节点间的灵活切换，从而实现了按需调度切换 GPU 算力以融合实现训练与推理服务，并且提高了 GPU 算力的利用率。

- 25 图 4 示出了根据本公开一实施例可用于实现上述异构服务器系统的使用方法的计算设备的结构示意图。如前所述，这个实现本公开的方法的计算设备可以是由该异构服务器系统中的各计算节点和计算资源节点兼任，或是该系统中独立于这些节点之外的其它计算设备，也可以是该系统外部的计算设备。

参见图 4，计算设备 400 包括存储器 410 和处理器 420。

- 30 处理器 420 可以是一个多核的处理器，也可以包含多个处理器。在一些实施例中，处理器 420 可以包含一个通用的主处理器以及一个或多个特殊的协处理器，例如图形处理器（GPU）、数字信号处理器（DSP）等等。在一些实施例中，处理器 420 可以使用定制的电路实现，例如特定用途集成电路（ASIC，Application Specific Integrated Circuit）或者现场可编程逻辑门阵列（FPGA，Field Programmable Gate Arrays）。

- 35 存储器 410 可以包括各种类型的存储单元，例如系统内存、只读存储器（ROM），和永久存储装置。其中，ROM 可以存储处理器 420 或者计算机的其他模块需要的静态数据或者指令。永久存储装置可以是可读写的存储装置。永久存储装置可以是即使计算机断电

后也不会失去存储的指令和数据的非易失性存储设备。在一些实施方式中，永久性存储装置采用大容量存储装置（例如磁或光盘、闪存）作为永久存储装置。另外一些实施方式中，永久性存储装置可以是可移除的存储设备（例如软盘、光驱）。系统内存可以是可读写存储设备或者易失性可读写存储设备，例如动态随机访问内存。系统内存可以存储一些或者
5 所有处理器在运行时需要的指令和数据。此外，存储器 410 可以包括任意计算机可读存储媒介的组合，包括各种类型的半导体存储芯片（DRAM，SRAM，SDRAM，闪存，可编程只读存储器），磁盘和/或光盘也可以采用。在一些实施方式中，存储器 410 可以包括可读和/或写的可移除的存储设备，例如激光唱片（CD）、只读数字多功能光盘（例如 DVD-ROM，双层 DVD-ROM）、只读蓝光光盘、超密度光盘、闪存卡（例如 SD 卡、min SD 卡、Micro-SD
10 卡等等）、磁性软盘等等。计算机可读存储媒介不包含载波和通过无线或有线传输的瞬间电子信号。

存储器 410 上存储有可执行代码，当可执行代码被处理器 420 处理时，可以使处理器 420 执行上文述及的异构服务器系统的使用方法。

上文中已经参考附图详细描述了根据本发明的异构服务器系统及其使用方法。

15 此外，根据本发明的方法还可以实现为一种计算机程序或计算机程序产品，该计算机程序或计算机程序产品包括用于执行本发明的上述方法中限定的上述各步骤的计算机程序代码指令。

或者，本发明还可以实施为一种非暂时性机器可读存储介质（或计算机可读存储介质、或机器可读存储介质），其上存储有可执行代码（或计算机程序、或计算机指令代码），当
20 所述可执行代码（或计算机程序、或计算机指令代码）被电子设备（或计算设备、服务器等）的处理器执行时，使所述处理器执行根据本发明的上述方法的各个步骤。

本领域技术人员还将明白的是，结合这里的公开所描述的各种示例性逻辑块、模块、电路和算法步骤可以被实现为电子硬件、计算机软件或两者的组合。

附图中的流程图和框图显示了根据本发明的多个实施例的系统和方法的可能实现的
25 体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分，所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标记的功能也可以以不同于附图中所标记的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意，框图和/或流
30 程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在
35 最好地解释各实施例的原理、实际应用或对市场中的技术的改进，或者使本技术领域的其

它普通技术人员能理解本文披露的各实施例。

权 利 要 求 书

1. 一种异构服务器系统，包括：
第一计算节点，被配置为提供第一服务；
第二计算节点，被配置为提供第二服务；以及
5 计算资源节点，包括交换机和连接到所述交换机的计算处理单元，
其中，所述计算处理单元用于执行所述第一服务或所述第二服务的至少部分计算任务，
其中，所述交换机连接到所述第一计算节点和所述第二计算节点，并且能在第一状态
与第二状态之间切换，其中在所述第一状态下所述交换机将所述计算处理单元连接到所述
10 第一计算节点，而在所述第二状态下所述交换机将所述计算处理单元连接到所述第二计算
节点。
2. 根据权利要求1所述的异构服务器系统，其中，
所述交换机为 PCIe 交换机，
所述计算处理单元经由 PCIe 线缆连接到所述交换机的下行端口，
15 所述第一计算节点和所述第二计算节点经由 PCIe 线缆分别连接到所述交换机的第一
端口和第二端口，并且在所述第一状态下所述第一端口被设为所述交换机的上行端口且所
述第二端口被关闭，而在所述第二状态下所述第二端口被设为所述交换机的上行端口且所
述第一端口被关闭。
- 20 3. 根据权利要求2所述的异构服务器系统，其中，
所述计算资源节点还包括基板管理控制器，所述交换机在第一状态与第二状态之间的
切换是通过所述基板管理控制器更改所述交换机的固件来实现的。
- 25 4. 根据权利要求2所述的异构服务器系统，其中，
所述计算资源节点还包括基板管理控制器，所述交换机还包括内部处理器，所述交换
机在第一状态与第二状态之间的切换是通过如下来实现的：
所述交换机被配置为使所述基板管理控制器与所述内部处理器通信；
所述内部处理器获取并保存所述下行端口的 PCIe 拓扑结构；
在所述第一状态下，所述基板管理控制器将所述第一端口配置为所述交换机的上行端
30 口且关闭所述第二端口，并且所述交换机将所述 PCIe 拓扑结构提供给所述第一计算节点；
以及
在所述第二状态下，所述基板管理控制器将所述第二端口配置为所述交换机的上行端
口且关闭所述第一端口，并且所述交换机将所述 PCIe 拓扑结构提供给所述第二计算节点。
- 35 5. 根据权利要求1或2所述的异构服务器系统，其中，

所述异构服务器系统包括多个第二计算节点，
所述计算资源节点包括多个交换机和多个计算处理单元，
其中，所述多个计算处理单元被分成多组，并且每组分别连接到所述多个交换机中的一个，

5 其中，所述第一计算节点连接到所述多个交换机中的至少两个交换机；

其中，所述多个第二计算节点中的每一个连接到所述多个交换机中的至少一个交换机，
并且所述第二计算节点连接的交换机的数量是根据所述第二服务所需的计算处理单元数量、以及计算处理单元与交换机的连接架构而计算得出的；并且

10 其中，所述第一计算节点连接的交换机的数量大于所述第二计算节点连接的交换机的数量。

6. 根据权利要求 1 所述的异构服务器系统，其中，

所述计算处理单元为 GPU；并且/或者

所述第一计算节点和所述第二计算节点分别包括 CPU 和内存；并且/或者

15 所述第一服务为人工智能的训练服务；并且/或者

所述第二服务为人工智能的推理服务。

7. 根据权利要求 1 或 2 所述的异构服务器系统，其中，

所述计算资源节点还包括：

20 第一接口，用于将所述第一计算节点连接到所述交换机的第一端口；以及/或者

第二接口，用于将所述第二计算节点连接到所述交换机的第二端口；以及/或者

存储器接口，用于将存储器连接到所述交换机的第三端口。

25 8. 一种使用根据权利要求 1 至 7 中任何一项所述的异构服务器系统执行计算任务的方法，包括：

确定交换机所连接的计算处理单元用于第一服务还是第二服务；

在交换机所连接的计算处理单元用于第一服务的情况下，将所述交换机配置为第一状态；以及

30 在交换机所连接的计算处理单元用于第二服务的情况下，将所述交换机配置为第二状态。

9. 根据权利要求 8 所述的方法，其中，

所述交换机为 PCIe 交换机，

所述计算处理单元经由 PCIe 线缆连接到所述交换机的下行端口，

35 所述第一计算节点和所述第二计算节点经由 PCIe 线缆分别连接到所述交换机的第一

端口和第二端口，

所述计算资源节点还包括基板管理控制器，

将所述交换机配置为第一状态的步骤包括：

- 5 通过所述基板管理控制器将所述交换机的固件设置为第一固件，所述第一固件将所述第一端口设为所述交换机的上行端口从而连接到所述下行端口，且关闭所述第二端口，
并且/或者，将所述交换机配置为第二状态的步骤包括：
通过所述基板管理控制器将所述交换机的固件设置为第二固件，所述第二固件将所述第二端口设为所述交换机的上行端口从而连接到所述下行端口，且关闭所述第一端口。

- 10 10. 根据权利要求 8 所述的方法，其中，

所述交换机为 PCIe 交换机，

所述计算处理单元经由 PCIe 线缆连接到所述交换机的下行端口，

所述第一计算节点和所述第二计算节点经由 PCIe 线缆分别连接到所述交换机的第一端口和第二端口，

- 15 所述计算资源节点还包括基板管理控制器，所述交换机还包括内部处理器，

将所述交换机配置为第一状态的步骤包括：

配置所述交换机以使所述基板管理控制器与所述内部处理器通信；

通过所述内部处理器获取并保存所述下行端口的 PCIe 拓扑结构；以及

- 20 通过所述基板管理控制器将所述第一端口配置为所述交换机的上行端口且关闭所述第二端口，并且通过所述交换机将所述 PCIe 拓扑结构提供给所述第一计算节点，

并且/或者，将所述交换机配置为第二状态的步骤包括：

配置所述交换机以使所述基板管理控制器与所述内部处理器通信；

通过所述内部处理器获取并保存所述下行端口的 PCIe 拓扑结构；以及

- 25 通过所述基板管理控制器将所述第二端口配置为所述交换机的上行端口且关闭所述第一端口，并且通过所述交换机将所述 PCIe 拓扑结构提供给所述第二计算节点。

11. 根据权利要求 8 所述的方法，其中，

所述异构服务器系统包括多个第二计算节点，

所述计算资源节点包括多个交换机和多个计算处理单元，

- 30 所述多个计算处理单元被分成多组，并且每组分别连接到所述多个交换机中的一个，
所述第一计算节点连接到所述多个交换机中的至少两个交换机；

所述多个第二计算节点中的每一个连接到所述多个交换机中的至少一个交换机，并且
所述第二计算节点连接的交换机的数量是根据所述第二服务所需的计算处理单元数量、以及
计算处理单元与交换机的连接架构而计算得出的；

- 35 所述第一计算节点连接的交换机的数量大于所述第二计算节点连接的交换机的数量；

并且

所述确定交换机所连接的计算处理单元用于第一服务还是第二服务的步骤包括：

根据所述异构服务器系统要提供的第一服务和第二服务的数量，确定所述多个计算处理单元中分别用于第一服务和第二服务的计算处理单元的数量，并且

- 5 根据计算处理单元与交换机的连接架构、在第一计算节点、第二计算节点与交换机之间的连接架构、以及各服务所需的计算处理单元的数量，将所述多个计算处理单元分别分配给第一服务和第二服务。

12. 一种计算设备，包括：

- 10 处理器；以及

存储器，其上存储有可执行代码，当所述可执行代码被所述处理器执行时，使所述处理器执行如权利要求 8 至 11 中任何一项所述的方法。

13. 一种计算机程序产品，包括可执行代码，当所述可执行代码被电子设备的处理器
15 执行时，使所述处理器执行如权利要求 8 至 11 中任何一项所述的方法。

14. 一种非暂时性机器可读存储介质，其上存储有可执行代码，当所述可执行代码被电子设备的处理器执行时，使所述处理器执行如权利要求 8 至 11 中任何一项所述的方法。