

仿真装置、仿真系统及其仿真方法、存储介质

本申请要求于 2022 年 5 月 31 日递交的中国专利申请第
5 202210613143.4 号的优先权，在此全文引用上述中国专利申请公开的内
容以作为本申请的一部分。

技术领域

本公开的实施例涉及一种仿真装置、仿真系统及其仿真方法、非瞬时性
10 计算机可读存储介质。

背景技术

随着人工智能（人工智能，Artificial Intelligence）的发展，算法模型的参
数量剧增，对算力的需求越来越大。对于传统硬件架构（例如，CPU（中央
15 处理器，central processing unit）/GPU（图形处理器，graphics processing unit）），
由于其在架构设计阶段考虑了不同业务需求之间的平衡，导致其在 AI 应用
上提供的算力有限。由此，领域专用加速器（Domain-Specific Accelerator，
DSA）应运而生。DSA 的核心思想同样是使用专用的硬件做专用的事情，DSA
是满足一个领域（Domain）内的应用，而非一个固定的应用，因此，DSA 能
20 够满足灵活性与专用性的折衷。

发明内容

提供该内容部分以便以简要的形式介绍构思，这些构思将在后面的具体
实施方式部分被详细描述。该内容部分并不旨在标识要求保护的技术方案的
25 关键特征或必要特征，也不旨在用于限制所要求的保护的技术方案的范围。

本公开至少一个实施例提供一种仿真装置，该仿真装置用于对神经网络
处理器进行仿真且包括：代理模块，被配置为与模拟所述神经网络处理器的
对象模型进行通信并在所述仿真装置中作为所述对象模型的代理；管理模块，
被配置为对所述仿真装置进行管理；互连模块，被配置为将所述代理模块和
30 所述管理模块通信连接。所述仿真装置接收主机发送的任务，通过所述代理

模块将与所述任务相关的工作信息发送给所述对象模型，通过所述代理模块接收所述对象模型对所述工作信息处理后所返回的反馈信息，并将所述反馈信息提供给所述主机。

5 本公开至少一个实施例还提供一种仿真系统，包括：根据本公开任一实施例所述的仿真装置、对象模型和主机，所述主机被配置为获取所述任务并将所述任务发送至所述仿真装置；所述对象模型被配置为对所述工作信息进行处理以得到所述反馈信息。

10 本公开至少一个实施例还提供一种应用于本公开任一实施例所述的仿真系统的仿真方法，包括：通过所述主机发送任务；通过所述仿真装置接收并解析所述任务，以将与所述任务相关的工作信息发送给所述对象模型；所述对象模型对所述工作信息进行处理，以得到所述反馈信息；通过所述仿真装置将所述反馈信息提供给所述主机。

15 本公开至少一个实施例还提供一种非瞬时性计算机可读存储介质，其中，所述非瞬时性计算机可读存储介质存储有计算机可执行指令，所述计算机可执行指令被处理器执行时实现根据上述任一实施例所述的仿真方法。

附图说明

20 结合附图并参考以下具体实施方式，本公开各实施例的上述和其他特征、优点及方面将变得更加明显。贯穿附图中，相同或相似的附图标记表示相同或相似的元素。应当理解附图是示意性的，原件和元素不一定按照比例绘制。

图 1A 为本公开至少一个实施例提供的一种仿真装置的硬件架构的示意图；

图 1B 为本公开至少一个实施例提供的另一种仿真装置的硬件架构的示意图；

25 图 2 为本公开至少一个实施例提供的一种仿真系统的硬件架构的示意图；

图 3 为本公开至少一个实施例提供的一种仿真方法的示意性流程图；

图 4 为本公开至少一个实施例提供的一种非瞬时性计算机可读存储介质的示意图；

图 5 为本公开至少一个实施例提供的一种电子设备的硬件结构示意图。

具体实施方式

下面将参照附图更详细地描述本公开的实施例。虽然附图中显示了本公开的某些实施例，然而应当理解的是，本公开可以通过各种形式来实现，而且不应该被解释为限于这里阐述的实施例，相反提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是，本公开的附图及实施例仅用于示

5 例性作用，并非用于限制本公开的保护范围。

应当理解，本公开的方法实施方式中记载的各个步骤可以按照不同的顺序执行，和/或并行执行。此外，方法实施方式可以包括附加的步骤和/或省略执行示出的步骤。本公开的范围在此方面不受限制。

10 本文使用的术语“包括”及其变形是开放性包括，即“包括但不限于”。术语“基于”是“至少部分地基于”。术语“一个实施例”表示“至少一个实施例”；术语“另一实施例”表示“至少一个另外的实施例”；术语“一些实施例”表示“至少一些实施例”。其他术语的相关定义将在下文描述中给出。

需要注意，本公开中提及的“第一”、“第二”等概念仅用于对不同的装置、模块或单元进行区分，并非用于限定这些装置、模块或单元所执行的功

15 能的顺序或者相互依存关系。

需要注意，本公开中提及的“一个”、“多个”的修饰是示意性而非限制性的，本领域技术人员应当理解，除非在上下文另有明确指出，否则应该理解为“一个或多个”。

20 本公开实施方式中的多个装置之间所交互的消息或者信息的名称仅用于说明性的目的，而并不是用于对这些消息或信息的范围进行限制。

经过最近几年的发展，AI 硬件加速器的架构设计日趋复杂。从最初单核共享内存，逐步发展到目前的同构众核-分布式内存，AI 硬件加速器中的核的设计也由简单的超长指令字（Very Long Instruction Word, VLIW）结构

25 发展成图灵完备的指令集架构（ISA, Instruction Set Architecture），指令驱动的数据访问的两种实现架构，即 Master-Slave（数据块）和 Data-Streaming（数据流），并存。对于 AI 硬件加速器的前期的架构设计探索以及后期的验证等，一个完善的用于对 AI 硬件加速器进行仿真的仿真系统变得不可或缺，也即，目前 AI 硬件加速器缺少一个简单高效的仿真框架，用于架构设计和功能验

30 证。

AI 编译器涉及到的优化策略主要包括算子融合、拆分、IO/计算并行调度等，这些优化策略构成了一个巨大的优化搜索空间。每一种优化策略最终在底层加速器上执行的效率是不同的，需要代价模型(cost-model)来评估每一种优化策略，以及减小优化搜索空间。代价模型需要足够精确，同时不能引入过多编译阶段的开销。目前，开源全系统仿真平台包括 gem5 模拟器，gem5 模拟器的缺点主要包括：系统庞杂、不是为 AI 加速器建模而设计、执行效率低等，而且，gem5 模拟器作为代价模型，其开销太大。因此，目前 AI 编译器需要高性能的 cost-model 进行评估。

本公开至少一个实施例提供一种仿真装置。该仿真装置用于对神经网络处理器进行仿真，且包括：代理模块、管理模块和互连模块。代理模块被配置为与模拟神经网络处理器的对象模型进行通信并在仿真装置中作为对象模型的代理；管理模块被配置为对仿真装置进行管理；互连模块被配置为将代理模块和管理模块通信连接。仿真装置接收主机发送的任务，通过代理模块将与任务相关的工作信息发送给对象模型，通过代理模块接收对象模型对工作信息处理后所返回的反馈信息，并将反馈信息提供给主机。

本公开的实施例提供的仿真装置建模了神经网络处理器的全系统仿真平台，可以有效地实现对神经网络处理器的架构设计和探索以及功能验证，加速神经网络处理器的结构和功能设计过程，从而可以避免由于硬件而限制神经网络处理器的发展。而且，该仿真装置的结构简单，便于实现。

本公开至少一个实施例还提供一种仿真系统及其仿真方法、非瞬时性计算机可读存储介质。

下面结合附图对本公开的实施例进行详细说明，但是本公开并不限于这些具体的实施例。为了保持本公开实施例的以下说明清楚且简明，本公开省略了部分已知功能和已知部件的详细说明。

图 1A 为本公开至少一个实施例提供的一种仿真装置的硬件架构的示意图，图 1B 为本公开至少一个实施例提供的另一种仿真装置的硬件架构的示意图，图 2 为本公开至少一个实施例提供的一种仿真系统的硬件架构的示意图。

本公开的实施例提供的仿真装置可以用于对神经网络处理器进行仿真，该神经网络处理器可以采用硬件形式实现，例如，可以实现为基于程序设计

语言的芯片。例如，该神经网络处理器可以用于实现卷积运算、矩阵运算等。

如图 1A 和图 1B 所示，在本公开的一些实施例中，该仿真装置 100 可以包括代理模块 110、管理模块 120 和互连模块 130。如图 2 所示，代理模块 110 被配置为与模拟神经网络处理器的对象模型 300 进行通信并在仿真装置 100 中作为对象模型的代理，即对象模型 300 用于仿真神经网络处理器的功能；管理模块 120 被配置为对仿真装置 100 进行管理；互连模块 130 被配置为将代理模块 110 和管理模块 120 通信连接。需要说明的是，虽然图 2 所示的仿真装置 100 为图 1A 所示的仿真装置，但是，在仿真系统 1000 中的仿真装置 100 也可以为图 1B 所示的仿真装置。

例如，管理模块 120 可以对代理模块 110 进行管理，以控制代理模块 110 执行相应的功能。

例如，如图 2 所示，仿真装置 100 接收主机 200 发送的任务（task），通过代理模块 110 将与任务相关的工作信息发送给对象模型，通过代理模块 110 接收对象模型对工作信息处理后所返回的反馈信息，并将反馈信息提供给主机 200。

例如，仿真装置 100 通过虚拟仿真平台实现，例如，仿真装置 100 可以实现为 QEMU（Quick EMUlator）虚拟平台（virt platform）等。

例如，管理模块 120 通过虚拟中央处理器实现，例如，在一些实施例中，虚拟中央处理器可以为 RISC（Reduced Instruction Set Computer RISC）-V（V 表示为第五代 RISC）核。

例如，任务可以为任何需要由神经网络处理器执行的任务，例如，目标识别、矩阵运算（例如，矩阵乘法运算）等。

例如，代理模块 110 与主机 200 之间可以通过至少一种方式通信，例如，至少一种方式可以包括套接字方式、共享存储方式和/或消息队列方式通信。如图 2 所示，在一些实施例中，代理模块 110 与主机 200 通过套接字方式、共享存储方式和消息队列方式通信。套接字方式可以为 Unix 域套接字（UDS，Unix Domain Socket）方式。共享存储方式表示通过共享存储文件的方式进行通信，例如，共享存储文件可以为主机 200 中的/dev/shm/ivshmem 文件。ivshmem 文件可以实现在不同 QEMU 的进程间共享由主机 200 创建的内存区域。消息队列（message Queue）方式 MQ1 可以采用先进先出（First Input

First Output, FIFO) 的方式传递消息。

例如, 在代理模块 110 与主机 200 之间, 通过共享存储方式实现数据量较大的数据之间的通信, 而通过套接字方式和/或消息队列方式实现数据量较小的数据之间的通信。

5 例如, 代理模块 110 与对象模型 300 之间也可以通过至少一种方式通信, 例如, 代理模块 110 与对象模型 300 之间通过消息队列方式通信, 如图 2 所示, 在一些实施例中, 代理模块 110 与对象模型 300 通过消息队列方式通信, 例如, 代理模块 110 通过消息队列方式 MQ2 将与任务相关的工作信息发送给对象模型 300, 对象模型 300 通过消息队列方式 MQ3 将反馈信息发送给
10 代理模块 110。类似地, 消息队列方式 MQ2 和消息队列方式 MQ3 可以采用 FIFO 的方式传递消息。

例如, 对象模型 300 可以通过可插拔的方式连接至仿真装置 100, 也就是说, 仿真装置 100 可以对不同的对象模型进行仿真, 例如, 不同的对象模型可以模拟不同的神经网络处理器。

15 例如, 对象模型 300 可以加速神经网络处理器的运算速度, 节省运算时间, 提高运算效率。

例如, 如图 2 所示, 主机 200 可以从外部设备获取任务, 并通过套接字方式将任务发送给仿真装置 100。

20 例如, 神经网络处理器的参数以及与任务相关的输入和/或输出等信息通过共享存储方式由主机 200 和仿真装置 100 访问以进行读取或写入等操作。例如, 与任务相关的输入和/或输出可以根据任务的类型决定, 例如, 在一些实施例中, 任务可以为对图像中的目标物体进行识别, 并将标注了目标物体的图像反馈到主机 200 中, 此时, 输入的图像则可以为与任务相关的输入, 在通过对象模型 300 对图像进行识别并标注了目标物体的图像可以为与任务
25 相关的输出。

例如, 仿真装置 100 可以将反馈信息通过消息队列方式发送(例如, 同步)至主机 200。

30 例如, 在一些实施例中, 如图 1B 所示, 仿真装置 100 还包括可以输入输出模块 140, 输入输出模块 140 可以与主机 200 进行交互, 例如, 通过套接字方式与主机 200 进行通信, 以接收任务。例如, 输入输出模块 140 被配

置为将任务发送给代理模块 110 或管理模块 120。例如，输入输出模块 140 可以包括缓存器（buffer）等，以用于存储任务。

例如，如图 1A 和图 1B 所示，代理模块 110 的地址空间可以包括代理寄存器空间 Re 和模型空间 Mem1。

5 例如，代理寄存器空间 Re 用于定义神经网络处理器的寄存器，神经网络处理器的寄存器都是存储器映射（memory mapped）的，memory mapped 指的是设备的寄存器和内存统一编址，即用内存地址来定位设备的寄存器，从而神经网络处理器的寄存器被定义在内存空间中，可以通过内存地址的方式被定位从而实现读写。

10 例如，代理寄存器空间 Re 可以通过消息队列方式与主机 200 和对象模型 300 进行通信。

例如，模型空间 Mem1 用于存放神经网络处理器的参数以及与任务相关的输入和/或输出等。

15 例如，在一些实施例中，模型空间 Mem1 由代理模块 110 与主机 200 共享，此时，例如，模型空间 Mem1 通过共享存储方式与主机 200 进行通信。

20 例如，在一些实施例中，如图 2 所示，模型空间 Mem1 可以被映射为对象模型中的存储空间 Mem2。由此可知，模型空间 Mem1 和存储空间 Mem2 为同一地址空间，且实际的存储被映射到主机 200 中的共享存储文件中。也就是说，模型空间 Mem1 实际上可以被仿真装置 100、主机 200 和对象模型 300 访问，例如，主机 200 通过将神经网络处理器的参数以及与任务相关的输入和/或输出等信息写入到模型空间 Mem1，则对象模型 300 可以直接读取该写入到模型空间 Mem1 中的与任务相关的输入和/或输出等信息以进行相关处理。

25 例如，在一些实施例中，如图 2 所示，代理模块 110 的地址空间还可以包括配置指令空间 Mem3。配置指令空间 Mem3 用于存放神经网络处理器的寄存器的配置指令以及与任务相关的控制指令。

30 例如，在一些实施例中，代理模块 110、管理模块 120 和互连模块 130 可以挂载到高速串行计算机扩展总线（PCIe, peripheral component interconnect express）系统中。PCIe 系统可以包括以下几种设备类型：根复合体（RC, root complex）、桥（Bridge）、Switch、末端设备（Endpoint）等，根复合体是 CPU

和 PCIe 总线之间的接口，Bridge 提供了与其他总线（如 PCI 或 PCI-x，甚至是另一个 PCIe 总线）的接口，有时也被称为转发桥接；Switch 提供扩展或聚合能力，并允许更多的设备连接到一个 PCIe 的端口，Switch 可以充当包路由器，根据地址或其他路由信息识别给定包需要走哪条路径，是一种 PCIe 转
5 PCIe 的桥；Endpoint 处于 PCIe 总线系统的拓扑结构中的最末端，一般作为总线操作的发起者（initiator，类似于 PCI 总线中的主机）或者终结者（Completers，类似于 PCI 总线中的从机）。

例如，在一些实施例中，互连模块 130 可以为 PCIe 系统中的根复合体。

例如，在一些实施例中，代理模块 100 可以为 PCIe 系统中的 endpoint。
10 例如，代理模块 100 可以包括多个基地址寄存器（base address register, BAR）空间，多个基地址寄存器包括第一基地址寄存器空间和第二基地址寄存器空间。在一些实施例中，多个基地址寄存器空间可以包括 BRA0 空间~BRA5 空间，第一基地址寄存器空间可以为 BRA0 空间，第二基地址寄存器空间可以为 BRA3 空间。

15 例如，代理寄存器空间 Re 对应于第一基地址寄存器空间，例如 BRA0 空间，配置指令空间 Mem3 对应于第二基地址寄存器空间，例如 BRA3 空间。基地址寄存器空间指的是 PCIe 的 BAR 空间；仿真装置作为主机（HOST）的设备，通过 PCIe 接口与 HOST 相连；HOST 要访问仿真装置，需要将仿真装置上的空间映射到 PCIe 的 BAR 空间里；如果仿真装置上有多个独立的可
20 访问空间，PCIe 提供了多个 BAR 空间以供映射。

例如，配置指令空间 Mem3 被映射到由代理模块 110 和主机 200 共享的共享存储文件 ivshmem 且共享存储文件 ivshmem 位于主机 200。也就是说，在仿真装置 100 中，对 BAR3 空间的所有访问，都被转发到主机 200 上的共享存储文件 ivshmem，而主机 200 通过写该共享存储文件 ivshmem 以实现给
25 仿真装置 100 中的代理模块 110 发送配置指令。

例如，在本公开的一些实施例中，如图 2 所示，对于模型空间 Mem1、存储空间 Mem2 和配置指令空间 Mem3，其实际的存储被映射到主机 200 中的共享存储文件中，需要说明的是，模型空间 Mem1 和存储空间 Mem2 可以映射到共享存储文件中的同一位置，而配置指令空间 Mem3 映射到共享存储文件的位置与模型空间 Mem1/存储空间 Mem2 映射到共享存储文件中的位
30

置不同。

例如，代理模块 110 被配置为：基于配置指令空间 Mem3 中的内容，将代理寄存器空间 Re 和模型空间 Mem1 中的内容提供至对象模型 300 以配置和调度对象模型 300 以及进行工作模拟。需要说明的是，当模型空间 Mem1
5 可以被映射为对象模型中的存储空间 Mem2 时，对象模型 300 可以直接访问模型空间 Mem1 而获取需要的数据，例如，与任务相关的输入等数据。

例如，主机 200 可以通过套接字方式给代理模块 110 拉中断，即主机 200 发送的由对象模型 300 执行的任务可以通过中断方式被执行。

例如，在一些实施例中，代理模块 110 可以包括中断寄存器，主机向中
10 断寄存器写入与任务相关的通知信息，以通过中断方式通知管理模块 120 来执行任务。例如，主机 200 中的内存管理器（runtime）的进程，通过套接字方式与仿真装置 100 之间的通信的内容是：内存管理器对仿真装置 100 中的代理模块 110 的中断寄存器进行写操作，以写入与任务相关的通知信息；代理模块 110 的中断寄存器被写入与任务相关的通知信息之后，会给管理模块
15 120 拉中断；代理模块 110 的中断拉给管理模块 120 之后，会执行代理模块 110 的中断处理程序（属于代理模块 110 的驱动程序的一部分），而中断处理程序在内核态（当一个进程因为系统调用陷入内核代码中执行时处于内核运行态（内核态），此时，特权级最高），因此，需要通知用户态（当一个进程在执行用户自己的代码时处于用户运行态（即用户态），此时，特权级最低）
20 的调度程序，例如，可以采用信号（signal）的方式通知用户态的调度程序接收新的任务（即主机 200 发送的任务）。

例如，如图 1A 和图 1B 所示，仿真装置 100 还可以包括软件模块 150，软件模块 150 包括应用程序 App 和虚拟机系统（Guest OS），应用程序 App 包括多种工具 Tools、库文件 Library、调度器 Scheduler 等，虚拟机系统可以
25 包括操作系统内核和驱动器。

例如，管理模块 120 中运行有操作系统内核，例如，操作系统内核可以为 Linux 5.2 内核等。

例如，代理模块 110 的驱动程序被加载到操作系统内核以被执行，例如，代理模块 110 的驱动程序可以通过内核模块（kernel module）的方式加载到
30 内核。内核模块（kernel module）是操作系统的一个概念；驱动程序一般作

为内核模块被操作系统加载。

本公开至少一个实施例还提供一种仿真系统。

例如，如图 2 所示，仿真系统 1000 可以包括仿真装置 100、主机 200 和对象模型 300。需要说明的是，关于仿真装置 100、主机 200 和对象模型 300 之间的通信方式可以参考上述对于仿真装置 100 的实施例中的描述，重复之处不再赘述。

例如，主机 200 被配置为获取任务并将任务发送至仿真装置 100。例如，主机 200 包括内存管理器，内存管理器与仿真装置 100 进行通信以传输任务至仿真装置 100 并从仿真装置 100 接收对象模型 300 对与任务相关的工作信息处理后所返回的反馈信息。

例如，对象模型 300 被配置为对工作信息进行处理以得到反馈信息。

例如，在一些实施例中，对象模型 300 用于仿真硬件加速器（例如，神经网络处理器）的功能，且可以采用 SystemC 语言进行建模。SystemC 是由一组 C++ 类库所组成的建模平台，加入了一个仿真核，可以在系统级、行为描述级和寄存器转换级支持硬件建模。

又例如，对象模型 300 也可以采用 Verlog 语言进行建模。

例如，对象模型 300 的抽象层次为算法级 (ALM)、系统结构级 (SAM)、事务级 (TLM) 和寄存器传输级 (RTL) 中的至少一个。

例如，在一些实施例中，对象模型 300 可以包括矩阵 (Matrix) 执行单元、向量 (Vector) 执行单元等执行单元，片上静态随机存取存储器 (Static Random-Access Memory, SRAM)，SRAM 控制器等存储相关单元，以及微控制单元 (Microcontroller Unit, MCU) 等。

例如，在一些实施例中，对象模型 300 可以对神经网络处理器中的至少部分读取存储流水线和计算流水线进行建模，并针对神经网络处理器中的每个操作统计平均执行时间。例如，以 CPU 举例，计算流水线表示整型执行单元、浮点执行单元等；相应地，读取存储流水线表示 IO 操作，IO 操作表示 LOAD (加载)/SAVE (存储) 执行单元，通常也被称为 LSU。

例如，每个操作的平均执行时间可以表示每个阶段（例如，阶段表示 pipeline stage）的执行周期 (cycle) 数，即执行每个操作的周期 (cycle) 数。例如，通过对每个操作统计平均执行时间，可以计算得到近似周期精确的模

型。

为了提高对象模型 300 作为 cost-model 的执行效率，对象模型 300 支持两种模式：功能模式/性能模式；在性能模式下，只模拟 pipeline（例如，pipeline 表示交换数据的管线段的线性通信模型）上的延迟，读取存储流水线和计算流水线对应的实际操作并不执行，在性能模式下可以得到一个准确的模拟 cycle 数，相对于功能模式，其执行速度提高了一个数量级。

在本公开的实施例中，利用 SystemC 建模对象模型 300 的读取存储流水线和计算流水线中的延迟，该对象模型 300 满足 AI 编译器对高性能代价模型的需求，从而对象模型 300 可以作为 AI 编译器的代价模型，而作为 AI 编译器的 cost-model，能够明显提高 resnet50 网络在芯片上的推理时间性能指标，例如可以减少推理时间。

本公开至少一个实施例还提供一种应用于仿真系统的仿真方法，仿真系统可以为本公开任一实施例提供的仿真系统，例如，图 2 所示的仿真系统 1000。

图 3 为本公开至少一个实施例提供的一种仿真方法的示意性流程图。

如图 3 所示，该仿真方法可以包括以下步骤 S10~S13。

步骤 S10：通过主机发送任务。

步骤 S11：通过仿真装置接收并解析任务，以将与任务相关的工作信息发送给对象模型。

步骤 S12：对象模型对工作信息进行处理，以得到反馈信息。

步骤 S13：通过仿真装置将反馈信息提供给主机。

例如，以图 2 所示的仿真系统 1000 为例，步骤 S10 由主机 200 实现，在步骤 S10 中，主机 200 可以从外部设备接收任务，并将任务发送至仿真装置 100。例如，在一些实施例中，该任务可以由用户通过外部设备发送至主机 200 中。

例如，步骤 S11 由仿真装置 100 实现，例如，仿真装置 100 中的管理模块 120 可以解析任务，然后，仿真装置 100 中的代理模块 110 将与任务相关的工作信息发送至对象模型 300。

例如，步骤 S12 由对象模型 300 实现，在步骤 S12 中，对象模型 300 可以对工作信息进行处理，以得到反馈信息，该反馈信息可以被发送至仿真装置 100。例如，反馈信息可以包括对象模型 300 对工作信息进行处理之后的

结果,例如,当任务为对图像中的目标进行识别时,该反馈信息可以包括该图像中具有该目标的概率等。

例如,步骤 S13 由仿真装置 100 实现,在步骤 S13 中,仿真装置 100 通过代理模块 110 接收对象模型 300 对工作信息处理后所返回的反馈信息。

5 关于仿真方法可以实现的技术效果可以参考上述仿真装置和仿真系统的实施例中的相关描述,重复之处不再赘述。

本公开至少一个实施例还提供一种仿真装置,该仿真装置可以包括一个或多个存储器和一个或多个处理器。应当注意,上述仿真装置的组件只是示例性的,而非限制性的,根据实际应用需要,该仿真装置还可以具有其他组
10 件,本公开的实施例对此不作具体限制。

例如,一个或多个存储器用于非瞬时性地存储有计算机可执行指令;一个或多个处理器被配置为运行所述计算机可执行指令。计算机可执行指令被一个或多个处理器运行时实现根据本公开任一实施例所述的仿真方法中的一个或多个步骤。关于该仿真方法的各个步骤的具体实现以及相关解释内容
15 可以参见上述仿真方法的实施例,重复之处在此不作赘述。

例如,处理器和存储器之间可以直接或间接地互相通信。

例如,处理器和存储器可以通过网络进行通信。网络可以包括无线网络、有线网络、和/或无线网络和有线网络的任意组合。处理器和存储器之间也可以通过系统总线实现相互通信,本公开对此不作限制。

20 例如,处理器和存储器可以设置在服务器端(或云端)。

例如,处理器可以控制仿真装置中的其它组件以执行期望的功能。处理器可以是中央处理器(CPU)、图形处理器(GPU)、网络处理器(NP)等;处理器还可以为具有数据处理能力和/或程序执行能力的其它形式的处理单元,例如,数字信号处理器(DSP)、专用集成电路(ASIC)、现场可编程门
25 阵列(FPGA)、张量处理单元(TPU)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。中央处理元(CPU)可以为 X86 或 ARM 架构等。

例如,存储器可以为计算机可读介质,且可以包括一个或多个计算机程序产品的任意组合,计算机程序产品可以包括各种形式的计算机可读存储介质,例如易失性存储器和/或非易失性存储器。易失性存储器例如可以包括随
30

机存取存储器 (RAM) 和/或高速缓冲存储器 (cache) 等。非易失性存储器例如可以包括只读存储器 (ROM)、硬盘、可擦除可编程只读存储器 (EPROM)、便携式紧致盘只读存储器 (CD-ROM)、USB 存储器、闪存等。在所述计算机可读存储介质上可以存储一个或多个计算机可读指令，处理器可以运行所述

5 计算机可读指令，以实现仿真装置的各种功能。在存储介质中还可以存储各种应用程序和各种数据等。

关于仿真装置可以实现的技术效果可以参考上述仿真方法的实施例中的相关描述，重复之处不再赘述。

图 4 为本公开至少一个实施例提供的一种非瞬时性计算机可读存储介质的示意图。例如，如图 4 所示，在非瞬时性计算机可读存储介质 40 上可以非暂时性地存储一个或多个计算机可执行指令 401。例如，当计算机可执行指令 401 由处理器执行时可以执行根据本公开任一实施例所述的仿真方法中的一个或多个步骤。

10

例如，该非瞬时性计算机可读存储介质 40 可以应用于上述仿真装置中。

15 例如，非瞬时性计算机可读存储介质 40 可以包括上述仿真装置中的存储器。

例如，关于非瞬时性计算机可读存储介质 40 的说明可以参考仿真装置的实施例中对于存储器的描述，重复之处不再赘述。

下面参考图 5，图 5 示出了适于用来实现本公开实施例的电子设备 500 的结构示意图。该电子设备 500 可以为终端设备（例如，计算机）或处理器等，并可用于执行上述实施例的仿真方法。本公开实施例中的电子设备可以包括但不限于诸如移动电话、笔记本电脑、数字广播接收器、个人数字助理 (Personal Digital Assistant, 简称 PDA)、平板电脑 (Portable Android Device, 简称 PAD)、便携式多媒体播放器 (Portable Media Player, 简称 PMP)、车载终端（例如车载导航终端）、可穿戴电子设备等等的移动终端以及诸如数字

20 TV、台式计算机、智能家居设备等等的固定终端。图 5 示出的电子设备仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

如图 5 所示，电子设备 500 可以包括处理装置（例如中央处理器、图形处理器等）501，其可以根据存储在只读存储器 (ROM) 502 中的程序或者从存储装置 508 加载到随机访问存储器 (RAM) 503 中的程序而执行各种适当的动作和处理。在 RAM 503 中，还存储有电子设备 500 操作所需的各种程

30

序和数据。处理装置 501、ROM 502 以及 RAM 503 通过总线 504 彼此相连。输入/输出 (I/O) 接口 505 也连接至总线 504。

通常，以下装置可以连接至 I/O 接口 505：包括例如触摸屏、触摸板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置 506；包括例如
5 液晶显示器 (Liquid Crystal Display, LCD)、扬声器、振动器等的输出装置 507；包括例如磁带、硬盘等的存储装置 508；以及通信装置 509。通信装置 509 可以允许电子设备 500 与其他设备进行无线或有线通信以交换数据。虽然图 5 示出了具有各种装置的电子设备 500，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。

10 特别地，根据本公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在非暂态计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码，以执行根据上文所述的仿真方法中的一个或多个步骤。在这样的实施例中，该计算机程序可以通过通信装置 509 从网
15 络上被下载和安装，或者从存储装置 508 被安装，或者从 ROM 502 被安装。在该计算机程序被处理装置 501 执行时，可以使得处理装置 501 执行本公开实施例的仿真方法中限定的上述功能。

需要说明的是，在本公开的上下文中，计算机可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。计算机可读介质可以是计算机可读信号
20 介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是，但不限于：电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、
25 随机访问存储器 (RAM)、只读存储器 (ROM)、可擦式可编程只读存储器 (EPROM 或闪存)、光纤、便携式紧凑磁盘只读存储器 (CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中，计算机可读信号
30 号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了

计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：电线、光缆、RF（射频）等等，或者上述介质的任意合适的组合。

上述计算机可读介质可以是上述电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。

10 可以以一种或多种程序设计语言或其组合来编写用于执行本公开的操作的计算机程序代码，上述程序设计语言包括但不限于面向对象的程序设计语言，诸如 Java、Smalltalk、C++，还包括常规的过程式程序设计语言，诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户
15 计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络（包括局域网(LAN)或广域网(WAN)）连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。

附图中的流程图和框图，图示了按照本公开各种实施例的系统、方法和
20 计算机程序产品的可能实现的体系架构、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，该模块、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，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示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。也要注意的，框图和/或流程图中的每个方框、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。

描述于本公开实施例中所涉及到的单元可以通过软件的方式实现，也可
30 以通过硬件的方式来实现。例如，单元的名称在某种情况下并不构成对该单

元本身的限定。

本文中以上描述的功能可以至少部分地由一个或多个硬件逻辑部件来执行。例如，非限制性地，可以使用的示范类型的硬件逻辑部件包括：现场可编程门阵列（FPGA）、专用集成电路（ASIC）、专用标准产品（ASSP）、片上系统（SOC）、复杂可编程逻辑设备（CPLD）等等。

5 第一方面，根据本公开的一个或多个实施例，一种仿真装置，用于对神经网络处理器进行仿真，且包括：代理模块，被配置为与模拟所述神经网络处理器的对象模型进行通信并在所述仿真装置中作为所述对象模型的代理，管理模块，被配置为对所述仿真装置进行管理，互连模块，被配置为将所述代理模块和所述管理模块通信连接，其中，所述仿真装置接收主机发送的任务，通过所述代理模块将与所述任务相关的工作信息发送给所述对象模型，通过所述代理模块接收所述对象模型对所述工作信息处理后所返回的反馈信息，并将所述反馈信息提供给所述主机。

15 根据本公开的一个或多个实施例，所述代理模块与所述主机之间通过套接字方式、共享存储方式和/或消息队列方式通信。

根据本公开的一个或多个实施例，所述代理模块的地址空间包括代理寄存器空间、模型空间，所述代理寄存器空间用于定义所述神经网络处理器的寄存器，且所述神经网络处理器的寄存器通过内存地址的方式被定位；所述模型空间用于存放所述神经网络处理器的参数以及与所述任务相关的输入和/或输出。

根据本公开的一个或多个实施例，所述模型空间被映射为所述对象模型中的存储空间。

根据本公开的一个或多个实施例，所述模型空间由所述代理模块与所述主机共享。

25 根据本公开的一个或多个实施例，所述代理模块的地址空间还包括配置指令空间，所述配置指令空间用于存放所述神经网络处理器的寄存器的配置指令以及与所述任务相关的控制指令。

根据本公开的一个或多个实施例，所述代理模块包括多个基地址寄存器空间，所述多个基地址寄存器包括第一基地址寄存器空间和第二基地址寄存器空间，所述代理寄存器空间对应于所述第一基地址寄存器空间，所述配置

指令空间对应于所述第二基地址寄存器空间。

根据本公开的一个或多个实施例，所述配置指令空间被映射到由所述代理模块和主机共享的共享存储文件且所述共享存储文件位于所述主机。

根据本公开的一个或多个实施例，所述代理模块被配置为，基于所述配置指令空间中的内容，将所述代理寄存器空间和所述模型空间中的内容提供至所述对象模型以配置和调度所述对象模型以及进行工作模拟。

根据本公开的一个或多个实施例，所述代理模块与所述对象模型之间通过消息队列方式通信。

根据本公开的一个或多个实施例，所述管理模块中运行有操作系统内核，所述代理模块的驱动程序加载到所述操作系统内核。

根据本公开的一个或多个实施例，所述代理模块包括中断寄存器，所述主机向所述中断寄存器写入与所述任务相关的通知信息，以通过中断方式通知所述管理模块来执行所述任务。

根据本公开的一个或多个实施例，仿真装置还包括输入输出模块，所述输入输出模块与所述主机之间通过套接字方式通信，以接收所述任务，所述输入输出模块被配置为将所述任务发送给所述代理模块或所述管理模块。

根据本公开的一个或多个实施例，所述仿真装置通过虚拟仿真平台实现，所述管理模块通过虚拟中央处理器实现。

第二方面，根据本公开的一个或多个实施例，一种仿真系统，包括：根据本公开任一实施例所述的仿真装置、所述对象模型和所述主机，其中，所述主机被配置为获取所述任务并将所述任务发送至所述仿真装置；所述对象模型被配置为对所述工作信息进行处理以得到所述反馈信息。

根据本公开的一个或多个实施例，所述对象模型的抽象层次为算法级、系统结构级、事务级或寄存器传输级。

根据本公开的一个或多个实施例，所述对象模型对所述神经网络处理器中的至少部分读取存储流水线和计算流水线进行建模，并针对所述神经网络处理器中的每个操作统计平均执行时间。

第三方面，根据本公开的一个或多个实施例，一种应用于本公开任一实施例所述的仿真系统的仿真方法，包括：通过所述主机发送任务；通过所述仿真装置接收并解析所述任务，以将与所述任务相关的工作信息发送给所述

对象模型；所述对象模型对所述工作信息进行处理，以得到所述反馈信息；通过所述仿真装置将所述反馈信息提供给所述主机。

5 第四方面，根据本公开的一个或多个实施例，一种非瞬时性计算机可读存储介质，其中，所述非瞬时性计算机可读存储介质存储有计算机可执行指令，所述计算机可执行指令被处理器执行时实现根据本公开任一实施例所述的仿真方法。

10 以上描述仅为本公开的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本公开中所涉及的公开范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述公开构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开中公开的（但不限于）具有类似功能的技术特征进行互相替换而形成的技术方案。

15 此外，虽然采用特定次序描绘了各操作，但是这不应理解为要求这些操作以所示出的特定次序或以顺序次序执行来执行。在一定环境下，多任务和并行处理可能是有利的。同样地，虽然在上面论述中包含了若干具体实现细节，但是这些不应被解释为对本公开的范围的限制。在单独的实施例的上下文中描述的某些特征还可以组合地实现在单个实施例中。相反地，在单个实施例的上下文中描述的各种特征也可以单独地或以任何合适的子组合的方式实现在多个实施例中。

20 尽管已经采用特定于结构特征和/或方法逻辑动作的语言描述了本主题，但是应当理解所附权利要求书中所限定的主题未必局限于上面描述的特定特征或动作。相反，上面所描述的特定特征和动作仅仅是实现权利要求书的示例形式。

对于本公开，还有以下几点需要说明：

25 （1）本公开实施例附图只涉及到与本公开实施例涉及到的结构，其他结构可参考通常设计。

（2）在不冲突的情况下，本公开的实施例及实施例中的特征可以相互组合以得到新的实施例。

30 以上所述仅为本公开的具体实施方式，但本公开的保护范围并不局限于此，本公开的保护范围应以所述权利要求的保护范围为准。

权利要求书

- 1、一种仿真装置，用于对神经网络处理器进行仿真，包括：
代理模块，被配置为与模拟所述神经网络处理器的对象模型进行通信并
5 在所述仿真装置中作为所述对象模型的代理；
管理模块，被配置为对所述仿真装置进行管理；
互连模块，被配置为将所述代理模块和所述管理模块通信连接；
其中，所述仿真装置接收主机发送的任务，通过所述代理模块将与所述
任务相关的工作信息发送给所述对象模型，通过所述代理模块接收所述对象
10 模型对所述工作信息处理后所返回的反馈信息，并将所述反馈信息提供给所
述主机。
- 2、根据权利要求1所述的仿真装置，其中，所述代理模块与所述主机之
间通过套接字方式、共享存储方式和/或消息队列方式通信。
- 3、根据权利要求1或2所述的仿真装置，其中，所述代理模块的地址空
15 间包括代理寄存器空间和模型空间，
所述代理寄存器空间用于定义所述神经网络处理器的寄存器，且所述神
经网络处理器的寄存器通过内存地址的方式被定位；
所述模型空间用于存放所述神经网络处理器的参数以及与所述任务相
关的输入和/或输出。
- 20 4、根据权利要求3所述的仿真装置，其中，所述模型空间被映射为所述
对象模型中的存储空间。
- 5、根据权利要求3或4所述的仿真装置，其中，所述模型空间由所述代
理模块与所述主机共享。
- 6、根据权利要求3~5任一项所述的仿真装置，其中，所述代理模块的地
25 址空间还包括配置指令空间，
所述配置指令空间用于存放所述神经网络处理器的寄存器的配置指令
以及与所述任务相关的控制指令。
- 7、根据权利要求6所述的仿真装置，其中，
所述代理模块包括多个基地址寄存器空间，所述多个基地址寄存器包括
30 第一基地址寄存器空间和第二基地址寄存器空间，

所述代理寄存器空间对应于所述第一基地址寄存器空间，所述配置指令空间对应于所述第二基地址寄存器空间。

8、根据权利要求6所述的仿真装置，其中，所述配置指令空间被映射到由所述代理模块和所述主机共享的共享存储文件且所述共享存储文件位于
5 所述主机。

9、根据权利要求6所述的仿真装置，其中，所述代理模块被配置为，基于所述配置指令空间中的内容，将所述代理寄存器空间和所述模型空间中的内容提供至所述对象模型以配置和调度所述对象模型以及进行工作模拟。

10、根据权利要求1~9任一项所述的仿真装置，其中，所述代理模块与
10 所述对象模型之间通过消息队列方式通信。

11、根据权利要求1~10任一项所述的仿真装置，其中，所述管理模块中运行有操作系统内核，所述代理模块的驱动程序加载到所述操作系统内核。

12、根据权利要求1~11任一项所述的仿真装置，其中，所述代理模块包括中断寄存器，所述主机向所述中断寄存器写入与所述任务相关的通知信息，
15 以通过中断方式通知所述管理模块来执行所述任务。

13、根据权利要求1~12任一项所述的仿真装置，还包括输入输出模块，
其中，所述输入输出模块与所述主机之间通过套接字方式通信，以接收
所述任务，

所述输入输出模块被配置为将所述任务发送给所述代理模块或所述管
20 理模块。

14、根据权利要求1~13任一项所述的仿真装置，其中，所述仿真装置通过虚拟仿真平台实现，所述管理模块通过虚拟中央处理器实现。

15、一种仿真系统，包括：

根据权利要求1~14任一项所述的仿真装置、所述对象模型和所述主机，
25 其中，所述主机被配置为获取所述任务并将所述任务发送至所述仿真装置；

所述对象模型被配置为对所述工作信息进行处理以得到所述反馈信息。

16、根据权利要求15所述的仿真系统，其中，所述对象模型的抽象层次为算法级、系统结构级、事务级或寄存器传输级。

30 17、根据权利要求15或16所述的仿真系统，其中，所述对象模型对所

述神经网络处理器中的至少部分读取存储流水线和计算流水线进行建模，并针对所述神经网络处理器中的每个操作统计平均执行时间。

18、一种应用于权利要求 15~17 任一项所述的仿真系统的仿真方法，包括：

5 通过所述主机发送任务；

通过所述仿真装置接收并解析所述任务，以将与所述任务相关的工作信息发送给所述对象模型；

所述对象模型对所述工作信息进行处理，以得到所述反馈信息；

通过所述仿真装置将所述反馈信息提供给所述主机。

10 19、一种非瞬时性计算机可读存储介质，其中，所述非瞬时性计算机可读存储介质存储有计算机可执行指令，所述计算机可执行指令被处理器执行时实现根据权利要求 18 所述的仿真方法。