

一种数据存储管理装置及处理核

5 技术领域

本发明涉及到处理核技术领域，尤其是涉及到一种数据存储管理装置及处理核。

背景技术

随着科学技术的发展，人类社会正在快速进入智能时代。智能时代的重要特点，就是人们获得数据的种类越来越多，获得数据的量越来越大，而对处理数据的速度要求越来越高。

芯片是数据处理的基石，它从根本上决定了人们处理数据的能力。从应用领域来看，芯片主要有两条路线：一条是通用芯片路线，例如 CPU 等，它们能提供极大的灵活性，但是在处理特定领域算法时有效算力比较低；另一条是专用芯片路线，例如 TPU 等，它们在某些特定领域，能发挥较高的有效算力，但是面对灵活多变的比较通用的领域，它们处理能力比较差甚至无法处理。

由于智能时代的数据种类繁多且数量巨大，所以要求芯片既具有极高的灵活性，能处理不同领域且日新月异的算法，又具有极强的处理能力，能快速处理极大的且急剧增长的数据量。

发明内容

本发明提供一种数据存储管理装置及处理核，可以消除由于 Cache 访问失效带来的计算效率的下降，提高了程序效率的可控性。

本发明的第一方面提供了一种数据存储管理装置，包括：至少两个随机

存取存储器 RAM；控制单元，用于接收指令，根据所述指令生成并发送控制信号；直接存储器访问控制器 DMAC，用于根据所述控制信号实现对所述随机存取存储器 RAM 中数据的存取。

5 本发明实施方式提供的数据存储管理装置，接收并响应来自外部的处理单元发送的指令，从外部的存储单元中读取数据，使得外部的处理单元在执行程序时能直接从数据存储管理装置中读取执行程序所需的数据，外部处理单元无需通过高速缓存 Cache 从外部存储单元中取数，消除了由于 Cache 访问失效带来的计算效率的下降，提高了程序效率的可控性。

10 优选的，直接存储器访问控制器 DMAC，用于根据所述控制信号实现对所述 RAM 中数据的存取，包括：所述直接存储器访问控制器 DMAC，用于根据所述控制信号从外部存储单元中读取数据，并将所述数据存储在该所述控制信号指示的所述随机存取存储器 RAM 中；或所述直接存储器访问控制器 DMAC，用于根据所述控制信号从所述控制信号所指示的所述随机存取存储器 RAM 中读取数据，并将所述数据存储在外部的存储单元中。

15 优选的，控制信号所指示的所述随机存取存储器 RAM 的个数为一个或多个。

优选的，所有的所述随机存取存储器 RAM 的地址与所述外部存储单元的地址统一编址。或，所有的所述随机存取存储器 RAM 的地址统一编址。

20 优选的，所述直接存储器访问控制器 DMAC 的访问地址范围为所述随机存取存储器 RAM 的地址段和外部的存储单元的地址段。

根据本发明的另一方面，还提供了一种处理核，包括处理单元、存储单元和第一方面提供的存储管理装置；所述处理单元，用于发送指令，所述指令用于指示所述存储管理装置实现对所述存储单元中数据的存取；所述处理单元，还用于从任意所述随机存取存储器 RAM 中读取执行程序所需的数据。

25 优选的，其中指令包括取数指令和存数指令；处理单元用于发送取数指令，所述取数指令用于指示所述数据存储管理装置从所述存储单元中取出数据并将所述数据存储至所述取数指令指示的随机存取存储器 RAM 中。

优选的，直接存储器访问控制器 DMAC 用于在完成所述取数指令后，发送存储完成的信号；所述处理单元，用于根据所述存储完成的信号，发出新的取数指令，并从所述取数指令指示的随机存取存储器 RAM 中读取数据。

优选的，在同一时间点，所述处理单元与所述直接存储器访问控制器
5 DMAC 访问不同的所述随机存取存储器 RAM。

优选的，所述在同一时间，所述处理单元与所述直接存储器访问控制器 DMAC 访问不同的所述随机存取存储器 RAM，包括：所述至少两个随机存取存储器 RAM 包括第一随机存取存储器 RAM 和第二随机存取存储器 RAM；在第一时间，所述处理单元从第一随机存取存储器 RAM 中读取第一数据，所述直接
10 存储器访问控制器 DMAC 向所述第二随机存取存储器 RAM 中写入从所述存储单元中取出的第二数据；在第二时间，所述处理单元从所述第二随机存取存储器 RAM 中读取所述第二数据，所述直接存储器访问控制器 DMAC 向所述第一随机存取存储器 RAM 中写入从所述存储单元中取出的第三数据。

优选的，第一随机存取存储器 RAM 为包括多个 RAM 的第一组随机存取存储器 RAM，第二随机存取存储器 RAM 为包括多个随机存取存储器 RAM 的第二
15 组随机存取存储器 RAM。

当第一随机存取存储器 RAM 为第一组随机存取存储器 RAM，第二随机存取存储器 RAM 为第二随机存取存储器 RAM 组时，在同一时间，处理单元和直接存储器访问控制器 DMAC 分别访问一个组内的各个随机存取存储器 RAM，
20 或者，在同一时间，处理单元和直接存储器访问控制器 DMAC 访问分属于两个组的 RAM。

根据本发明的第三方面，提供了一种芯片，包括一个或多个第二方面提供的处理核。

根据本发明的第四方面，提供了一种卡板，包括一个或多个第三方面提
25 供的芯片。

根据本发明的第五方面，提供了一种电子设备，包括一个或多个第四方面提供的卡板。

根据本发明的第六方面，提供了一种数据存储管理方法，控制单元，接收指令，根据所述指令生成并发送控制信号；直接存储器访问控制器 DMAC 根据所述控制信号实现对所述随机存取存储器 RAM 中数据的存取。

根据本发明的第七方面，提供了一种电子设备，包括：存储器，用于存储计算机可读指令；以及一个或多个处理器，用于运行所述计算机可读指令，使得所述处理器运行时实现前述第六方面中的任一所述数据存储管理的方法。

根据本发明的第八方面，提供一种非暂态计算机可读存储介质，该非暂态计算机可读存储介质存储计算机指令，该计算机指令用于使计算机执行前述第六方面中的任一所述数据存储管理的方法。

根据本发明的第九方面，提供一种计算机程序产品，其中，包括计算机指令，当所述计算机指令被计算设备执行时，所述计算设备可以执行前述第六方面中的任一所述数据存储管理的方法。

本发明实施方式提供的数据存储管理装置，接收并响应来自外部的处理单元发送的指令，从外部的存储单元中读取数据，使得外部的处理单元在执行程序时能直接从数据存储管理装置中读取执行程序所需的数据，外部处理单元无需通过高速缓存 Cache 从外部存储单元中取数，消除了由于 Cache 访问失效带来的计算效率的下降，提高了程序效率的可控性。

附图说明

图 1 是现有技术中处理核中读取数据的示意图；

图 2 是根据本发明一实施方式的数据存储管理装置的结构示意图；

图 3 是根据本发明一实施方式的处理核的结构示意图；

图 4 是根据本发明一实施方式的神经网络的结构示意图；

图 5 是根据本发明一实施方式的处理核的结构示意图。

图 6 是根据本发明一实施方式的处理核进行神经网络计算的时序图；

图 7 是根据本发明一实施方式的数据存储管理方法的流程示意图。

具体实施方式

为使本发明的目的、技术方案和优点更加清楚明了，下面结合具体实施方式并参照附图，对本发明进一步详细说明。应该理解，这些描述只是示例性的，而并非要限制本发明的范围。此外，在以下说明中，省略了对公知结构和技术的描述，以避免不必要地混淆本发明的概念。

在神经网络计算中，经常会用到多核或者众核的芯片。此处多(众)核架构中的核，都有一定独立处理能力，并且带有比较大的核内存储空间，用于
10 存储自身的程序、数据和权重。

单核的基础计算能力的发挥，决定了整个芯片计算神经网络的能力。而单核的基础计算能力的发挥，由单核的计算单元的理想计算能力及存储访问效率决定。

不同的存储单元，其被访问速度是不一样的。一般来说，访问寄存器的
15 速度最快，一般访问一次用时几百ps(皮秒)；其次是静态随机访问存储器(Static Random Access Memory, SRAM)，一般访问一次用时在几百ps到几ns(纳秒)的范围内；再次是内存单元，也就是双倍速率同步动态随机访问存储器(Double Data Rate Synchronous Dynamic Access Memory, DDR SDRAM)，一般访问一次用时在几十到几百ns；最后是通过IO口访问的其他存储器，如
20 硬盘等，其访问速度缓慢，一般是ms(毫秒)级。

在神经网络处理场合，一般关注的是处理单元对内存单元的访问。众所周知，处理单元的速度非常快，其主频一般是几百MHz(兆赫兹)到几GHz(吉赫兹)，也就是ps到ns级，而内存单元的访问速度是几十ns级别，两者的速度有着较大的差异。如何解决处理单元和内存访问的速度差，有效发挥处理单元的
25 算力，是现代CPU设计的一个难点。

图 1 是一种处理核中读取数据的示意图。

如图 1 所示,该处理核中,处理单元(Processing Unit,PU)和存储单元 memory 之间插入高速缓存 Cache,PU 采用分层的、间接的方式访问 Memory,即 PU 直接访问的是 Cache,PU 通过 Cache 间接访问 Memory。Cache 是 Memory 的映射,其内容是内存内容的子集。Cache 没有独立的编制空间,Cache 的地址与访问的 memory 的地址相同。

例如,PU 在执行程序时,通过 Cache 从 Memory 中读取了一些数据,即 Cache 会保存了这部分数据,当在短时间内 PU 需要再次使用该部分数据时,则 PU 会直接从 Cache 中调用。

但是,对于 PU 执行的程序,Cache 是透明的,没有功能上的意义,也就是程序不能单独的访问 Cache,即程序认为 PU 是从 memory 中调取了数据,但是实际上 PU 是从 Cache 中调取的数据。

上述方案存在如下的缺陷:

(1) 由于神经网络计算中,用到的参数和数据量庞大,通常远超 Cache 的容量,而 Cache 基于数据的时间局部性特性和空间局部性特性而采取的降低访问失效率的措施将无法实现,从而大大降低处理单元的算力发挥。

(2) Cache 电路复杂,导致芯片设计的难度和芯片的成本较高。

下面将详细说明本申请一实施方式提供的数据存储管理装置。在本发明的描述中,需要说明的是,术语“第一”、“第二”、“第三”、“第四”仅用于描述目的,而不能理解为指示或暗示相对重要性。此外,下面所描述的本发明不同实施方式中所涉及的技术特征只要彼此之间未构成冲突就可以相互结合。

图 2 是根据本发明一实施方式的数据存储管理装置的结构示意图。

如图 2 所示,该数据存储管理装置包括:至少两个随机存取存储器 RAM、控制单元和直接存储访问控制器(Direct Memory Access Controller,DMAC)。

优选的,该数据存储管理装置可设置在处理核中。

其中,至少两个 RAM 包括 RAM_0、RAM_1...RAM_N。该数据存储管理装置设置的至少两个 RAM,各个 RAM 可以独立的并行的被访问。

可选的，所有的 RAM 的存储容量可以相同或者不同。

控制单元，用于接收指令，根据指令生成并发送控制信号 C_DMAC。其中，指令是位于数据存储管理装置外的处理单元 PU 发送的。

DMAC, 用于根据控制信号实现对所述 RAM 中数据的存取。

5 在一个实施例中，DMAC, 用于根据控制信号实现对 RAM 中数据的存取，包括：DMAC, 用于根据所述控制信号从外部存储单元 Memory 中读取数据，并将所述数据存储在该控制信号指示的所述 RAM 中。

10 在一个实施例中，DMAC 用于根据控制信号实现对 RAM 中数据的存取，包括：DMAC, 用于根据所述控制信号从所述控制信号所指示的所述 RAM 中读取数据，并将所述数据存储在外部的 memory 中。

优选的，所述控制信号所指示的 RAM 的个数为一个或多个。

优选的，所有的 RAM 的访问地址统一编址，更优选的，RAM 的访问地址连续的编制，降低程序控制的复杂度。

15 在一个优选的实施例中，所有的 RAM 的地址与所述外部存储单元的地址统一编址。

例如，该数据存储管理装置中含有两个 RAM，即 RAM_0 和 RAM_1。其中，RAM_0 的地址为 0000H-0FFFH，RAM_1 的地址为 1000H-1FFFH，外部的 memory 的地址为 2000H-FFFFH。

20 进一步优选的，DMAC 的访问地址为全地址范围，具体的是所有的 RAM 的地址段和外部的 memory 的地址段。

PU 的访问地址为所有的 RAM 的地址段。

25 在一个实施例中，DMAC 在根据所述控制信号从外部 memory 中读取数据，并将数据存储在该控制信号指示的所述 RAM 中后，向外部的处理单元发送存储完成的信号，存储完成的信号用于提示外部的处理单元可以从刚刚完成存储的 RAM 中读取数据。

图 3 是根据本发明一实施方式的处理核的结构示意图。

如图 3 所示，该处理核包括处理单元 PU、存储单元 memory 和上述实施

方式提供的数据存储管理装置。

PU 用于发送指令，指令用于指示数据存储管理装置实现对 memory 中数据的存取。

其中指令包括取数指令和存数指令。

- 5 所述处理单元，用于发送指令，所述指令用于指示所述数据存储管理装置实现对所述 memory 中数据的存取，包括：

PU 用于发送取数指令，取数指令用于指示数据存储管理装置从 memory 中读取数据，并将数据存储于取数指令所指示的 RAM 中。优选的，当数据完成在指令所指示的 RAM 中的存储后，DMAC 向 PU 发送存储完成的信号。

- 10 在一个实施例中，PU 还用于从任意 RAM 中读取执行程序所需的数据。

优选的，PU 用于在每次收到 DMAC 发出的存储完成的信号后，发出新的取数指令，并从刚刚完成存储的 RAM 中读取数据。

- 具体地，当 PU 在收到 DMAC 发出的存储完成的信号后，发出新的取数指令，然后从刚刚完成存储的 RAM 中读取数据，这样 DMAC 根据新的取数指令从
15 memory 中读取数据并存储于相应的 RAM 中，与 PU 从刚刚完成存储的 RAM 中取数以执行程序可以并行，提高运算的效率。

当然，还可以 PU 在收到 DMAC 发出的存储完成的信号后，先从刚刚完成存储的 RAM 中读取数据，然后在发出新的取数指令。

在一个实施例中，在同一时间点，所述 PU 与 DMAC 访问不同的所述 RAM。

- 20 具体地，至少两个 RAM 包括第一 RAM 和第二 RAM。在第一时间，PU 从第一 RAM 中读取第一数据，DMAC 向第二 RAM 中写入从 memory 中取出的第二数据。在第二时间，PU 从第二 RAM 中读取第二数据，DMAC 向第一 RAM 中写入从 memory 中取出的第三数据。

- 可选的，PU 与 DMAC 还可以同时访问同一个 RAM，RAM 串行的响应 PU 和
25 DMAC。

可选的，如果 RAM 为双端口 RAM，PU 与 DMAC 还可以同时访问同一个 RAM，双端口 RAM 并行的响应 PU 和 DMAC。

需要说明的是，上述第一 RAM 可以为包括多个 RAM 的第一组 RAM，第二 RAM 可以为包括多个 RAM 的第二组 RAM。

可选的，第一组 RAM 的 RAM 个数可以与第二组 RAM 的个数相同或不同。

当第一 RAM 为第一组 RAM，第二 RAM 为第二 RAM 组时，在同一时间，PU
5 和 DMAC 可以分别访问一个组内的各个 RAM，或者，在同一时间，PU 和 DMAC 访问分属于两个组的 RAM。

具体的，在第一时间，PU 从第一组 RAM 中读取第一数据，DMAC 向第二组 RAM 中写入从 memory 中取出的第二数据；在第二时间，PU 从第二组 RAM 中读取第二数据，DMAC 向第一组 RAM 中写入从 memory 中取出的第三数据。

10 可选的，第一 RAM 为第一组 RAM，第二 RAM 为第二 RAM 组时，所述处理单元或 DMAC 也可以分时访问同一个组内的各个 RAM。

本发明实施方式提供的处理核，PU 从 RAM 中取数和 DMAC 将存储在 memory 的数据存储在 RAM 中能够并行处理，进一步能够提高处理核的算力，更加适用于神经网络的运算。另外，处理核中无需设计复杂的 Cache 电路，节省了
15 处理核的成本也降低了芯片设计的难度，并且由于处理核中无需设计 Cache 电路，处理单元无需通过高速缓存 Cache 从外部 memory 中取数，消除了由于 Cache 访问失效带来的计算效率的下降，处理核可以直接从存储管理装置的 RAM 中调用数据，提高了程序效率的可控性。

图 4 是根据本发明一实施方式的神经网络的结构示意图。

20 如图 4 所示，神经网络为 2 层，第一层的输出结果作为第二层的输入，第二层的输出结果为整个神经网络的输出。

图 5 是根据本发明一实施方式的处理核的结构示意图。图 5 所示的处理核用于实现图 4 所示的神经网络的计算。

如图 5 所示，该处理核包括数据存储管理装置、处理单元和存储单元。

25 其中，数据存储管理装置包括 RAM_0、RAM_1、DMAC 和控制单元。

假设每一层神经网络的参数和数据均小于单块 RAM 的容量，即 RAM_0 和 RAM_1 均可容下一层神经网络计算的参数和数据。

设置在同一时间点, PU 和 DMAC 访问不同的 RAM, 进而实现 PU 执行程序与 DMAC 进行存储数据并行, 从而优化计算和存储效率。例如第一时间点 PU 访问 RAM_0, DAMC 访问 RAM_1。在第二时间点, PU 访问 RAM_1, DAMC 访问 RAM_0, 如此循环。

5 在一个具体的实施例中, PU 发送指令 `lls_dis`, 控制单元接收指令, 并产生控制信号 `C_DMAC` 并发送给 DMAC, DMAC 根据指令从 memory 中读取指令所指示的数据, 并将数据存储至指令所指示的 RAM_0 中, 当 DMAC 将数据存储完成后, 向 PU 发送存储完成的信号; PU 接收到存储完成的信号, 发出新的指令用于指示 DMAC 从 memory 中读取新的数据存储至 RAM_1 中, 然后从 RAM_0
10 中读取数据。当 DMAC 将数据存储至 RAM_1 后, 发送存储完成的信号, PU 再次发出指令, 然后从 RAM_1 中读取数据, 再次发出的指令指示 DMAC 从 memory 中读取新的数据存储至 RAM_0 中。这样, DMAC 存储数据和 PU 读取数据实现了并行处理, 从而计算和存储都可以最大的发挥效率。

图 6 是根据本发明一实施方式的处理核进行神经网络计算的时序图。

15 如图 6 所示, 在 t_0 时, 该 PU 从 RAM_0 中读取数据时, 即 RAM_0 被 PU 执行神经网络第一层的程序所占用, 同时 DMAC 将从 memory 中读取的数据写入到 RAM_1 中; 在 t_1 时, 当 RAM_1 被 PU 执行神经网络第二层的程序所占用, 同时 DMAC 将从 memory 中读取的数据写入到 RAM_0 中。通过设置 PU 和 DMAC 访问地址的不同, 无需 PU 从 memory 中取数, 能够降低程序的复杂度, 提高
20 处理核的算力。

应当理解的是, 可以设置 PU 在 t_2 时执行的程序与在 t_1 时执行的程序相同, 即在 t_2 时也进行相同的神经网络的第一层的计算, 或者, 设置 PU 在 t_2 时执行的程序与在 t_1 时执行的程序不相同, 本发明不以此为限。

根据本发明的一个实施方式, 提供了一种芯片, 包括一个或多个上述实
25 施方式提供的处理核。

根据本发明的一个实施方式, 提供了一种卡板, 包括一个或多个上述实施方式提供的芯片。

根据本发明的一个实施方式，提供了一种电子设备，包括一个或多个上述实施方式提供的卡板。

图 7 是本发明一实施方式提供的数据存储管理方法，该方法包括：步骤 S101-步骤 S102；

5 其中，步骤 S101，控制单元接收指令，根据所述指令生成并发送控制信号。

步骤 S102，直接存储器访问控制器 DMAC 根据所述控制信号实现对所述 RAM 中数据的存取。

DMAC 根据所述控制信号实现对所述 RAM 中数据的存取，包括：DMAC 根据
10 所述控制信号从外部 memory 中读取数据，并将所述数据存储在该控制信号指示的 RAM 中。

DMAC 根据所述控制信号实现对所述 RAM 中数据的存取，包括：DMAC 根据所述控制信号从所述控制信号所指示的 RAM 中读取数据，并将所述数据存储
在外部 memory 中。

15 可以理解的是，本实施方式提供的数据存储管理方法由本发明前述实施方式提供的数据存储管理装置执行，因此，相同特征在此不做雷同描述。

本发明一实施方式提供了处理核处理数据的方法流程示意图。

该方法包括：步骤 S201-步骤 S202，

步骤 S201，处理单元发送取数指令；

20 步骤 S202，数据存储管理装置根据所述取数指令从存储单元中读取数据，并将数据存储至取数指令指示的数据存储管理装置的 RAM 中。

在一个优选的实施例，数据存储管理装置在将数据存储至取数指令指示的数据存储管理装置的 RAM 中后，发出存储完成的信号。

所述处理单元在每次收到 DMAC 发出的存储完成的信号后，发出新的取数
25 指令，在从刚刚完成的取数指令所指示的 RAM 中读取数据。

在一个实施例中，在同一时间点，处理单元与直接存储器访问控制器 DMAC 访问不同的所述 RAM。

具体地，在第一时间，处理单元从第一 RAM 中读取第一数据，DMAC 向第二 RAM 中写入从 memory 中取出的第二数据。在第二时间，PU 从第二 RAM 中读取第二数据，DMAC 向第一 RAM 中写入从 memory 中取出的第三数据。

可以理解的是，本实施方式提供的处理核处理数据的方法由本发明前述
5 实施方式提供的处理核执行，因此，相同特征在此不做雷同描述。

根据本发明的一个实施方式，提供了一种电子设备，包括：存储器，用于存储计算机可读指令；以及一个或多个处理器，用于运行所述计算机可读指令，使得所述处理器运行时实现前述实施方式的数据存储管理的方法。

根据本发明的一个实施方式，提供一种非暂态计算机可读存储介质，该
10 非暂态计算机可读存储介质存储计算机指令，该计算机指令用于使计算机执行前述实施方式的数据存储管理的方法。

根据本发明的一个实施方式，提供一种计算机程序产品，其中，包括计算机指令，当所述计算机指令被计算设备执行时，所述计算设备可以执行前述实施方式的数据存储管理的方法。

应当理解的是，本发明的上述具体实施方式仅仅用于示例性说明或解释
15 本发明的原理，而不构成对本发明的限制。因此，在不偏离本发明的精神和范围的情况下所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。此外，本发明所附权利要求旨在涵盖落入所附权利要求范围和边界、或者这种范围和边界的等同形式内的全部变化和修改例。

20