

存内处理芯片

交叉引用

[0001] 本申请要求于 2024 年 10 月 9 日提交中国专利局、申请号为 202411396760.9、申请
5 名称为“存内处理芯片”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

[0002] 本申请实施例涉及半导体领域，特别涉及一种存内处理芯片。

背景技术

[0003] 随着人工智能、大数据的不断发展，各种应用场景对于算力的需求不断提高。然而，
10 由于主流计算架构采用的是存算分离的冯诺依曼架构，而存储器的带宽增长速度远远赶不上
处理器算力提升速度，因此会存在由于带宽不足而导致的计算系统实际算力受限的存储墙问
题。已有的存内处理芯片是将存储芯片中的部分存储单元替换为计算单元实现的，可以不改
变整个芯片的封装方式，但存在容量损失和算力不高的问题。

发明内容

15 [0004] 本申请实施例提供一种新的存内处理芯片的架构。
[0005] 根据本申请一些实施例，本申请实施例提供一种存内处理芯片，包括接口电路，存储
电路以及计算单元；所述接口电路与所述存储电路之间具有第一传输路径；所述存储电路与
所述计算单元之间具有第二传输路径；所述接口电路与所述计算单元之间具有第三传输路径，
所述第三传输路径与所述第一传输路径相互独立。
20 [0006] 在一些实施例中，所述存内处理芯片包含多个所述存储电路和多个所述计算单元，每
一所述计算单元对应至少一个所述存储电路，不同所述存储电路具有对应的所述第一传输路
径和所述第二传输路径，不同的所述存储电路通过对应的所述第一传输路径与所述接口电路
连接。
[0007] 在一些实施例中，所述第一传输路径至少用于传输模型数据，所述第二传输路径至少
25 用于传输模型数据，所述第三传输路径至少用于传输初始数据和目标数据，所述计算单元基
于所述模型数据对所述初始数据进行计算处理，以得到所述目标数据。
[0008] 在一些实施例中，所述第一传输路径还用于传输所述初始数据和所述目标数据，所述
第二传输路径还用于传输所述初始数据和所述目标数据。
[0009] 在一些实施例中，所述第一传输路径和所述第二传输路径的带宽大于所述第三传输路
30 径，所述第三传输路径传输所述初始数据和所述目标数据的优先级高于所述第一传输路径和
所述第二传输路径。
[0010] 在一些实施例中，所述存内处理芯片还包括：选通电路，所述存储电路通过所述选通

电路与所述第一传输路径或所述第二传输路径连接。

[0011] 在一些实施例中，所述选通电路还用于记录对应的所述存储电路的刷新信息，且用于在所述接口电路与所述存储电路断开连接之前或者重新连接之后，将记录的所述刷新信息发送给所述接口电路。

- 5 [0012] 在一些实施例中，所述存内处理芯片还包括：存储控制器，设置于对应的存储电路和对应的选通电路之间，或者，设置于对应的选通电路和计算单元之间，所述存储控制器至少用于读取所述存储电路中的数据。

[0013] 在一些实施例中，所述存内处理芯片还包括：模式控制电路，用于控制所述存储电路与所述第一传输路径连接，或者，控制所述存储电路与所述第二传输路径连接。

- 10 [0014] 在一些实施例中，所述模式控制电路还用于所述第三传输路径导通或断开。

[0015] 在一些实施例中，所述模式控制电路处于所述接口电路所在区域中。

[0016] 在一些实施例中，所述存内处理芯片中所述存储电路与所述计算单元之间的总带宽大于所述存内处理芯片与外部处理器之间的带宽。

附图说明

- 15 [0017] 一个或多个实施例通过与之对应的附图中的图片进行示例性说明，这些示例性说明并不构成对实施例的限定，除非有特别申明，附图中的图不构成比例限制。

[0018] 图 1 至图 7 为本申请实施例提供的存内处理芯片的结构示意图；

[0019] 图 8 为本申请实施例提供的计算系统的结构示意图。

具体实施方式

- 20 [0020] 下面将结合附图对本申请的各实施例进行详细的阐述。然而，本领域的普通技术人员可以理解，在本申请各实施例中，为了使读者更好地理解本申请而提出了许多技术细节。但是，即使没有这些技术细节和基于以下各实施例的种种变化和修改，也可以实现本申请所要求保护的技术方案。

[0021] 图 1 至图 7 为本申请实施例提供的存内处理芯片的结构示意图。

- 25 [0022] 参考图 1，存内处理芯片包括：接口电路 11，存储电路 12 以及计算单元 13，接口电路 11 与存储电路 12 之间具有第一传输路径 TR1，存储电路 12 与计算单元 13 之间具有第二传输路径 TR2，接口电路 11 与计算单元 13 之间具有第三传输路径 TR3，第三传输路径 TR3 与第一传输路径 TR1 相互独立。

- [0023] 本公开中，计算单元 13 分别与接口电路 11 和存储电路 12 连接，第一传输路径 TR1、
30 第二传输路径 TR2 以及第三传输路径 TR3 均可用于数据传输，第三传输路径 TR3 和第一传输路径 TR1 相互独立指的是，第一传输路径 TR1 的可用状态和第三传输路径 TR3 的可用状态之间没有必然关联，两者可以同时处于可用状态或不可用状态，也可以一者处于可用状态而另一者处于不可用状态，取决于对应的控制信号。每一传输路径的可用状态至少与两个因素有

关：1.该传输路径内部的导通或关断，例如内部驱动器的导通或关断，若传输路径本身关断，则传输路径不可用；2.该传输路径与数据输入电路或数据输出电路是否数据连通，若传输路径与数据输入电路或数据输出电路的数据通路断开，则传输路径不可用。数据输入电路与传输路径的输入端直接连接，数据输出电路与传输路径的输出端直接连接。

- 5 [0024] 如此，计算单元 13 在完成计算处理之后无需通过第二传输路径 TR2 传输数据以及无需通过存储电路 12 进行数据的转发，而可以直接通过独立的第三传输路径 TR3 将处理后的目标数据通过接口电路 11 输出，如此，有利于直接提高数据传输效率以及间接提高存内处理芯片的数据处理效率；此外，由于设置有两条相互独立的传输路径，计算单元 13 可以在同一时刻通过其中一条传输路径接收数据，以及通过另一传输路径输出数据，如此，有利于简化时
10 序和进一步提高数据传输效率。

[0025] 可以理解的是，数据传输和信号传输是不同的概念，在本申请中，信号传输指的是控制信号的传输，数据传输指的是模型数据和常规数据的传输，信号传输的目的是为了辅助数据传输。

[0026] 以下将结合附图对本申请实施例进行更为详细的说明。

- 15 [0027] 本公开中，接口电路 11 不仅可以接收和输出数据，还可以承担部分信号处理功能。接口电路 11 的信号处理功能可以根据存内处理芯片中其他电路的实际架构进行调整。存储电路 12 可以视为存储芯片，包含易失型存储芯片和非易失型存储芯片等，存储电路 12 包括用于存储数据的存储单元和用于控制数据存储的外围电路。

- [0028] 计算单元 13 用于基于目标规则对输入数据进行计算处理，并将处理后的目标数据进行输出；其中，计算处理既可以是对目标数据进行运算，以从中提炼所需要的其他类型信息，
20 例如视频不同时刻中不同位置的颜色，也可以是对目标数据进行简化，以减小传输所需的带宽，对于被简化的数据，后续可能还需要进行一定程度地或者完全的还原，以便于用于提炼其他类型的信息。本申请并不对计算单元的功能进行任何限制，计算单元的功能可以根据实际应用场景适应性调整。此外，本申请也不对计算单元的类型进行任何限制，计算单元 13 既
25 可以使用传统算术逻辑单元实现，也可以使用新型的 CIM (Compute in Memory) 存内计算单元，例如 SRAM 存内计算单元。

- [0029] 在一些实施例中，参考图 2，存内处理芯片包含多个存储电路 22 (Bank) 和多个计算单元 23 (PU)，每一计算单元 23 对应至少一个存储电路 22，不同存储电路 22 具有对应的
30 的第一传输路径 TR1 和第二传输路径 TR2，不同的存储电路 22 通过对应的第一传输路径 TR1 与接口电路 21 连接。

[0030] 在图 2 所示示例中，每一计算单元 23 对应一个存储电路 22，每个存储电路 22 具有对应的第一传输路径和第二传输路径，不同存储电路 22 对应的传输路径相互独立。不同的第一传输路径的第一端用于连接不同的存储电路 22，不同的第一传输路径的第二端均用于连接接口电路 21，以用于接收或输出数据。不同的第二传输路径的第一端用于连接不同的存储电

路 22，不同的第二传输路径的第二端用于连接不同的计算单元 23。当每一计算单元 23 对应至少两个存储电路 22 时，不同的第二传输路径的第二端可用于连接相同的计算单元 23。可以理解的是，在一些场景中，还可以一部分计算单元 23 对应一个存储电路 22 以及另一部分计算单元 23 对应至少两个存储电路，如此，有利于提高存内处理芯片内部的算力分配，避免
5 算力浪费。

[0031] 存储电路 22 和计算单元 23 均具有对应的连接模块 24，存储电路 22 和计算单元 23 通过对应的连接模块 24 进行连接，连接方式包括 TSV (Through Silicon Vias，硅穿孔) 和混合键合 (Hybrid bonding) 中的至少一种，进而进行数据传输和信号控制。

[0032] 进一步地，存内处理芯片可以由在垂直方向上堆叠的多个芯片构成，芯片之间的通信
10 协议为私有协议，多个芯片被封装在一起，构成存内计算芯片。在一些实施例中，参考图 2，存内处理芯片包含在垂直方向堆叠的计算芯片和存储芯片，计算单元 23 和接口电路 21 均设置在计算芯片中，存储电路 22 设置在存储芯片中。可以理解的是，存内处理芯片还可以由在垂直方向上堆叠的至少 1 个存储芯片和至少 1 个计算芯片构成，例如两个存储芯片或者两个计算芯片，理论上可以 N 个存储芯片和 M 个计算芯片，M 和 N 为大于 1 的自然数。

[0033] 在其他实施例中，存内处理芯片包括在垂直方向堆叠的接口芯片、计算芯片和存储芯片，接口电路 21 设置于接口芯片中，即接口电路 21 和计算单元 23 设置于不同的芯片中，在三个芯片相堆叠的情况下，可以将任一芯片设置于中间位置。在一些场景下，可以将输入输出的数据量较大的芯片设置于中间位置，例如预存数据量较大的模型数据的存储芯片，和/或，将因数据传输或数据计算而导致发热量较高的芯片设置于非中间位置，即设置于外侧，
20 例如需要进行计算的计算芯片或者需要频繁进行输入输出的接口芯片。

[0034] 重新参考图 1，在一些实施例中，第一传输路径 TR1 至少用于传输模型数据，第二传输路径 TR2 至少用于传输模型数据，第三传输路径 TR3 至少用于传输初始数据和目标数据，计算单元 13 基于模型数据对初始数据进行计算处理，以得到目标数据。

[0035] 模型数据也可以理解为模型权重数据，用于表征计算处理的目标规则。接口电路 11
25 在接收到模型数据之后，可先存储于存储电路 12 中，计算单元 13 可通过第二传输路径 TR2 从存储电路 12 中读取模型数据，以及通过接口电路 11 和第三传输路径 TR3 接收初始数据，进而基于模型数据对初始数据进行计算处理，获得处理后的目标数据。目标数据可以通过第三传输路径 TR3 和接口电路 11 进行输出。

[0036] 可以理解的是，“初始数据”指未处理的数据，“目标数据”指处理后的数据，而无论是初始数据和目标数据，区别仅在于是否进行了计算处理，但都是相对于模型数据而言的常规数据。为了表达上的简洁，后续部分描述将使用“常规数据”同时指代初始数据和目标数据。
30

[0037] 在一些实施例中，接口电路 11 可以先接收模型数据再接收初始数据，如此，能够在存储电路 12 中预存模型数据错开时序，实现接口电路 11 采用相同的输入输出端口 (以

下简称为 IO 端口) 依次接收模型数据和初始数据, 简化 IO 端口; 在另一些实施例 5 中, 接口电路 11 中用于接收模型数据和初始数据的端口不同。需要说明的是, 即便接口电路 11 采用相同的 IO 端口接收和输出模型数据和常规数据, 接口电路也可以通过不同的传输参数进行数据传输, 传输参数包括连续传输长度 (burst length) 和同一时刻使用的端口数量。例如, 接口电路 11 可以利用 16 个端口传输模型数据, 且利用 16 个端口中的 8 个端口传输常规数据; 以及传输模型数据时连续传输长度为 16, 而传输常规数据时连续传输长度为 8。

[0038] 在一些场景下, 模型数据的体量一般远远大于初始数据的体量, 前者通常为 GB 级别, 后者通常为 MB 级别, 因此可以采用存储电路 12 预存模型数据以加快载入速度, 而不存储初始数据。

10 [0039] 在一些实施例中, 第一传输路径 TR1 还用于传输初始数据和目标数据, 第二传输路径 TR2 用于传输初始数据和目标数据。如此, 初始数据还可以通过第一传输路径 TR1 和第二传输路径 TR2 输入至计算单元内 13, 目标数据既可以通过第三传输路径输出, 也可以通过第二传输路径和第一传输路径进行输出。如此, 有利于保证在第三传输路径损坏或拥堵 (缓冲空间不足) 时, 通过第一传输路径和第二传输路径输入初始数据以及输出目标数据。

15 [0040] 在一些实施例中, 第一传输路径和第二传输路径的带宽大于第三传输路径的带宽, 第三传输路径传输初始数据和目标数据的优先级高于第一传输路径和第二传输路径。

[0041] 需要注意的是, 目标数据的数据量与初始数据的数据量没有绝对的大小关系, 取决于计算单元的计算规则。在一些实施例中, 若目标数据的数据量大于初始数据的数据量或者说若目标数据的数据量大于目标阈值, 则可以通过第一传输路径和第二传输路径传输至少部分
20 目标数据, 从而提高传输目标数据的效率。其中, 既可以是超出目标阈值的部分目标数据由第一传输路径和第二传输路径传输, 也可以是所有目标数据均由第一传输路径和第二传输路径传输。

[0042] 在初始数据通过第一传输路径输入之后, 且传输至第二传输路径 TR2 之前, 既可以存储于存储电路 12 内的存储单元中, 也可以暂存于存储电路 12 内的缓冲器或锁存器之类的
25 缓冲电路中, 在将初始数据传输至第二传输路径 TR2 之后, 存储于存储单元中的初始数据可以保留也可以清除 (存储单元的数据清除指的是存储单元的电位被调整为预充电状态, 不表征任何数据, 或者, 存储单元被调整为可被新数据覆盖的状态), 暂存于缓冲电路中的初始数据通常清除 (缓冲电路的清除指的是缓冲电路中的数据被调整会默认电位状态)。

[0043] 在目标数据通过第二传输路径输出时, 既可以暂存于存储电路 12 的缓冲电路中, 也可以存储于存储电路的存储单元中。传输至存储电路 12 中的目标数据既可以存储于存储电路
30 12 中且暂不输出, 也可以直接通过第一传输路径 TR1 和接口电路 11 进行输出。若存储于存储电路 12 中且暂不输出, 则可以存储于存储电路 12 的存储单元中, 以通过刷新保证目标数据的准确存储。在将处理后的目标数据传输至第一传输路径 TR1 之后或者通过接口电路 11 输出之后, 存储于存储单元中的目标数据可以保留也可以清除。每一存储电路 12 可以在接收到输

出指令或者自身存储的未输出的目标数据量达到第一预设阈值后输出之前未输出的目标数据，或者，在所有存储电路 12 存储的未输出的总目标数据量达到第二预设阈值后输出之前未输出的目标数据。

[0044] 在又一些实施例中，第二传输路径 TR2 用于单向传输初始数据，单向传输初始数据
5 指的是只允许自存储电路 12 向计算单元 13 传输初始数据，而不允许从计算单元 13 向存储电路 12 传输目标数据。如此，有利于简化时序控制，避免第二传输路径因指令错误同时进行输入输出而发生冲突，同时保留了因第三传输路径拥堵而暂时通过第一传输路径和第二传输路径输入待处理的初始数据的可能。

[0045] 在一些实施例中，参考图 3，存内处理芯片还包括：选通电路 35，存储电路 32 通过
10 选通电路 35 与第一传输路径 TR1 或第二传输路径 TR2 连接。也就是说，选通电路 35 包含选通器的功能，在同一时刻，仅有计算单元 33 和接口电路 31 中的一者能够与存储电路 32 连接。如此，有利于控制数据传输流向，将初始数据和模型数据的数据存储操作和初始数据的数据计算操作分离，避免数据存储操作和数据计算操作并行而发生冲突。

[0046] 在一些实施例中，选通电路 35 还用于记录对应的存储电路 32 的刷新信息，且用于
15 在接口电路 31 与存储电路 32 断开连接之前或者重新连接之后，将记录的刷新信息发送给接口电路 31。刷新信息包含常规刷新操作（即按序刷新）的当前刷新行，当前刷新行可以根据刷新计数器的计数值得到。刷新信息还可以包括行锤地址信息，行锤地址信息可以包括达到行锤刷新阈值的被攻击行或受害行地址，以及包括尚未达到行锤刷新阈值但正在进行地址累计的被攻击行地址。若接口电路 31 具有内存管理功能，则接口电路 31 可以根据接收到的刷新
20 信息控制存储电路 32 的内部刷新，若接口电路 31 不具有内存管理功能，则可以由与接口电路 31 连接的外部电路（例如 CPU）控制存储电路 32 的内部刷新。

[0047] 其中，选通电路 35 既可以通过内设的独立的计数器，记录对应的存储电路 32 的刷新信息，也可以直接获取存储电路 32 内刷新计数器的信息来实现刷新信息的记录。

[0048] 需要说明的是，若选通电路 35 在接口电路 31 与存储电路 32 断开连接之前发送刷新
25 信息，则接口电路 31 或外部电路可以根据刷新信息和目标数据是否通过接口电路 31 输出控制存储电路 32 与接口电路 31 的连接时机，以保证存储电路 32 内部的及时刷新。若存内处理芯片中包含多个存储电路 32，由于每个存储电路 32 均具有对应的一个选通电路 35（不同存储电路 32 对应的选通电路 35 不同），因此，接口电路 31 或外部电路会接收到多个存储电路 32 的刷新信息，进而根据不同存储电路 32 的刷新信息，控制接口电路 31 与不同存储电
30 路 32 的连接顺序，进而对不同存储电路 32 依次刷新。

[0049] 在一些实施例中，参考图 4，存内处理芯片还包括：存储控制器 46（LMC(s)，local memory controller(s)），设置于对应的存储电路 42 和对应的选通电路（未示意）之间，或者，设置于对应的选通电路和计算单元 43 之间，存储控制器 46 至少用于读取存储电路 42 中的数据。可以理解的是，无论存储控制器 46 的具体位置位于哪里，它都会被用于在部分时段控

制存储电路 42。

[0050] 存储控制器 46 与存储电路 42 之间的接口主要包括数据总线、地址/指令总线以及测试模式总线。存储控制器 46 所需要具备的基本功能和对应内容包括：指令转译模块，用于接收计算单元 43 或接口电路 41 传输的指令并转为私有协议下存储电路 42 可识别的指令，并根据私有接口的时序要求进行指令调度；数据处理模块，用于进行 ECC 错误检测和修正，用于将数据进行异步和同步的转换，用于数据缓存，以及用于数据接收与发送等；内存管理模块，用于对存储电路进行刷新控制，用于对存储电路进行模式控制，用于混合键合或硅穿孔的冗余控制与修复，用于对存储电路进行测试模式控制等。

[0051] 具体地，参考图 5，当存储控制器 56 设置于选通电路 55 和计算单元 53 之间时，接口电路 51 通过选通电路 55 与存储电路 52 连接，存储控制器 56 处于第二传输路径上，此时，若选通电路 55 与第一传输路径连接且与第二传输路径断开，则存储控制器 56 无法起到控制存储电路 52 的作用，如此，接口电路 51 需要起到部分控制存储电路的功能，即接口电路 51 与存储控制器 56 具有重复的功能。重复的功能包含存储电路指令译码，数据处理以及内存处理等。

[0052] 在一些实施例中，存储控制器 56 设置于选通电路 55 和计算单元 53 之间时，选通电路还用于通过第二传输路径向存储控制器 56 发送存储电路 52 的刷新信息，以便于存储控制器 56 中的内存管理模块基于刷新信息对存储电路 52 进行刷新控制。

[0053] 参考图 6，当存储控制器 66 设置选通电路 65 和存储电路 62 之间时，无论存储电路 62 与第一传输路径连接还是与第二传输路径连接，存储控制器 66 均会保持与存储电路 62 的连接，如此，接口电路 61 的功能可以不与存储控制器 66 的功能重复。在此场景下，接口电路 61 的主要功能包含指令解码与转译，数据处理功能，以及通过片上总线标准（例如 AXI，Advanced extensible Interface）与存储控制器 66 连接。

[0054] 此外，无论是在图 5 还是图 6 中，不同存储电路均通过主数据通路 datapath 与接口电路连接，接口电路同一时刻可以连接至少一个存储电路。在采用图 5 所示连接方式时，接口电路可以采用 JEDEC 标准的 DRAM 接口，在采用图 6 所示连接方式时，接口电路可采用 JEDEC 标准的 DRAM 接口或者采用其他通用总线，其他通用总线标准包括 CXL（Compute Express Link）和 UCIE（Universal Chiplet Interconnect Express，通用芯片快连）。除了接口电路与外部电路之间的连接协议采用公有标准以外，存内处理芯片内部不同电路之间的数据传输协议均可采用私有协议，例如接口电路与存储控制器之间，接口电路与存储电路之间，接口电路与计算单元之间，计算单元与存储控制器之间。

[0055] 需要说明的是，在图 4 所示实施例中，存储电路 42 采用 Bank(s) 进行标记，存储控制器 46 采用 LMC(s) 进行标记，计算单元 43 用 PU 进行标记，其表征存储电路 42 与存储控制器 46 一一对应，每个计算单元 43 对应至少一个存储电路 42 和至少一个存储控制器 46。而在图 5 和图 6 所示实施例中，存储电路采用 Bank 进行标记，存储控制器采用 LMC 进行标

记，计算单元用 PU 进行标记，其表征每个计算单元对应一个存储电路和一个存储控制器，不同存储电路并行设置，存储电路与选通电路一一对应。可以理解的是，每个计算单元 43 对应的存储电路 42 和存储控制器 46 的数量越多，每个计算单元 43 的最大带宽也就越大，同时，每个计算单元 43 占据的面积可能越大。进一步的，计算单元 43 的数量决定了整个存
5 内处理芯片的最大内部带宽。

[0056] 此外，在图 4 中，采用两条虚线分别连接存储电路 42 和存储控制器 46 表征图 5 和图 6 所示的两种实施例，即接口电路 41 可以通过选通电路与存储电路 42 连接但不与存储控制器 46 连接，也可以依次通过选通电路存储控制器 46 与存储电路 42 连接。

[0057] 在一些实施例中，参考图 7，存内处理芯片还包括：模式控制电路 77，模式控制电路
10 77 用于控制存储电路 72 与第一传输路径 TR1 连接，或者，控制存储电路 72 与第二传输路径 TR2 连接。换句话说，模式控制电路 77 通过控制选通电路 75 实现传输路径的切换。在图 7 中，实线表征的是数据流，虚线表征的是控制流，且以存储控制器 76 位于选通电路 75 和存储电路 72 之间作为示例，在此场景下，接口电路的部分核心功能是将外部的标准接口的指令转译为存储控制器 76 和模式控制电路 77 可识别的指令。

[0058] 在图 7 中，通过设置模式控制电路 77，使得存内处理芯片能够具有完全独立的两种
15 工作模式，且两种工作模式互不干扰，当存储电路 72 与第一传输路径 TR1 连接而与第二传输路径 TR2 断开时，存内处理芯片的功能为正常的存储芯片，此时可以作为常规存储芯片使用或者在此场景下写入模型数据，当存储电路 72 与第一传输路径 TR1 断开而与第二传输路径 TR2 连接时，存内处理芯片的功能至少包含计算功能。模式控制电路 77 在识别指令后可以向
20 计算单元 73 发出控制指令。控制信号可以用于控制计算单元 73 进行计算处理的时序，以及控制计算单元 73 从休眠模式调整为工作模式等。

[0059] 在一些实施例中，在图 5 所示实施例中，接口电路 51 与存储控制器 56 的重复功能还可以包括控制指令译码，接口电路 51 用于接收外部输入的指令并进行译码，存储控制器 56 用于接收模式控制电路输出的控制指令并进行译码，译码的结果至少可以表征选通电路是否
25 与第一传输路径断开并与第二传输路径连通。存储控制器 76 还可以具有休眠状态和使能状态，当选通电路与第一传输电路连通且存储控制器无需传输数据时，存储控制器处于休眠状态，当选通电路与第二传输电路连通时，存储控制器处于使能状态。

[0060] 根据在前的描述可知，当存储电路 72 与第一传输路径 TR1 断开而与第二传输路径 TR2 连接时，计算单元 73 既可以通过第二传输路径 TR2 接收初始数据，也可以通过第二传输路
30 径接收预存在存储电路 72 中的初始数据。在计算单元 73 完成计算处理之后，目标数据可以通过第三传输路径 TR3 进行输出，也可以通过第二传输路径 TR2 存储于存储电路 72 中。存储于存储电路 72 中的目标数据可以在选通电路 75 (MUX) 与第一传输路径 TR1 连接后输出或不输出。存储于存储电路 72 中的目标数据在通过第一传输路径 TR1 传输后可以清除或不清除。可以理解的是，计算单元 73 通过第二传输路径 TR2 将目标数据存储于存储电路 72 中和

通过第三传输路径 TR3 和接口电路 71 输出目标数据可以同时进行。

[0061] 在一些实施例中，模式控制电路 77 还用于控制第三传输路径 TR3 导通或断开。例如当选通电路 75 与第一传输路径 TR1 连接时，控制第三传输路径 TR3 断开，使得计算单元 73 不与接口电路 71 或存储电路 72 连接，且在选通电路 75 与第二传输路径 TR2 连接时，控制
5 第三传输路径 TR3 导通，使得计算单元 73 可以通过第三传输路径 TR3 接收初始数据和输出目标数据。

[0062] 在一些实施例中，第三传输路径 TR3 上还设置有全局缓冲器 78，用于接收初始数据和输出目标数据，可以对初始数据和目标数据进行时序调整，以等待计算单元完成计算处理操作或者等于接口电路完成数据输出操作。

10 [0063] 在一些实施例中，参考图 2，模式控制电路 27 处于接口电路 21 所在区域中。此外，在一些实施例中，存内处理芯片包括存储芯片和计算芯片，存储电路 22 位于存储芯片上，接口电路 21 和计算单元 23 处于计算芯片上。

[0064] 图 8 为本申请实施例提供的计算系统的结构示意图，计算系统至少包括外部处理器和上述任意实施例中的存内处理芯片。参考图 8，在一些实施例中存内处理芯片中存储电路 82
15 和计算单元 83 之间的总带宽大于存内处理芯片 84 与外部处理器之间的带宽。外部处理器包括中央处理器 85 (CPU)。需要说明的是，在存内处理芯片 84 本身的算力较低的情况下，还可以与外部的图形处理器 (Graphics Processing Unit, GPU) 和/或神经处理单元 (Neural Processing Unit, NPU) 进行协同计算；在存内处理芯片 84 本身算力较高的情况下，可以由存内处理芯片 84 独立完成人工智能计算。

20 [0065] 需要声明的是，本申请中描述的“连接”指的是可进行数据传输或者信号通信。

[0066] 本领域的普通技术人员可以理解，上述各实施方式是实现本申请的具体实施例，而在实际应用中，可以在形式上和细节上对其作各种改变，而不偏离本申请的精神和范围。任何本领域技术人员，在不脱离本申请的精神和范围内，均可作各自更动与修改，因此本申请的保护范围应当以权利要求限定的范围为准。

权 利 要 求 书

1. 一种存内处理芯片，包括接口电路，存储电路以及计算单元；
- 所述接口电路与所述存储电路之间具有第一传输路径；
- 所述存储电路与所述计算单元之间具有第二传输路径；
- 5 所述接口电路与所述计算单元之间具有第三传输路径，所述第三传输路径与所述第一传输路径相互独立。
2. 根据权利要求 1 所述的存内处理芯片，其中，所述存内处理芯片包含多个所述存储电路和多个所述计算单元，每一所述计算单元对应至少一个所述存储电路，不同所述存储电路具有对应的所述第一传输路径和所述第二传输路径，不同的所述存储电路通过对应的所述第一传输路径与
- 10 所述接口电路连接。
3. 根据权利要求 1 或 2 所述的存内处理芯片，其中，所述第一传输路径至少用于传输模型数据，所述第二传输路径至少用于传输模型数据，所述第三传输路径至少用于传输初始数据和目标数据，所述计算单元基于所述模型数据对所述初始数据进行计算处理，以得到所述目标数据。
- 15 4. 根据权利要求 3 所述的存内处理芯片，其中，所述第一传输路径还用于传输所述初始数据和所述目标数据，所述第二传输路径还用于传输所述初始数据和所述目标数据。
5. 根据权利要求 4 所述的存内处理芯片，其中，所述第一传输路径和所述第二传输路径的带宽大于所述第三传输路径，所述第三传输路径传输所述初始数据和所述目标数据的优先级高于所述第一传输路径和所述第二传输路径。
- 20 6. 根据权利要求 1 或 2 所述的存内处理芯片，其中，所述存内处理芯片还包括：选通电路，所述存储电路通过所述选通电路与所述第一传输路径或所述第二传输路径连接。
7. 根据权利要求 6 所述的存内处理芯片，其中，所述选通电路还用于记录对应的所述存储电路的刷新信息，且用于在所述接口电路与所述存储电路断开连接之前或者重新连接之后，将记录的所述刷新信息发送给所述接口电路。
- 25 8. 根据权利要求 6 所述的存内处理芯片，其中，所述存内处理芯片还包括：存储控制器，设置于对应的存储电路和对应的选通电路之间，或者，设置于对应的选通电路和计算单元之间，所述存储控制器至少用于读取所述存储电路中的数据。
9. 根据权利要求 6 所述的存内处理芯片，其中，所述存内处理芯片还包括：模式控制电路，用于控制所述存储电路与所述第一传输路径连接，或者，控制所述存储电路与所述第二传输路径连接。
- 30 传输路径连接。

10. 根据权利要求 9 所述的存内处理芯片，其中，所述模式控制电路还用于所述第三传输路径导通或断开。
11. 根据权利要求 9 所述的存内处理芯片，其中，所述模式控制电路处于所述接口电路所在区域中。
- 5 12. 根据权利要求 1 所述的存内处理芯片，其中，所述存内处理芯片中所述存储电路与所述计算单元之间的总带宽大于所述存内处理芯片与外部处理器之间的带宽。