(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 104572528 A
(43)申请公布日 2015.04.29

(21)申请号 201510040704.6

(22)申请日 2015.01.27

(71)申请人 东南大学
地址 214135 江苏省无锡市新区菱湖大道99号

(72)发明人 李冰 姜伟 徐寅 刘勇 赵霞 王刚 董乾

(74)专利代理机构 江苏永衡昭辉律师事务所 32250
代理人 王斌

(51)Int.Cl.
G06F 13/30(2006.01)
G06F 12/08(2006.01)

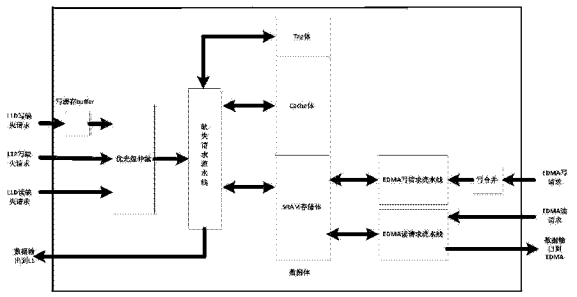
权利要求书1页 说明书5页 附图4页

(54)发明名称

一种二级Cache对访问请求的处理方法及系统

(57)摘要

本发明公开了一种二级Cache对访问请求的处理方法。本发明针对现有二级Cache访问请求技术的不足，对其进行了改进，利用三条并行的流水线对L1D读缺失请求、L1D写缺失请求、L1P读缺失请求、EDMA读请求、EDMA写请求这五种访问请求分别进行流水线处理，并利用缓存机制将多个写偏移地址相同的请求合并为一个以加速多个写请求的处理速度，通过比较偏移地址来减少侦听次数以提高EDMA的读数据能力。本发明还公开了一种二级Cache对访问请求的处理系统。相比现有技术，本发明对访问请求的处理效率更高，且结构简单、可移植性强，可在大多数微处理器芯片中都可应用。



CN 104572528 A

1. 一种二级 Cache 对访问请求的处理方法，其特征在于，利用三条并行的流水线：第一～第三流水线对访问请求进行并行处理，第一流水线负责处理一级数据 Cache 的读缺失请求、写缺失请求以及一级程序 Cache 的读缺失请求，第二流水线负责处理 EDMA 的读请求，第三流水线负责处理 EDMA 的写请求；为第一～第三流水线分别设置与其一一对应的第一～第三缓存单元，第一缓存单元用于缓存一级数据 Cache 的写缺失请求并将所缓存的多个行偏移地址相同的写缺失请求合并为一个，第二缓存单元用于缓存侦听到的数据并在 EDMA 发送数据的同时将侦听到的最新数据写回二级 Cache 的对应数据体中，第三缓存单元用于缓存 EDMA 的写请求并将所缓存的多个行偏移地址相同的写请求合并为一个。

2. 如权利要求 1 所述二级 Cache 对访问请求的处理方法，其特征在于，所述第一流水线为包括：地址解析、读取 Tag 体数据、Tag 值及状态位比较、读取数据体、数据发送这五级的五级流水线，第二流水线为包括侦听、侦听数据返回或读数据体、数据发送这三级的三级流水线，第三流水线为包括侦听、数据写回、EDMA 数据写入这三级的三级流水线。

3. 如权利要求 2 所述二级 Cache 对访问请求的处理方法，其特征在于，第二流水线在处理 EDMA 的读请求时，根据读请求的地址对一级 Cache 进行数据侦听并将所侦听数据的行偏移地址暂存；对于之后的每一个读请求，首先将该读请求所请求的行偏移地址与暂存的行偏移地址进行比较，如果相同，则不对一级 Cache 进行数据侦听，如果不同，则对一级 Cache 进行数据侦听并将该读请求所请求的行偏移地址暂存。

4. 一种二级 Cache 对访问请求的处理系统，其特征在于，包括用于对访问请求进行并行处理的三条并行的流水线：第一～第三流水线，以及与第一～第三流水线一一对应设置的第一～第三缓存单元；第一流水线负责处理一级数据 Cache 的读缺失请求、写缺失请求以及一级程序 Cache 的读缺失请求，第二流水线负责处理 EDMA 的读请求，第三流水线负责处理 EDMA 的写请求；第一缓存单元用于缓存一级数据 Cache 的写缺失请求并将所缓存的多个行偏移地址相同的写缺失请求合并为一个，第二缓存单元用于缓存侦听到的数据并在 EDMA 发送数据的同时将侦听到的最新数据写回二级 Cache 的对应数据体中，第三缓存单元用于缓存 EDMA 的写请求并将所缓存的多个行偏移地址相同的写请求合并为一个。

5. 如权利要求 4 所述二级 Cache 对访问请求的处理系统，其特征在于，所述第一流水线为五级流水线，依次包括：地址解析模块、读取 Tag 体数据模块、Tag 值及状态位比较模块、读取数据体模块、数据发送模块；第二流水线为三级流水线，依次包括侦听模块、侦听数据返回或读数据体模块、数据发送模块；第三流水线为三级流水线，依次包括侦听模块、数据写回模块、EDMA 数据写入模块。

6. 如权利要求 5 所述二级 Cache 对访问请求的处理系统，其特征在于，第二流水线的侦听模块中包括侦听偏移地址比较与暂存单元，用于暂存所侦听数据的行偏移地址并将新到来读请求所请求的行偏移地址与暂存的行偏移地址进行比较，如果相同，则侦听模块不对一级 Cache 进行数据侦听，如果不同，则侦听模块对一级 Cache 进行数据侦听并将新到来读请求所请求的行偏移地址暂存至侦听偏移地址比较与暂存单元。

7. 如权利要求 4～6 任一项所述二级 Cache 对访问请求的处理系统，其特征在于，该系统还包括用于对同时到来的多个缺失请求进行处理优先级判断的优先级仲裁模块，处理优先级从高到低依次为：一级程序 Cache 的读缺失请求、一级数据 Cache 的写缺失请求、一级数据 Cache 的读缺失请求。

一种二级 Cache对访问请求的处理方法及系统

技术领域

[0001] 本发明涉及一种二级 Cache 对访问请求的处理方法及系统，属于微处理器技术领域。

背景技术

[0002] Cache（高速缓冲存储器） 技术的广泛应用，很好的解决了存储墙问题对微处理器性能提升的限制，而超大规模集成电路技术的发展，也使得片上集成大容量 Cache 成为可能，这在很大程度上降低了 Cache 的失效率。一级 Cache 分为程序 Cache（L1P）和数据 Cache（L1D），二级 Cache 为数据和程序共享 Cache（L2），当一级 Cache 缺失时便会向二级 Cache 发出缺失请求。EDMA 是直接存储器访问控制器，连接着二级 Cache 和外存，用于实现大片数据从内存到外存的搬移。

[0003] 由于一级 Cache 容量有限，并且随着内核处理速度的提高及数据通路的增加，一级 Cache 的缺失率增加，采用流水线处理这些缺失请求能有效减少因缺失造成的大量时间阻塞。由于 EDMA 的读写请求为猝发传输，一次可发送最多 8 个请求，传统的串行处理将阻塞大量的时钟周期。

[0004] 现有的二级 Cache 处理访问请求的方法有串行处理方法和单一流水线处理方法，串行处理方法面对多个请求连续性的到来时将导致浪费大量时间，而单一的流水线一定程度上能缓解这种时间上的阻塞，但是整条流水线周期较长，依旧无法满足高性能处理器的处理要求。

发明内容

[0005] 本发明所要解决的技术问题在于克服现有技术所存在的二级 Cache 对访问请求的处理效率较低的问题，提供一种更高效的二级 Cache 对访问请求的处理方法及系统。

[0006] 本发明具体采用以下技术方案解决上述技术问题：

一种二级 Cache 对访问请求的处理方法，利用三条并行的流水线：第一～第三流水线对访问请求进行并行处理，第一流水线负责处理一级数据 Cache 的读缺失请求、写缺失请求以及一级程序 Cache 的读缺失请求，第二流水线负责处理 EDMA 的读请求，第三流水线负责处理 EDMA 的写请求；为第一～第三流水线分别设置与其一一对应的第一～第三缓存单元，第一缓存单元用于缓存一级数据 Cache 的写缺失请求并将所缓存的多个行偏移地址相同的写缺失请求合并为一个，第二缓存单元用于缓存侦听到的数据并在 EDMA 发送数据的同时将侦听到的最新数据写回二级 Cache 的对应数据体中，第三缓存单元用于缓存 EDMA 的写请求并将所缓存的多个行偏移地址相同的写请求合并为一个。

[0007] 优选地，所述第一流水线为包括：地址解析、读取 Tag 体数据、Tag 值及状态位比较、读取数据体、数据发送这五级的五级流水线，第二流水线为包括侦听、侦听数据返回或读数据体、数据发送这三级的三级流水线，第三流水线为包括侦听、数据写回、EDMA 数据写入这三级的三级流水线。

[0008] 进一步地，第二流水线在处理 EDMA 的读请求时，根据读请求的地址对一级 Cache 进行数据侦听并将所侦听数据的行偏移地址暂存；对于之后的每一个读请求，首先将该读请求所请求的行偏移地址与暂存的行偏移地址进行比较，如果相同，则不对一级 Cache 进行数据侦听，如果不同，则对一级 Cache 进行数据侦听并将该读请求所请求的行偏移地址暂存。

[0009] 一种二级 Cache 对访问请求的处理系统，包括用于对访问请求进行并行处理的三条并行的流水线：第一～第三流水线，以及与第一～第三流水线一一对应设置的第一～第三缓存单元；第一流水线负责处理一级数据 Cache 的读缺失请求、写缺失请求以及一级程序 Cache 的读缺失请求，第二流水线负责处理 EDMA 的读请求，第三流水线负责处理 EDMA 的写请求；第一缓存单元用于缓存一级数据 Cache 的写缺失请求并将所缓存的多个行偏移地址相同的写缺失请求合并为一个，第二缓存单元用于缓存侦听到的数据并在 EDMA 发送数据的同时将侦听到的最新数据写回二级 Cache 的对应数据体中，第三缓存单元用于缓存 EDMA 的写请求并将所缓存的多个行偏移地址相同的写请求合并为一个。

[0010] 优选地，所述第一流水线为五级流水线，依次包括：地址解析模块、读取 Tag 体数据模块、Tag 值及状态位比较模块、读取数据体模块、数据发送模块；第二流水线为三级流水线，依次包括侦听模块、侦听数据返回或读数据体模块、数据发送模块；第三流水线为三级流水线，依次包括侦听模块、数据写回模块、EDMA 数据写入模块。

[0011] 进一步地，第二流水线的侦听模块中包括侦听偏移地址比较与暂存单元，用于暂存所侦听数据的行偏移地址并将新到来读请求所请求的行偏移地址与暂存的行偏移地址进行比较，如果相同，则侦听模块不对一级 Cache 进行数据侦听，如果不同，则侦听模块对一级 Cache 进行数据侦听并将新到来读请求所请求的行偏移地址暂存至侦听偏移地址比较与暂存单元。

[0012] 优选地，该系统还包括用于对同时到来的多个缺失请求进行处理优先级判断的优先级仲裁模块，处理优先级从高到低依次为：一级程序 Cache 的读缺失请求、一级数据 Cache 的写缺失请求、一级数据 Cache 的读缺失请求。

[0013] 相比现有技术，本发明具有以下有益效果：

本发明采用并行流水线处理多个请求，可以有效地提高整个存储系统的性能进而提高整个微处理器的处理速度；旁路机制及缓存单元的引入提高了整个流水线的性能；由于侦听访问耗费大量时间，本发明通过比较偏移地址来减少侦听次数提高了 EDMA 的读数据能力；写合并的引入进一步加快了对多个请求的处理能力，其结构简单、可移植性强，可在大多数微处理器芯片中都可应用。

附图说明

[0014] 图 1 为本发明二级 Cache 对访问请求的处理系统的总体系统框图；

图 2 为具体实施方式中写缓存 buffer 的暂存与合并原理示意图；

图 3 为具体实施方式中缺失请求流水线的结构框图；

图 4 为具体实施方式中缺失请求流水线的流水线处理流程图；

图 5 为具体实施方式中 EDMA 读请求流水线的结构框图；

图 6 为具体实施方式中 EDMA 写请求流水线的结构框图。

具体实施方式

[0015] 下面结合附图对本发明的技术方案进行详细说明：

本发明针对现有二级 Cache 访问请求技术的不足，对其进行了改进，利用三条并行的流水线对 L1D 读缺失请求、L1D 写缺失请求、L1P 读缺失请求、EDMA 读请求、EDMA 写请求这五种访问请求分别进行流水线处理，并利用缓存机制将多个写偏移地址相同的请求合并为一个以加速多个写请求的处理速度，通过比较偏移地址来减少侦听次数以提高 EDMA 的读数据能力。

[0016] 图 1 示出了本发明二级 Cache 对访问请求的处理系统的总体结构。

[0017] 如图 1 所示，本发明二级 Cache 对访问请求的处理系统包括：

写缓存 buffer，用于缓存多个写缺失请求，并将多个写偏移地址相同的请求合并为一个；

优先级仲裁模块，用于对同时访问的多个缺失请求进行优先级的判断；

缺失请求流水线，用于对 L1D 读、写缺失及 L1P 读缺失进行流水线处理；

数据体存储模块，用于存储所需要的 Tag 值、状态位及数据；

EDMA 读请求流水线，用于对 EDMA 的读请求进行处理；

EDMA 写请求流水线，用于对 EDMA 的写请求进行处理；

EDMA 写合并处理模块，用于将多个写偏移地址相同的 EDMA 写请求合并为一个请求。

[0018] 其中，写缓存 buffer 用于对连续到来的写缺失请求进行缓存，并将其中的写偏移地址相同的请求合并为一个。图 2 显示了写缓存 buffer 的暂存与合并的原理。如图 2 所示，写缓存 buffer 为 4 个 64bit 的缓存空间，当连续有 4 个写缺失请求(假设偏移地址分别为 100h、101h、100h、100h) 时，每个请求可暂存于相应的一个 64 位 buffer 中，若不采用写合并则写缓存 buffer 最多可以缓存 4 个写缺失请求，若采用写合并，则可以将偏移地址都为 100h 的写请求合并为一个写请求，因为偏移地址为 100h 的写请求都需写入二级 Cache 的同一行，因此不必分多次写入，通过写合并可一次写入。采用写合并策略的写缓存 buffer 最多可以缓存 16 个写缺失请求。

[0019] 优先级仲裁模块用于对同时到来的多个缺失请求进行优先级判断，其中 L1P 读缺失的优先级最高，L1D 的读缺失优先级最低，L1D 读缺失请求只有在写缓冲 buffer 为空时才可接收其请求。其具体处理过程如下：

步骤 1、接收多个缺失请求；

步骤 2、首先判断是否是 L1D 写缺失请求，若是则进行步骤 3，然后继续处理下一个请求，若不是则进行步骤 4；

步骤 3、将写缺失请求暂存于写缓存 buffer 中；

步骤 4、判断是否为 L1D 读缺失请求，若不是说明请求为 L1P 读缺失请求，则直接将其送入流水线处理，若是则进行步骤 5；

步骤 5、判断写缓存 buffer 是否为空，即判断写缺失请求是否执行完毕，若执行完毕则进行请求处理，若不为空则延迟 L1D 读缺失请求，直至写缓存 buffer 为空。

[0020] 缺失请求流水线用于对 L1D 读、写缺失及 L1P 读缺失进行流水线处理；其包括五级，若一个请求从第一级进入第二级之后，下一个请求便可进入第一级执行，而不必等到一

个请求全部执行结束再执行下一请求，缺失请求流水线的基本结构如图 3 所示，包括地址解析、读取 Tag 体数据、Tag 值及状态位比较、读取数据体数据、数据发送这五级模块。

[0021] 其中地址解析模块通过对请求地址的解析，判断出请求位于 Cache 中还是 SRAM 中还是外存中，每个存储体都对应一段特定的存储地址，通过请求地址便可判断出位于哪一存储体；

读取 Tag 体数据模块根据请求中的行索引值从 Tag 存储体中将 Tag 值及状态标志位读出；

Tag 值及状态位比较模块将读出的 Tag 值与请求的 Tag 进行比较，判断是否命中，若命中判断状态标志位是否有效，有效则可从 Cache 体中读取数据；

读取数据体模块负责从相应的数据存储空间读取数据，其中包括三种途径，从 Cache 体中读取、从 SRAM 中读取、通过 EDMA 读取；

数据发送模块用于完成将读取的数据发送给 L1 的操作；

缺失请求流水线对缺失请求的处理过程如图 4 所示，具体如下：

步骤 1、地址解析模块对请求的地址进行解析，可请求的数据可能位于三个部分，包括 L2 的 Cache 部分、SRAM 部分及外存部分；

步骤 2、地址解析模块判断地址中的行索引值是否属于 Cache 空间，若不属于则进行步骤 3，若属于则进行步骤 4；

步骤 3、地址解析模块判断地址中的行索引值是否属于 SRAM 空间，若不属于则进行步骤 8，若属于则进行步骤 9；

步骤 4、读取 Tag 体数据模块根据行索引值从 Tag 存储体中读取 Tag 值及相应的状态标志位；

步骤 5、Tag 值及状态位比较模块将请求的 Tag 值与从 Tag 存储体中读取的 Tag 值进行比较，若相同则进行步骤 6，若不同则进行步骤 8；

步骤 6、Tag 值及状态位比较模块判断状态标志位是否有效，有效则进行步骤 7，无效则进行步骤 8；

步骤 7、读取数据体模块从命中的 Cache 体中读出命中的数据；

步骤 8、读取数据体模块通过 EDMA 从外存中读取所需要的数据；

步骤 9、读取数据体模块从片内 SRAM 中读取相应的数据；

步骤 10、数据发送模块将读取的数据发送给请求方。

[0022] 数据体存储模块用于存储相应的数据信息，其包括 Tag 存储体、Cache 存储体、SRAM 存储体。其中 Tag 存储体用来存储 Cache 行对应的 Tag 值及状态标志位。

[0023] EDMA 读请求流水线用于对 EDMA 的猝发读请求进行流水线处理。图 5 为 EDMA 读请求流水线的结构框图，如图 5 所示，该流水线共分为 3 级流水线，包括侦听模块、侦听数据返回或读数据体模块、数据发送模块，其中侦听模块中包括侦听偏移地址比较与暂存单元。该流水线对读请求处理的过程如下：

侦听模块可一次性接收 EDMA 猝发传来的 8 个读请求，首先对第一个请求进行 L1 的数据侦听，侦听偏移地址比较与暂存单元将其侦听地址进行暂存，然后读取 L1 对应行的 dirty 位，若 dirty 位为 1 表示此行数据被修改过，需要写回最新数据，若 dirty 位为 0 表示此行数据未被修改过，则无需写回。对于第 2~8 个请求，侦听偏移地址比较与暂存单元首先

将此次请求的偏移地址与暂存列表中的偏移地址进行比较，若偏移地址相同则不需要再次进行侦听。当 8 次请求处理结束需对偏移地址暂存列表清空。例如，8 次请求的偏移地址分别为 100h、100h、101h、102h、101h、103h、100h、104h，对第一次请求将侦听 L1 中偏移地址为 100h 的行，并将偏移地址 100h 暂存，第二次请求与暂存表中的暂存地址相同则不再进行侦听，第三次请求的偏移地址为 101h，与暂存表中不同，则需侦听，同时将 101h 暂存于暂存表中，以此类推，当第 8 次请求执行结束，将暂存表全部清零。

[0024] 若执行了侦听操作并且 dirty 位为 1，则第二级的侦听数据返回或读数据体模块读取 L1 中的最新数据，将最新的数据从 L1 数据体中读出。读出后进入流水线第三级的数据发送模块，将 L1 返回的数据写入数据发送缓存，执行数据发送的同时将 L1 的最新数据写回 L2 数据体。

[0025] 若无需进行侦听操作或者 dirty 位为 0，则进行第二级流水线中的读取 L2 数据体，从 L2 的数据体中读取对应行中的数据。读取数据后发送到第三级流水线中的数据缓存，然后进行数据发送。

[0026] EDMA 写请求流水线用于对 EDMA 猝发的多个写请求进行流水线处理。其结构如图 6 所示，共分为三级 ：

侦听模块，根据请求地址对对应的 L1 行进行数据侦听，若 dirty 位为 1 则进入第二级流水线，若 dirty 位为 0 则进入第三级流水线 ；

数据写回模块，将侦听的 L1 最新数据写回到 L2 对应的数据体中 ；

EDMA 数据写入模块，将写请求的数据写入 L2，若对应的 L2 行中已有数据，需先将已有数据写回到外存 ；

写合并模块，将多个 EDMA 写偏移地址相同的请求合并为一个请求，其合并原理与写缓存 buffer 的合并原理相同。

[0027] 本发明提供了一种二级 Cache 对访问请求的流水线处理方法，采用三条流水线处理 5 类请求，流水线机制能有效提高了处理性能，写缓存和写合并机制的引入能进一步加强对请求的处理能力，本发明还采用减少侦听次数的方法提高 EDMA 的读请求处理时间，以上方法还可用于到 EDMA 的写请求、二级 Cache 的行替换，数据一致性维护等多个方面。

[0028] 本发明提供了一种二级 Cache 对访问请求的流水线处理系统，该系统可以应用大部分微处理器芯片上，可提高整个存储系统的性能，对提高微处理器芯片的处理能力意义重大。

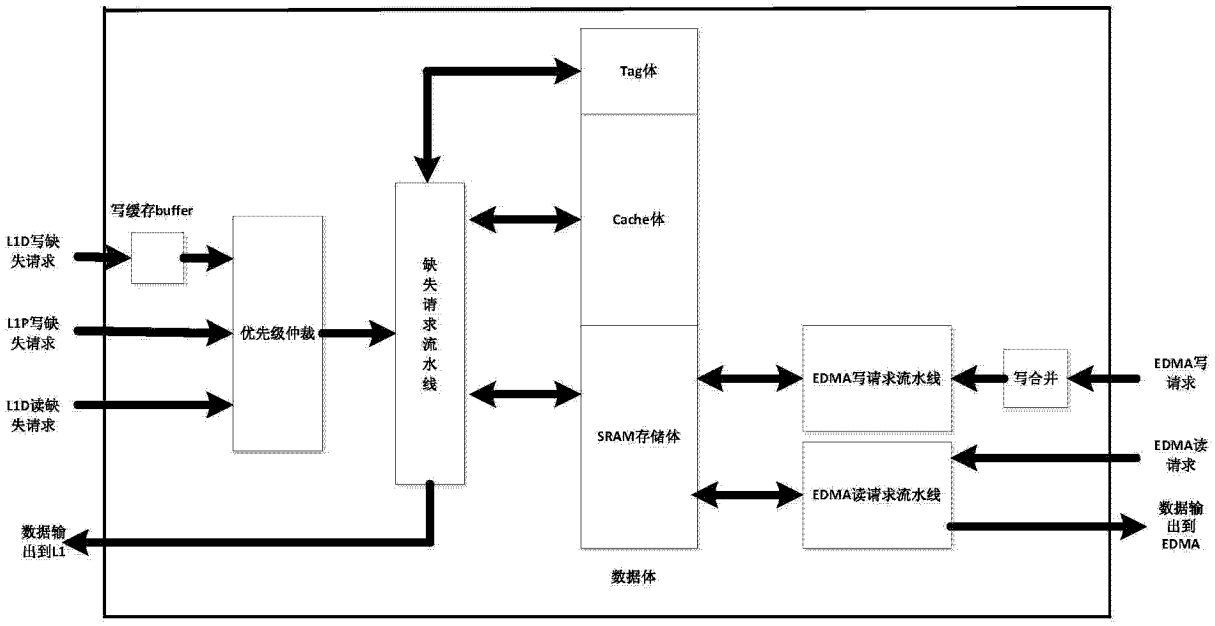

Tag体
Cache体
SRAM存储体
数据体
写缓存buffer
L1D写缺失请求
L1P写缺失请求
L1D读缺失请求
优先级仲裁
缺失请求流水线
数据输出到L1
EDMA写请求流水线
写合并
EDMA写请求
EDMA读请求
EDMA读请求流水线
数据输出到EDMA

图 1

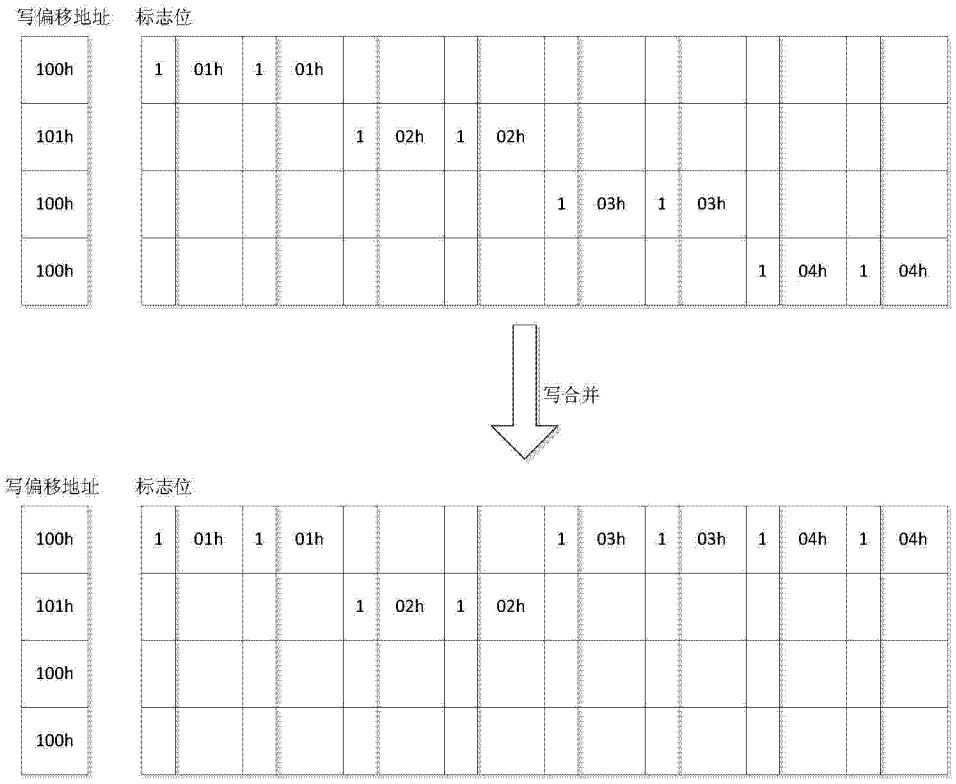

写偏移地址
标志位
100h 1 01h 1 01h
101h 1 02h 1 02h
100h 1 03h 1 03h
100h 1 04h 1 04h
写合并
写偏移地址
标志位
100h 1 01h 1 01h 1 03h 1 03h 1 04h 1 04h
101h 1 02h 1 02h
100h
100h

图 2

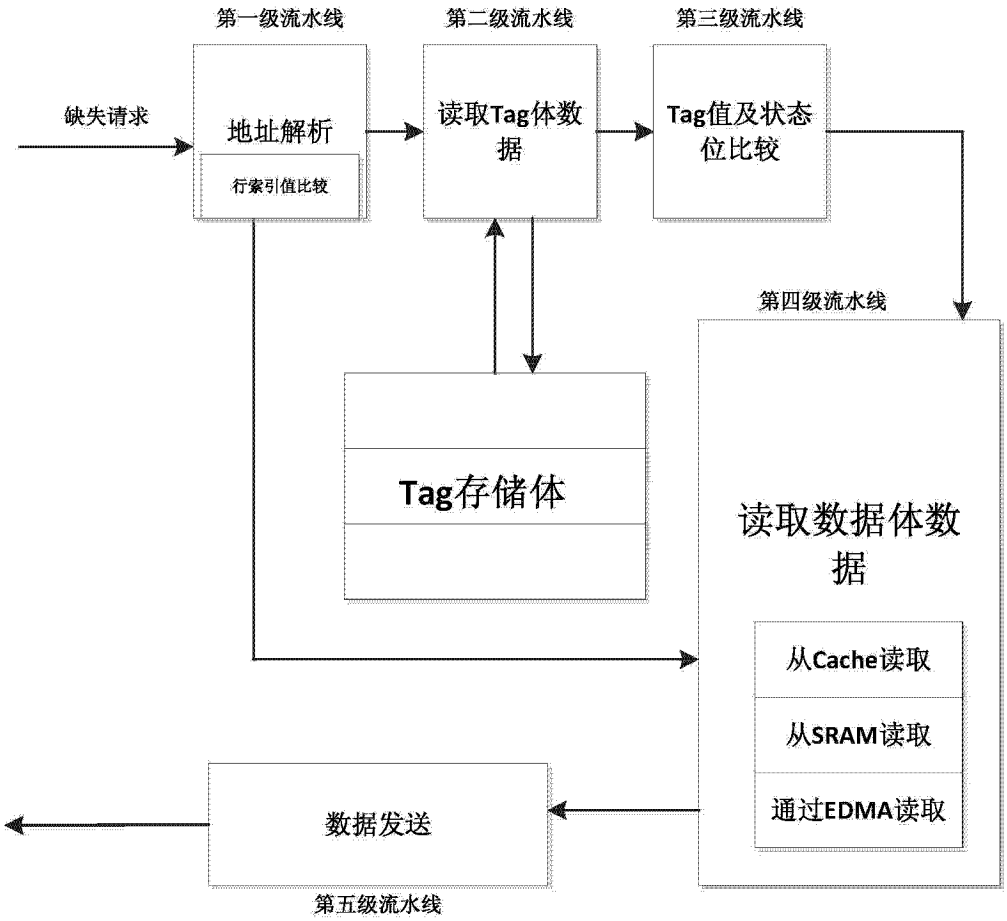

第一级流水线
第二级流水线
第三级流水线
缺失请求
地址解析
行索引值比较
读取Tag体数据
Tag值及状态位比较
第四级流水线
Tag存储体
读取数据体数据
从Cache读取
从SRAM读取
通过EDMA读取
数据发送
第五级流水线

图 3

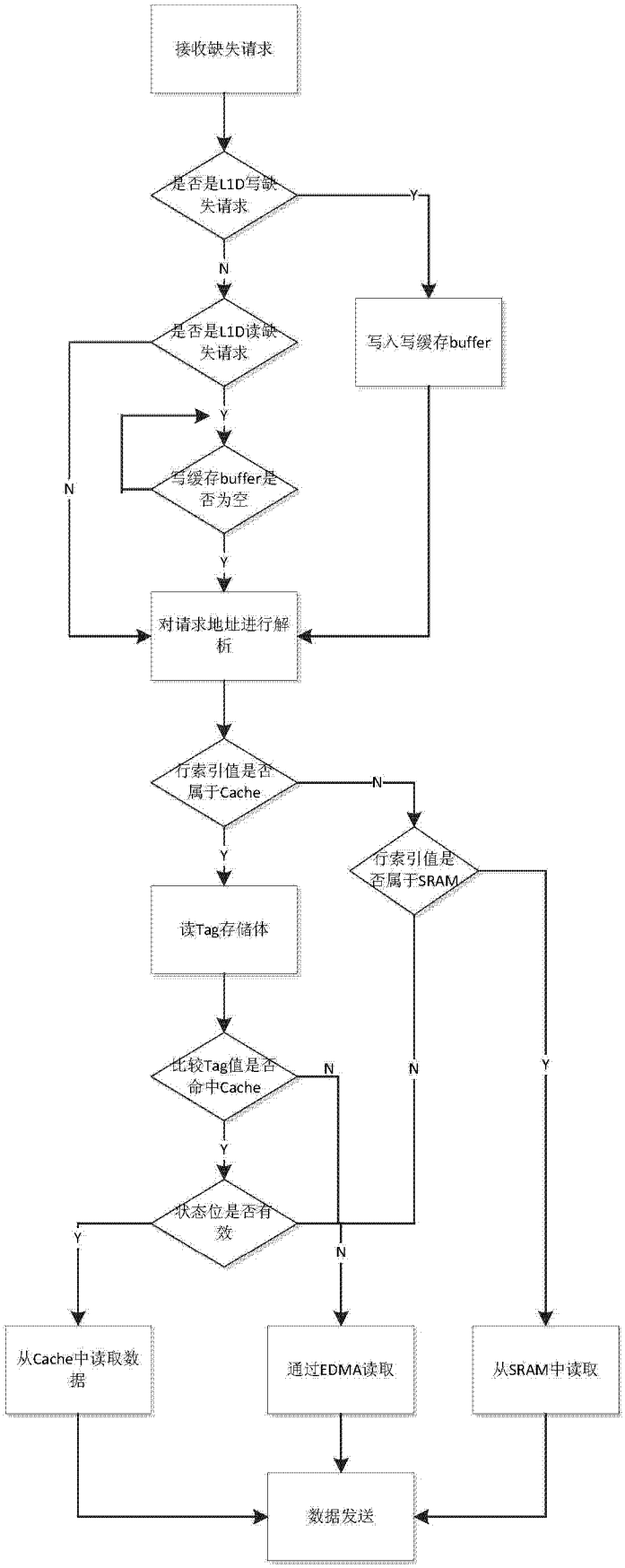

接收缺失请求
是否是L1D写缺失请求
Y
N
是否是L1D读缺失请求
写入写缓存buffer
Y
写缓存buffer是否为空
N
Y
对请求地址进行解析
行索引值是否属于Cache
N
Y
行索引值是否属于SRAM
读Tag存储体
比较Tag值是否命中Cache
N
N
Y
Y
状态位是否有效
Y
N
从Cache中读取数据
通过EDMA读取
从SRAM中读取
数据发送

图 4

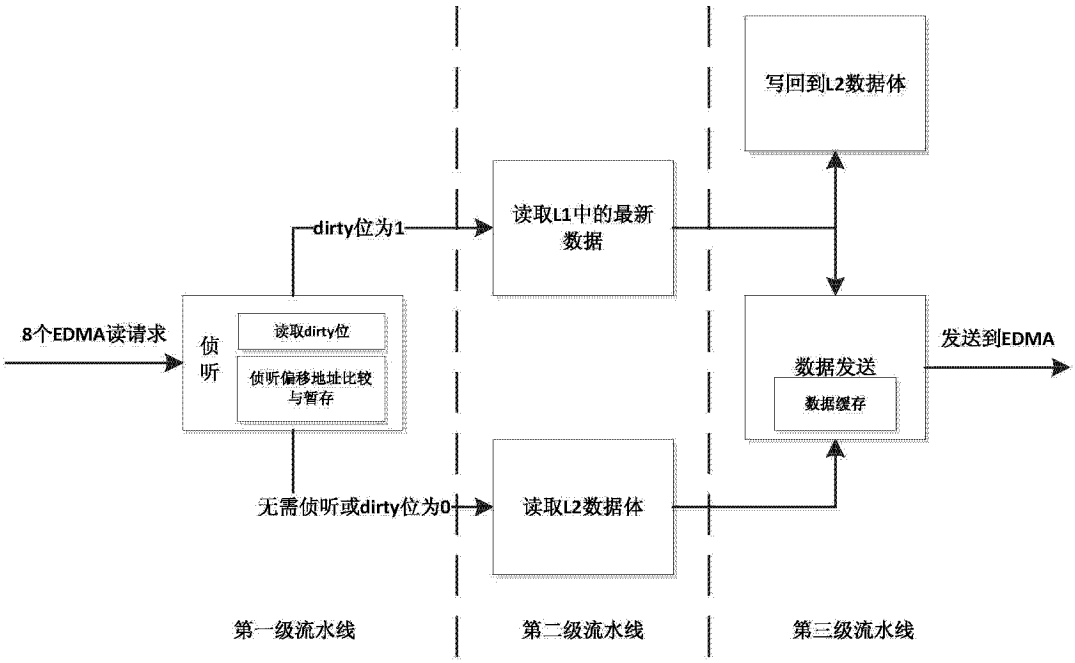

写回到L2数据体
读取L1中的最新数据
dirty位为1
8个EDMA读请求
侦听
读取dirty位
侦听偏移地址比较与暂存
数据发送
数据缓存
发送到EDMA
无需侦听或dirty位为0
读取L2数据体
第一级流水线
第二级流水线
第三级流水线

图 5

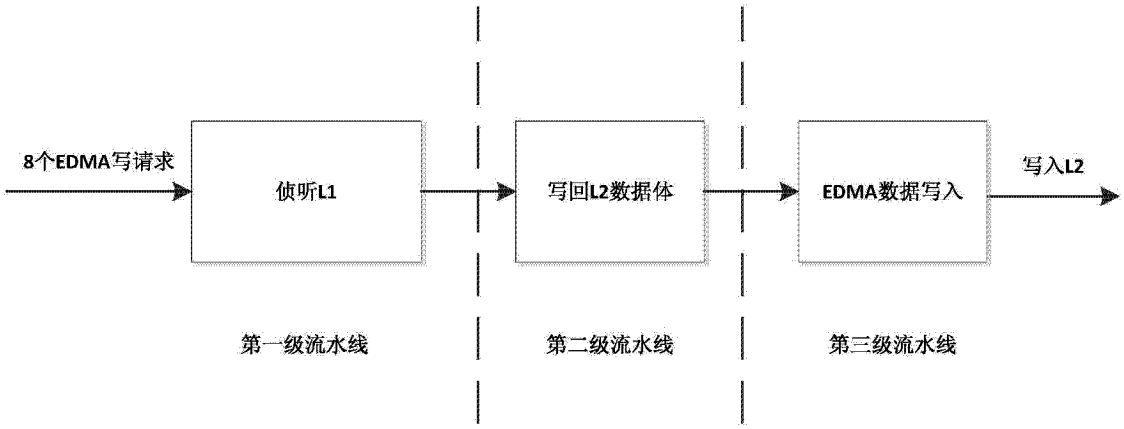

8个EDMA写请求
侦听L1
写回L2数据体
EDMA数据写入
写入L2
第一级流水线
第二级流水线
第三级流水线

图 6