第 35 卷 第 4 期 计 算 机 工 程 2009 年 2 月
Vol.35 No.4 Computer Engineering February 2009
·工程应用技术与实现· 文章编号：1000—3428(2009)04—0238—03 文献标识码：A 中图分类号：TP333

多核处理器片上 Cache 的选择性复制策略

郭惠芳，姜鲲鹏，赵荣彩，姚 远
(解放军信息工程大学计算机科学与技术系，郑州 450002)

摘 要：在统计分析 8 个典型测试程序的模拟运行的基础上，提出在多核处理器的私有 L1 上对部分只读共享数据进行复制以加快访问速度，对读-写共享数据采用“原地通信”策略，以减少一致性开销，针对其中的“写无效化共享标记”问题提出一种更简单的解决方法。从模拟实验的统计分析可知，采用这 2 种策略可获得比传统策略更小的平均访问时延和更高的空间利用率。

关键词：对称多处理器；选择性复制；高速缓冲

Selective Replication Policy in CMP Cache Hierarchy

GUO Hui-fang, JIANG Kun-peng, ZHAO Rong-cai, YIAO Yuan
(Department of Computer Science and Engineering, PLA Information Engineering University, Zhengzhou 450002)

【Abstract】 Based on the statistical analysis of eight benchmarks simulation, this paper proposes selective replication for some read-only shared data, and in-situ communication for write-read shared data. The former means to improve cache average access speed, while the latter's purpose is to reduce coherence overhead. It extends a more simple method to solve the ping-pong problem. Statistical analysis of simulation shows these two schemes achieve less access latency and better space utilization compared with conventional cache scheme.

【Key words】 Symmetric Multiprocessor(SMP); selective replication; cache

1 概述

多核处理器(CMP)逐渐成为处理器发展的主流，它的性能优劣很大程度依赖于其片上存储层次及共享策略的设计，当今典型商业多核处理器采用的是私有 L1 Cache 和共享 L2 Cache。私有 L1 主要提供每个处理器核高速的存储访问，而共享的 L2 主要提供较大的片上存储空间，这种结构类似于对称多处理器(Symmetric Multiprocessor, SMP)的结构，但 CMP 与 SMP 的最大区别是核间通信开销远远小于处理器之间的通信开销，核间的数据传送比从共享的 L2 Cache 中取数快。另一方面，CMP 这种特殊的结构也使解决读-写共享数据在各私有 L1 Cache 中常出现的“乒乓”效应成为可能。本文介绍多核处理器上可行的存储结构，论述采用选择性复制策略的原因，给出了选择性复制策略的具体实现方法。

2 多核处理器的存储结构

2.1 私有 Cache 的多核处理器

为每个处理核分配一个 2 层或 3 层的私有 Cache 模块，与 Intel 双核处理器 Montecito 的架构一样，每个处理器核访问私有的 L2 Cache 可以不用通过共享的片上网络，并且访问 L1 与查询 L2 标记表可以同步进行。即使 L1 访问失效，替换也只涉及自己私有的 L2 模块，不需定位远程 L2 模块。这种设计对数据访问局部性较好的应用是有利的，但由于不加限制地复制存储块，因此片上存储空间的利用率不高，数据的频繁读-写共享会导致“乒乓”效应[1]。

2.2 共享 L2 Cache 的多核处理器

L1 Cache 每核私有，L2 Cache 为片上所有处理器核共享，但一般保证每个核有一个与它接近的 Cache 模块；而核访问不同的 Cache 模块的时延是与距离相关的，距离越近，时延越小。这种结构静态地将地址映射至不同的 Cache 模块上形成了一个非均匀访问(NUCA)[2]的共享 L2 Cache，现在大部分小规模多处理器都是这种结构。L2 Cache 的空间利用率较高，但 L1 之间仍然可能存在读-写共享时的“乒乓”效应，L2 数据的平均访问时延较长。

2.3 选择性复制的共享 Cache 多核处理器

在私有和共享策略之间考虑增加受控制的复制策略，可以提高片上存储空间的利用率，加快对部分只读共享数据块的访问速度。在不引起过多一致性代价的前提下尽量缩短数据访问的时延，使经常使用的只读共享数据尽量放在离访问者最近的 Cache 模块上，避免对反复读、写共享的数据块进行复制，这会带来沉重的一致性代价。对读-写共享的数据尽量保持一个数据块备份，访问时增加适当的远程写，与反复的作废和调入相比，代价较小。

3 采用选择性复制的依据

SMP 的多个处理器由片外总线连接，其延迟相对较长，如果一个处理器上的 Cache 不命中，就去访问片外的存储器，而不会通过访问另一个处理器上的 Cache 得到这批数据(特殊的协议除外)。现在处理器核之间的通信代价大大降低，核间 Cache 的直接访问或迁移成为可能，而且 SMP 中每个处理器中的 Cache 都只为一个处理器服务，存储空间的压力不大，而多核处理器的片上 Cache 空间需同时支持多个核的访存，其片上存储空间成为紧缺资源。

完全的共享能提供较大的访存空间，但空间的增加带来了访问时延的增加。完全的私有能提供最快的直接访问速度，

作者简介：郭惠芳(1970－)，女，博士研究生，主研方向：多核处理器体系结构，编译技术；姜鲲鹏，讲师；赵荣彩，教授、博士生导师；姚 远，副教授

收稿日期：2008-07-27 **E-mail：**too_ghf@163.com

但多个核间的多个备份使存储空间利用率降低，所以，多核处理器需要在最快访存速度和有效空间利用率间寻找一个折中，使其性能达到一个最佳点。纯粹的每核私有或多核共享的 Cache 策略都无法同时满足这两方面的要求，所以，选择性的复制是一种较好的中间策略。

下文通过对测试程序的访问空间及访问次数进行统计分析来决定是采用复制还是共享的策略[3]。

3.1 模拟方法

设模拟一个 8 个核的处理器，存储结构采用一个 16 MB 单模块、16 路组相联的共享 Cache，在其上分别执行商业和科学计算负载，模拟器使用 Wisconsin 大学的 GEMS，它对 Simics 进行了扩展，提供了 CMP 存储层次的时间模型。包括片内及片外的网络及其一致性协议。测试集采用典型的商业计算应用：apache, jbb, oltp, zeus 及科学计算方面 SPECOMP 中的 4 个程序：Apsi, Art, Galgel 和 Mgrid。

3.2 模拟运行结果

一般将 CMP 片上数据块的共享分为 3 种类型[3]：(1)只有一个处理器核访问；(2)有多个处理器核以只读方式共享访问；(3)有多个处理器以可读可写的方式共享访问，并且至少有一个写。表 1 是对 8 个测试程序运行的一个统计。

表 1 L2 Cache 的空间利用分析[3]

测试程序		单独请求		只读共享		读-写共享			
		请求比率/(%)	空间比率(%)	请求比率/(%)	空间比率/(%)	平均共享数	请求比率/(%)	空间比率/(%)	平均共享数
商业应用	Apache	11	55	48	17	3.7	41	28	3.0
	Jbb	55	91	44	10	3.5	1	<1	2.4
	Oltp	4	52	72	20	4.3	24	28	3.8
	Zeus	17	76	59	9	3.0	24	15	2.3
科学计算	Apsi	99	>99	<1	<1	7.3	<1	<1	2.8
	Art	49	62	51	38	3.0	<1	<1	2.7
	Galgel	<1	84	<1	<1	4.0	99	16	5.3
	Mgrid	96	98	4	2	2.3	<1	<1	2.2

如表 1 所示，在 4 个商业应用中有 44%~72%的请求是只读共享请求，而它们所占的空间比例是 9%~20%，读-写共享数据请求占 24%~41%。而对于科学计算测试程序，只有 Art 的只读请求占了 51%，其余都在 4%以下；只有 Galgel 对读-写共享块的请求占了 99%，其余都小于等于 1%。只读共享请求的空间局部性很好，如果这部分请求块全部复制，空间开销将达到 20%~100%，也是不能承受的，如果对这部分数据采用选择性复制[4]，一方面可以有选择地对那些常访问的存储块进行备份，使常访问的数据块能就近访问，另一方面由于多个备份是只读共享，因此没有写导致的一致性代价。对于读-写共享块，由于每次写一个备份都会造成其他备份的无效，当其他处理器核随后读时会发生失效；且反复的写、读会引发 Cache 的“乒乓”效应，大大影响其性能，因此对读-写共享块采用不复制策略比较适合。

4 多核 Cache 共享策略的实现

通过对典型应用的分析可知，针对不同类型的共享数据块采取选择性复制策略有利于提高访问速度和空间利用率。

4.1 L2 上选择性复制策略的实现

如果在 L2 上采用上述策略，那么结构如下：L1 采用每核私有，以保证每个核的直接访问速度；L2 采用选择性复制的多模块结构。

对于只读共享的数据来说，在每核 L1 和 L2 上都存在同一数据块的备份，设置多个备份的目的是提高经常访问的速度，而在 L1 上的多个备份已经可以起到这个作用，那么在 L2 上的多个备份只有在发生 L1 失效时才可能有用，这个失效率只有 2%~5%，所以，2 个层次上的重复备份是不必要的，它对速度的提高影响并不大，但对整个片上 Cache 空间利用率影响较大，空间利用率下降直接导致片外访存的次数增加。

从以上分析可看出，在 L2 层采用选择性复制有可能得不偿失。

4.2 L1 上选择性复制策略的实现

在 L1 上采用选择性复制策略的优势就是在保证 L1 访问速度的基础上提供一定的远程共享，减少多个备份的一致性开销；而 L2 着重保证片上容量的最大化，所以，在 L2 级每个存储块都只有一个备份，当发生一次 L1 失效后，它会将数据从 L2 调入 L1 中，在后续的访问中就可在 L1 中得到，保证大部分的访存速度是快的。由于 L1 是每个核直接访问的，它必须提供较快的访问速度，因此不能采用过于复杂的策略。

4.2.1 存储结构

如图 1 所示，假设芯片内部 4 个处理器核对应 4 个 L1 模块，L1 模块之间通过快速的片上网络和 L1 控制器连接，以提供对一部分远程访问控制及一致性维护的支持。而 L2 采取多模块非均匀访问的共享模式。每个核有一个小型的私有标记表，记录该处理器核最近使用的存储块的位置及状态，在处理器核与多个 L1 模块之间的总线上采用一个简单的总线侦听策略。

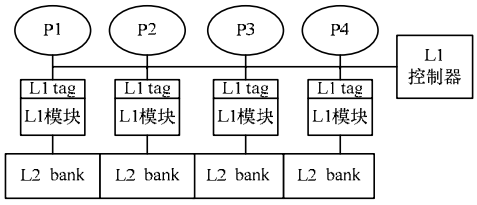

图 1 4 核处理器结构示意图

4.2.2 选择性复制策略

私有的 L1 策略采用的是不加限制地复制数据块，每个核根据自己的需要备份存储块，其空间利用率不高。另一方面，多个备份在读写共享时一致性开销增加。选择性复制对备份进行控制，对只读共享的数据才能进行复制，对读-写共享的数据只在片上放置一个备份。下面首先讨论只读数据的“选择性复制”[4-5]，然后讨论对读-写共享数据进行“原地通信”(In-Situ Communication, ISC)的策略[5]。

每个私有标记表中有一项指向对应的数据块，当读一个数据块失效，如果片上另一个 L1 中有该数据块，则这次读可以从远程的 L1 中获得，并在本地的标记表中增加一项指向远程的存储块，即多个私有的标记项可指向同一个数据块。因为在一些商业应用中，许多数据块被调入 Cache 后再没有被重用，所以在第 1 次读访问时，并不拷贝数据块到本地，而只是备份一个标记项。如果这个处理器核第 2 次使用该数据块，这个数据块才被调入本地 L1，以备将来再次使用，并将本地标记项指向自身的数据块。由于本地已存在该数据块的标记项，因此第 2 次访问不会发生失效，时间开销会更小。

由于一个数据块可能对应多个私有标记表项，因此如果发生数据块被替换，问题就可能出现：数据块已被替换，但另一处理器核的标记表中还可能存在指向这个数据块的标记项，这时读写该块将发生错误。为了解决这个问题，可在替换一个共享数据块之前发送一个总线替换(BusRepl)事务，使所有指向该数据块的标记项作废。

4.2.3 原地通信策略

如果采用私有的作废式 L1 策略，每次的写都将使其他核的读备份作废，而后续的读就不会命中，又需从写的 L1 中获取该数据块，放在本地 L1 中，但后续的写又将使其无效……。如此反复既浪费了空间又造成了较长的读时延和一致性开销。

在多核处理器中可用“原地通信”方式，即对读－写共享的数据块在片上只提供一个备份，避免发生反复的一致性开销。读和写的核各有一个标记项指向同一个数据块，一般一个写后会跟多个读，所以，可使该数据块离读的核近一些。

在 MESI 作废协议[1]中，一旦发生写，就会使其余 L1 中的备份无效，如果在多核处理器的 L1 上使用 MESI 协议，每次写将会使其他指向数据块的标记作废，导致“原地通信”策略失败，解决这一问题的方案是使用文献[5]提出的增加一个新的“C 状态”。下面提出另一种更简单的解决方法。

MESI 回写作废式协议在总线上有一个附加的“共享”信号(S)，在发生总线读事务时缓存控制器用它确定是否有其他缓存持有该数据，从而决定是将装入的存储块置为“E”还是“S”状态。现在为了支持多个标记共享一个已修改的数据块，在总线上增加一个类似的“脏”信号(D)，用它确定是否有其他缓存当前持有同一个已修改的数据块，当写不命中时，可以通过这个信号了解当前是否在其他缓存中存在该数据块并处于“W”状态。这对于解决“写作废”的“乒乓”效应有重要的作用。

当发生写不命中时，通过总线上的信号“S”和“D”可知在其他缓存中是否存在该数据块以及它处于“W”还是“S”状态，如果“S”信号有效，即存在至少一个处于“S”状态的数据块，根据前面讨论的选择性复制策略，可能还存在多个私有标记表指向同一个数据块的情况，这时，采取与传统方案一致的方法，将数据块调入本地缓存，并使其余标记及数据备份作废；如果在发生写不命中时“D”信号有效，即该块在其他缓存中处于“W”状态，则这种状态的数据块在 L1 中只有一个备份，其余共享者都只有标记项指向它。这时，控制器就可以采取与传统不一样的策略：不在本地调入该块，而是远程写入，并在本地的标记表中加入一项，指向远程的数据块，对其他共享者无影响，保证了多核中读-写共享数据的高效性。

5 结束语

多核处理器已广泛进入市场，但其发展空间还很大，如何使片上的存储层次结构提供足够的运行空间和运行速度还需深入的研究。本文仅对最近提出的“选择性复制”或“受控复制”策略进行了讨论，并针对多核处理器讨论了该策略的实现细节，对 MESI 协议进行了补充，在私有 L1 策略基础上加入对部分只读共享数据的复制以加快访问速度，同时最大化其片上的空间利用率以提高命中率；而对读-写共享数据采用“原地通信”策略，并针对“写无效化共享标记”问题提出了一种新的解决方法，减少了访问读-写共享数据常发生的“乒乓”效应，提高了平均访问速度，同时又节约了存储空间。

参考文献

[1] Culler D E, Singh J P, Gupta A. 并行计算机体系结构[M]. 2 版. 李晓明, 译. 北京: 机械工业出版社, 2002.

[2] Kim Changkyu, Burger D, Keckler S W. An Adaptive, Non Uniform Cache Structure for Wire Delay Dominated on Chip Caches[C]// Proc. of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems. San Jose, California, USA: [s. n.], 2002: 211-222.

[3] Beckmann B, Marty M, Wood D. Balancing Capacity and Latency in CMP Caches[Z]. (2006-02-02). http://www.cs.wisc.edu.

[4] Garg R, El-Moursy A, Dwarkadas S, et al. Cache Design Options for a Clustered Multithreaded Architecture[Z]. (2005-05-05). ftp://ftp.cs.rochester.edu.

[5] Chishti Z, Powell M, Vijaykumar T. Optimizing Replication, Communication, and Capacity Allocation in CMPs[C]//Proceedings of the 32nd Annual International Symposium on Computer Architecture. Madison, Wisconsin, USA: [s. n.], 2005: 357-368.

(上接第 237 页)

图 4 显示在视频码率不变的情况下，视频缓存随网络带宽的变化情况，从图中可以看出，网络带宽越大，视频缓存越小，从而保证延时越小。而如果网络带宽越小，视频缓存越大，可以保证视频数据不丢失，从而保证视频质量。图 5 显示在网络带宽不变的情况下，缓存随视频码率的变化情况。从中可以看出，缓存分配算法对视频码率变化的跟随性好，在视频帧突然增大时可以保证数据不丢失。

7 结束语

本文针对目前嵌入式系统资源有限的情况，提出了提高视频监控系统实时性的方案。此方案结合了目前的 VBR 视频流量预测和网络带宽预测算法。本文视频流量预测算法采用了实时性好、算法简单的线性规划视频预测模型。网络带宽检测算法采用了数据包对算法，此算法简单实用，适合本方案。提出平衡网络带宽和视频流量的视频缓存分配算法，此算法计算简单，对网络带宽和视频流量变化的跟随性好，适合在嵌入式视频监控系统中采用。本文提出的视频监控系统实时性方案可以保证较短的视频传输延迟，并且节约系统内存资源。

参考文献

[1] 周 健, 戴梅萼, 余震建, 等. 远程实时视频传输的自适应技术[J]. 清华大学学报: 自然科学版, 2004, 44(7): 966-968, 973.

[2] 张占军. 无线多媒体网络中端到端自适应 QoS 保证[J]. 计算机学报, 2004, 27(8): 1064-1073.

[3] 刘亚伟, 卢燕飞, 冯玉珉. 宽带 IP 网中 VBR 视频流量的建模与预测[J]. 铁道学报, 2004, 26(6): 55-61.

[4] 刘晓颖, 戴琼海, 刘晓东. 智能集成 VBR MPEG 视频流量预测模型[J]. 电子学报, 2006, 34(5): 383-386.

[5] Lv Jun, Li Xing, Ran Congsen, et al. Network Traffic Prediction and Fault Detection Based on Adaptive Linear Model[J]//Proceedings of the 2004 IEEE International Conference on Industrial Technology. [S. l.]: IEEE Press, 2004: 880-885.

[6] Wang Yaqin, Chen Yue, Qin Minggui, et al. Dynamic Traffic Prediction Based on Traffic Flow Mining[C]//Proceedings of the World Congress on Intelligent Control and Automation. Dalian, China: [s. n.], 2006: 6078-6081.