文章编号:1001-2486(2011)02-0145-05

一种新型的 Free-memory 众核处理器片上通信接口*

郭御风,李 琼,窦 强,张 磊,刘 路
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:高性能计算机系统越来越多采用集群系统,集群系统的性能极大地依赖于通信接口。基于片外SRAM保存地址变换表的用户级通信方法,极大地增加了芯片和系统的设计复杂度和成本。在传统基于I/O总线的HCA基础上,提出并实现了一种新型的Free-Memory的众核处理器片上通信接口,去掉了本地存储器接口,通过高效的cache管理策略降低地址变换开销。测试结果表明我们实现的通信接口在降低了芯片和系统实现复杂度和成本的同时,还获得了比Infiniband的QDR HCA更好的通信带宽和延迟。

关键词:众核处理器;通信接口;用户级通信;Infiniband;地址变换表;无本地存储器
中图分类号:TP301 **文献标识码**:A

A Novel Free-memory Communication Interface of Many Cores Processor

GUO Yu-feng, LI Qiong, DOU Qiang, ZHANG Lei, LIU Lu
(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: High-performance computing has witnessed a tremendous growth and acceptance over the last decade, primarily due to the availability of clusters. The performance of these clusters hinges upon the communication interface. User level communication based on storing address translation table in off-chip SRAM has deeply increased the design complex and cost of chip and system. The paper put forward and implemented a novel Free-Memory communication interface of many cores processor which differs from traditional HCA based on I/O bus, without local memory interface and reduced cost of address translation by efficient cache management method. Experimental results show the communication interface which we implemented not only can reduces the design complex and cost of chip and system, but also can achieve better bandwidth and latency than infiniband QDR HCA.

Key words: many cores processor; communication interface; user level communication; Infiniband; address translation table; free memory

在过去的几年内,通过提升主频来提高处理器性能变得越来越困难,能获得的性能提升也越来越有限[1],提高执行并行度成为提高处理器性能的主要手段。多核结构将传统的复杂的多流出结构精简成分布式的多个处理器核,利用多核并行处理来实现高计算性能和高吞吐率。多核处理器的吞吐率性能一直在持续攀升,维持了摩尔定律[2]的发展。

高性能微处理器作为现代高端计算机系统的核心和引擎,被广泛地应用到高性能大规模集群系统中。每年Top500机器中,集群系统所占的比例越来越大[3]。众核处理器集成了越来越多的高性能处理器核,多个处理器核共享同一个通信接口,并对通信接口带宽和通信延时提出了更高的需求。

当前高性能商用互联系统主要有Myrinet[4]、Quadrics QsNet[5]、Infiniband[6]等。目前基于Infiniband的通信接口卡被广泛应用,Mellanox[14]、Qlogic[12]和Myricom[13]三家都有各自的HCA卡。它们都以HCA接口卡的形式连到服务器的PCI Express总线或PCI总线。虽然随着I/O总线性能的不断提升,通信接口的延迟和带宽都得到了很大的改善,但I/O总线仍是制约通信接口性能发挥的关键因素之一。

用户级通信把硬件通信资源虚拟化,允许用户进程直接访问通信接口,减少了通信操作的软件层开销。通信接口通常采用基于地址变换表的虚实地址转换机制实现用户进程间的直接通信,

* 收稿日期:2010-09-09
基金项目:核高基重大专项——面向科学计算的多核CPU资助项目(2009ZX01028-002-002)
作者简介:郭御风(1979—),男,讲师。

由通信接口硬件完成虚实地址转换,地址转换通过查询地址转换表实现,系统中必须维护一个地址转换表。目前主流的通信接口通常采用外带一个 SRAM 用于保存地址转换表。由于地址转换表保存在通信接口的本地存储器中,这样带来的好处是查表无需访问主存,而是直接从本地存储器中读取,大大减少了地址转换的延迟,却对芯片和系统的设计带来了很大的复杂性和困难:

1)存储器接口大量的引腿,提高了芯片封装设计的复杂度和成本,增大了芯片的封装大小;

2)存储器接口通常需要多种特殊电压[7],增大了芯片和 PCB 板电源系统的设计复杂度;

3)存储器接口设计比较复杂,逻辑量也比较大,将会大大增大芯片的面积和功耗;

4)PCB 板设计要相应地增加存储颗粒或者 DIMM 槽,造成外围电路设计很繁杂;

5)外接本地存储器将会大大增加通信接口的设计成本和使用成本。

Mellanox 的 InfiniHost Ⅲ Ex[15] 实现了 Memfree 的通信方式,去掉了本地存储器接口,降低了芯片的设计复杂度,但仍然采用 I/O 接口卡的形式,通信延迟较长。

1 众核处理器片上通信接口结构

图 1 为片上高性能通信接口 PNI 的硬件结构图,PNI 直接集成到众核处理器芯片 HMCP 片内。下面我们对 PNI 的主要功能模块进行介绍。

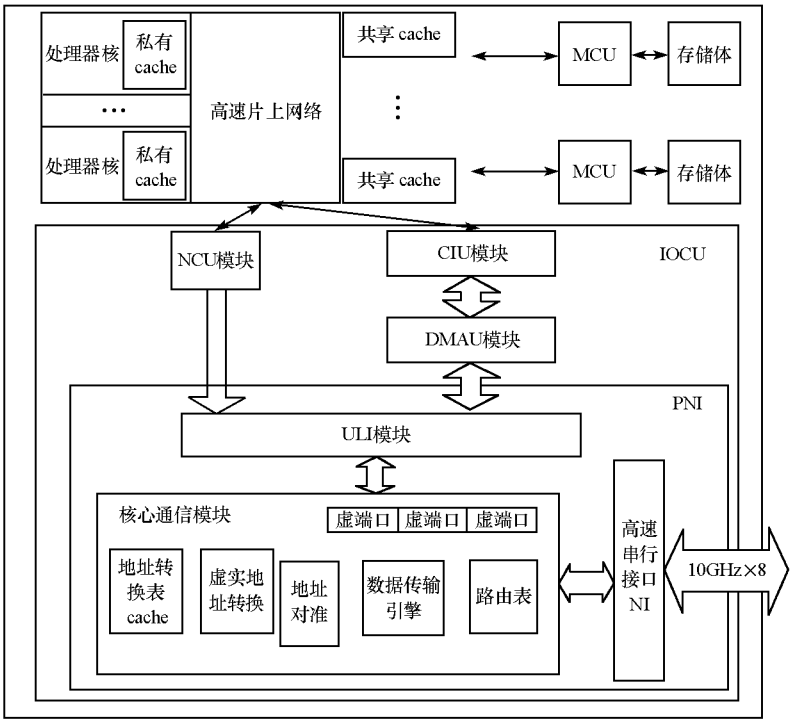

图 1 众核处理器片上通信接口结构图

Fig. 1 Communication interface's architecture of many cores chip

1)ULI 模块:上层接口模块,是 PNI 与 PIO 控制模块、DMA 控制模块的接口单元;负责从 PIO 控制模块接收寄存器访问请求,经过译码和分派后发送到目的地,并向 DMA 控制模块返回寄存器读响应;负责把 DMA 读写访存请求和中断请求发送给 DMA 控制模块,并接收 DMA 控制模块返回的 DMA 读响应。

2)核心通信模块:为通信接口的核心模块,负责实现高性能的通信机制,主要包括虚端口、数据传输引擎、虚实地址转换模块、地址拆分和对准模块及路由表等。虚端口对硬件资源进行了虚拟化,支持用户进程直接操作硬件,并提供了一种基于描述符的通信请求处理方式;数据传输引擎支持短报文 MP 和大块数据 RDMA 两种传输方式;虚实地址转换通过查找虚实地址变换表,实现用户进程的直接虚地址数据传输;地址拆分和对准模块负责实现地址的拆分和对准,支持数据的非对准传输;路由表用于实现自适应路由。

3)高速串行接口 NI:负责与系统高速互联网连接,通过查找路由表实现自适应路由,支持 8lane 的 10Gbps 串行链路,外部连接光电转换模块实现高速光互连,双向带宽为 160Gbps。

4)地址映射表 cache:为了地址转换表的快速查找,片内设计了大容量的 cache,对地址转换表

项进行预取和缓存。

2　Free-memory 的用户级通信方法

2.1　实现思想

Free-memory 用户级通信方法的主要思想是:通信接口不外接片外存储器,降低芯片设计和 PCB 的设计复杂度和成本;把用户级通信的虚实地址转换表保存在主存中,查表时直接访问主存;由于访问内存延迟较长,因此为通信接口硬件设计大容量的 cache,用于保存常用的地址转换表项,通过高效的 cache 管理策略、快速的 cache 失效机制以及访存延迟优化方法,隐藏访问主存中地址转换表带来的开销,以获得和外接本地存储器相当的通信带宽和延迟性能指标。优化主要采用下面两种手段:

1)多种策略优化访存路径延迟:把通信接口集成在处理器片内,消除 I/O 总线延迟对访存延迟的影响;

2)高效的地址转换表 cache 的管理策略:高效的 cache 管理策略可以大大降低 cache 的失效率,使得大部分的查表操作都能命中 cache,无需访问主存,降低地址转换开销。

2.2　实现算法

由于通信接口没有了本地存储器,虚实地址变换表被直接放在了主存中,因此一旦需要查表,通信接口必须访问主存,由于访存路径比较长,因此必须采取有效的方法隐藏地址转换表的访问延时,才能实现高性能的用户级通信。我们采取了两种策略提高虚实地址转换效率。通过优化访存路径减少实际访存延迟;设计地址转换表 cache,并通过高效的 cache 管理策略,尽可能地使得查表操作都能命中 cache,减少真正的访存次数,隐藏查询地址转换表的访问延迟。

I/O 总线延迟是传统基于 I/O 总线的通信接口访存延迟的重要组成部分,目前主流的 PCI Express 总线一次来回的延时在 200ns 左右。我们采用 SoC 的设计,把通信接口硬件逻辑直接集成到众核处理器芯片内,去掉 I/O 总线,大大降低了通信接口访问主存的延迟。

在片内通信接口硬件里设计了一个 128KB 大小的 ATT cache(Address Translate Table cache),用于保存主存中地址转换表的副本,一旦地址命中 cache,无需访问存储器,直接可以从 cache 中取地址转换信息。128KB 可以对应 16K 个地址转换表项,如果页大小为 8KB,那么可以实现 128MB 空间的虚实映射。ATT cache 分成 2048 组,每组 64 字节,由 8 个 64 位组成,每一个地址转换表项为 64 位,这样每组可以保存 8 个地址转换表项。由于处理器 cache 行大小为 64 字节,因此 cache 的申请和替换都以一组为单位,一次更新 8 个 64 字节。由于该 cache 用途比较单一,因此可以简化设计,设计原则是以尽量简单的方式实现 cache 的高效管理,同时使得查表操作能够尽可能的命中 cache,达到隐藏访存延迟的效果。

图 2 为 ATT cache 结构图。cache 以组为单位组织,每一组带有一个有效位和 Tag 位。ATT cache 的分配和查找都是采用物理地址的低位为索引,其中地址的[2:0]对应 8 字节的地址转换表项,地址[5:3]为组内 offset,用于选择组内 8 个表项的某一项,地址[16:6]对应组号。这种方式使得 cache 的分配和查找都非常简单,操作效率很高。ATT cache 一次分配一组,对应 8 个地址转换表项,每次从内存读一个 cache 行 64 字节,同时保存到 cache 中,这种做法有两个好处:

1)相当于每次预取了 7 个转换表项,通过预取提高了 cache 的命中率,如果 RDMA 数据传输粒度比较大,后续的地址转换就能命中 cache,减少了失效率;

2)由于 DMA 读每次可以从存储器中返回一个整 cache 行,读 8 字节和读 64 字节的延时相同,在不增加开销的情况下,实现了 cache 预取。

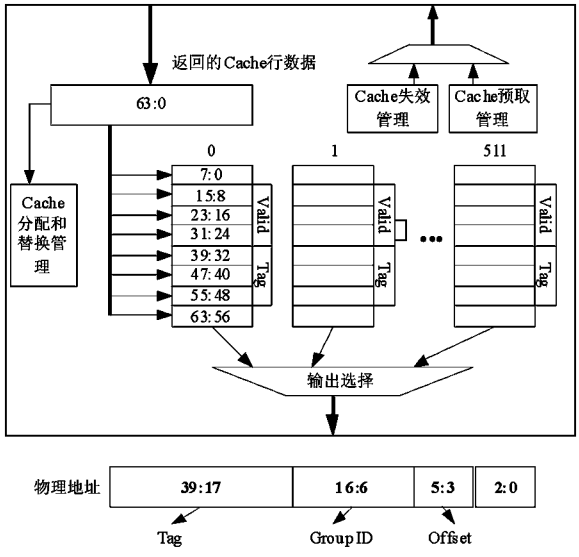

图 2　ATT cache 结构图

Fig.2　Architecture of ATT cache

cache 替换也是以组为单位,一次替换 8 个地址转换表项。cache 查询时只要看地址[16:6]对应的组号的 Tag 是否匹配,如果匹配,则 cache 命中,简化了 cache 查询操作。每一组 8 个地址转换

表项对应 64KB 虚存空间,当一次 RDMA 长度比较大时,地址转换表项可能落在多个组,通信接口在处理描述符时知道 RDMA 传输的长度,因此 ATT Cache 支持智能预取,当同一个 RDMA 传输长度跨多个组时,支持提前预取组,这样可以大大减少 cache 的失效率。

128KB 大小的 cache 相对于几亿个晶体管的处理器芯片来说非常小,带来的面积和功耗开销也非常小,因此该方法带来的面积和功耗开销相对于所能提供的性能好处来说是可以接受的。

3 性能评测

3.1 测试环境

测试采用自主设计的众核处理器芯片 HMCP。HMCP 包括 8 个处理器核,共享 8MB 的二级 cache,每个处理器核 1MB。支持一个 8 通道的 PCI Express 2.0 接口和一个片上高性能通信接口 PNI。我们基于 HMCP 芯片设计了结点测试板,处理器芯片通过 8 通道的 PCI Express 2.0[11] 接口连接 PLX8648 Switch 芯片,Switch 连出 3 个 PCI Express 插槽。另外,我们还设计了高性能通信接口卡 PDP,支持用户级通信,主机接口为 16 通道的 PCI Express 2.0 接口,网络接口为 10GB/s 的光纤接口,PDP 的双向通信峰值带宽为 9.8GB/s。PDP 卡和 PNI 除了不支持 Free-memory 外,其他通信机制相同。测试采用经过一级路由点到点连接。测试分两步进行。

首先测试 Free-memory 的性能。把两块 PDP 卡插到 Xeon5600 服务器的 PCI Express 插槽上,同样经过一级路由构成点到点连接。分别测试 Xeon 服务器通过带有本地存储器 PDP 直连和两个 HMCP 片上通信接口直连两种情况下的带宽和延迟。带宽测试采用单向和双向测试程序 bibw 和 sibw,测试不同粒度的 RDMA 的通信带宽;延迟测试采用 PingPong 方式,目的结点收到消息后再向源结点返回一个同样的消息,由源结点计算延迟。

然后测试 HMCP 片上通信接口和 Qlogic 基于 Infiniband 的 QDR[8] 卡的 MPI 通信延迟和带宽。QDR 卡通过 8 通道的 PCI Express 链路连到 HMCP 主板上,通过 Infiniband 电缆同样经过一级路由构成点到点连接。采用 MPI 程序[10] 对比了 PNI 和 QDR 的带宽和延迟性能。测试程序使用 Ohio State University 的 MVAPICH 项目[9] 组发布的延迟和带宽测试程序,用于测试 MPI 任务间点对点消息传递的单向延迟、单向带宽和双向带宽。

3.2 测试结果

图 3 为两种通信接口实现方法的带宽测试结果,从图中可以看出实现的 Free-memory 方法带宽基本没有损失,可以获得和带有本地存储器相当的带宽性能。图 4 为通信接口 RDMA 延迟测试结果。由于第一次地址转换查表通常会失效,必须读主存地址转换表,因此延迟会增大,但对后续同一页的转换都会命中 cache,延迟基本被隐藏。从测试结果我们可以看出,对于大小为一个 cache 行以下粒度的通信,Free-memory 方法延迟会略微增大一点,由于优化了访存路径,延迟增大也不明显,随着通信粒度的增大,Free-memory 方法几乎可以获得和带本地存储器相当的通信延迟。

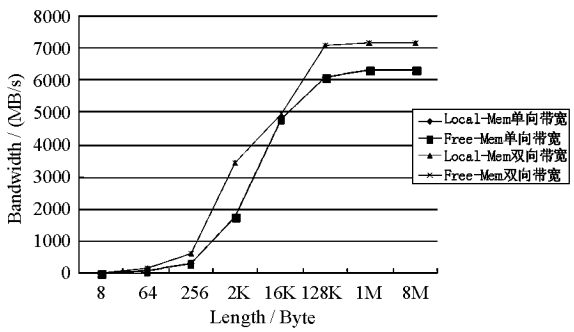

图 3 带宽测试结果

Fig.3 Results of bandwidth test

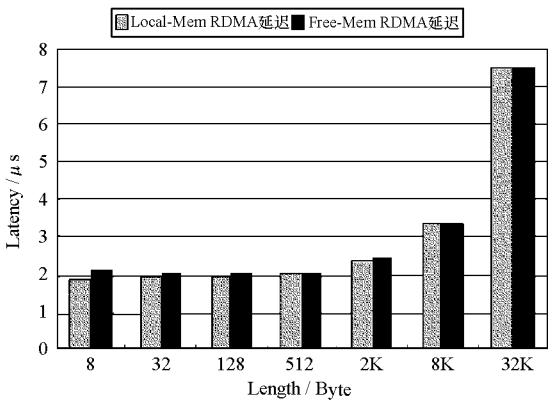

图 4 延迟测试结果

Fig.4 Results of latency test

图 5 和图 6 为 PNI 和 QDR MPI 测试程序通信带宽和点对点消息传递延迟测试结果。从图 5 中可看出 PNI 的带宽结果要远远高于 QDR 的带宽测试结果,峰值情况下接近两倍。从图 6 中可以看出,PNI 的 MPI 消息传递延迟也要小于 QDR 的延迟,最好情况延迟要减低 20%。因此,我们设计的片上高性能 I/O 通信接口的带宽和延迟都要优于 Infiniband QDR 卡的带宽和延迟。

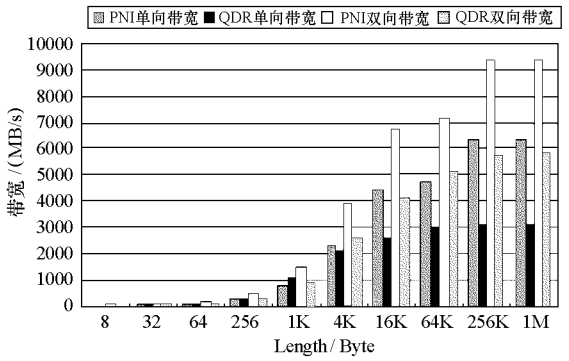

图 5　MPI 带宽测试结果
Fig.5　Results of MPI bandwidth test

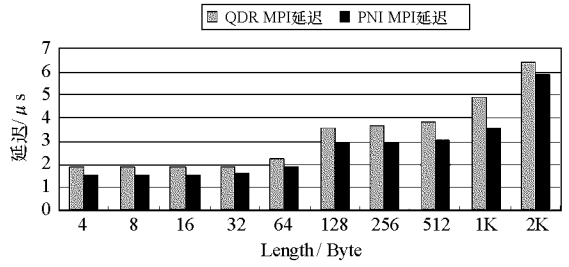

图 6　MPI 延迟测试结果
Fig.6　Results of MPI latency test

4　结束语

在传统基于 I/O 总线卡方式的通信接口的基础上,提出并实现了众核处理器片上高性能通信接口,由于减少了 I/O 总线的影响,大大降低了通信操作延迟,并有效提高数据传输带宽。为减少通信接口外接存储器带来的硬件和系统设计的复杂度和成本,提出了 Free-memory 的用户级通信方法,把虚实地址转换表直接保存到内存中,并在通信接口设计一个高速 cache,通过高效的 cache 管理策略和访存延迟优化策略隐藏地址转换查表延迟,测试结果表明实现的 Free-memory 的众核处理器片上通信接口获得了比 Infiniband QDR 更高的带宽和更低的延迟。

参 考 文 献:

[1] Gelsinger P P. Microprocessors for the New Millennium-challenges, Opportunities and New Frontiers[C]//Presentation on the IEEE International Solid-state Circuits Conference, 2001:22 – 23.

[2] Hennessy J L, Patterson D A. Computer Architecture. A Quantitative Approach[M]. Morgan Kaufmann, 3rd edition. 2002: 3 – 15.

[3] Top500 List[R]. http://www.top500.org.2009.

[4] Liu J, Chandrasekaran B, Wu J, et al. Performance Comparison of MPI Implementations over Infiniband, Myrinet and Quadrics[C]// Super Computing 2003 Conference, Phoenix, AZ, November, 2003:58 – 62.

[5] Beecroft J, Addison D, et al. QsNetII: Defining High-performance Network Design[M].IEEE Micro, 25(4), 2005:34 – 47.

[6] Infiniband Trade Association. Infiniband Architecture Specification[S]. http://www.infinibandta.com. 2004.

[7] JEDEC. JEDEC Publishes DDR2 Standard[EB/OL]. http://www.jedec.org/Home/press/press_release/jedec_publishes_DD2Std.pdf, 2003.

[8] Qlogic Corp. QDR HCA[EB/OL]. http://www.qlogic.com/Products/adapters/Pages/InfiniBandAdapters.aspx.2010.

[9] MPICH2[CP]. http://www.mcs.anl.gov/research/projects/mpich2.2010.

[10] Message Passing Interface Forum. MPI: A Message-passing Interface Standard Version 2.1 [S/OL]. http://www.mpi-forum.org/docs/mpi21-report.pdf., 2008.

[11] PCI-SIG. PCI Express Base Specification Revision2.0[S]. http://www.pci-sig.com.,2009.

[12] Brightwell R, Doerfler D, Underwood K D. A Preliminary Analysis of the InfiniPath and XD1 Network Interfaces[C]//Workshop on Communication Architecture for Clusters, 2006.

[13] Myricom Inc. MPICH-MX[CP].http://www.myri.com/scs/download-mpichmx.html.2009.

[14] Mellanox Technologies. ConnectX Architecture[R]. http://www.mellanox.com/products/connectx architecture.php.2009.

[15] Mellanox technology. InfiniHost Ⅲ Ex MemFree Mode Performance[R]. http://www.mellanox.com/pdf/whitepapers/PCIxVsMemfree_WP_100.pdf.2009.