文章编号: 1001- 2486(2010) 04- 0094- 06

一种多核微处理器互连接口的设计与性能分析*

周宏伟[1,2], 邓让钰[1], 窦 强[1], 齐树波[1], 沈长云[2]
(1. 国防科技大学 计算机学院, 湖南 长沙 410073; 2. 中国舰船研究院, 北京 100191)

摘 要: 并行是提高计算机性能最主要的方法, 随着集成电路生产工艺的不断发展, 除了在单个芯片内集成更多的处理器核外, 通过集成高速互连网络接口构建多路并行系统一直是提高高性能计算机并行性的主要方式。提出了一种面向多核微处理器的互连接口的设计方案, 基于精简的 PCI-E 总线协议, 采用高速串行数据传输技术, 支持 Cache 一致性报文和大块数据传输报文, 能够用于实现 4 个处理器的直接互连。模拟结果表明, 优化设计的互连接口每个接口能够实现 64Gbps 的双向最大有效带宽, 最小传输延迟为 120ns, 能够较好平衡不同报文类型对带宽和传输延时的要求。

关键词: 多核处理器; 互连; PCI-E
中图分类号: TP302. 1 文献标识码: A

Design and Performance Analysis of an Interconnect Interface for Multi-Core Microprocessor

ZHOU Hong-wei[1,2], DENG Rang-yu[1], DOU Qiang[1], QI Shu-bo[1], SHEN Chang-yun[2]
(1. College of Computer, National Univ. of Defense Technology, Changsha 410073, China;
2. China Ship Research & Development Academy, Beijing 100191, China)

Abstract: Parallelism is the most important way to improve the performance of computer. With the development of the integrated circuits' manufacture process, besides integrating more processor cores into one processor chip, building multi-way parallelism system through high-speed interconnect interface is the main method to increase the parallelism of high-performance computer. A design scheme of an interconnect interface for multi-core microprocessor was proposed. The proposed interface was based on a simplified PCI Express bus protocol and adopted the technology of high-speed serial data transferring. Cache coherence packet and large block data transfer packet were all supported. The interface can be used for connecting four processor nodes directly. Simulation results show that the max valid bandwidth per interface can reach 64Gbps and the minimum transfer delay is 120ns. The balance of the bandwidth and the transfer delay is reached, meeting the requirement of transferring different type of packets.

Key words: multi-core processor; interconnect; PCI-E

随着集成电路生产工艺的不断发展, 处理器芯片的集成度越来越高, 能够在片内集成多个处理器核以及互连网络接口。通过专门设计的处理器互连接口, 多个多核处理器可以相互连接构成对称多处理计算机系统(Symmetric Multi-Processing, SMP), 满足日益增加的并行计算需求。目前主流的商用处理器, 例如 Intel 公司的 IA- 32 架构"Nehalem"、IA- 64 架构"Tukwila"处理器、AMD 公司的 K8L 处理器等, 均具备高速直连接口。其中, Intel 公司的处理器直连接口采用快速通路互连(QuickPath Interconnect, QPI)

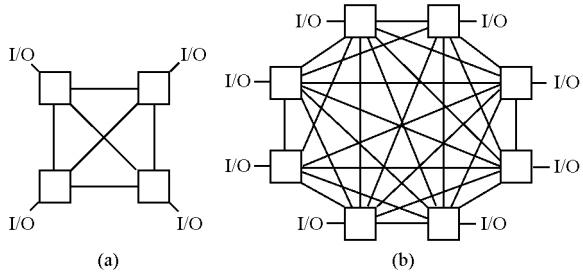

图 1 四处理器及八处理器直连拓扑结构
Fig. 1 Direct link topology for four cores and eight cores

* 收稿日期: 2009- 10- 09
基金项目: 国家"863"计划项目(2009AA01Z124)
作者简介: 周宏伟(1980—), 男, 助理研究员, 博士后。

总线接口[1], AMD 公司采用 3.0 版本的超传输(HyperTransport, HT)总线接口[2]。图 1 为采用 HT3.0 连接多个 Opteron 多核处理器的示意图, 图 1(a)为 4 个处理器直连示意图, 每个处理器使用 3 个 HT 接口用于处理器直连, 另外一个 HT 接口用于连接各种 I/O 设备; 图 1(b)为 8 个处理器直连示意图, 每个处理器使用 7 个 HT 接口用于处理器直连。

QPI 和 HT3.0 均将串行总线和并行总线技术相结合, 物理层基于低压差分信号(Low-Voltage Differential Signaling, LVDS)传输技术, 以点到点(Point-to-Point)方式进行高速链接。LVDS 技术的原理是在物理接口上均使用一对线路的电压差值来表示二进制数据, 特点是可以使用非常低的运行电压, 转换速度快, 同时具有很强的抗干扰能力。由于数据发送和接收可以双向同时进行, 因此拥有很高的传输效率, 优于传统的总线技术。除了 QPI 和 HT 总线, 目前在商业化领域比较成熟的高速总线还有用于高速 I/O 的 PCI Express 总线[3-4]、InfiniBand 总线[5], 另外用于千兆互连网络的 XAUI(10 Gigabit Attachment Unit Interface)也使用了串行传输技术[6]。刘涛等对 HT 链路层进行了研究并使用 FPGA 实现[7], David 等使用 FPGA 设计实现了一种通用的低延迟 HT 核[8]。相对于 QPI 和 HT 总线协议, PCI-E 总线更加成熟和公开。

1 多处理器互连接口体系结构

通过高速处理器互连接口, 多个处理器可以非常方便地实现互连。为了达到更高的带宽, 避免对共享资源的竞争, 目前主流多处理器间链路普遍使用点到点的直连链路。若有 N 个处理器, 则每个处理器芯片内部需要集成 $N-1$ 套直连接口, 最大支持 N 路处理器间的直连。点到点的直连方式虽然有效减少了各节点间的互连距离, 降低了总线冲突, 提高了总线带宽, 但其缺点是当需要互连的处理器节点较多时, 每个处理器所需要设计的直连接口的数目将线性增加, 互连通路的数目则成指数增加, 互连的开销将非常大, 处理器直连接口的设计面临巨大挑战。因此, 一般来说, 直接互连的处理器的数目小于 8 个, 最常见的是 2 到 4 个。

在多处理器互连体系结构中, 处理器互连接口一般分为多个层次, 但与传统的 Internet 中使用的 OSI 7 层协议不同。传统的 OSI 7 层协议由于协议层次多, 各层协议功能重复、复杂, 不能满足处理器互连的需要, 所以处理器间的互连一般采用功能精简、结构合理的互连网络协议, 其协议往往只包含传输层、数据链路层和物理层, 每个层次及其部件的主要功能为:

(1)传输层: 负责对处理器请求报文与响应报文的仲裁、处理器报文与传输层报文的格式转换、虚通道管理和错误报文处理等。

(2)链路层: 负责链路自管理、传输层报文的 CRC 校验、传输层报文打包为链路层报文、链路层报文到物理微包的转换、物理微包的发送和接收控制等[9]。

(3)物理层: 负责物理微包的发送和接收控制。可以采用基于源同步的并行传输技术, 也可以采用基于 LVDS 技术的串行数据传输技术。

2 多核处理器互连接口设计

基于点到点方式的互连拓扑结构能够避免总线冲突, 最大程度地提高总线带宽。虽然这种方式的扩展性差, 互连的成本随着节点数目的增加呈指数级数增加, 但是对于主流的多路并行系统来说, 紧耦合的处理器的个数一般为 2~4 个, 因此采用点到点方式的互连结构的复杂度相对来说可以接受。考虑到 PCI-E 协议是目前比较成熟的商业化高速 I/O 协议, 本文在此基础上提出了一种精简 PCI-E 总线协议的互连接口设计方案, 其体系结构示意图如图 2 所示。互连接口最大能够支持 4 路处理

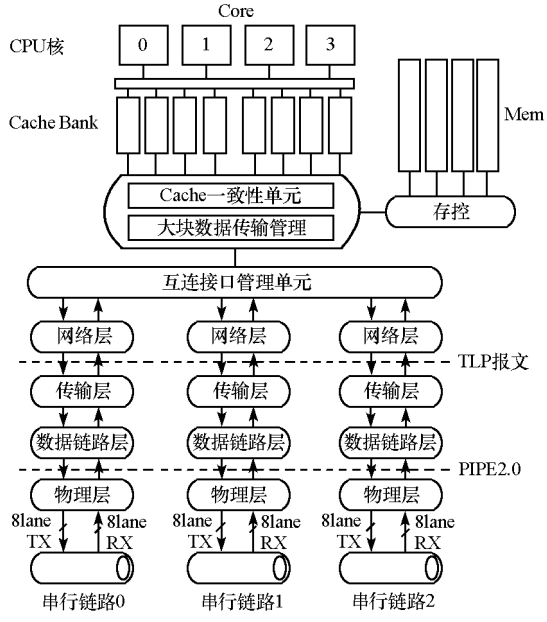

图 2 多核处理器互连接口示意图

Fig. 2 Interconnect interface of multi-cores processor

器直连,支持4处理器之间的Cache共享及Cache一致性(Cache Coherence, CC)协议,支持处理器间的大块数据传输(Large Block Transfer, LBT)。为了支持4路处理器直连,设计了3组处理器直连接口,每组接口均采用串行链路技术,基于精简PCI-E协议。每组直连接口除了包括传输层,数据链路层和物理层以外,额外增加一个网络层用于处理器的CC报文或LBT报文与PCI-E协议报文间的转换。

Cache一致性单元是一致性事务处理单元,主要用于实现多处理器Cache一致性。通过在Cache一致性单元中保持二级Cache的目录实现多个处理器间的Cache一致性,处理器内部多个核心间的Cache一致性由保存在二级Cache中的目录实现,每个一级Cache的标识(tag)都被保存在二级Cache的目录中。Cache一致性单元发送和接收的报文为CC报文,最大负载为一个二级Cache块的容量。大块数据传输单元支持多个处理器内存间的数据搬移。通过填写位于LBT中的描述符寄存器,告知LBT需要传输的大块数据的起始地址、数据长度、目标节点号、目标起始地址等,LBT会发起一个发送节点到目标节点内存间的数据传输操作,根据数据长度计算发送最大负载报文的次数,当所有数据发送完毕,LBT通过中断方式通知CPU块数据传输结束。由于处理器间采用点到点的连接,且数据链路层采用滑动窗口重传机制,因此能够保证每个LBT请求中块数据传输的顺序性。互连接口管理单元(Link Manage Unit, LMU)的主要功能为:对处理器的CC、LBT请求或响应封装为网络层报文(CC报文或者LBT报文),根据目标节点号对报文进行路由,发送到正确的直连通路中,对发送到网络层的报文进行流控等。

网络层是直连接口与处理器存储层次的接口层,该层负责将处理器的CC报文或者LBT报文与传输层协议(Transfer Layer Protocol, TLP)报文进行格式转换。为了提高带宽利用率,本文精简了传统PCI-E网络层报文格式,原先传送一个4DW(1DW= 1DoubleWord= 32 bit)长度的网络层报文,需要额外传输5DW的TLP报文头数据,而本文的精简设计将网络层报文的最小长度优化到2DW,且通过将根联合体(RootCombo, RC)和端点设备(EndPoint, EP)的地址空间设置为64位全地址空间,利用TLP报文头的地址字段存放网络层报文数据,减少了TLP报文的实际数据负载长度,达到进一步提高带宽利用率的目的。

传输层完成虚通道管理和流控,报文采用TLP报文格式。本文设计实现了两条虚通道,采用公平轮转的虚通道仲裁策略。TLP报文的发送和接收缓冲位于该层,报文发送过程中各种信用的管理也在本层完成。由于互连接口中仅传送CC报文和LBT报文,设计时精简了PCI-E的报文类型,将CC报文和LBT两种报文均封装为报告写事务(Posted Write Transaction)报文,该类报文在TLP层传输时,发送方不需要接收方自动返回应答。对于其他类型的事务报文不需要支持以简化设计。

数据链路层主要实现链路管理和差错控制,TLP报文在该层被封装成链路层协议(Link Layer Protocol, LLP)报文。数据链路层设计有报文重传缓冲,实现基于滑动窗口的重传协议,确保数据传输的可靠性。该层中还设计有链路训练(Link Training)模块,负载对链路结构侦测、链路参数配置、链路握手、速度协商等。数据链路层和物理层的接口为标准的PIPE 2.0(PHY Interface for the PCI Express Architecture, Version 2.0)接口。

物理层使用支持PIPE2.0接口的Serdes IP核实现,支持数据的串并转换与并串转换、数据编码和解码、数据加扰与解扰、LVDS信号发送和接收等。在物理层,通过绑定两个拥有4对传输线(4 lane)的Serdes IP实现每个直连接口拥有8对传输线(8 lane),单向带宽的理论最大值达到40Gbps。

3 互连结构配置及模拟环境构建

根据PCI-E总线协议,位于PCI-E总线两端的设备一个必须设置为RC模式,而对应的另一个设备必须设置为EP模式。根据该特性,2路和4路SMP系统的互连结构设置如图3所示。为了充分发挥直连接口的性能,本文配置两节点互连环境下3条互连通路可以同时使用,提高总带宽,降低拥塞概率。

为了对互连接口进行功能验证和性能分析,构建了一个双节点直连的模拟环境,如图4所示。该模拟环境为每个节点构造了直连通路的层次化结构。在网络层之上,设计两个激励模块,用于随机产生网络层报文,每个激励模块代表一个虚通道,根据网络层协议的设计,虚通道0专门用于产生请求报文,虚通道1专门用于产生响应报文。之所以在网络层上加载模拟激励,是因为网络层报文层次高,接口协议简洁,更加容易构造激励,且便于在接收方进行正确性检查。为了使模拟能够自动进行并易于查错,设计了发送与接收报文的自动比较机制,用于自动地将接收方收到的网络层报文与发送方发送出的网

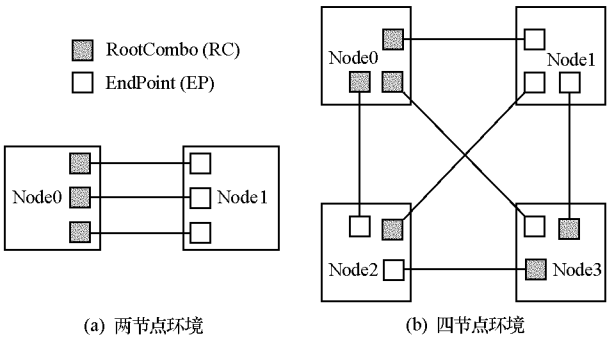

图 3 2- 4 路SMP 系统互连结构配置
Fig. 3 Interconnect configuration for 2-way and 4-way SMP system

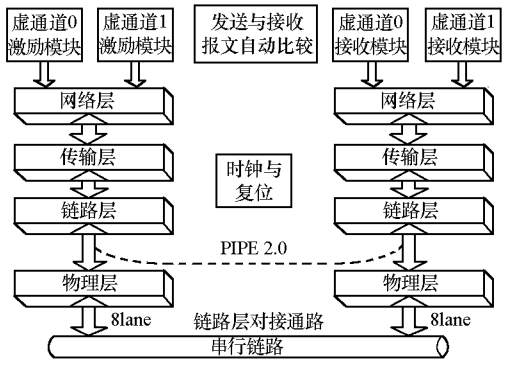

图 4 双节点直连接口模拟环境
Fig. 4 Simulation environment for direct link interface of two processors

络层报文进行比较, 检测报文是否正确地从发送端传输到了接收端。通过该环境, 可以监控直连接口中各层次报文的发送和接收情况, 分析直连接口各层协议的工作过程, 模拟和评价直连接口的实际性能。

4 模拟及结果分析

4.1 性能分析指标

(1) 最大有效带宽

定义

最大有效带宽率= 需要传输的有效数据/(实际传输的有效数据+ 额外传输的数据)

其中, 额外传输数据包含为了传输有效数据而额外增加的报文信息, 例如附加在报文中的 CRC 数据、信用信息、报文序列号和应答序列号等。一般来说, 一个报文的数据负载越大, 则传输该报文的有效带宽率越大。

当网络层报文通过TLP 报文进行传输时, 其额外开销为 5DW(TLP 报文头的开销), 因此当被传输的网络层报文长度为 4DW 时, 最大有效带宽率为 4/(4+ 5) = 44.44%。传输长度分别为 20DW、32DW、64DW 和 128DW 的网路层报文时, 最大有效带宽率分别为 80.0%、86.5%、92.8%和 96.2%。本文的物理层采用了 8 lane 的设计, 即具有 8 个单向 5Gbps 的串行链路, 共 40Gbps 的单向理论最大带宽。传输长度为 4DW、20DW、32DW、64DW 和 128DW 的网络层报文时, 最大单向有效带宽分别为: 17.8、32.0、34.6、37.1 和 38.5Gbps, 平均为 32Gbps, 双向传输平均为 64Gbps。

本文将网络层报文的报文头长度从 4DW 精简到 2DW, 且将这 2DW 数据放在 TLP 报文头的地址字段中, 减少了 TLP 报文的实际数据负载长度。优化后传输原长度为 4DW 的网络层报文头时只需要传送一个 5DW 的TLP 报文即可, 最大有限带宽率被改善到 4/5 即 80%, 传输长度为 20DW、32DW、64DW 和 128DW 的网络层报文时最大有效带宽率分别被改善到 95.2%, 97.0%, 98.5%和 99.2%。

(2) 传输延迟

定义报文第一个微包从发送方网络层发出到最后一个微包到达接收方网络层所经过的时间:

传输延迟= 发送方传输层和链路层延迟+ 物理层传输延迟+ 接收方传输层和链路层延迟

传输延迟除了与互连接口各层次的逻辑级数、工作频率有关, 还与互连接口网络层的重传 Buffer 容量、接收 Buffer 的容量和流控信用有关, 设计不合理会造成报文不能连续发送或者物理实现代价增大。

4.2 模拟结果分析

(1) 传输延迟分析

本文对不同配置情况下的互连接口的性能进行了试验, 获得了网络层报文通过互连接口的传输延迟。假设重传 Buffer 和接收 Buffer 足够大以排除由于这两个 Buffer 容量不足而额外增加的阻塞时间。模拟时发送方连续发送网络层报文, 不同长度的网络层报文的传输延迟如图 5 所示。接收方的传输层采用存储转发(Store and Forward, SF) 策略时传输层和链路层的传输延迟与网络层报文的长度有关, 采用

直通(Cut-Through, CT)策略时传输层和链路层的延迟与报文长度无关。图中每一对柱状堆叠图左边柱状图为采用SF策略时的传输延迟的构成,右边为采用CT策略时的传输延迟的构成。发送方和接收方的物理层Serdes IP核的传输时间平均为25ns,因此报文通过互连接口物理层的总延迟为50ns。当打开双虚通道时,由于两个虚通道要轮转使用物理通道,因此会增加额外的通道等待时间,该时间随着网络层报文长度的增加逐渐增大。由图5可见,采用直通策略时的传输延迟要比采用存储转发策略时的延迟小,特别是对于长报文更加明显。对于长度为1DW的最短报文,传输延迟为120ns;对于长度为128DW的长报文,采用SF策略的传输延迟为372ns,而采用CT策略时为238ns,相差比较明显。Cache一致性报文对传输延时要求较高,CC报文最大长度为16DW,采用SF策略和CT策略时的最大传输延迟分别为148ns和134ns。由于实际工作中互连接口会发送各种长度的报文,因此给出了采用SF策略和采用CT策略时的平均传输延迟,分别为189ns和150ns。

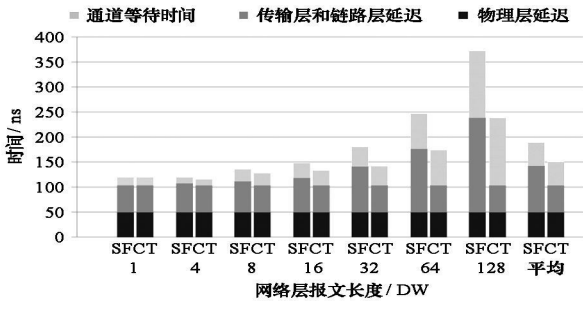

图5 报文传输延迟分析

Fig. 5 Analysis of packet transfer delay

(2)单通道下接收Buffer容量对性能的影响

前面对于传输延迟的分析没有考虑重传Buffer和接收Buffer的容量对传输延时的影响。本文模拟接收Buffer的容量对报文发送性能的影响,其中固定重传Buffer为2KB,单通道工作模式下不同容量(2KB, 4KB, 8KB)接收Buffer对性能的影响的模拟结果如图6所示。对于传统的PCI-E传输层设计,除了有报告事务(Posted Transaction)报文,还有非报告事务(Non-Posted Transaction)报文和完成事务(Completion Transaction)报文等,接收Buffer要为各事务类型的报文分配独立的信用和缓存空间,当接收Buffer容量为2KB时报告事务报文头的信用默认分配为8,数据信用默认分配为48(每个信用对应分配4DW的报文缓存空间),当接收Buffer容量为4KB时这两类信用默认值分别为16和96,容量为8K时分别为32和192。由于本文设计的互连接口仅需要支持报告写事务,因此对接收Buffer的信用进行了优化设置,将除报告事务外的其他报文的信用设置为最小,尽可能增大报告事务的信用。对于容量为2KB的接收Buffer,优化后报告事务头信用和数据信用分别设置为32和92,已经到达了2KB容量接收Buffer可以支持的最大信用值。优化的2KB容量接收Buffer的模拟结果如图6中2KB_opt系列所示。

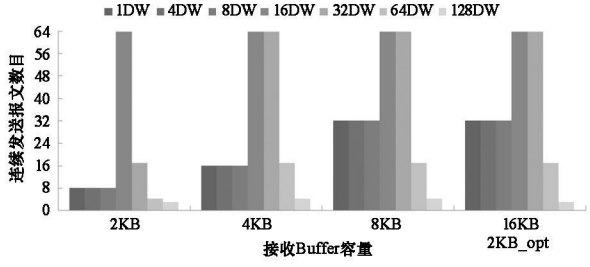

图6 单通道模式下接收Buffer容量对性能影响

Fig. 6 Influence of receive buffer's size on packet transmitting performance when only single channel is used

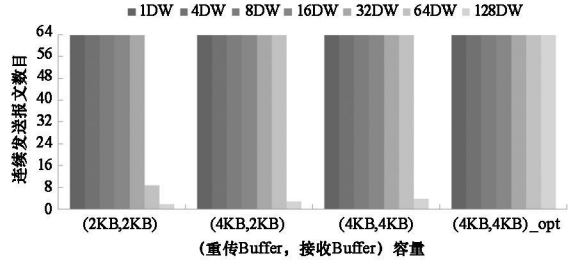

图7 双通道模式下接收与重传Buffer容量对性能影响

Fig. 7 Influence of receive buffer and reply buffer's size on packet transmitting performance when double channels are used

分析图6的结果可以得出如下结论:(1)对于短报文(小于16DW),能否连续发送取决于报文头信用,随着接收Buffer容量增大,由于报文头信用越大,因此连续发送的短报文数目越多(图中柱形图的高度超过64时,表示可以一直连续发送);(2)对于长报文(大于16DW),能否连续发送取决于报文数据信用,随着接收Buffer容量增加,报文数据信用增大,能够连续发送的报文数目越多;(3)当接收Buffer容量一定,报文长度从小到大变化时,起初连续发送的报文数目受限于报文头信用,当报文长度增大到一定程度,报文数据信用逐渐制约连续发送报文的数目,因此随着报文长度的进一步增大,连续发送的报文的数目反而逐渐减小。例如长度为32DW的报文,接收Buffer为2KB时,连续发送17个报文后由于没

有数据信用而停顿, 当接收 Buffer 增大到 4KB 及以上时, 数据信用的增加使得该长度的报文可以一直连续发送。如果使用优化后的 2KB 容量接收 Buffer(2KB _ opt) , 除了发送 128DW 长报文外, 发送其它长度的报文时的连续发送能力与使用未作信用优化的 8KB 容量接收 Buffer 时的连续发送能力相同。受到数据信用的限制, 2KB _ opt 配置下仅能连续发送 3 个长度为 128DW 的报文, 而 8K 配置下可连续发送 4 个。

(3) 双通道下重传 Buffer 和接收 Buffer 容量对性能的影响

由于双通道共享重传 Buffer, 因此重传 Buffer 容量也成为制约互连性能的重要因素。本文对双虚通道情况下重传 Buffer 和接收 Buffer 容量对连续发送报文数目的影响进行模拟, 设置每个虚通道具有独立的接收 Buffer 且容量相同, 模拟结果如图 7 所示。

由于两个虚通道轮转时会产生等待延时, 因此每个虚通道发送报文的频率会变慢。模拟结果表明在接收 Buffer 和重传 Buffer 容量均为 2KB 时, 长度小于 32DW 的报文均可以连续发送。对于长度为 64DW 和 128DW 的报文来说, 不能连续发送的首要原因是重传 Buffer 满, 然后是接收 Buffer 满。当重传 Buffer 容量从 2KB 增加到 4KB, 接收 Buffer 容量固定为 2KB 不变时, 长度为 64DW 的报文从不能连续发送变为可以连续发送。当继续增加接收 Buffer 容量到 4KB, 由于报文数据信用不够使得 128DW 长度的报文仍不能连续发送。通过对两个 Buffer 容量均为 4KB 时的报告事务信用进行优化设置, 增加报告事务报文数据信用所占比例, 数据信用从 96 优化为 192, 128DW 长度的报文终于可以连续发送, 结果如图 7 中(4K, 4K) _ opt 系列所示。

由于长报文主要属于 LBT 报文, 用于在不同节点的内存之间搬移数据, 因此报文发送频率不像 CC 这样的短报文高, 对传输延迟的要求较低, 更关注的是最大有效带宽。综合以上模拟结果数据, 在考虑平衡硬件开销和性能的情况下, 重传 Buffer 和接收 Buffer 均设置为 2KB, 且报告事务的报文头信用和数据信用分别优化设置为 32 和 92, 是比较合理的配置, 这样既可以满足 CC 报文的低传输延时要求, 又可以满足 LBT 报文的高带宽要求。

5 结 论

本文提出了一种基于串行传输技术的多核处理器互连接口方案, 采用精简的 PCI-E 协议用于互连接口的网络层和传输层, 高速 Serdes IP 核作为物理层。该互连接口提供对处理器 Cache 一致性协议报文以及大块数据传输报文的支持, 能够在最多 4 个全互连的节点间实现 Cache 一致性, 以及在各节点的内存中进行大块数据搬移。通过构建互连模拟环境, 验证了提出的互连接口方案的正确性, 并测试了互连接口的性能。模拟结果表明, 本文所提出的互连接口优化设计可以实现平均 64Gbps 的双向最大有效带宽, 采用直通策略时平均传输延迟为 150ns, 对于负载为 1DW 的短报文的传输延时仅为 120ns。考虑到 CC 报文对传输延迟要求较高, 而 LBT 报文对数据带宽要求较高, 本文通过模拟得到了能够较好平衡这两种报文传送需求的重传 Buffer 和接收 Buffer 容量优化配置方案。下一步我们将继续精简互连接口的层次, 优化互连协议, 在提高带宽的同时进一步降低传输延时。

参 考 文 献:

[1] Intel Inc, 320412-001US. An Introduction to the Intel QuickPath Interconnect[S]. Intel Inc, 2009: 3-21.

[2] The HyperTransport Consortium, White Paper HTG-WP02. HyperTransport I/O Technology Overview-An Optimized Low-latency Board-level Architecture[S]. The HyperTransport Consortium, 2004: 3-21.

[3] Ravi B, Don A, Tom S. PCI Express System Architecture [M]. USA: MindShare Inc, 2003.

[4] 张峰. 高速信号传输技术综述[J]. 信息技术快报. 2008, 6(2): 1-5.

[5] 杨晓东, 陆松, 牟胜梅. 并行计算机体系结构——技术与分析[M]. 北京: 科学出版社, 2009.

[6] Cadence Inc. I-IPA01-0140-USR Rev 14. 10 Gigabit Ethernet XAUI/XGXS Physical Coding Sublayer User Guide[S]. Cadence Inc, 2007: 6-32.

[7] 刘涛, 刘光明, 郭御风. 高性能 I/O 互连协议 HyperTransport 链路接口的研究与实现[J]. 计算机工程与科学. 2006, 28(4): 50-53.

[8] David S, Alexander G, Ulrich B. A Versatile, Low Latency HyperTransport Core[C]// Proceedings of the ACM/SIGDA 15th International Symposium on Field Programmable Gate Arrays, 2007: 45-52.

[9] 田颖瑞, 周宏伟. 微处理器直连接口链路层重传机制的设计与验证[D]. 长沙: 国防科技大学, 2009.