第 34 卷 第 5 期 计 算 机 工 程 2008 年 3 月
Vol.34 No.5 Computer Engineering March 2008
·开发研究与设计技术· 文章编号：1000—3428(2008)05—0271—03 文献标识码：A 中图分类号：TP393

cPCI 架构共享内存通信平台机制的实现

陈 勇，许 芳，袁 春，徐火生
(武汉数字工程研究所，武汉 430074)

摘 要：提出一种 cPCI 系统架构下实现主从式多 CPU 间共享内存通信的平台级分布软件机制。论述该机制中共享内存访问互斥、通信性能保证、心跳机制等关键问题及其解决方案。研究实时高效、健壮性、移植性、用户及系统规模扩展等 cPCI 平台级的应用需求，实现了共享内存通信软件中间件，在通信网关、智能通信平台等多个项目中得到实际应用。

关键词：cPCI 架构；共享内存；通信平台

Implementation of Shared Memory Based Communication Platform Under cPCI Architecture

CHEN Yong, XU Fang, YUAN Chun, XU Huo-sheng
(Wuhan Digital Engineering Institute, Wuhan 430074)

【Abstract】 A shared memory based communication platform software under cPCI architecture is proposed. Several key issues, including mutual exclusive access of the shared memory, communication performance assurance, heartbeat mechanism etc., are discussed. The corresponding solutions are described. The stable operation of communication gateway based on this design validates that, it is a platform level communication mechanism, meets the requirement of applications which demand high efficiency, real time capability, good portability and easy to extend.

【Key words】 cPCI architecture; shared memory; communication platform

1 概述

cPCI 结合了 PCI 总线电气协议标准与成熟的欧式插卡工业组装技术，是 PCI 总线的工业“加固”。它既吸收了 PC 机商用技术的成果，又能适应工控实时应用的多方面要求。cPCI 规范自制定以来，历经了多个版本，PICMG 2.x 是目前 cPCI 的主流规范。cPCI 总线的电气性能和特点包括：可达 512 MB 带宽的数据传输速率，高可用，高性能和高可靠性，扩展灵活，可实现标准化的设备平台，这使它成为新一代通信、工控、军用信息装备等领域中载板级处理器总线平台的理想选择。PICMG 系列标准规定了 cPCI 架构下总线功能、热插拔(hot swap)、桥接、系统管理等方面的硬件规范，在软件方面则需要更多地研究并实现满足该架构功能特点的技术，以组成实用系统。例如在实时操作系统方面，MontaVista 等厂家联盟研究发展了载板级 Linux(CGL)系列实时操作系统软件标准。有了硬件平台和操作系统，设备和应用开发商仍然需要研究实现热插拔管理、系统冗余和热切换(HA)等软件功能，而本文研究的 cPCI 架构下多 CPU 间的共享内存通信机制也是类似的、必需的平台级技术。cPCI 系统中各 CPU 间能否具有一个健壮、高效、使用和扩展方便的基础通信平台，直接影响着整个系统的功能和性能。

cPCI 技术利用硬件实现了处理器总线的隔离，在各 CPU 保持具有各自独立存储器空间的情况下，实现多个处理器间有效的数据交换。有关 cPCI 不透明桥的工作原理、使用及配置方法，可参考文献[1-3]。

2 设计原理

2.1 系统结构

本共享内存通信机制支持 cPCI 标准规定的主流设备架构，其系统结构及实现原理如图 1 所示。

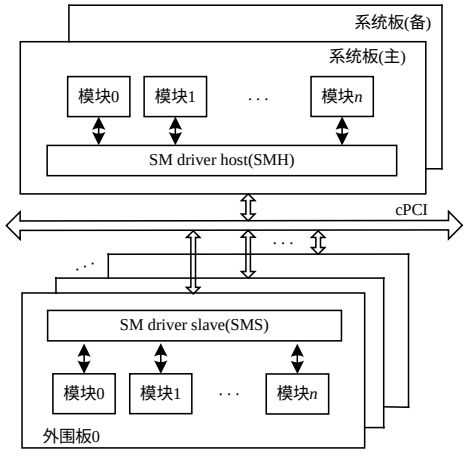

图 1 系统结构

cPCI 总线两侧分别对应有系统板和外围板。系统板处于 cPCI 总线的 HOST 侧(主侧)，作为系统的主控制板卡可以有 1 块或 2 块，如有 2 块系统板可实现系统冗余和热切换等 HA 功能。可以有多块不同种类的自带 CPU 的外围板(从侧)，它们完成系统定义的各自的功能。

作为平台层次的通信机制，本方案支持数量可增减的多个用户模块。这些用户模块有各自的特点和功能，例如可以进一步实现基于共享内存(SM)的 IP OVER PCI 功能等。

基金项目：国家部委预研基金资助项目
作者简介：陈 勇(1975－)，男，工程师、硕士，主研方向：指控网络，通信技术；许 芳、袁 春，工程师、硕士；徐火生，研究员
收稿日期：2007-05-13 **E-mail：**cy_hello@hotmail.com

2.2 核心对象定义

共享内存通信数据的接收、缓存、转发机制，决定了对用作共享内存的物理内存的组织管理方式，在机制实现中，主从两侧软件都设计并使用了一套共享内存的逻辑核心对象，其定义及工作方式说明如下：

(1)共享内存簇(SM Cluster)：用于管理 1 个或多个共享内存接口实体，每个接口实体的配置可以不同。

(2)共享内存接口实体(SM Instance)：每个外围板和系统板之间的共享内存物理接口对应一个共享内存接口实体。这样，在外围板上一般由 1 个共享内存簇管理 1 个和系统板通信的共享内存接口实体，而系统板上一般是 1 个共享内存簇管理和外围板数量对应的多个共享内存接口实体。通信在外围板上的共享内存接口实体和系统板上与之对应的共享内存接口实体之间端对端(peer to peer)进行。每个共享内存接口实体包括 2 个共享内存队列：高优先级队列和低优先级队列。主从两侧分别对应设置有高低 2 个优先级的扫描任务不断扫描这 2 个队列，以及时接收共享内存通信消息。

(3)共享内存队列(SM Queue)：存在于系统的共享内存段，是主从两侧 CPU 都能访问的区域，其物理位置是各外围板上物理内存的一段空间。每个队列包含两个共享内存缓存区，对某一侧的 CPU 来说，1 个用于收，1 个用于发，对用于供主侧接收数据的缓存区，在主侧具有只读性质，并且对从侧来说它为发缓冲区，具有只写性质。同样，另一个缓冲区对主侧是只写的发缓冲区，对从侧是只读的收缓冲区。系统中每个外围板有 2 个共享内存队列，对应了高低 2 个优先级，这样就有 4 个缓冲区，或者说 4 个共享内存通道。

(4)共享内存缓存区(SM Buffer)：每个共享内存缓存区由头部和数据部 2 部分组成。缓存区头部主要包括 3 项内容：收侧 CPU 读缓冲区时使用的缓存区数据头指针，发侧 CPU 写缓冲区时使用的缓存区数据尾指针，以及总是由从侧设置、供主侧读取的缓存区初始化标志。数据区用于存放共享内存通信消息。

(5)共享内存通信消息(SM Message)：是一种 2 层协议数据单元(layer-2 PDU)，包含了地址域(用户 ID)，长度域、数据域，以及可选的检验域。考虑到通信性能需要，共享内存通信消息在缓存区中使用 4 B 对齐存放方式。

(6)用户 ID(User ID)：由用户模块注册，在共享内存通信消息中存放于地址域，用于识别该消息属于哪个用户模块。扫描任务接收到一条消息时，搜索注册表中对应的用户模块，找到后根据其注册的数据传送方式将数据发送给该模块。

按照上述逻辑核心对象的定义，本设计共享内存(区域)的逻辑分布如图 2 所示。对于每块外围板，只包括图中虚线以上部分及用户注册表 Sap_reg 一份。同时，图中虚线以上部分的一个共享内存接口实体、2 个共享内存队列、4 个共享内存缓存区，也是系统板上对应于每块外围板都会具有的相对独立的一组逻辑核心对象。

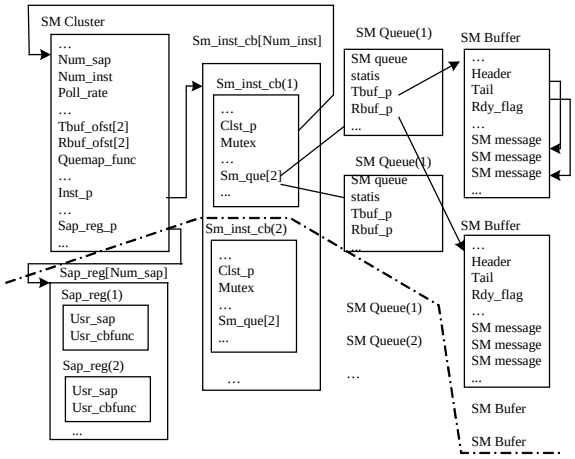

图 2　共享内存(区域)的逻辑分布

2.3 共享内存通信机制设计

上文提到，每次通信都是端对端进行的，以用 1 块系统板和 1 块外围板相对应的共享内存通信为例，本设计的原理如图 3 所示。

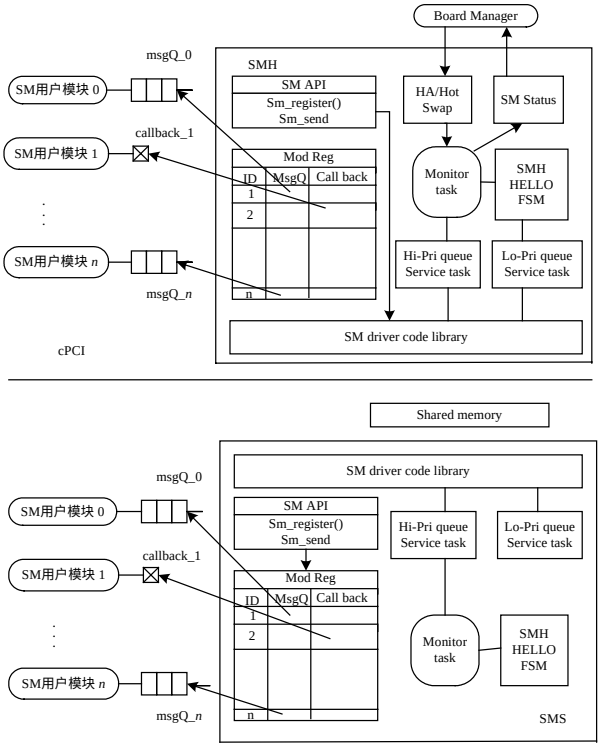

图 3　本设计原理框图

由图 3 可见，因为要实现的是双机、多机通信，所以本机制在功能分布上，主从两侧非常类似和对称。两侧的用户模块使用 SM 通信的主要接口是相似的数据收发接口。发送数据时使用 SM API 中的发送接口函数。接收数据则可根据用户选择，使用为其准备的消息队列或用户模块注册的回调函数。

主从两侧都各有 2 个共享内存队列扫描任务，用于实时扫描接收共享内存中的通信消息。在图 3 中，在系统板侧和外围板侧各有 2 个分别为高低优先级的接收队列实时扫描任务，正是对应了共享内存区的 4 个缓冲区通道。在多个外围板存在的情况下，主侧扫描任务负责扫描与各个外围板对应的、相应优先级的、主侧收共享内存缓存区，而外围板的扫描任务只扫描该板和系统板之间的从侧收共享内存缓存区。

下面以主侧高优先级扫描任务的主要执行步骤为例，说明通信消息扫描接收机制：

```
for (;;){
(1)semTake (clst_p->sem_tmr[que_id], WAIT_FOREVER);
(2)for (inst_id = 0; inst_id < clst_p->cfg.num_inst; inst_id++){
(3)while (clst_p->is_polling[que_id])     {
(4)       semTake (inst_p->mutex, WAIT_FOREVER);
          if (SM_ERR_NONE != sm_que_rcv ()){
                semGive (inst_p->mutex);
(5)                break;
        }
          sm_data_rcvd();
(6)               semGive (inst_p->mutex);
        }
        }
```

其中，(1)试图获得扫描信号量，该信号量由扫描定时器服务

程序释放，决定扫描频度。(2)针对多块外围板的扫描循环。(3)检查可扫描标志，该标志由扫描定时器服务程序设置，可决定每次扫描的时间。(4)保证对外围板访问互斥性。(5)扫描该外围板 sm instance 对应的队列，无消息或异常时释放该 instance 互斥信号量，跳出该板扫描循环，可进行后续外围板的扫描。(6)如果有待收消息，则接收该通信消息，释放该 instance 互斥信号量，进入该队列中下一条消息的扫描接收循环。

另外，主从两侧各有一个监控任务，主要负责实现两侧相互配合的心跳握手状态机，实时监控共享内存接口的运行情况。它们接收和处理上层管理模块的控制消息，同时向上层模块报告运行情况。同时，本设计中该同步状态机完成对热切换、热插拔等系统功能的支持。

SM 核心代码库是两侧公用的，提供通用的、基于共享内存的端到端通信功能支持。

3 关键问题的研究

3.1 分布式数据访问互斥问题

总线式共享内存技术在很多总线系统中都已有应用，比如VME总线、cPCI总线等。研究较多的一个关键问题是，对共享内存区域数据多CPU访问的互斥性。一般设计的共享内存体系结构如图 4 所示[2]，从主板的系统内存中或者专门的内存卡设备中预留出一部分内存作为共享内存，所有的节点都能够访问该共享内存，从而实现数据共享。

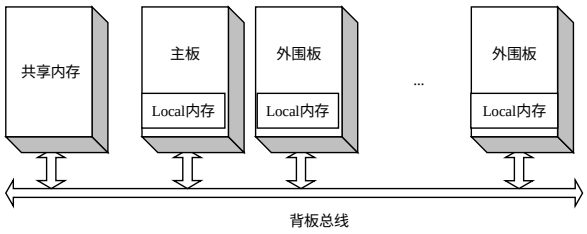

图 4 一般 SM 逻辑体系结构

该实现方案和对称多处理器系统 SMP 有相似之处，从使用上看，各节点基本对等，不存在主从特征。这样就必须设计一套适应该特点的比较复杂的互斥机制来保证数据访问的互斥性。文献[2]介绍了通过软件实现 TAS(Test And Set)及自旋锁(Spin Lock)，设计使用 SM 信号量机制，实现多 CPU 对共享内存互斥访问。文献[4]介绍了一些可以实现 TAS 原子操作的系统硬件，如在总线上支持 RMW 周期并保证该周期不可分的双口 RAM 共享内存，或者为一些处理器主板设置一个标记来防止主板释放它得到底板总线直到该标记被清除。

本文方案在一定程度上避免了上述一般系统中的复杂的互斥处理。如图 3 所示，用作共享内存的物理内存存在于各个外围板上，这样只会有 2 个 CPU 可能试图同时访问某块共享内存区，而不是一般系统中的各个 CPU 对等的访问机会，极大减少了对共享内存区访问的竞争和冲突概率。其次，将每个共享内存队列进一步划分为 2 个共享内存缓存区，避免了收发使用单一缓存区和多 CPU 竞争读或竞争写，从而不需要用复杂的数据结构管理缓存区，也避免了复杂的使用维护机制。

3.2 共享内存通信性能的保证

作为实时性能要求很强的通信、工控、军事指控等领域的解决方案，必须十分重视系统的通信效率和性能保证。同时要考虑系统扩展后通信能力和性能的变化。在本设计中，重点考虑以下 3 点：(1)2.3 节说明的扫描任务执行步骤可知，采用了定时器、信号量及扫描标志等相结合的方法，形成扫描频度和每次扫描时间均可控可调的状态。这样既能防止扫描任务占据 CPU 时间资源过多，又能通过分析和验证，调整扫描时间和频度，形成尽量保证通信实时性能的折中方案。在外围板数量为 10 块以内，系统板主频为 700 MHz，用户模块数量为 27 个(包括 IP OVER PCI 模块等实时性要求较高的模块)的应用规模下，经验证调整的优选参数为：扫描频度 10 ms，每次最长扫描时间 5 ms。(2)为了实现系统关键信息通信的实时性，采用区别服务(diff serv)的思想，设立 2 种优先级的共享内存队列，对于实时性要求很高的用户，注册使用高优先级队列。(3)将共享内存设置在外围板上，保证外围板数量较多时或再扩展时，共享内存相应自然增多，可以保证通信对共享内存容量的动态需求。

3.3 心跳机制研究及故障诊断恢复设计

系统长期不间断运行对主从两侧保持持续的活动状态监测、方便的故障诊断等方面提出了很高的要求，系统健壮性要求系统具有一定的系统故障恢复能力。已有方案通常由主处理器节点维护心跳，各从节点检查该心跳。本系统则采用从侧各单板启动类似心跳机制，供主侧检查跟踪和反馈，从侧再不断跟踪主侧相关信息的方式，配合以有穷状态机实现该功能。这样系统板和外围板都能及时知道对方的工作状态，都能够及时记录和试图恢复系统的故障。本设计中跟心跳机制相关的状态机主要部分如图 5 所示，其中从侧由 OUTSYNC 状态向 SYNCING 状态跃迁的(e1/a3+a2)路径是系统因故障失去同步后可能的恢复路径。

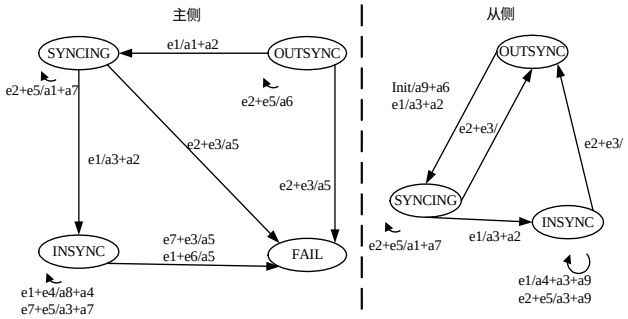

图 5 心跳机制相关的状态机主要部分

图 5 中，两侧初始状态都为 OUTSYNC；状态跃迁过程中 e 是条件，a 是动作。列举如下：e1 表示收到对侧 HELLO 消息；e2 表示定时器超时；e3 表示超过时间阈值；e4 表示检查心跳序号正常；e5 表示未超过时间阈值；e6 表示检查心跳序号不正常；a1 表示启动 SYNC2 定时器；a2 表示回发同步 HELLO 消息；a3 表示启动 HELLO 定时器；a4 表示将心跳序号增 1；a5 表示报告 SM 错误；a6 表示启动 SYNC 定时器；a7 表示重发前次 HELLO 消息；a8 表示清除超时计数；a9 表示向对侧发送 HELLO 消息。

3.4 心跳机制研究及故障诊断恢复设计

系统高可用 HA 方面，重点讨论 2 块系统板之间实现冗余和热切换。在正常工作状态，只有一块系统板作为主用板，通过共享内存与外围板通信。另一块作为备份板，启动部分共享内存模块功能并做实时数据同步，但不做数据收发工作。在收到管理模块做热切换的命令后，系统侧的共享内存通信主体迅速由原先的主用板切换为备用板。热插拔功能则要求系统能够动态安装和卸载外围板。

HA 和热插拔功能是 cPCI 相对 PCI 的重要优势所在，是

（下转第 276 页）