ISSN 1000-0054 清华大学学报 (自然科学版) 2002年 第 42卷 第 1期 15/38
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2002, Vol. 42, No. 1 52-55

面向路由器的分布式实时操作系统 HEROS

徐 恪, 吴建平, 喻中超, 徐明伟

(清华大学 计算机科学与技术系, 北京 100084)

摘 要: 分布式实时操作系统是高性能分布式路由器的控制核心。为了保证路由器系统的整体性能和安全性,设计并实现了实时分布式操作系统(highly efficient router operating system, HEROS)。HEROS基于微内核体系结构,其多任务内核实现了基于优先级的抢先式调度,高效率的任务间同步和通信原语,实时的中断处理和高效的内存管理机制。为了更好地服务于分布式路由器体系结构, HEROS实现了一种基于消息传递的分布式任务通信机制和面向网络协议的高性能的缓冲管理机制。基于HEROS的高性能安全路由器原型系统已经设计完成。

关键词: 分布式操作系统; 路由器; 消息传递

中图分类号: TP 393

文章编号: 1000-0054(2002)01-0052-04 文献标识码: A

HEROS router-oriented distributed real-time operating system

XU Ke, WU Jianping, YU Zhongchao, XU Mingwei

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract A distributed real-time operating system is the kernel controlling high performance distributed routers. This paper presents a distributed real-time operating system named HEROS (Highly Efficient Router Operating System) whose purpose is to ensure the whole performance and security of routers. Based on micro-kernel architecture, the multitasking kernel of HEROS implements a priority-based preemptive scheduler, high efficiency intertask synchronization and communication routines, real-time interrupt services and high efficiency memory management. To support distributed router architecture, HEROS implements a distributed message mechanism based on a message-passing model and network-oriented high performance buffer management.

Key words distributed operating system; routers; message-passing

Internet的飞速发展,对Internet互连的核心设备-路由器的性能提出了越来越高的要求。目前,路由器的体系结构已经从单处理机结构转向了分布式多处理的结构[1],而实时分布式操作系统是路由器的控制核心,它的稳定性和可靠性直接关系到路由器系统的整体性能和安全性。长期以来,中国在研制和开发自主产权的路由器时,都基于国外厂商开发的实时操作系统(如pSOS, VxWorks等),这样虽然减少了开发投入,加快了开发进度,但是产品的安全性和稳定性很难得到保证,也不可能在操作系统级对系统性能进行优化。

为了能够完全掌握路由器核心技术,并真正保证系统的安全性,清华大学在承担"八六三"重点攻关项目-"高性能安全路由器"时,自主设计和实现了分布式实时操作系统HEROS(highly efficient router operating system)。HEROS基于微内核体系结构,其多任务内核实现了基于优先级的抢先式调度,高效率的任务间同步和通信原语,实时的中断处理和高效的内存管理机制。为了更好地服务于分布式路由器体系结构, HEROS实现了一种基于消息传递的分布式任务通信机制和面向网络协议的高性能的缓冲管理机制。

1 HEROS的总体结构

HEROS的总体结构如图1所示。HEROS是一种松散耦合的多处理机体系结构,整个系统由多个具有自治能力的节点组成,每个节点都有自己的CPU,内存和输入输出设备(包括网络接口设备),每个节点上都运行单处理机的多任务HEROS内核,该内核实现了常用实时操作系统的主要功能。在各个HEROS内核之上是一种全局的分布式消息通信机制,它可以透明地在各个节点的任务之间传递消息。用户的程序,也就是高性能安全路由器中的IP协议和上层软件就基于该分布式消息机制构造。

收稿日期: 2001-01-02

基金项目: 国家"八六三"高技术项目(863-306-ZD-07-01)

作者简介: 徐恪(1974-),男(汉),江苏,助理研究员。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

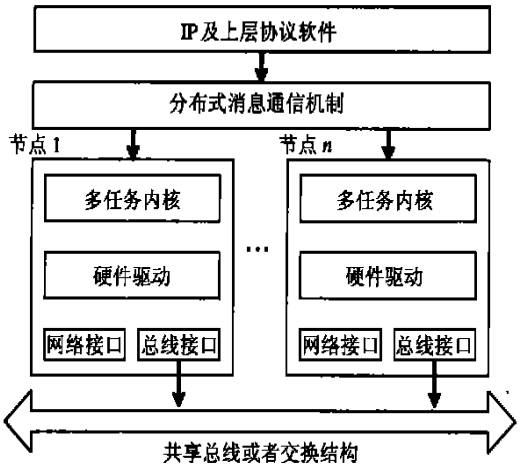

图 1 HEROS的总体结构

2 HEROS中的多任务内核

本节将讨论 HEROS多任务内核中的多任务机制,任务间通信和中断处理,它们是操作系统的核心功能.

2.1 HEROS中的任务调度

多任务操作系统都需要适当的任务调度算法为准备好的任务分配 CPU时间, HEROS支持两种任务调度算法,基于优先级的抢先式调度算法和轮转调度算法.

使用基于优先级的抢先式调度时,每个任务都有一个优先级,内核把 CPU 分配给处于就绪状态的任务中的优先级最高的任务.

当系统中同时存在多个相同的最高优先级的任务时,任务调度有两种选择. 在缺省状态下,当其中某个任务运行时,如果它不主动让出 CPU, 那么它就会一直运行下去,这会导致其它的任务没有机会执行. 另一种选择是在这些相同优先级的任务之间进行轮转调度. 轮转调度利用时间片来保证相同优先级的任务得到相同的处理时间.

2.2 HEROS中的任务间通信

HEROS中的任务间通信机制主要包括共享内存,信号量和消息队列.

共享内存 在实时操作系统中最常见的任务间通信方式就是共享内存. 由于实时操作系统的各个任务都在同一个地址空间里运行,因此相互之间共享数据结构很容易实现.

信号量 当多个任务共享同一个数据结构时,需要某种同步和互斥机制来保证数据的一致性. 信号量就提供了这种机制. HEROS中提供了两种信号量: 二值信号量和计数信号量.

在多任务实时操作系统中,由于采用了抢先式的调度算法,常常会造成优先级逆转,这时优先级高的任务必须等待优先级低的任务完成之后才能执行. 为了解决这一问题, HEROS内核中采用了优先级继承算法. 该算法的基本思想是,当高优先级的任务由于申请信号量而阻塞时,当前正在使用该信号量的低优先级的任务要继承高优先级任务的优先级直到它释放该信号量对应的共享资源为止.

消息队列 消息队列就是 HEROS中主要的任务间的信息交互方式. 单节点的 HEROS内核中实现了单节点的消息队列机制,后面我们将会看到,分布式的消息队列机制实际上就是对该单节点的消息机制进行扩展而得到的.

消息队列是独立于任务的对象,它可以缓存一定量的变长的消息. 任何一个任务或者中断服务程序都可以向消息队列发送消息,任何一个任务 (为了保证中断服务程序的效率,我们规定中断服务程序不能接收消息)也都可以从消息队列接收消息. 多个任务可以同时使用同一个消息队列. 消息队列是一种单向的通信机制,因此,如果两个任务之间需要双向的数据通信,则需要建立两条消息队列.

2.3 HEROS的中断处理

硬件中断处理是实时操作系统中的一项关键任务,因为外部事件正是通过中断通知给操作系统的. 为了加速处理, HEROS中的中断处理程序在特殊的上下文中运行,这样,执行中断处理时就不需要执行任务上下文切换. 在中断服务程序中有些系统调用不能使用,比如不能使用信号量,因此也不能使用内存分配原语 (因为要用到信号量),也不能调用 printf()等输入输出原语. 存在这些限制是因为实时操作系统要求中断实时响应,为了满足这一要求,中断服务程序必须比较短. 类似于内存分配等耗时的操作,应该尽量在应用任务中完成.

3 HEROS中的分布式消息通信机制 DMQ

随着对路由器性能要求的提高,路由器的体系结构也从单处理器体系结构发展成多处理器分布式体系结构. 为了适应分布式处理的需要,对 HEROS的单处理器内核进行了扩展,实现了分布式的消息队列机制 DMQ(distributed message queue).

操作系统对分布式处理的支持有两种方式,一种是实现真正的分布式操作系统[2],在这种分布式操作系统中,整个系统只有一个单一的映象,用户看起来就好象面对的是一个单处理器的操作系统,这

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

种系统实现复杂,需要实现任务的动态分配和迁移机制,需要实现分布式的文件系统等等,而且进程间通信的性能也可能会受到影响。另一种方式就是在现有的单处理器操作系统内核的基础上进行扩展,使其能部分支持分布式处理[3]。这样的优点是开发周期短,系统易于维护,且性能较好。

考虑到在路由器的软件处理过程中,并不需要真正的单一系统映象的分布式操作系统的支持,而是只需要多节点之间的分布式通信机制,因此我们选择采用第二种方式实现 DMQ。

3.1 DMQ的组成

DMQ的组成如图 2所示。

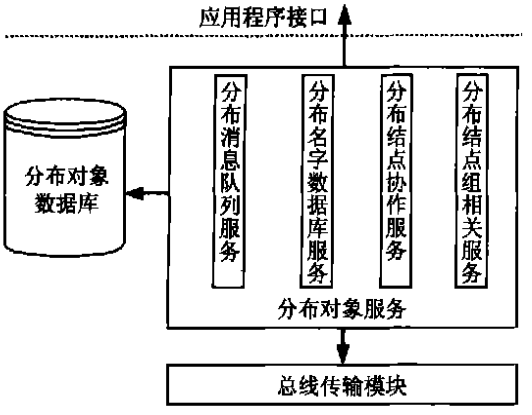

图 2 分布式消息队列服务的组成

DMQ由一组服务和分布式对象数据库组成。DMQ提供的服务包括:

分布式消息队列服务: 该服务处理来自远程节点的分布式消息队列消息。

分布式名字数据库服务: 该服务处理来自远程节点的分布式名字数据库消息。

分布式节点组相关服务: 该服务处理来自远程节点的分布式消息队列组数据库消息。包括组消息队列服务和组选举协议服务(用于选择单一的组ID号)。

分布式节点协作服务: 该服务处理节点状态维护协议,保证系统中的各个节点运行状态一致。当某个节点失效时,可以及时通知其它节点,以免影响其它节点的正常运行。

DMQ维护的分布式对象数据库包括如下几部分:

分布式名字数据库: 分布式名字数据库用于保存系统中的全局对象的名称,值和类型。在多节点系统中的每个节点上都有分布式名字数据库,这样可以保证任何节点上的任务都能够找到系统中的全局对象。

分布式组数据库: 该数据库保存分布式消息队列组的信息。

分布式节点数据库: 分布式节点数据库维护系统中的节点的信息和当前的状态。

总线传输模块负责通过总线收发消息,它又分为发送模块和接收模块。发送模块通过原语进行调用,而接收模块则监听总线,当有消息到来时将其放入临时的缓冲区等待任务接收。

3.2 分布式名字数据库

使用分布式消息机制时,需要解决的一个重要问题是如何寻找分布式消息队列。考虑到系统可靠性和性能,我们选择采用分布式名字数据库的方式实现消息队列寻址。

分布式名字数据库可以用于按照名称保存全局对象。比如可以为一个分布式消息队列取一个唯一的名称。分布式名字数据库中每行包括名称,值,类型这 3项。它可以根据名称取得值和类型,也可以根据值和类型取得名称。系统中的每个节点都有一份分布式名字数据库的拷贝。当某个节点修改了本地的名字数据库时,DMQ将会自动地把修改广播到整个系统中。这样可以保证系统中各个节点的名字数据库处于一致的状态。

当任务需要在系统中共享某个值时,它可以在分布式名字数据库中增加一项。增加之后,整个系统中的任务(无论处于哪个节点)都可以使用该值对应的名字来访问该值。

4 面向网络协议的高性能的缓冲管理机制

作为专用的路由器操作系统,HEROS必须对面向协议的缓冲进行有效的管理,影响缓冲管理性能的关键是缓冲内存管理。实践证明,直接通过核心内存分配器 KMA(kernel memory allocator)的kmalloc接口获得缓冲内存的方式对路由器应用来说是很慢的。因此,必须采取专门的手段管理缓冲内存。

缓冲内存管理有两种主要的方式。一种是独立于 KMA的,通常的做法是通过如 kmalloc之类的系统调用预先从 KMA割出一部分内存空间,此后缓冲内存都从该空间获取和释放,而与 KMA无关;另一种方法是与 KMA集成的方式,将缓冲内存视为一种核心内存对象,KMA对系统的全部核心内存(包括所有核心内存对象)进行统一管理。如 Mach-OSF/1的 zone分配器和 Solaris 2.4的 Slab

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分配器[4]都实现了对核心对象的管理。

这两种实现方案各有优劣。第一种方案的特点是实现和调试都比较简单,缺点是不够灵活,内存的使用效率不高,而且一次性给缓冲内存预留多大空间难以确定;第二种方案可以避免上述问题,其特点是较为灵活,有利于整个系统的内存使用效率,但是实现较复杂,需要编写核心内存管理代码。

HEROS的缓冲内存管理采用的是独立于KMA的实现方式,但同时引入了与KMA的一定的交互。其特点是实现简单,快速,同时尽可能保证了内存空间的使用效率。HEROS在系统启动之初从系统内存空间中先分配一部分缓冲池(buffer pool),此后的绝大部分缓冲块内存均从缓冲池中分配。缓冲池又分配几种大小的空闲缓冲块链表,各缓冲块的大小是通过实际网络上各包大小的分布确定的。分配策略是搜索当前buffer-pool中的最佳匹配内存块,由于空闲缓冲块的大小的种类比较少,所以分配的过程很快。与其它的实现方式(如Buddy系统[5])不同,HEROS在对缓冲内存管理时并不在小的缓冲块不够用的时候,分裂大的缓冲块以满足缓冲内存的要求,而是直接用大的缓冲块满足小内存缓冲块的分配请求。虽然从表面上看,有可能造成一定的内存浪费,但是一方面各种缓冲块大小的分布是依据网络数据包长度分布优化的,所以出现这种情况的概率很小;另一方面,由于HEROS的缓冲内存管理避免了内存的分裂和合并过程,因此相当快。总而言之,这种缓冲内存管理方式是由路由器的应用决定的。

5 结束语

本文介绍了面向高性能安全路由器的分布式实时操作系统HEROS。目前,基于HEROS的千兆以太驱动和SDH 155 M驱动已经设计完成,均达到了较高的性能(千兆以太驱动速度达到了900 Mbyte,基本上实现了线速)。基本的节点间转发测试证明,基于我们实现的缓冲管理和分布式消息机制实现的IP转发引擎每秒可以转发8万个分组(64 byte),已经超过了CISCO高档路由器7000系列的转发能力,证明我们设计和实现的HEROS操作系统具有良好的性能。

参考文献 (References)

[1] 徐恪, 熊勇强, 吴建平. 宽带IP路由器体系结构分析[J]. 软件学报, 2000, 11(2): 179-186.
XU Ke, XIONG Yongqiang, WU Jianping. Analysis of broadband IP router architecture [J]. J Software, 2000, 11(2): 179-186. (in Chinese)

[2] Tokuda H, Clifford W M. ARTS: a distributed real-time kernel [J]. Oper Sys Review, 1989, 23(3): 29-53.

[3] Ernesto F V, Antonio N C. An operating system extension for a multiprocessor [J]. J Sys Arch, 1998, 45(5): 341-361.

[4] Jeff B. The Slab Allocator: An object-caching kernel memory allocator [A]. Glen Dudek Conference Proceeding of USENIX Summer 1994 Technical Conference [C]. Boston, USA: USENIX Association, 1994. 87-98.

[5] Peterson J L, Norman T A. Buddy systems [J]. Commu ACM, 1977, 20(6): 421-431.

书 讯

《计算机图形学基础》

该书是作者在多年教学工作的基础上,参考了国内外最新版本的教材,为高校本科生、研究生学习“计算机图形学基础”课程编写的教材,在编写的过程中,力求体现下述原则:

1) 根据教学大纲要求,着重介绍计算机图形学的基本内容。

2) 尽量反映1990年代以来计算机图形学的最新发展所需要的基本知识,为掌握最新成果准备条件。

3) 贯彻理论联系实际原则,为学生上机实践提供指导。全书共分9章。第1章为绪论,以后7章分别论述了图形系统,二维图形生成技术,图形的裁剪及几何变换,图形学中的交互技术与用户接口,三维形体的表示和输出流水线,消隐及真实感图形生成技术。增加了第9章数字图像处理技术,着重介绍与计算机图形学密切相关的图像处理技术。每章之后附有习题,该书的最后给出了上机实验题。

该书由清华大学出版社出版。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net