第30卷 第20期 计 算 机 工 程 2004年10月
Vol.30 №20 Computer Engineering October 2004
·软件技术与数据库· 文章编号：1000—3428(2004)20—0081—03 文献标识码：A 中图分类号：TP311.52

多级反馈队列调度策略在Linux中的应用和实现

黄 斌
(上海交通大学计算机科学与工程系, 上海 200030)

摘 要：Linux操作系统作为日益流行的服务器操作系统，目前已得到广泛应用。该文分析了当前在Linux系统中进程调度策略的不足，探讨了在Linux系统中对多级反馈队列调度策略的应用和实现，提出了对Linux 内核的修改方法。

关键词：Linux；进程调度；多级反馈队列；内核

Application and Implementation of Multi-level Feedback Queue Scheduling Policy in Linux

HUANG Bin
(Computer Science & Engineering Department, Shanghai Jiaotong University, Shanghai 200030)

【Abstract】 Linux system is a popular OS, which is widely used for multi-purpose servers. After researching and analyzing the Linux scheduling policy, this paper discusses the application and implementation of multi-level feed back queue scheduling method in current Linux kernel.

【Key words】 Linux; Process scheduling; Multi-level feedback queue; Kernel

目前，Linux运用最多的两个领域是服务器和嵌入式系统。作为多功能的服务器操作系统，Linux必须跟踪系统中每个进程及其资源使用状况，以便在进程间实现各种资源的公平合理分配。如果系统有一个进程独占了大部分物理内存或者CPU的使用时间，这种情况造成了对系统中其它进程的不公平。在作为服务器操作系统时，Linux进程调度还必须确保进程的平均响应时间和平均周转时间尽可能短。目前，普通Linux进程的调度策略为动态优先级调度，该策略采用简化的调度算法，能较为有效地对进程进行调度。但是，在实际应用中，此策略还存在一些不足。多级反馈队列调度策略是一种较公平的进程调度策略，能兼顾交互、批处理和CPU占用型进程。多级反馈队列调度在Linux系统中的实现有助于改进作为服务器操作系统的 Linux 系统的进程调度策略。

1 当前普通Linux进程的调度策略

Linux系统中所有进程具有关联性，除了初始化进程外，所有进程都有一个父进程。新进程不是被创建，而是被复制，或者从以前的进程克隆而来。

对于普通进程，Linux采用基于时间片轮转的动态优先级调度，选择进程的依据就是进程计数器(counter)的大小。进程创建时，静态优先级(priority)被赋一个初值，可以为0～70之间的数字，一般内核创建新进程时分配给进程的缺省为20，这个数字同时也是counter的初值，就是说进程创建时二者是相等的。字面上看，priority是“优先级”，counter是“计数器”的意思，然而实际上，它们表达的是同一个意思——进程的“时间片”。priority代表分配给该进程的时间片，counter表示该进程剩余的时间片。在进程运行过程中，counter不断减少，而priority保持不变，以便在counter变为0的时候(该进程用完了所分配的时间片)对counter重新赋值。当一个普通进程的时间片用完以后，并不马上用priority对counter进行赋值，只有所有处于可运行状态的普通进程的时间片(p->counter==0)都用完了以后，才用priority对counter重新赋值，这个普通进程才有了再次被调度的机会。这说明，在普通进程运行过程中，counter的减小给了其它进程得以运行的机会，直至counter减为0时才完全放弃对CPU的使用，这就类似优先级在动态变化，所以称之为动态优先级调度策略。

2 Linux动态优先调度策略存在的问题

总体来说，Linux的动态优先级调度策略有一定的优点，如实现方便、算法简单、系统开销较小、能在一定程度上体现进程之间的公平性等。Linux的进程调度策略对于一个仅仅运行数十个程序的工作站来说是有效的。但是，在作为服务器时，尤其在大用户量的交互系统中还是存在以下一些缺点：

(1)对于系统负担较重的情况下，预定义的时间片过大。但如果定得过小，处理器在进程间的切换工作过于频繁，其系统调度所占的时间比例将增大，使系统的效率降低。

(2)不能很好地考虑系统的吞吐量。尤其是不能减少短进程的平均等待时间。

(3)不能得到较好的输入输出设备的利用率。

(4)不能因交互用户需及时响应而照顾输入输出型进程。

3 多级反馈队列进程调度策略的方法

多级反馈队列，是对时间片轮转调度算法的发展，其不必估计进程运行时间的大小。目前，在流行的操作系统(如Unix、Windows2000)中都用其作为基本的进程调度策略。其在克服了时间片轮转调度算法缺点的同时，按进程运行情况来动态地考虑进程的性质(以输入输出为主还是以计算为主)进行相应调度。在建模(M/M/1排队系统分析)分析和模拟测试中，都显示出其有很好的表现。它是根据进程执行情况反馈信息而对进程队列进行组织并调度进程，因而得此名。

作者简介：黄 斌(1975－)，男，硕士生，主研方向为人工智能、系统软件等

收稿日期：2003-09-02 **E-mail：**hb_a@163.net

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

多级反馈队列的具体描述如下(如图1所示)。

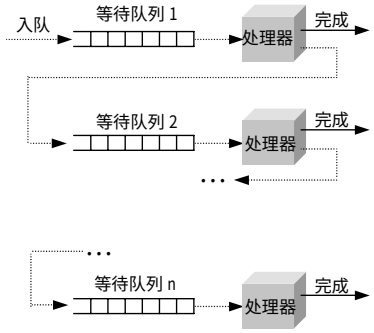

图1 多级反馈队列

(1)有多个进程就绪队列，每个队列对应一个调度级别，各级都有不同的运行优先级。第1级队列的优先级最高，以下的各级队列的优先级依次降低。

(2)每个就绪队列中的进程具有相同的时间片长度。优先级最高的第1级队列中的进程的时间片最小，随着队列的级别增加其进程的优先级降低，但时间片长度相应增加。增加的幅度可以根据系统的情况制定。通常优先级每高一级，时间片增加一倍。

(3)除最后一级外，各级队列按先来先服务原则排序后被调度。

(4)调度方法：当一个新进程进入系统后，它被加入第1级就绪队列的末尾。该队列中的进程按先来先服务原则分给处理器，并运行一个对应于该队列的时间片。如果进程在这个时间片中完成了其全部工作或因等待某些事件(如等待输入输出)而主动放弃了处理器，那么该进程或撤离系统(任务完成)或进入相应的等待队列，从而离开当前就绪队列。如果进程使用完了整个时间片后，还未完成而仍需运行(也未产生输入输出要求)，那么该进程被抢占处理器，同时将它放入下一级就绪队列的队尾。

(5)第1级进程就绪队列为空后，调度程序才去调度第2级就绪队列中的进程。其调度方法同前。当第1、第2级队列均为空时才去调度第3级队列中进程……。当前面各级队列均为空时，调度程序才调度最后第n级队列中的进程。第n级(最低级)队列中的进程是采用时间片轮转(Round Robin)方法进行调度。

(6)当比运行进程更高级别的队列中到来一个新的进程时，它将抢占运行进程的处理器，而被抢占的进程回到原队列的末尾。

4 在Linux内核中实现多级反馈队列进程调度

在Linux进程调度中每个进程都有一个task_struct结构，task_struct在内核中的形式就是通常所说的PCB，它是对进程控制的唯一也是最有效的手段。计算进程当前优先级的是goodness()函数。真正执行进程调度的函数是schedule()。要在Linux中实现多级反馈队列进程调度，必须修改的部分有kernel的部分代码、task_struct结构和schedule以及goodness函数。

需修改的文件清单如下：

include/linux/ 中的文件

sched.h；interrupt.h；tqueue.h；wait.h；locks.h；spinlock.h；smp_lock.h；timer.h；signal.h

kernel/ 中的文件

sched.c； softirq.c；fork.c；itimer.c；signal.c

以下是实现的要点：

(1)在kernel中的初始化代码中必须创建确定数量的进程就绪队列，数量可以根据处理器的能力大小确定。我们暂定为8(0~7)级，就绪队列过多会增加系统的开销，这对一些低端服务器不利。前7级队列为FIFO型，最后一个队列是时间片轮转型。Linux的定时器提供10ms的调度粒度，因此，时间片基数可定为10ms，各级的时间片大小按$10*2^n$ 计算(n为进程队列序号，$0 \le n \le 7$)。除此之外，还应创建一个等待队列。

(2)在task_struct结构中，需加入的成员变量有进程当前所在的就绪队列号Qnumber和进程在当前队列中的等待时间WaitTime。内核中的就绪进程队列结构如图2所示。

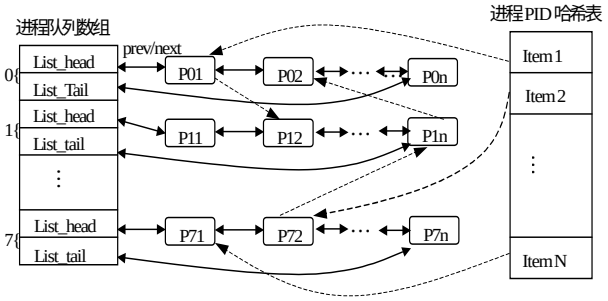

图2 就绪进程队列结构图

(3)对schedule()函数按如下原则修改：

1)先检查第8级队列，如果有进程的WaitTime大于等待的极限时间maxWaitTime时，修改相应变量的值，然后将该进程重新移到第1级队列的末尾以避免饥饿现象。

2)检查第1级队列，如果其中有就绪进程，则按照队列中排列的顺序(FIFO)进行调度，直到此队列空为止。对规定时间片内未运行完的进程，修改其相应的变量值，放入下一级队列。接着按此方法依次检查第2级到第7级队列。对第8级队列，按时间片轮转法(Round Robin)运行，未运行完的进程仍排入第8级队列的末尾。在每次调度后都要做上述1)步骤。

3)如果有运行进程因为I/O、进程间通信等事件被阻塞后，将运行进程放入等待队列，当进程重新就绪时，修改相应变量值，然后将其放入高一级的队列的末尾(本来在第1级的仍放在第1级末尾)。

4)当一个比当前运行进程级别高的就绪队列中有进程到来时，当前进程就被抢占，而运行新到来的进程。修改被抢占进程的剩余时间片值counter(其值为当前所在队列的时间片常量减去已运行的时间片数)等相关参数，将其放到原就绪队列的末尾。

(4)对于goodness()函数，因为进程的优先级和时间片都为固定值，为减少系统开销，可以不用此函数。

5 测试和对比

修改内核后，如果在x86平台上可以直接对内核进行编译、调试和安装。我们测试的软件环境为2.4.17版本的Linux内核、Apache Web服务器；硬件环境为：IBM xSeries 235服务器和普通PC；网络环境为100Mbps以太局域网。对比中采用自己编写的客户端仿真程序模拟有500~3 000个客户同时请求HTTP服务时的情况。结果如图3、图4所示。

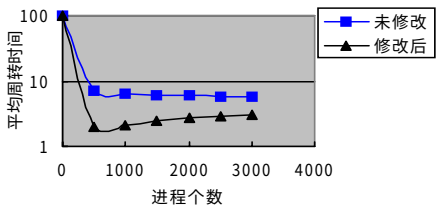

图 3 平均周转时间对比

图3 平均周转时间对比

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

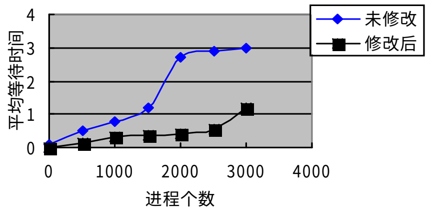

图4 平均等待时间对比

从中可以看到，修改后的Linux系统具有比原来更短的平均周转时间和平均响应时间，从而说明了在大用户量的系统中，多级反馈队列调度策略确实要优于Linux原来的基于时间片轮转的动态优先级调度策略。

6 总结与展望

Linux是一个开放源码、有广泛应用前景的多用途操作系统。作为有希望替代Unix的服务器操作系统，其还有一些性能上的问题亟需克服。本文在研究目前已有的普通Linux进程调度策略的基础上，提出了用多级反馈队列调度策略替代Linux现有的普通进程调度策略，从而使Linux系统具备良好的大用户量时的分时响应性能。这种多级反馈队列调度的策略虽然比Linux原有的进程调度策略复杂，但是可以有效地提高Linux服务器的性能，从而减小与高端Unix服务器的差距，为Linux更广泛的应用打下一个良好的基础。随着更多新的进程调度策略在Linux上实现，这种现代操作系统在服务器应用领域必将会有光辉的前景。

参考文献

1 Aivazian T.Linux Kernel 2.4 Internals.http://www.linuxdoc.org,2002-08

2 Stallings W.Operating Systems:Internals and Design Principles (Fourth Edition). Prentice Hall, Inc.,2002

3 尤晋元.Unix高级编程技术.上海:上海科技文献出版社,1994

4 毛德操,胡希明.Linux内核源代码情景分析.杭州:浙江大学出版社,2001

5 Torvalds L.Linux Kernel Source 2.4.17.http://www.kernel.org,2002

6 屠 祁,屠立得.操作系统基础(第三版). 北京:清华大学出版社,2002

7 黄千平,陈洛资.计算机操作系统.北京:科学出版社,1989

☆☆☆

(上接第21页)

的：从最初大量的原始统计数据数据(计算机领域：50万条，经济领域：37万条)，经过层层过滤后得到最终的结果(统计数据见表5)。从表5可以看出该系统能够帮助我们发现有益的新词、新概念。考虑到分词的效率和歧义性，我们只将通用的新词可以加入通用分词词典中，而其它的新词可加入领域词典中。

表5 新词发现结果示例

领域	计算机领域		经济领域	
类型	单字组合词	多字组合词	单字组合词	多字组合词
数量	3 000	3 010	1 586	3 550
例子	“死机”、“按钮”、“字节”、“字段”、“总机”、“声卡”、“站点”、“页眉”、“鼠标”、“网吧”、“黄页”、“师姐”等	“即插即用”、“北大方正”、“批处理”、“结束符”、“硬拷贝”、“局域网”、“差错 率”、“制造业”、“神经计算模型”、“演示版”、“计算机网络”等	“下岗”、“树叶”、“影碟”、“造假”、“韩国”、“招标”、“中标”、“转账”、“征婚”、“诊所”、“执着”、“庄家”等	“安全感”、“八达岭长城”、“领导班子”、“包装箱”、“保健品”、“保障体制”、“项目经理”、“个体户”、“菜篮子”、“侏罗纪公园”等

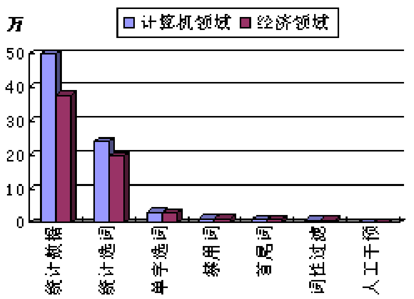

图2 单字词过滤过程结果显示

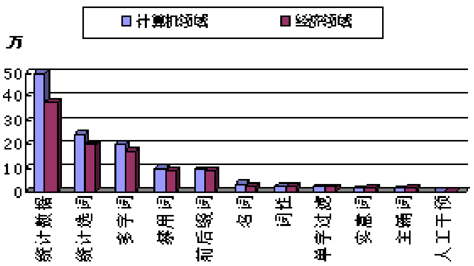

图3 多字词过滤过程结果显示

6 结束语

本文根据汉语的特点，采用基于概率统计和规则相结合的方法对新词的自动获取进行了有益尝试。这是一种适用于从大规模语料库中发现新知识的方法。该方法的不足之处是：使用统计公式进行统计选词的结果依赖于语料库的发散程度，而规则过滤的效率依赖于规则知识的获取和完备。下面的工作将尝试使用数据挖掘的一些算法(例如关联规则)，以及Markov概率统计模型来提高系统的效率。

致谢 该研究工作得到首都信息发展有限公司的李蕾博士、郭祥昊博士、王楠和伏小妹等人的帮助和建议。这里表示深深谢意！

参考文献

1 Schutze H,Hull D,Pederson J.A Comparison of Classifiers and Document Representations for the Routing Problem. In Croft . (Eds.), Proceedings of SIGIR-95, 15th ACM International Conference on Research and Development in Information Retrieval,New York: ACM Press,1995:229-237

2 Tan Chademeng ,Wang Yuanfang, Lee Chando.The Use of Bigrams to Enhance Text Categorization.Information Processing and Management, 2002,38: 529-546

3 Lai Yusheng,Wu Chunghsien.Meaningful Term Extraction and Discriminative Term Selection in Text Categorization via Unknown-word Methodology.ACM Transaction on Asian Language Information Processing, 2002,1:34-64

4 Lewis D.Feature Selection and Feature Extraction for Text Categorization. In Proceedings of a Workshop on Speech and Natural Language , San Mateo, CA: Morgan Kaufmann,1992:212-217

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net