

95-97

状态检测

空闲机 管理软件

系统

计算机科学1998 Vol. 25 No. 4

(24)

PC 机环境中的状态检测和空闲机选择*)

State checking and Selecting an Idle Machine in PC Environments

王庆凯

(吉林建筑工程学院 长春130021)

胡亮

(吉林大学计算机科学系)

TP360.6

TP315

摘 要 With the development of computer technology, PC cluster is very popular especially in distributed computing. This cluster includes many PCs connected by network, total computing power is managed by cluster management software. This paper introduces the problem of selecting an idle machine in Windows 95 or Linux environment

关键词 Cluster, Distributed computing, Scheduling

1 引言

现有的机群管理软件主要用于工作站机群,不能在 PC 机群上运行。因此,针对我国国情,开发 PC 机群上的管理软件是十分有价值的。

机群管理软件的前提是存在一个能正确反映系统当前负载情况的负载指标,机群系统的一个主要目标就是提高系统性能,即缩短作业的平均响应时间。一个作业的响应时间依赖于其所运行的主机上的负载,负载越重,其运行时间越长。因此,理想的负载指标应该与作业响应时间有一个单调关系。一个作业的执行需要各种计算机资源,主要包括 CPU 周期、主存空间以及磁盘空间。机群系统的另一个重要目标是均匀且充分地利用整个系统的资源。这个目标与上面提到的目标是一致的。资源的使用越平衡,作业的响应时间就越短。

理想的体现系统负载情况的负载指标应当满足这些条件:1)测量开销低,这样可以频繁测量以确保信息最新。2)能体现所有竞争资源上的负载。3)各个负载指标在测量及控制上彼此独立。

Ferrari^[2]建议使用各种资源队列长度的线性组合作为负载指标,但其条件是假定系统处于稳定状态,并且要求资源的排队规则是 FCFS,PS 或 LCF-SPR,但实际的系统并不满足这些条件。Zhou^[3]建议使用 CPU 队列长度作为负载指标,发现 CPU 队列长度和磁盘队列长度分别与 CPU 类作业和 I/O 类作业有密切关系,但资源队列长度与资源利用率并

没有紧密的关系。Bonomi^[4]使用处理机上活动的进程数的瞬时信息和周期地搜集到的平均 CPU 运行队列等信息的组合作为负载指标,并且进行了实际测量,性能上有所改进,但仍不能正确反映资源利用率。Kunz^[5]通过实验比较了运行队列任务数、系统调用率、CPU 进程切换率、CPU 利用率、空闲内存大小和1分钟负载平均值六个单项负载指标以及它们的“与”及“或”组合发现,效果最好的是单项指标中的运行队列任务数,任何其它单项指标或其组合都不比它好,但实验条件有很大局限性。Banawan 和 Zahorjan^[6]使用预言(oracle)模拟方法比较了作业到达数、作业离开数、瞬时队列长度、平均队列长度、平均响应时间、平均伸展系数(作业从到达到离开的时间与在此阶段收到的服务时间之比)、利用率、未完成的工作量等负载指标,得到的结论是使用瞬时队列长度、利用率、平均队列长度、平均响应时间和平均伸展系数中任何一个都可以极大地改进系统性能。Stumm^[7]指出 CPU 队列很少与 CPU 利用率有关系,特别是在交互式环境中。Mehra 和 Wah 使用了比较元(comparator)神经网络,该网络能学习预测一个作业的相对执行时间,但该作业必须只使用在其到达之前便观察到的资源利用率模式。他们的负载指标包含如 CPU、磁盘、网络及主存等主要被争用资源的信息,但这种方法只适用于重复执行的小作业,随着执行时间的增多,其准确性便下降了。

很多分布计算系统使用进程数作为负载指标,

*)吉林省科委应用基础项目资助。王庆凯 副教授。胡亮 博士生,讲师。

但该指标没有区分进程的性质和大小,实际系统中一个大进程对资源的占用量及其运行时间可能比几个小进程对资源的占用量总和还要多。I/O 类作业和 CPU 类作业对资源的使用情况也有很大区别,一个长时间运行的 I/O 类作业占用很少的 CPU 周期,而一个短的 CPU 类作业在运行时可以占用全部的 CPU。若表示系统负载情况的 CPU 队列长度将系统进程也考虑进去,则更不准确,因为系统进程很少占用资源。例如一个工作繁忙的 PC/586 系统开机 2 小时其 Swap 进程和 Inetd 进程才分别使用 3.5 秒和 1 秒的时间,因此单独使用队列长度作为负载指标是不合适的,但直接使用资源利用率,即 CPU 利用率和 I/O 利用率却很直观、准确。只有当资源利用率达到 100% 时, CPU 队列长度对作业响应时间影响才是主要的。

在作业执行所需的资源中,通信带宽主要在诸如发送信号、等待信息到达等同步及通信延迟方面影响作业的响应时间,其变化性较大,这里不考虑。另外,现代个人计算机 PC486、PC586 软硬件技术对运行的作业提供足够的内存,使得内存很少争用。

我们在 Windows95 环境下衡量一个机器的负载指标包括两个参数, CPU 利用率、CPU 队列长度。一般情况下使用 CPU 利用率,只有当网络中所有主机的利用率都在 90% 以上时,才使用队列长度作为负载指标。这两个指标能够很好地描述主机上的负载情况。

2 主机负载状态的测量及空闲机的选择方法

选择 CPU 利用率, CPU 队列长度作为本地负载状态信息,这些状态信息不断更新,如果更新的周期太长,则表示的信息是过时的,所以要求采样的周期越短越好。另一方面,获得这些参数值所占用的系统资源应当尽可能少,系统的采集不能太频繁,否则会增加系统的开销。因此采样系统的周期应作适当的选择,这是测量主机状态时,需要选择的第一个重要参数。在我们的环境中所选的采样周期为 1 秒,这时采样所占用的系统资源不到整个系统资源的千分之一。

系统的状态用当前一段时间系统状态的平均值来表示,在测量主机状态时,第二个需要选择的参数是这段时间的长短。如果选择得太长,则它不能很好

地反映系统当前的状态,如果选择得太短,则不能消除小作业所产生的高峰现象。在我们的环境中所选值为 10 秒,这样,系统的状态可以用最近的 10 个采样样本的平均值来表示。

在机群系统中,从全局调度的结构来看,主要有集中式和分散式两种。集中式方法中有一个中央服务员,负责收集整个系统的状态信息并作出全局调度决策,其它各主机向该服务员周期地发送有关本地负载信息的信件,报告它们的负载信息;该中央服务员维持一个有关机群系统中各主机负载情况的全局状态表,为其它有请求的节点分配空闲机。这种调度方式的优点是通信开销较少,能支持较大的规模,但缺点是坚韧性较差,一旦中央服务员发生故障则整个机群系统便会陷入瘫痪。在分散式方案中,没有集中的中央服务员,每个主机必须不断地记录整个系统其它机器的状态,或在需要时查询其它主机的状态。它的优点是可靠性高,易扩展,任何一个机器的瘫痪不会影响整个系统的正常工作,缺点是通信开销大,所支持的规模较小。目前,绝大多数系统采用分散式的调度结构。

在我们的 PC 机群系统中空闲机的选择采用分散式调度结构,每个机器上的调度服务员都维持一个主机状态信息表,记录网络中所有主机的状态信息。调度服务员根据主机的状态信息为并行任务分配合适的处理机,它们之间的关系如图 1 所示。

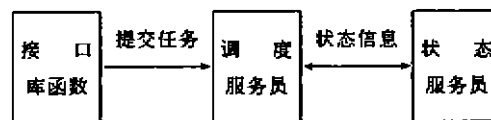


图1 调度系统结构

主机状态信息表用下面数据结构表示:

```
struct hostinfo{
    struct hostinfo * link; /* 下一个主机的信息指针 */
    float   cpu.u; /* cpu 利用率 */
    float   cpu.q; /* cpu 队列长度 */
    int     occupied; /* 该主机是否被占用 */
    int     time; /* 多长时间状态未被修改 */
    int     ntask; /* 已分配多少任务在该机上 */
}
```

当主机状态服务员发现某主机负载发生变化时,主机状态服务员通知调度服务员修改主机状态表中的信息。当系统中有等待执行的任务时,调度服务员从主机状态表中选择一个 CPU 利用率最低的空闲主机,在该机上执行该等待任务,直到系统中没有空闲的主机或者没有等待执行的任务。