高性能集群系统中资源负载量化的研究*)

袁立强　徐炜民
(上海大学计算机工程与科学学院　上海200072)

摘　要　负载均衡一直是高性能集群系统中资源分配追求的一个主要目标，能否有效地进行负载识别则直接关系到负载均衡的最终实现。本文针对计算资源和通信资源这两类作业争用的主要资源，分别讨论了其负载的衡量以及负载的量化方法，并通过具体的实验来加以说明。

关键词　集群系统，资源分配，负载均衡，负载量化

Researches on Quantification of Resource Load in High Performance Cluster System

YUAN Li-Qiang　XU Wei-Min
(Shanghai University School of Computer Engineering and Science, Shanghai 200072)

Abstract　Load balancing is a desired target for resource allocation in high performance cluster system , and to get a well-balanced load is determined directly by load identification . This paper mainly discusses a way to quantize computing and communication resource load , and also gives detailed experiments for illustration .

Keywords　Cluster system, Resource allocation, Load balancing, Load quantification

1　引言

随着计算机体系结构以及通讯技术的发展，在并行处理技术中，采用高带宽(High Bandwidth)、低延时(Low Latency)的通讯网络，以高性能的工作站集群系统作为并行计算的平台已经越来越多地引起了人们的重视[2]。作为集群系统的重要组成部分，集群管理系统可以统一地进行任务调度和资源分配，保证用户公平合理地使用集群系统，从而提高整个系统的利用率[3]。

负载平衡一直是集群系统中进行资源分配时追求的一个主要目标。一个可以正确反映当前系统负载情况的负载指标对一个成功的负载平衡系统来说是至关重要的，文[1]指出：理想的、体现系统负载情况的负载指标应该满足以下条件：(a)测量开销低，这意味着以频繁测量确保最新信息；(b)能体现所有竞争资源上的负载；(c)各个负载指标在测量控制上彼此独立。基于以上三点，使用资源利用率作为负载平衡系统的负载指标是一个合理且有效的选择[1]。

人们通常将资源的负载情况定性地划分成轻载、中等和重载三种，如何将这类定性问题在资源利用率这个负载指标上定量化，将是我们实现整个系统负载平衡中极为关键的一步。

2　计算资源和通信资源负载的量化

面向任务来分配处理机是当前集群管理系统中进行资源分配时使用得较多的一种方法，它追求的是系统的负载平衡性，即尽可能提高每个处理机的利用率[4]。不同类型的并行任务对资源的需求和占用不尽相同。计算资源和通信资源是作业执行时争用的两类主要资源，因此，对这两类资源当前负载的正确识别直接关系到资源分配的成功与否。文[1]指出：使用资源利用率作为负载平衡系统的负载指标。在本文中，我们将CPU利用率和系统网络带宽的利用率作为计算资源和通信资源负载的衡量指标。

资源负载通常被定性地划分成轻载、中等和重载三种。在使用资源利用率来衡量负载状况时，如何给上述定性问题合理地界定一个数值范围，即有效地解决负载的量化问题，将直接关系到整个系统负载均衡的最终实现。针对集群系统中的计算资源和通信资源，我们在这里分别讨论了其负载量化的方法，并给出特定平台上的实验结果。实验的硬件平台是由上海大学和清华大学合作研发的自强2000高性能集群式计算机系统，软件平台是Redhat Linux 6.2。

2.1　计算资源负载的量化

CPU是系统的主要资源，也是作业争用的主要对象，CPU利用率对不同类型的任务响应时间的影响是不同的[1]。计算密集型任务和通讯密集型任务是所有并行任务中的两个对立面，在不同的CPU利用率下，通过综合分析上述两类任务的平均响应时间，我们得出一种计算资源负载的量化结果。

表1　一个结点两个测试进程时间增加值

CPU% / 基准	10%	20%	30%
IS. B. 1	0.35%	10.8%	17.45%
IS. B. 2	5.1%	18.39%	23.45%
IS. C. 8	11.49%	27.83%	39.01%
EP. A. 1	9.53%	20.47%	29.68%
EP. B. 4	10.79%	21.46%	29.42%
EP. C. 4	10.59%	21.22%	29.45%
EP. B. 9	9.5%	20.96%	29.74%

*)本课题得到上海市教委重大科技项目——"网格技术E研究院"的支持。袁立强　硕士，主要从事高性能集群式计算机系统方面的研究；徐炜民　教授，主要研究领域为：计算机系统结构、高性能计算、并行处理等。

表2 一个结点一个测试进程时间增加值

基准 \ CPU%	50%	60%	70%	80%
IS.B.1	0%	9.35%	20.57%	33.79%
IS.B.4	0%	8.89%	17.38%	33.34%
IS.C.8	0%	16.5%	33%	48.53%
EP.B.1	0%	10.07%	22.88%	32.11%
EP.A.4	1.1%	11.1%	25.55%	32.22%
EP.B.4	2.5%	10.28%	21.67%	31.47%
EP.C.9	0.04%	9.9%	22.26%	30.55%

我们选取了NASA开发的一套并行基准包NPB(NAS Parallel Benchmark)[7]作为测试程序,其中,IS[7]和EP[7]分别代表了NAS套件的两个对立面:通信密集型应用和处理器密集型应用[7]。表1和表2给出了在不同的CPU%负载环境下分别运行这两类作业的响应时间的增幅情况。由于在我们的集群系统中每个节点都是包含两个处理器的SMP架构,这就意味着每个结点同时能处理两个进程。在不同的CPU%负载下,我们为每个结点分配的进程个数是不同的:当CPU%足够小时,我们可以一次性地为该结点同时分配两个进程,而当CPU%处在一定范围内时,我们只能为其分配一个进程;当CPU%达到某一程度时,我们便不再为该结点分配任务。我们通过在一个结点上同时运行两个IS或EP进程来测试可以同时分配两个进程的CPU%的范围,又通过在一个结点运行一个IS或EP进程来测试只分配一个进程的CPU%的范围。表1和表2分别给出了这两组测试的所有结果,其中,表1中的数据是相对于CPU%负载环境为0时同时运行两个IS或EP所用时间的增幅,而表2中的数据则是相对于CPU%负载环境为0时只运行一个IS或EP所用时间的增幅,所列数据均是我们5次测得的平均值。

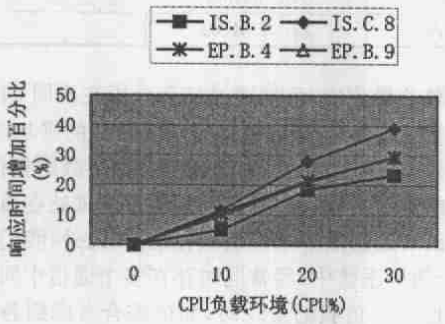

图1 一个结点两个测试进程时间增长曲线

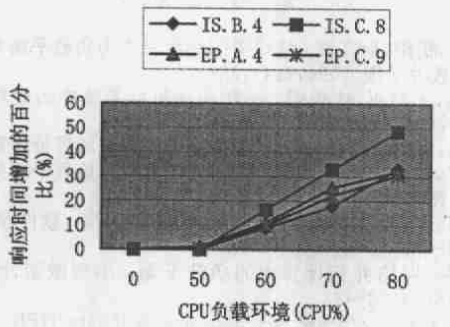

图2 一个结点一个测试进程时间增长曲线

为了更直观地表述,我们在图1中给出了表1中EP和IS测试的边缘值,从该图可以看出,EP基准测试出的曲线几乎重合且被夹在IS基准测试出的曲线之间,这说明在相同的CPU%负载环境下,不同规模和大小的通信密集型程序所受影响的差异要大于计算密集型程序;另外,当CPU%负载环境为10%时,两类应用程序响应时间的增加不多,均接近于或小于10%这个经验值[1],而当CPU%负载环境超过10%时,响应时间的增长大大增加,当为20%时在8个结点上运行C类的IS基准时增幅达到了27.83%,在30%处更是接近了40%。因此,我们选择了当CPU利用率处在0~10%之间时,给该结点同时分配两个进程。图2同样给出了表2中EP和IS测试的边缘值,在相同的CPU%负载环境下,不同规模和大小的通信密集型程序所受影响的差异要大于计算密集型程序同样在该图中得到了体现。从图中可以看出,CPU%为50%时两类基准程序的响应时间与CPU%为0(即空载)时基本相同(本文中CPU%值并不是单个CPU利用率,而是将两个CPU作为一个整体时计算出来的利用率,这样,当CPU%为50%时,逻辑上可以理解成只用了一半的CPU资源,即只完全使用了一个CPU,而另一个处于空闲,尽管微观上操作系统对双CPU进行了调度,并未让其中一个始终空闲);而当CPU%到达60%时,增幅最大的同样是在8个结点上运行C类的IS基准程序,达到了16.5%,但考虑到其他基准的增幅都小于或者接近于10%,且从总体上来说,所有基准测试的平均值是接近于10%这个经验值的,因此,我们选取了10%~60%作为只给结点分配一个进程的负载范围。

由此,我们可以得出,基于我们特定的自强2000系统,当一个结点CPU利用率处在0~10%内时,我们可以为其分配两个进程运行;在10%~60%之间只为其分配一个进程;当CPU利用率超过60%时,我们建议不再为其分配进程。

2.2 通信资源负载的量化

与仅仅使用处理器的频率来衡量计算能力不同的是,决定一个结点通信能力的因素更多,也更复杂。在这些因素中,物理网卡厂商公布的最大传输速度是结点通信能力在理论上的峰值,而总线速度、主存容量等硬件因素也同时影响和决定着结点的通信能力;除硬件上的制约外,软件层对通信能力的影响同样深刻而巨大,编程人员通常是通过使用各类通信中间件来实现进程和结点之间的通信的,这些中间件中的通信原语所能达到的通信效率在很大程度上决定着整个并行程序的通信性能。譬如,Myricom公司提供的基于GM的MPI实现MPICH-GM[8],其通信速率比起直接使用Myrinet网卡要慢上15MB/s[8]。因此,同时考虑通信硬件和软件因素,并以此来衡量网络负载,是十分有必要的一件工作!各个层次对结点通信能力的影响见图3。

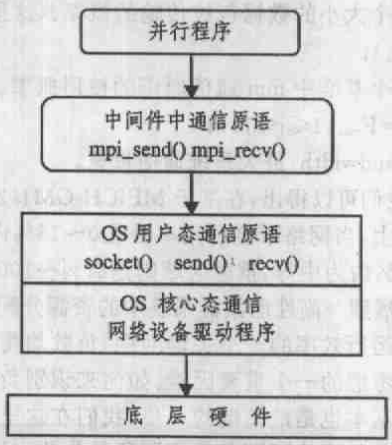

图3 通信过程中的软、硬件层

在本文中，我们将网络带宽的利用率作为通信资源负载的衡量指标。我们所述的网络带宽并不局限于网卡所能达到的通信带宽，而是综合考虑通信软、硬件因素后得到的系统的通信带宽。在我们特定的集群系统上，借助于 MPICH-GM 这个中间件，讨论了同时考虑通信硬件和软件因素来对通信资源负载进行量化的一种方法，并给出了实验结果。

由于 MPI[10]通信原语处在整个通信流程的最顶层，因此，处于其下层的所有因素对通信性能的影响必将在 MPI 通信原语层得到体现。这样，MPI 通信原语层的通信效率实际上也是整个系统通信效率的直接反映。为了得到 MPI 通信原语的通信效率，我们引进了 PALLAS 公司开发的一组测试基准套件，称为 PMB(Pallas MPI Benchmark)[9]。PMB 开发的目的就是用于测试不同的 MPI 实现在不同的软、硬件平台上的运行效率，在微观上则表现为对 MPI 各个通信原语进行测试，最后给出其通信的带宽值[9]。

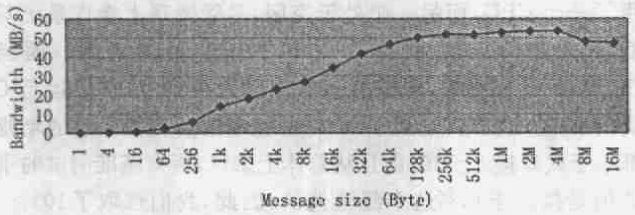

图4 PingPong 基准在 Myrinet 网络上的测试结果

图4是在 Myrinet 高速网上运行 PMB 中 PingPong[9]这个 benchmark 的测试结果，从该图可以看出，通讯包的大小直接影响着通讯带宽值，在我们这个系统中，当包的大小达到 4MB 时通讯带宽达到了最大值，此后便开始下降。我们得到的最大带宽值是54MB/s(432Mbps)，这个结果的获得应该综合以下的软、硬件因素来加以解释：我们系统中的 PCI 总线是32位，频率为33MHZ，因此，理论上插在 PCI 总线上的网卡所能达到的最大通信带宽为132MB/s，我们继而又使用了 Myricom 公司提供的 gmdebug[8]工具进行测试，发现 PCI 总线读、写数据的速率分别为76MB/s 和 85MB/s；而使用 MPICH-GM 通信要比直接使用 Myrinet 网卡通信慢上 15MB/s[9]，所以，我们使用 MPICH-GM 所能达到的系统带宽峰值为61MB/s。由以上分析，我们测试得到的54MB/s 是一个正常工作的结果。从这组测试同样也可以看出，对 MPI 通信原语的性能测试完全能够折射出整个系统通信性能！另外，PingPong 作为一个最经典的点对点通信的基准程序，通信时不受到其他外在条件的干扰，因此，它所能达到的最大带宽值也必定高于其他基准中 MPI 通信原语所能达到的最大带宽值(有关这一点，在我们其后的测试值中已经得到体现)。由此，我们可以得出，基于 MPI 程序的 Myrinet 网络轻载上限为15%。我们计算轻载上限的方法为：

Light_Load＝(Max_bandwidth－Max_mpibandwidth)/Max_bandwidth

各参数的说明如下：

Max_bandwidth ：最大系统通信带宽；

Max_mpibandwidth ：MPI 通信原语所能使用的最大带宽。

表3 PMB MPI-1中基准在 Myrinet 网络上的测试结果

PingPong	PingPing	Sendrecv	Exchange	Allreduce	Reduce_scatter
25.53 MB/s	14.89 MB/s	14.91 MB/s	7.73 MB/s	11.43 MB/s	20.93 MB/s
Allgatherv	Alltoall	Bcast	Reduce	All_gather	
11.24 MB/s	13.02 MB/s	19.75 MB/s	5.98 MB/s	13.09 MB/s	

表3给出了在 Myrinet 高速网上的各个基准(包括所有点对点通信原语和组通信原语)的测试结果，所有给出的结果值都是基于不同大小的包所得到的一组平均值，我们使用所有这些结果来得到 Myrinet 网络上重载的下限值为81%，网络重载下限值的计算方法如下：

Heavy_Load＝$(\Sigma(\Sigma BW_i \times F_i) \times F_j)$/ Max_bandwidth

各参数说明如下：

BW_i：第 i 个大小的数据包传输时获得的通讯带宽($1 \leqslant i \leqslant 24$)；

F_i：第 i 个大小的数据包被传输的概率。(这里，$F_1 = \cdots = F_i = \cdots = F_{max}$)；

F_j：第 j 个基准中 mpi 通信原语的使用概率。(这里，$F_1 = \cdots = F_j = \cdots = F_{max}$，$1 \leqslant j \leqslant 11$)；

Max_bandwidth：最大系统通信带宽。

由此，我们可以得出，在基于 MPICH-GM1.2.5..10的自强2000系统上，当网络带宽利用率处在0～15%内时为轻载，在15%～81%内为中等，重载区域则是81%～100% 。

结论及展望 高性能集群系统中的资源分配是关系到整个集群系统运行效率的一个关键问题，负载均衡是进行资源分配时必须考虑的一个重要因素。如何来识别负载是实现负载均衡的最基本也是最关键的一步。我们在这里分别讨论了计算资源负载和通信资源负载识别和量化的一种方法，在计算资源负载的量化中，我们通过综合分析在不同 CPU%负载环境下计算密集型和通讯密集型两类任务的平均响应时间，来得到最终的量化结果。而对于通信资源负载的量化，我们建议综合考虑通信的软、硬件因素，并通过测试处在通信最顶层的通信原语的通信效率来得出量化结果。此外，我们还应该看到，在一个集群系统中，经常同时存在多个通信中间件，因此，在进行通信资源负载的量化时，如何综合考虑到各个通信中间件的通信效率，便成为我们下一步工作的主要方向。

参 考 文 献

1 鞠九滨，杨鲲，徐高潮. 使用资源利用率作为负载平衡系统的负载指标. 软件学报，1996，7(4)：238～243
2 温钰洪，王鼎兴，沈美明. 一种同构机群系统中的处理机分配算法. 软件学报，1997，8(3)：161～169
3 叶庆华，肖利明，祝明发. 机群作业管理系统的评价体系
4 林成江，李三立. 一种可适应的分布式动态负载平衡策略及其仿真. 计算机学报，1995，18(10)：721～729
5 齐红，鞠九滨. 工作站网络上协作任务的调度. 软件学报，1998，9(1)：14～17
6 廖湘科. 网络并行计算中的负载平衡. 小型微型计算机系统，1995，16(9)：32～33
7 NPB 2.4 http://www.nas.nasa.gov/Software/NPB
8 MPICH-GM-1.2.5..10，gm-1.5.2 http://www.myri.com
9 PMB MPI-1 http://www.pallas.com
10 MPI Forum. MPI: A message-passing interface standard. International Journal of Supercomputing Applications and High Performance Computing，1994，8(3-4)