文章编号:1007 – 6492(2001)01 – 0006 – 04

构建高性能集群计算机系统的关键技术

王文义,张 影

(郑州工业大学电气信息工程学院,河南 郑州 450002)

摘 要:高性能计算机技术是衡量一个国家科技水平及综合国力的重要标志之一,目前世界上一些发达国家都在争相投入巨额资金对它进行开发和研究.PC 集群计算机就是最廉价的高性能计算机.着重讨论了集群计算机系统构建中的一些关键技术,如可扩展性、可用性、资源管理、负载平衡和并行程序设计环境等,并给出了一个基于 MPI 环境的并行程序设计实例,同时,也根据集群系统的特点提出了它在不同领域中的实用意义.

关键词:集群计算机系统;并行计算;可用性;可扩展性

中图分类号:TP 303 文献标识码:A

1 集群计算机系统

集群计算机是指利用高速通信网络将一组高档工作站或 PC 按某种结构连接起来,在并行程序设计及可视化人机交互集成开发环境支持下,统一调度,协调处理,实现高效并行处理的系统.由于集群计算机具有投资风险小,可扩展性好,可继承现有软硬件资源和开发周期短、容易编程等突出特点[1],目前已很快成为并行处理的热点和主流,据专家预测"未来的高性能计算机和超级服务器都将基于集群结构".

集群系统中的结点可以按 3 种方式实现连接:

(1) 无共享方式.指结点之间通过 I/O 总线连接,在大多数集群中都采用这种方式.

(2) 共享磁盘方式. 较小规模的商用性集群常常采用这种连接方式,其优点是当某个结点出现故障时, 其它结点可以代替其工作.

(3) 共享存储器方式. 这是一种新型的连接方式. 在其结构中,互连系统与每个结点中的存储总线相连,而在其它两种结构中,互连系统则是与结点的 I/O 总线相连.

集群中的结点数越多,则系统的整体处理能力也就越强,但结点数的增多受限于消息传递的通信速度和容量.如果采用 16 端口的 100MB/s 的快速以太网交换机作为网络互连,由于集群的结点连接是独占端口,所以独享 100MB/s 带宽,对该集群系统,它的通信容量为(100 MB/s × 16)/2 = 800 MB/s,其中除以 2 是因为通信端口总是成对工作的.

对于一个理想集群系统的体系结构,可用的结点有工作站、PC 机、SMP 服务器,甚至超级计算机.结点的操作系统是多用户、多任务和多线程的系统,如 Linux 等.结点可以是同构的,也可以是异构的.其中可用性基础设施层提供高可用性服务,单一系统映像基础设施层提供单一系统映像服务,最上层的 3 类子系统则用来支持该集群系统的可用性.

2 构建集群系统的关键技术

不同集群系统构建的难易程度也不同.对于廉价的集群系统,仅需将一定数量的高档 PC 机通过快速以太网进行互连,并辅之以某些相关的免费自由软件,如 Linux,PVM,MPI 等,即可得到一个性价比很不错的并行计算环境.而对于精心设计的高性能集群系统,则还需要着重考虑其它一些问题.

2.1 可扩展性及其设计原理

如果能够通过增加系统资源以满足不断增长的对性能和功能的要求,或者能够通过减少系统

收稿日期:2000 – 11 – 10;修订日期:2001 – 01 – 10

作者简介:王文义(1947 –),男,河南省洛阳市人,郑州工业大学教授,主要从事算法分析和并行处理方面的研究.

资源以降低成本,则称这样的计算机系统是可扩展的.一个系统的可扩展性包含性能和功能、成本伸缩、可兼容性等几个方面[2,3].

系统的可扩展性一般是指:

(1) 资源可扩展性. 是指通过增加系统规模(即处理器数)、投入更多存储部件(高速缓存、主存、磁盘)以及增加软件等方法,使系统具有更高性能或更多功能.

(2) 应用可扩展性.要充分开发可扩展并行计算机的能力,应用程序也必须是可扩展的,即,当相同程序在一个可扩展系统上运行时,其性能也能够随系统规模扩大而成比例地得到改进.该性能可以用机器规模和问题规模的可扩展性来进行度量.

(3) 技术可扩展性.是指系统能够适应技术环境改变的程度.它可进一步分为:代可扩展性、空间可扩展性以及异构可扩展性.

设计可扩展高性能计算机是一个复杂的工程过程,它大致包括 4 个设计原理:

(1) 独立原理.该原理要求我们应努力使系统中的各个组成部分相互独立.如果无法达到要求,则应尽量使相关程度减至最小并使相关性尽量清晰.这里的组成部分包括硬、软件两方面.采用独立原理的一个好处是使独立扩展(增量扩展)成为可能;另一好处是使异构可扩展性成为可能.

(2) 平衡设计原理.该原理要求我们应努力最小化任何性能瓶颈.应避免不平衡系统的设计,因为在这种系统中,一个慢速的部件将会导致整个系统性能下降,即使其它部件的速度再高也无济于事.此外,还应避免单点失效,即一个部件的失效将会引起整个系统崩溃.

(3) 可扩展性设计原理.即在设计一个可扩展系统时,应该从一开始就将可扩展性作为主要目标,而不是设计完成后再来考虑它.可扩展性设计的两种流行方法是过渡设计和向后兼容性设计.

(4) 时延隐藏原理.是指利用计算来隐藏通信时延,也就是说,能够保证即使是在长时延不可避免的情况下系统也能达到高性能.其基本思想之一是使计算和通信在时间上重叠.可以通过 4 种互补的方法进行时延隐藏:预取技术、分布式一致性高速缓存、非严格的存储器一致性模型、多线程处理器.

2.2　可用性和可用性技术

在设计健壮、高可用的系统时,必需同时考虑可靠性、可用性及可维护性 3 个因素.而其中又以可用性标准最为重要,它同时结合了可靠性和可维护性两个概念.

2.2.1　可用性概念

系统的可靠性可表示为平均无故障时间(MTTF:mean time to failure),即在系统或其部件发生故障前正常运行的平均时间.可维护性表示为平均故障修复时间(MTTR:mean time to repair),即用于修复系统和在修复后恢复正常工作状态所用的平均时间.系统的可用性则可定义为

$$\text{可用性} = \text{MTTF}/(\text{MTTF} + \text{MTTR}).$$

2.2.2　可用性技术

由可用性定义可知,提高系统可用性的基本方法有两种:增加 MTTF 或减少 MTTR.如今工作站的 MTTF 范围已经可以达到从几百小时到几千小时.但要再进一步提高 MTTF 将非常困难且开销很大.

多结点集群的 MTTF 要低于一个工作站的 MTTF,所以它比工作站发生故障的可能性要大.然而,如果能迅速处理这些故障,即减少系统的 MTTR,也同样可以提高系统的可用性.适用于集群系统的可用性技术主要有以下几个:

(1) 相互独立的冗余设备. 改善任何系统可用性的一个重要技术是使用冗余部件.当一个主要部件发生故障时,由另一个备用部件继续提供服务.此外主要部件和备用部件之间必须相互隔离,使得它们不会因为同一个原因而发生故障.

(2) 故障接管. 对于现在的商用集群来说,故障接管可能是最重要的性能需求.一个部件发生故障时,该技术允许系统的余留部分能继续提供原来由故障部件提供的服务.

(3) 恢复技术. 指为了接管一个已发生故障部件的工作负载所要做的动作.有后向和前向两种恢复技术.对前者,周期地为运行在集群中的进程在稳定存储设备中保存它的一个一致状态(即检查点).发生故障后,系统重组以与故障部件相隔离,恢复前一个检查点,然后继续正常的操作,整个过程称为卷回.在独立于应用程序的可移植方式下后向恢复较容易实现,并已被广泛运用.然而,卷回过程要有较大的时间开销,这在实时系统中是不能容忍的,这时就要考虑使用前向恢复技术.这种技术要求系统不是卷回到故障前的某个检查点而是利用故障诊断信息去重构一个有效的

系统状态,并继续执行下去.前向恢复将依赖于应用程序并且可能还需要额外的硬件设备支持.

2.3 高效的通信系统

通信子系统是并行计算机系统的重要组成部分,它完成系统中各结点之间的数据传递功能,因此通信性能的好坏将直接影响到并行计算的加速比和效率.这是因为并行计算时间是由各结点的CPU时间和结点间数据通信时间两部分组成,如果通信时间所占比例过大,则必然会使得并行计算的加速比下降,从而导致整个系统的效率下降.由于一般的集群系统往往是通过普通LAN互连而成,结点之间采用TCP/IP协议进行通信,所以存在低带宽和高延迟的问题.针对第一个问题的解决办法是采用新型高速网络如快速以太网,ATM,Myrinet等,来提高网络带宽.传统TCP/IP协议的多层次结构使得复杂的缓冲管理带来了很大的网络延迟和操作系统的额外开销.相应的解决办法是,在用户空间实现通信协议、精简通信协议、采用Active Message通信机制.前两种方法是针对传统通信协议在实现方法上进行的改进,而后一种方法则是一种全新的通信机制,能够更为有效地提高通信系统的性能.

2.4 并行程序设计环境

广义地说,并行程序设计环境应包括硬件平台、操作系统和并行程序语言、编程、编译、调试及性能分析工具等,狭义的并行程序设计环境则仅指系统核心之上的工具软件部分.作为一个并行程序的支撑境,至少应包括:① 并行语言支持或并行操作库函数支持;② 一种或多种并行编程模型.

我们知道,集群系统各结点间连接结构的区别取决于有无共享存储器的存在.如果系统中各结点间没有共享内存支持而只是通过消息传递机制来实现数据通信,那么消息传递就成为并行程序设计环境构造的基础.这种环境现在常用的有[4]:PVM,MPI,EXPRESS,Linda等.对于具有共享存储器的集群系统,则应采用共享变量模型来进行并行编程.需要注意的是,在前一种集群系统上也可以采用共享变量的并行编程模型,这时需要使用一种称为虚拟共享存储器的技术,利用它在基于分布存储器的集群系统中,实现物理上分布但逻辑上共享的存储系统.相应的支撑软件有ThreadMarks DSM,Midway DSM等[5].

利用MPI并行程序设计环境计算圆周率π的C++程序:

```
#include <math.h>
#include "mpi.h"
int main(int argc,char *argv[])
{
int n,rank,size, i;
double PI25DT = 3.14159265358979323846 2643;
double mypi,pi,h,sum,x;
MPI::Init(argc, argv);
size = MPI::COMM_WORLD.Get_size();
rank = MPI::COMM_WORLD.Get_rank();
While(1) {
if (rank == 0) {
cout << "Enter the number of intervals:
(0 quits)" << endl;
cin >> n;
}
MPI::COMM_WORLD.Bcast(&n, 1, MPI::
INT,0);
if (n == 0)
Break;
else {
h = 1.0/(double) n;
sum = 0.0;
for (i = rank + 1,i <= n;i += size) {
x = h * ((double) i - 0.5);
sum += (4.0/(1.0 + x * x));
}
mypi = h * sum;
MPI::COMM_WORLD.Reduce(&mypi,&pi,1,
MPI::DOUBLE,MPI::SUM,0);
if(rank == 0)
cout << "pi is approximately" <<pi
<< ",Error is" <<fabs(pi - PI25DT)
<< endl;
}
}
MPI::Finalize();
Return 0;
}
```

2.5 资源管理与负载平衡

如何有效地管理系统中的所有资源是集群系统的一个非常重要的方面,常用的并行编程环境PVM,MPI等对这方面的支持都比较弱,仅提供了统一的虚拟机.主要原因是结点的操作系统是单机系统,并不提供全局服务支持,同时也缺少有效

的全局共享方法.因此,就有必要在结点操作系统和并行编程环境之间加入一些中间件,即所谓的集群操作系统,来解决对系统中所有资源的调度,其中包括组调度、资源分配和并行文件系统等.

负载平衡也是并行处理中的一个重要问题,其解决的好坏将直接影响到系统的性能.负载平衡技术的核心是调度算法,即将各个任务比较均衡地分布到不同的处理结点进行并行处理以使各结点的利用率达到最大.除此之外,在设计负载平衡系统时,还需要考虑诸如决策时机、调度系统模式、负载指标的设计与收集、负载调度策略等问题.比较成熟的负载平衡系统有美国 Wisconsin – Madison 大学的 Condor 系统和加拿大 Platform 公司的 LSF 系统.它们的特点是只需对原有系统稍加改动,即可使之与并行程序设计环境结合起来,提供负载平衡功能.

3　结束语

集群计算机系统作为当前世界上并行处理的热点和主流,具有许多其它系统不可替代的优势:性价比高、可扩展性好、高可用性和高能用性.尤其是 PC 并行集群系统以它系统开发周期短、用户投资风险小、节约系统资源、用户编程方便等优点,非常适合我国国情,它的构建将给我国各行各业提供极为廉价的高性能并行计算资源,所以对我国的高性能科学计算、商业领域数据处理、互联网应用以及教育事业发展等都将具有重要而深远的意义.

参考文献:

[1] 黄　恺,徐志伟.可扩展并行计算技术、结构与编程[M].北京 机械工业出版社,2000.298 – 331.

[2] BUYYA Rajkumar. High Performance Cluster Computing: Architectures and Systems[M]. Englewood Ciffs: Prentice – Hill Inc,1999.409 – 433.

[3] 郑纬民,汤志忠.计算机系统结构[M].北京:清华大学出版社,1998.541 – 555.

[4] DIETZ Hank. Linux Parallel Processing HOWTO. http://yara.ecn.purdue.edu/pplinux/,1998 – 08 – 23.

[5] HWANG Kai. Advanced Computer Architecture[M]. New York: McGraw – Hill book Co,1993.

The Key Technologies to Deploy a High Performance Cluster Computer System

WANG Wen – yi, ZHANG　Ying

(College of Electrical & Information Engineering, Zhengzhou University of Technology, Zhengzhou 450002, China)

Abstract: High – performance computer technique is one of the important signs of iudging a country's science and technical level and synthetical national power. Now, some developed countries in world are positively investing huge funds for its development and study. PC cluster system is the cheappest one. The paper emphatically discusses some key technologies used for deploying a cluster system, such as scalability, availability, resource management, load balancing and parallel programming environment etc, and it also gives a MPI, parallel computing instance. At the same time, according to characteristics of a cluster system, it presents the practical significance of such a system in diverse fields.

Key words: cluster computer system; parallel computing; availability; scalability