实时集群系统设计与性能分析

王 丽[1],白 欣[2,3]
WANG Li[1],BAI Xin[2,3]

1.空军工程大学 工程学院,西安 710038
2.中国人民解放军 94569 部队,济南 250023
3.南京理工大学 弹道国防重点实验室,南京 210094
1.Engineering College,Air Force Engineering University,Xi'an 710038,China
2.Army 94569 of PLA,Ji'nan 250023,China
3.Nanjing Science and Engineering University,Nanjing 210094,China
E-mail:baix76@163.com

WANG Li,BAI Xin.Design and performance analysis of real-time cluster based on M/M/N model.Computer Engineering and Applications,2007,43(18):120-122.

Abstract: According to the demand of high availability and real-time characteristics of real-time cluster system,and with regard to the effect of the capability of cluster network transmission on the system real-time performance,this paper gives a design and construction project of a redundant real-time cluster system with high availability,discusses the realization of systematic parallel computing and the scheduling strategy of redundant group.Then,this paper models the M/M/N queueing model of the system,analyses the performance according as the model.The practical test proves that this system possesses the comparably high availability and real-time characteristics,and it can be used as the platform of periodical and high intensity multi-source floating-point information processing in the C2 area.

Key words: real-time cluster;availability;redundant;parallel computing;scheduling;M/M/N model;queueing theory;Command and Control(C2)

摘 要:设计和构建了一个高可用性冗余实时集群系统,讨论了系统并行计算的实现和冗余机组调度策略,建立了系统的M/M/N排队论模型,并依据此模型对系统进行了性能分析。经测试证明,系统具有较高的可用性和实时性,可作为周期性高强度多源浮点信息处理平台,用于军事指挥控制等实时性要求较高的领域。

关键词:实时集群;可用性;冗余;并行计算;调度;M/M/N 模型;排队论;指挥控制
文章编号:1002-8331(2007)18-0120-03 **文献标识码**:A **中图分类号**:TP302

1 引言

集群计算机系统是利用高速互联网络将一组工作站或 PC 计算机按照某种结构连接起来,在并行系统软件支持下,统一调度,协调处理,实现高效并行处理的系统。随着现代技术的发展和各种计算机部件的商品化,集群系统成为吸引人的并行计算与处理应用工具。集群系统目前主要应用于大型数据库、并行科学计算以及 Web 服务器[1,3,5]。实时集群系统有别于一般的事务处理和科学计算集群系统,它除了要发挥一般集群系统的并行计算能力外,还需要足够的系统反应时间。另外,实时集群系统应当具有较高的可用性,使得当运行实时任务的处理机出现故障时,系统仍能确保任务的正常完成[8,9]。文献[6,8]建立了一个双工集群系统模型, 并阐述了系统实现中的关键技术;文献[7]提出了轮转式任务调度策略,从直观上说明其算法提高了 CPU 利用率;文献[9]分析了硬实时集群系统中独立周期任务组和具有简单前趋约束关系的任务组的调度,但均未考虑集群系统的网络传输性能对系统实时性能的影响。根据实时集群系统的高可用性和实时性要求,文章设计和构建了一个高可用性冗余实时集群系统,讨论了系统并行计算的实现和冗余机组调度策略;最后,建立了系统的 M/M/N 排队论模型,并依据此模型对系统进行了性能分析。经实际测试证明,系统具有较高的可用性和实时性。

2 系统体系结构

高可用性冗余实时集群系统体系结构如图 1 所示。在系统中,为提高系统的可靠性和可用性,消除单一故障点,采用双机热备份冗余技术,即集群 A 和集群 B 互为备份。每个集群由管理员控制台(控制中心)、通信服务器(CS)、数据库服务器(DS)、计算单元(CU)及交换、互连网络设备组成。操作系统为内核经

基金项目:国家部委重点工程项目;国家科技创新基金资助项目(No.01C26226111003)。
作者简介:王丽(1975-),女,博士研究生,讲师,主要研究方向为航空武器系统、复杂系统故障诊断;白欣(1976-),男,博士,工程师,主要研究方向为高可用性实时集群系统与并行计算、通信保密与抗干扰技术,当前感兴趣的领域为制导弹药研究。

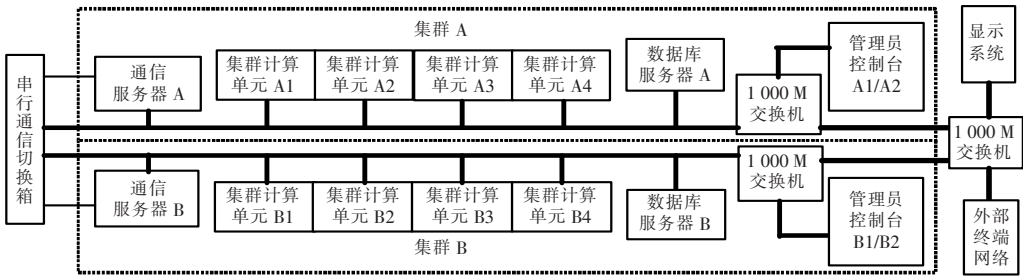

图 1　高可用性冗余实时集群系统

过改造的中软 Linux3.0，控制中心和各计算节点采用浪潮 NF300 服务器，并通过 1 000 M 交换机 Cisco 4912 和 1 000 M 交换机 Cisco 6009 互连构成。为避免集群单元硬件资源的消耗,并便于管理,除了管理员控制台外,集群各节点均为无头服务器。管理员控制台作为系统控制中心,负责集群系统各单元的节点信息收集、共享负载任务调度和集群管理的人机交互;通信服务器负责接收来自外部终端网络系统的原始数据；数据库服务器负责记录计算出的结果数据；计算单元则根据管理员控制台发来的任务处理信息,完成数据处理。

3　并行计算的实现

系统软件结构如图 2 所示。集群计算机是多处理器并行计算机系统的延伸。对于并行计算机而言,系统处理能力和效率直接与所执行的任务特征相关。对于并行度高的应用任务,并行计算机能够充分发挥多处理器的计算能力，实现高性能计算;对于并行度低的应用任务,并行计算机无法将任务拆分,以发挥多处理器的计算潜力。在集群系统中,计算单元间过度的数据(消息)依赖,也会使集群系统性能下降。

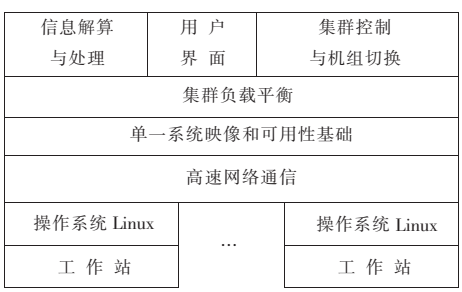

图 2　实时集群系统软件结构

集群作为高性能计算机系统,必须提供“单一系统映像”功能。在应用中,本实时集群系统中各节点驻留相同的系统应用软件,具有“单一系统映像”功能。各节点首先读取集群系统配置文件,通过与本机静态 IP 地址相比较,判读出节点属性,然后转去执行不同的功能软件模块。

集群系统在统一的高精度时统信号同步下工作。系统启动后,各节点广播“上线”(upline)信号,管理员控制台 A 和 B 则广播“心跳”(heartbeat)信号,并竞争产生主控制台(MC),主控制台所在的集群成为工作集群,备份控制台(BC)监测主控制台的“心跳”信号,在两个时统周期内主控“心跳”失常的情况下,则夺取主控制台的主控身份。在每个时统周期到来时,各节点向主控制台发送节点硬件和系统资源情况,主控制台根据各节点资源信息，在规定的时间内检测出集群系统的失效节点,从而调整负载平衡和任务分配软件,将失效节点的数据处理任务移交给其他可用节点完成。

对实时高性能集群计算机系统而言,建立满足系统实时性要求的负载平衡机制非常重要。实时集群计算机的特点是,有足够的冗余计算单元以保证系统可用性;足够的内存资源以确保实时应用程序驻留在内存而不会被交换到硬盘上。根据这一特点,通过建立、维护任务分配表的方法,所有计算节点读取任务分配表,确定本节点应处理的数据、处理方式、结果消息去向等信息,来实现实时集群系统的动态负载平衡。文献[9]具体论述了基于任务分配表的负载平衡机制。

4　冗余机组调度策略

为了保证系统的高可用性，系统采用双机热备份方式,每个集群计算机有自己的控制台,控制着集群计算的正常进行和冗余机组调度。调度通过驻留在控制台上的守护进程在每一个时统周期中检查失效节点完成。

主控在线时,在每一个时统周期中需判断的条件是:(1)通信服务器失效;(2)数据库服务器失效;(3)计算节点失效,指有效计算节点数少于系统设定的最低要求值。机组冗余调度策略伪码描述如下:

Step 1 等待时统同步信号的到来;

Step 2 收集来自各节点的工作状态及负载信息,并填入相应的数据结构;

Step 3 集群机组冗余调度条件的判断:

if ((主控制台发出机组冗余调度命令‖通信服务器失效‖数据库服务器失效‖计算节点失效)&&(备份机组工作正常))

调度备份机组,并拷贝相关数据;

else

继续工作。

5　系统 M/M/N 模型与性能分析

排队论(Queueing Theory)在计算机网络和计算机系统的性能评价中占有相当重要的地位。运用经典的肯达尔模型 A/B/n/S/Z,不难分析,文中研究的集群系统为多服务员系统,可建立如图 3 所示的 M/M/N 排队论模型。

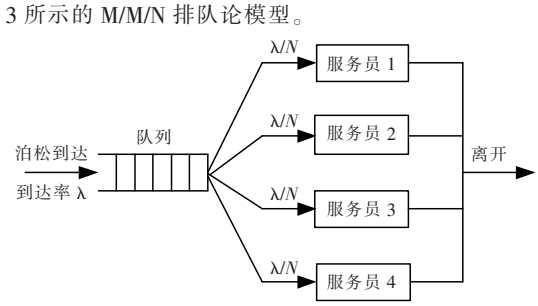

图 3　集群系统的 M/M/N 模型图

其中各符号的含义如下:

M:第一个 M 用于描述到达,表示泊松到达过程,到达时间间隔符合指数分布;第二个 M 用于描述服务,指具有指数分布

的服务时间。指数分布具有马尔可夫特性或称无记忆特性。

N:表示服务员的数目。此系统中 N=4。

所有服务员共享一个队列,如果一个顾客到达而至少有一个服务员是可以服务的,那么这个顾客就立即被交给服务员。如果所有服务员都忙,那么队列就开始形成。

队列分析的基本任务是:给出如下输入信息(概率分布):

(1)到达速率 λ

(2)服务速率 μ

求出如下输出信息(均值、方差):

(1)队列中顾客的数量 q

(2)排队时间 T_q

条件 在运用排队论评价系统的性能时,首先假设系统中顾客的到达速率服从泊松分布,即顾客的到达时间间隔服从均值为的负指数分布。顾客的服务时间是独立同分布的随机变量,且服从均值为 $1/\lambda$ 的负指数分布。服务规则为 $1/\mu$ 循环轮转 RR(Robin Round)方式。

该队列是一个生灭过程模型[2],其生灭速率为:

$$\lambda_k=\lambda,\quad k=0,1,2,\cdots$$

$$\mu_k=\begin{cases}k\mu & 0\leqslant k<N\\ N\mu & N\leqslant k\end{cases}$$

系统的状态图如图 4 所示。定义任一服务员的利用率 $\rho=\lambda/(N\mu)$,根据排队论理论[2]可以得到系统的平均顾客数量 q 为:$q=4\rho+\rho\dfrac{(4\rho)^4}{4!}\dfrac{\eta_0}{(1-\rho)^2}$,排队时间 T_q 为:$T_q=\dfrac{q}{\lambda}$。其中,$\eta_0=\left[\sum_{k=0}^{3}\dfrac{(4\rho)^k}{k!}+\dfrac{(4\rho)^4}{4!}\dfrac{1}{1-\rho}\right]^{-1}$,$\rho=\lambda/(4\mu)$。

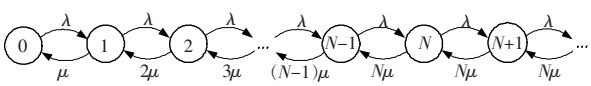

图 4 集群系统 M/M/N 队列的状态图

系统中顾客的平均服务时间 $1/\mu$ 与顾客在系统中花费的总时间 T_q 的仿真关系如图 5 所示,图中两条曲线分别表示 λ=40 和 λ=50 的情况。

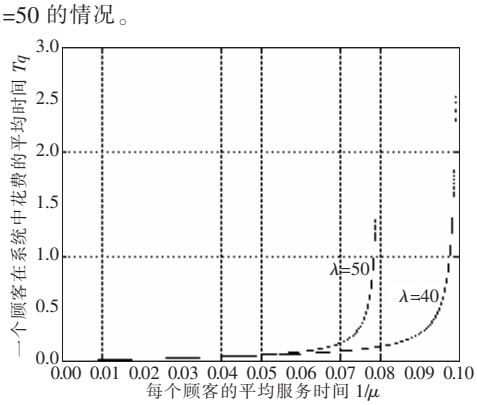

图 5 系统中顾客的服务时间与总时间的关系

由仿真结果可以看出:

(1)随着系统中顾客数量和任务强度的增加,系统中会出现拥塞情况的发生,从而顾客在系统中的延迟时间急剧增长,此时应依据 M/M/N 排队模型合理扩展集群系统的规模。

(2)针对同一顾客(任务),λ=40 条件下要比 λ=50 条件下在系统中花费的总时间要少,这也符合直观推理。

(3)系统正常工作即无拥塞情况发生时,每个顾客在系统中花费的时间不大于 100 ms,证明系统具有较高的实时处理能力,适合周期性、高强度多源信息处理系统。

6 结论

实时集群系统性能测试为:系统时延<100 ms,节点失效发现时延<70 ms,资源重组时延<100 ms,系统 I/O 处理时间<10 ms。可见,基于此 M/M/N 排队论模型的实时集群系统的事件响应能力达到了毫秒级细粒度。

(1)文章设计和构建了一个高可用性冗余实时集群系统,研究了集群系统并行计算的实现,提出了系统冗余机组调度策略,并给出了实现算法,满足了系统的高可用性要求;

(2)建立了集群系统的 M/M/N 队列模型,很好地表达了系统的竞争情况,从理论上验证了系统的实时性。对实时集群系统的研究和设计具有一定的指导意义。

文章基于 M/M/N 排队论模型,考虑集群系统的网络传输性能对系统实时性能的影响,所设计的冗余实时集群系统具有较高的可用性和实时性,适用于周期性、高强度、浮点多源信息处理系统,可用于复杂仿真、军事指挥控制、卫星测控、民用航空指挥控制和大型工业过程控制等实时性要求较高的领域,例如:空军指挥所指挥控制系统、武器仿真系统、飞行器模拟训练系统等。(收稿日期:2006 年 12 月)

参考文献:

[1] Kai H,Xu Zhi-wei.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.

[2] 林闯.计算机网络和计算机系统的性能评价[M].北京:清华大学出版社,2001.

[3] 陈国良.并行计算—结构,算法,编程[M].北京:高等教育出版社,1999.

[4] Chapin J,Chiu A,Hu R.PC Clusters for signal processing: an early prototype[J].IEEE,2000:525-529.

[5] Kim H J,Kim H S.Cost effective parallel processing for remote sensing applications[J].IEEE,1996:405-407.

[6] 白欣,左继章,向建军.基于 Linux 的实时指挥控制集群系统的方案研究[J].计算机工程与应用,2002,38(19):38-39.

[7] 孙英华,马军,许曰滨,等.多处理机容错系统中实时任务的轮转式调度算法[J].计算机工程与应用,2001,37(17):104-106.

[8] 杨文波,王志英,张春元,等.高性能分布式双工实时容错系统中的若干技术问题[J].小型微型计算机系统,2001,22(2):250-253.

[9] 毛羽刚,金士尧,张拥军.并行与分布硬实时系统的调度[J]. 计算机科学,1999,26(9):51-54.

[10] 白欣,宋博,左继章,等.测控集群系统随机 Petri 网模型与可用性分析[J].小型微型计算机系统,2005,26(6):979-982.