文章编号:1008-0570(2010)05-3-0006-02

高可用性系统技术研究

The Research on High availability Technology

(1.厦门大学;2.厦门理工学院) 许高攀[1,2] 曾文华[1]
XU Gao-pan ZENG Wen-hua

摘要: 主要从如何减少系统的平均修复时间 MTTR 和增加系统的平均无故障时间 MTTF 两方面来提高系统可用性。首先,根据可信性的定义公式提出提高可信性的措施,如自主运算、面向恢复计算等。最后,深入研究高可用性系统的关键技术,比如数据存储技术、系统软件技术、进程检查点和迁移技术、故障检测技术、冗余和备份技术。

关键词: 高可用性; 体系结构; 关键技术
中图分类号: TP302.1 **文献标识码:** A

Abstract: There were two ways to improve the availability, one of them is to decrease the mean time to repair (MTTR), the other is to increase the mean time to fail (MTTF). The measures of improving the availability were brought forward according to the formula of availability, taking the Autonomic Computing and Recovery-Oriented Computing for example. In the end, the key technologies of high availability were deeply studied, such as data storage, system software, process checkpoints, process transfering fault detecting, redundacy, backup technology and so on.

Key words: High Availability; Architecture; Key technology

技术创新

1 引言

在现代生活中,计算机系统被广泛地应用于各个方面。无论是在军事、金融、电信等关系到国计民生的关键性部门和行业,还是在平时的日常生活中,都广泛地使用计算机系统处理信息。随着计算机应用的不断深入,人们对计算机系统可用性(Availability)的要求越来越高。人们不仅希望能够保障关键业务数据信息的完整,而且希望网络应用能够不间断或者在最短的时间内自动恢复,这就是所谓的计算机系统的高可用性(High Availability)问题。

2 可用性问题的理论研究

从可用性的定义可以知,提高系统的可用性基本上有两种方法:增加 MTTF(Mean Time To Fail)或减少 MTTR(Mean Time To Repair)。增加 MTTF 要求增加系统的可靠性,而对于系统而言,当故障的产生难以进行有效的预测和消除时,通过快速故障恢复,降低平均修复时间(MTTR)也可以达到提高可用性的目的。如何减少系统恢复时间是提高系统可用性的一个重要课题。

考虑到计算机系统软硬件自身的错误在减少,由于人为因素带来的系统失效的情况成为主要原因,而这单靠系统结构方面的改善是无法解决的。因此研究者们把更大的注意力放在了提高系统的恢复能力上,希望能够提高计算机系统处理自身错误的能力。如 Jim Gray 提出的 Trouble-Free Systems 的概念,Butler Lampson 认为系统设计面临的挑战之一就是保持系统的总是可用,而且能够自适应环境的改变。John Hennessy 建议研究的目标应在可用性、可维护性、可扩展性上。IBM 公司提出了新的研究计划:自主运算(Autonomic Computing),把计算机系统看作一个可以自调节、自我管理、自我诊断的生物系统,其主要目标也是使计算机系统更加"聪明"而不是更加的快速。Dave Patterson、Kathy Kellick(UC Berkeley)、Armando Fox(Stanford)等领导的 Recovery-Oriented Computing(ROC)研究项目。他们认为硬件故障、软件 BUG、操作人员的误操作等都是要处理的存在的事实,而不是有要要解决的问题。ROC 更加关注于 MTTR 而不是 MTTF,通过减少系统的恢复时间来提供系统的高可用。同时考虑到管理人员大部分的工作都是在处理系统的失效,因此这也有助于减小 TCO (Total Cost of Ownership)。TCG 提出了 Trusted Computing 并制定了相应的规范,Trusted Computing 的核心是 TPM,它更多的是通过安全性来提高系统的可用性,防止系统被它人恶意篡改和使用。Dionysius Lardner 博士提出了 Dependable Computing,它主要侧重于通过冗余、NVP 等方式提高系统的可用性。2002 年 Bill Gates 提出 Trustworthy Computing,它与 Trusted Computing 类似也是通过安全性来提高系统的可用性,更多的是从微软企业自身的角度来思考问题,从操作系统上来提高安全性,从而实现高可用性。1993 年美国陆军学院的 Barnes 等人提出了 Survivability 的概念,它主要是从军事需求的角度来考虑的,当军事系统受损时,希望能够继续提供服务,它也牵涉到如何减少修复时间以提高系统可用性的问题。

3 高可用性系统的关键技术

对于可用性的问题,人们最初的策略是为了达到某种目标而针对具体的应用服务,例如程序设计者会想方设法地设计出健壮的应用程序,但是随着系统复杂性的提高和认知的加深,这种单一的方法无法真正达到目的。于是人们考虑如何在基础结构上保证高可用性,于是出现了磁盘镜像,RAID,以及集群技术等等,减少单点失效而保证高可用性,但是这并没有真正的解决问题,于是人们在思考如何合理计划,充分发挥现有技术的优势,并且能够融合即将出现的技术来共同组建一种高可用性的解决方案。计算机系统主要由软件、硬件和以软硬件为载体的数

许高攀: 讲师 博士生

据组成,在高可用性的研究领域也正是从这三个方面入手,加以综合考虑和运用来设计高可用性系统。

1)数据存储技术

现代的数据存储系统已经形成了融合文件存储服务和数据块存储服务的统一的存储网络, 可以为主机提高更好的数据服务。通过这种体系结构的变化,可以更为方便的实现数据的各种备份策略方法和数据的容错、容灾,提高数据的安全性能;可以方便地扩展存储设备, 提高存储系统的容量; 可以更有利于对存储系统的管理,提高可维护性;可以根据用户需求,采取有针对性的措施提供数据和文件服务,实现数据和文件服务的 QoS, 从而最终实现数据的高可用性。这些优势和特征显然是最初的数据存储系统所不具备的。数据存储系统现在正在朝着存储虚拟化的方向发展,试图为主机系统提供一个虚拟的、海量的存储池资源,其中可以根据需要容纳各种存储设备,从而更加方便数据的存储和管理,对可用性的保证也更为彻底。在存储领域还有一个很为重要的思想,就是借鉴自然进化的理论,设计进化的存储系统, 这种存储系统可以在不影响系统其他部件的情况下自我调节和更新,相对于以前的 RAID 技术、磁盘镜像技术,存储网络化、虚拟化和智能化方向的发展为高可用性系统的设计提供了更好的平台,数据的高可用性问题得到了更为彻底的解决。

2)系统级软件技术

目前高可用性的软件的研究主要涉及系统软件, 如文件系统,数据库系统中针对数据高可用性特性和需求的研究。举例来说, 现在许多文件系统中引入了日志或者记录的技术和数据库中事物处理的技术来保证系统数据的一致性和系统恢复的快速性, 典型的有 IBM 的 JFS 文件系统,ext3 文件系统,Veritas 的 VxFS 和 SGI 的 XFS 文件系统。通过对网络文件系统进行改造, 使之具有高性能,高可用性,可扩展的性能;这种文件系统除了具有分布式文件系统的特征以外,还充分利用了存储网络的技术; 用户需要对数据进行操作时, 实现访问元数据服务器来获得具体数据在存储网络中的位置信息, 然后直接从存储网络中获取所需的数据并对其进行操作。这样将元数据访问和具体数据的访问分开,从而充分利用存储网络的优势,保证数据的高可用性, 其中元数据可以通过元数据服务器的冗余和在文件系统中加入日志等特性保证元数据的高可用性, 而且可以提高数据的访问性能和数据的共享。

除了文件系统以外,人们还试图设计一种高可用的操作系统。对于现代通用的操作系统 UNIX、Windows 之类,都不能满足高可用的需求,于是人们开始设计一种全新的操作系统或者虚拟机系统来保证高可用性,增加对检查点和进程迁移恢复的支持。

3)进程检查点和迁移技术

在高可用性系统的实现中,针对应用软件的故障恢复问题, 可以采用检查点和进程迁移的技术来解决。所谓检查点,就是在一个事务结束,另一个事务即将开始的时候,对系统状态的一次快照。检查点技术是高可用性、进程迁移、负载平衡、系统管理和升级以及许多其它应用的基础。检查点的关键是透明性,其作用对象是进程。进程是运行在操作系统上的单位实体,一个进程是一个复杂的登记信息和资源信息的结合体,包含进程 ID 和其它统计信息、寄存器集、地址空间以及诸如打开的文件这样的资源。在现代操作系统中,进程拥有自己的用户空间,进程需要和操作系统内核通信,以及进程间的相互通信,在某个时刻进程所拥有的各种资源和登记信息形成进程在这一时刻的状态,进程的检查点技术就是记录这个状态信息。进程迁移分为本地恢复和远程恢复的方法,进程迁移是以检查点为基础,具体实现是恢复到检查点时刻的进程状态,减少故障对用户的影响。检查点和进程迁移已经被应用到许多高可用性系统或者软件中, 是实现应用服务的主流技术,其实现方法可以分为应用级、虚拟机级和内核级。

4)故障检测技术

在对计算机系统研究的过程中人们发现系统故障是不可避免的,目前没有一种技术可以彻底消除故障,因此高可用性系统的设计是以减少故障出现的概率和恢复故障的时间为目标的,为了达到这样的目标,故障检测技术显得尤为重要。只有有效地探测系统失效并正确恢复才能达到提高系统可用性的目的。故障检测根据针对的个体不同采取的策略也不一样,对于软件和单个硬件可以采用 agent 技术来检测在运行的过程中故障的出现情况,而对于一个节点(完整的计算机系统)来说,需要采取心跳技术,通过集群中的其他节点来检测故障。故障检测的关键是透明性和故障通知的及时性。

5)冗余和备份技术

在高可用性系统中一个关键的思想就是冗余,不管是软件、硬件还是数据,都需要采取冗余的策略,从而在故障时可以及时恢复,否则一切高可用性的措施和方法都失去意义。在高可用性系统中,I/O 路径、存储系统、CPU、应用程序、节点服务器都需要采取冗余的策略,从而可以实现热备份。

对于关键性的应用程序和数据,一旦系统崩溃,备份的数据就成了唯一的希望。当然对数据进行备份,也是出于保证数据的安全性、对系统信息做历史记录、在灾难发生时恢复系统等多方面考虑的。备份分为冷备份和热备份两种。在进行冷备份时, 系统管理员要首先发出一个停机通知, 然后停止服务并断开服务器的网络连接。然后安装备份设备、开始备份。待备份完毕后再连接网络、启动服务。一旦出现故障,还可以从容地发出一个通知,停止服务并开始数据恢复,甚至可以重新安装系统。然而, 随着计算机系统涉及到越来越多的关键业务应用, 这种备份方式的局限性越来越明显。数据的高可用性意味着 7×24 的不间断服务,数据访问的连续性必须得到保证。服务器出现的故障应尽量避免在客户端体现出来。因此,在线的数据热备份成为一项基本的要求。热备份就是在用户和应用服务正在更新数据时,系统也可进行备份。

4 小结

本文作者创新点: 主要从可用性的公式:Availability=MTTF/(MTTR+MTTF), 分别从如何减少 MTTR 和增加 MTTF 两个方面研究如何提高系统的可用性。

参考文献

[1]R. Bhagwan, S. Savage, and G. M. Voelker. Understanding availability. In Proc. of IPTPS, 2003.

[2]王晶,季新生,朱云志.嵌入式系统高可用性应用软件设计. 微计算机信息. 2005 ,8-2: 42-44.

[3]Yuan-Shun Dai, Tom Marshall and Xiaohong Guan.Autonomic and Dependable Computing: Moving Towards a Model-Driven Approach. Journal of Computer Science 2 (6): 496-504, 2006.

[4]David Oppenheimer edal. Roc1: hardware support for recovery oriented computing. IEEE Transactions on computers, Vol. 51,No. 2, Feb. 2002.pages 100-107.

(下转第 5 页)

技术创新

预处理。采用向量空间模型来表示每个文本向量,文本向量的特征值采用 TFIDF 函数表示。然后通过人工和机器交互方式选择类关联词,并利用类关联词形成初始聚类中心。

为了评价算法的效率和类关联词的作用,设计了两个程序,一个在循环中采用类关联词来约束聚类过程(记为 CAW),另一个没有采用类关联词(记为 NOCAW)的约束,即每一个待确定类别文档要和所有聚类中心进行相似度比较。实验中 NOCAW 方法需要 6 次迭代,而 CAW 方法只需 4 次,CAW 方法提高了算法执行的效率。为了验证方法的正确率,从每年随机选取了 50 个案例文档,共选取 200 个文档,通过阅读案例文档手工进行了类标号标注。实验显示,CAW 的总的分类正确率是 90%(见表 1),而 NOCAW 的总的分类正确率是 89%,CAW 比 NOCAW 的分类正确率稍高一些,说明类关联词在采用该算法时对提高分类正确率是有效的。

表 1 CAW 的分类正确率

类别	手工标注数	预测数量/正确数量	类别	手工标注数	预测数量/正确数量
版权(著作权)	110	111/107	厂商名称	5	3/2
商标权	45	40/37	植物新品种	2	2/2
专利权	25	27/25	货源标记	1	2/0
制止不正当竞争	6	5/1	原产地名称		2/0
商业秘密	6	7/6	集成电路设计权		1/0

对比 CAW 和 NOCAW 的分类一致率,显示两种方法对 1158 篇文档的 1110 篇文档的分类结果是一致的,分类一致率高达 95.85%。因为两种方法的初始聚类中心是一样的,说明初始聚类中心对最终结果有较大的影响,而聚类算法迭代过程中类关联词的作用相对较小。

4 结束语

提出的算法可以利用类关联词和 K-Means 聚类算法实现对文本文档的分类,算法执行效率和算法分类正确率较高。类关联词确定的初始聚类中心对于最终分类结果影响较大,在迭代过程中的约束对于最终的分类结果影响较小。

本文作者创新点:在 K-Means 文本聚类算法中引进类关联词,形成初始聚类中心,在聚类过程中利用类关联词的监督作用,使聚类形成的簇与分类类别一一对应,从而达到利用聚类算法进行文本分类的目的。

参考文献

[1]Macqueen J. Some methods for classification and analysis of multivariate observations[C]. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.Berkeley:University of California Press, 1967.

[2]Inderjit S. Dhillon D S M. Concept decompositions for large sparse text data using clustering[J]. Machine Learning. 2001, 42(1): 143-175.

[3]索红光 王玉伟. 一种用于文本聚类的改进 k_means 算法[J]. 山东大学学报(理学版). 2008, 43(1): 60-64.

[4]Paul S. Bradley U M F. Refining initial points for k-means clustering[C]. Proceedings of the 15th International Conference on Machine Learning (ICML98), 1998.

[5]行小帅潘进,焦李成. 基于免疫规划的 K_means 聚类算法[J]. 计算机学报. 2003, 26(5): 605-610.

[6]杨丽华,戴齐,杨占华. 文本分类技术研究[J]. 微计算机信息. 2006, 22(5-3): 209-211.

作者简介:韩红旗(1971-),男(汉族),河南省洛阳市人,北京理工大学,博士生,讲师,主要从事数据挖掘,信息系统研究;朱东华,(1963-),男,福建省三明市人,博士生导师,主要研究领域为数据挖掘,科技评价;汪雪锋 (1977-),男,湖北省荆门市人,博士,讲师,主要研究领域为科技评价,信息系统。

Biography:HAN Hong-qi (1971-),Male (the Han nationality), Henan, Beijing Institute of Technology, Doctor, Lecturer,Management Science and Engineering, Research direction: Data Mining & Information System.

(100081 北京 北京理工大学管理与经济学院) 韩红旗 朱东华 汪雪锋

(450011 河南省郑州市 华北水利水电学院管理与经济学院) 韩红旗

(School of Management and Economics, Beijing Institute of Technology, Beijing, 100081, China) HAN Hong-qi ZHU Dong-hua WANG Xue-feng

(School of Management and Economics, North China University of Water Conservancy and Electric Power, Zhengzhou 450011, China) HAN Hong-qi

通讯地址:(100081 北京理工大学管理与经济学院) 韩红旗

(收稿日期:2009.06.26)(修稿日期:2009.09.26)

(上接第 7 页)

[5]朱岩;分布式关键任务系统高可用性研究。哈尔滨工程大学硕士学位论文,2006.

[6]A. Adya et al. FARSITE: Federated, Available, and Reliable Storage for an Incompletely Trusted Environment. In Proc. of OSDI, 2002.

[7]Roberto Gioiosa, edal. Transparent, Incremental Checkpointing at Kernel Level: a Foundation for Fault Tolerance for Parallel Computers. SC'05 November 12-18, 2005, Seattle, Washington.

[8]周国峰;高可用性系统中检查点技术的研究与实现。华中科技大学硕士学位论文,2004。

作者简介:许高攀(1976-):福建人,厦门理工学院讲师,厦门大学智能科学系博士生,研究方向为可信计算、软计算技术及应用;曾文华,江苏人,厦门大学软件学院教授、博导。

Biography:XU Gao-pan(1976-), male, Fujian Province,lecturer of Xiamen University of Technology, doctoral student of Xiamen University, research on the dependability computing and soft computing technology & application. Dr.

(361005 厦门大学智能科学系) 许高攀

(361005 厦门理工学院计算机系) 许高攀

(361005 厦门大学软件学院) 曾文华

(Dept. Of Cognitive Science&Technology, Xiamen University, 361005, China) XU Gao-pan

(Dept. Of Computer Science&Technology, Technology of Xiamen University, 361005, China) XU Gao-pan

(Software of Xiamen University,361005,China)

ZENG Wen-hua

通讯地址:(361005 厦门市思明区厦港不见天 2 号 103 室) 许高攀

(收稿日期:2009.06.02)(修稿日期:2009.09.02)