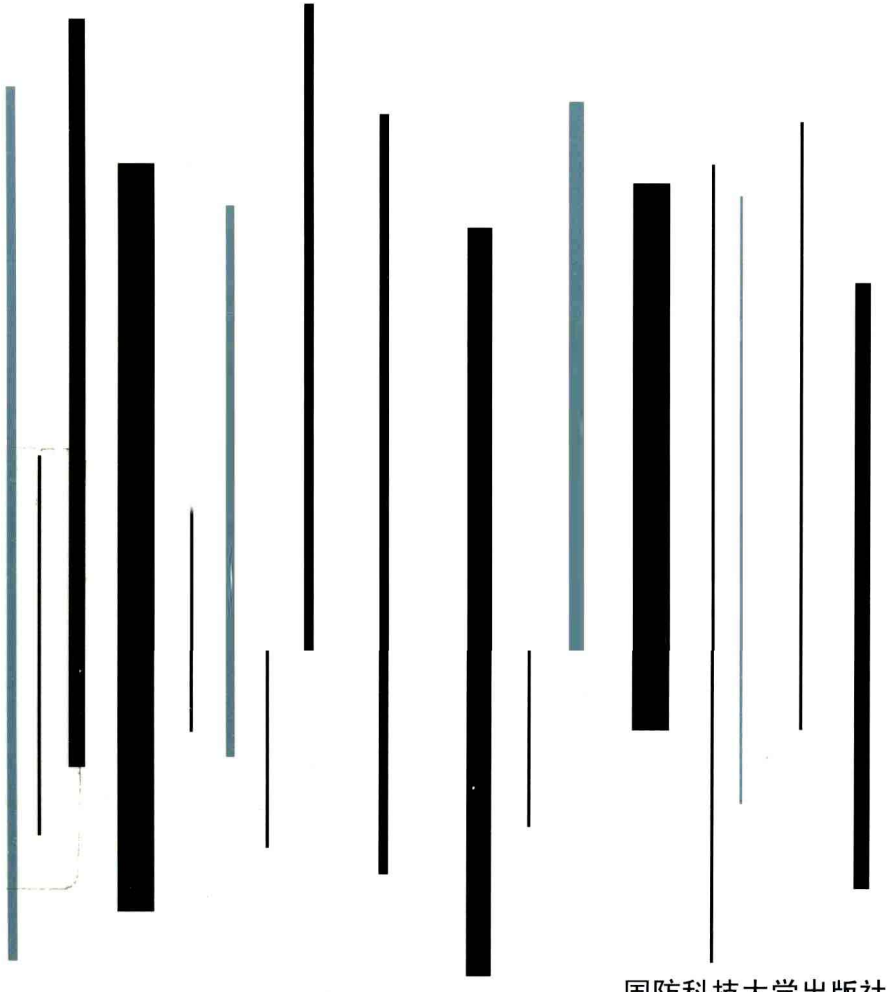
基于虚拟计算环境的内存共享技术

● 褚瑞　张一鸣　著 ●



国防科技大学出版社

基于虚拟计算环境的内存共享技术

褚　瑞　张一鸣

国防科技大学出版社
湖南·长沙

图书在版编目(CIP)数据

基于虚拟计算环境的内存共享技术/褚瑞,张一鸣著.—长沙:国防科技大学出版社,2010.12
ISBN 978-7-81099-810-9

Ⅰ.①基… Ⅱ.①褚… ②张… Ⅲ.①存贮器共享-技术 Ⅳ.①TP333

中国版本图书馆 CIP 数据核字(2010)第 205715 号

国防科技大学出版社出版发行
电话:(0731)84572640 邮政编码:410073
http://www.gfkdcbs.com
责任编辑:常春喜 责任校对:刘 梅
新华书店总店北京发行所经销
国防科技大学印刷厂印装
*
开本:850×1168 1/32 印张:6.75 字数:175 千
2010 年 12 月第 1 版第 1 次印刷 印数:1-600 册
ISBN 978-7-81099-810-9
定价:30.00 元

前　言

互联网资源的成长性、自治性和多样性，使得面向互联网的资源共享技术面临很大困难。虚拟计算环境是一种新型网络计算技术，它以互联网资源的自主化为基础，以按需聚合和自主协同为核心机制，在开放的网络基础设施之上，实现多种资源的共享与协同工作。

内存资源的共享是虚拟计算环境中一个重要而特殊的问题。一方面，虽然传统的网络内存技术能够有效共享集群中的内存资源，提高内存密集型或 I/O 密集型应用的性能，却不能适应松耦合的网络计算环境的特点；另一方面，虽然对数据、存储等资源的共享已有很成熟的研究，但内存资源又具有完全不同的特殊性质。本文主要借助虚拟计算环境的思想，试图把传统的网络内存技术扩展到基于公共网络的松耦合计算环境中，并分别针对内存资源的自主化方法、按需聚合与自主协同机制展开深入研究。

传统的网络内存技术通常面向紧耦合的集群环境，受可用内存容量限制较大，本文将内存资源共享的思想引入到松耦合的网络计算环境中，提出了基于虚拟计算环境的内存资源共享系统 iVCE/M，适用于具有大量小粒度、随机性的磁盘访问的应用。本文首先提出了 iVCE/M 的单一服务原则、单一

身份原则和指数探测原则，然后利用虚拟计算环境中资源自主化的思想，把 iVCE/M 中的所有节点分为五种，并定义其相互转换关系，及其交互的内存服务和代理服务机制。为适应资源的成长性、自治性和多样性，为各种节点分别设计了基本行为策略，并与虚拟计算环境的相应概念进行了对比。通过基于真实的气象应用程序运行状态的模拟，说明了 iVCE/M 在校园网等条件较好的网络环境下，能够有效提高内存密集型应用或 I/O 密集型应用的性能。

由于内存资源的特殊性，使传统的集中式、非结构化分布式和结构化分布式资源信息管理方法都很难适用。为体现虚拟计算环境所强调的无集中管理、无全局视图、根据实际需求动态获取资源信息的思想，本文结合内存资源共享的特殊性，提出了基于网络延迟的资源聚类机制，把 iVCE/M 中的节点划分为若干个组，确保在大概率情况下，只有同一组内的节点才需要进行内存资源的共享。从而有效降低了资源信息管理的问题规模。通过建立内存资源聚类的基本模型，给出了满足分布式要求的贪婪算法，并分析其不足之处；然后借鉴物理学中的力场和势能理论，建立了内存资源聚类的力场－势能模型，以考虑网络与磁盘延迟的多样性、动态性等多种因素；在相应的算法方面，通过定义节点的冲突集和对网络延迟的概率估计，提出了力场－势能聚类算法，并通过基于真实网络拓扑的模拟，对两种模型和算法的假阳性和假阴性错误、组内内存资源容量等重要指标进行了评估。

iVCE/M 的性能主要取决于内存页帧存取过程中的网络通

信开销，隐藏网络通信开销也是本文重点考虑的问题之一。预取技术在计算机系统中应用广泛，但由于受资源的限制等，难以对数据访问模式进行有效预测，对空间局部性较弱的应用预取效果有限。考虑虚拟计算环境中资源自主化的思想，本文提出了一种基于“推”模式的 iVCE/M 预取技术。内存节点不仅需要响应用户节点存取页帧的操作，还能自动分析用户节点在存取页帧过程中的模式，预测其后续操作，并主动把可能需要的页帧推向用户节点。借助数据挖掘领域中的序列模式挖掘技术，提出了相应的 I/O 模式预测算法，与一般的访问模式预测算法相比，其分析粒度更细，效果更好。本文还分析和证明了上述算法的时空开销及正确性，通过基于实际运行状态的模拟，对预取机制中的几个重要参数的影响进行了考察。

iVCE/M 的性能还可以从时间和空间两方面进行优化，在时间上，可以通过多个内存节点的并行传输，提高内存节点和用户节点之间的网络通信效率；在空间上，多个内存节点有时会保存大量完全相同的数据，可以将其进行合并，令 iVCE/M 保存更大的应用工作集。本文在时间方面，分析了传统的数据条块化技术、I/O 并行化技术，通过对网络传输进行建模，提出了内存页帧传输中时间相关的概念，并阐述了通过页帧交换形成自适应的并行传输的方法；在空间方面，提出了合并冗余页帧以保存更大的应用工作集的方法，由于各个内存节点是完全分布的，本文采用基于 Bloom Filter 技术的冗余检测机制与合并机制，其时间和空间开销都比较低，且

能够适应分布式计算的要求，具有很强的实用性。最后，采集了三种I/O密集型应用在不同配置下的实际运行状态，并据此进行了模拟评估。

本书由国家自然科学基金项目：基于内存资源的计算虚拟存储技术研究和国家重点基础研究发展计划（973）项目：虚拟计算环境聚合与协同机理研究资助出版。

作　者

2010.10.10

目　录

第一章 绪 论

随着计算技术和网络技术的飞速发展和广泛应用，两者之间逐渐融合而形成的网络计算技术，也在学术界和工业界的共同关注下得到了不断的进步[1]。从面向紧耦合、集中式、同构资源的集群计算(Cluster Computing)技术[2]，到近年来得到广泛关注的面向松耦合、广域分布式、异构资源的网格计算(Grid Computing)[3]和对等计算(Peer-to-Peer Computing)[4]技术，以及相关的企业计算(Enterprise Computing)[5]、普适计算(Pervasive Computing)[6]等技术，都可以看作是网络计算的分支。

为了使网络计算技术更好地适应互联网环境，从根本机制上解决目前互联网资源难以有效共享和综合利用的问题，本课题组在过去的研究工作中，提出了以按需聚合和自主协同为核心机制的面向互联网的虚拟计算环境(Internet-based Virtual Computing Environment，iVCE)的概念(简称虚拟计算环境)[7]。虚拟计算环境是一种新型网络计算技术，主要致力于多种互联网资源的共享与协同工作。其中，内存资源的有效共享，是虚拟计算环境中一个重要而特殊的问题。一方面，虚拟计算环境所面向的资源具有很强的成长性、自治性和多样性，令传统的内存共享技术难以适用；另一方面，内存资源又具有和数据、存储及其他资源不同的特性，使虚拟计算环境中的内存共享问题具有一定的挑战性。本书主要基于虚拟计算环境的总体思想与核心机理，对上述问题展开研究；而

本书的研究成果对于构造和谐的虚拟计算环境，也具有较大的理论意义和实践价值。

1.1 虚拟计算环境概述

互联网资源具有复杂的自然特性，从而使资源的有效共享和综合利用面临巨大的挑战。本课题组在对上述问题研究过程中，提出以网络资源的按需聚合和自主协同为核心，建立虚拟计算环境的思路[7]。虚拟计算环境试图在开放的网络基础设施之上，为终端用户或应用系统提供和谐、可信、透明的一体化服务。构建虚拟计算环境，对于释放互联网资源的巨大能力，提高国家在信息时代的综合国力和国际竞争力，促进人与网络的和谐发展都具有重要的战略意义。

本书的研究工作属于虚拟计算环境的重要组成部分，并充分体现了按需聚合与自主协同的思想。下文首先对虚拟计算环境的产生背景、基本概念及其核心机理进行概括性描述。

1.1.1 虚拟计算环境的背景

互联网是计算机技术与通信技术融合的产物。目前普遍认为，经过近 40 年的发展，互联网已成为信息社会的重要基础设施。近 10 年来，互联网在我国也得到了迅速的发展[8]，并以其强大的渗透力对传统行业带来了巨大的影响。互联网上汇集的各种资源，包括设备资源、信息资源和应用资源等，已成为国家资源的重要组成部分。随着国家信息化的推进，经济、行政、科研、教育、军事等各个领域都对互联网资源的共享和综合利用提出了迫切的需求。

从20世纪60年代，Internet的先驱Licklider教授提出和谐、安全、透明的网络计算环境这一目标开始[9]，在过去的40年里，伴随着网络技术的发展，共享网络资源的技术也在不断取得进步，同时带来新的挑战。从20世纪70年代至今一直在不断研究和开发的集群计算技术[2]，把一组通过高速局域网络连接的同构节点有效整合，形成单一的系统映象(Single System Image)；20世纪80年代，学术界提出了分布式操作系统和网络操作系统的思想[10]，试图通过扩展传统操作系统的资源管理机制，对网络上的资源实施有效管理；从20世纪90年代到21世纪伊始，企业计算、网格计算和对等计算等各种新型网络计算技术不断涌现。虽然上述技术具有不同的计算模型、关注不同的问题、适用不同模式的应用，但其基本思路主要都来源于传统资源管理机理在分布式环境中的拓展。传统计算机资源管理模式的成功是因为这类资源管理具有以下三个基本特点：一是可以明确被管资源边界；二是可以实施全局资源控制；三是可以建立统一的资源描述。然而，对于互联网资源管理，这三个基本特点不再存在，使得传统的计算机资源管理模式不适用于互联网资源，从而令互联网资源的有效共享与综合利用成为一个突出的、亟待解决的现实问题，必须在概念和方法上寻求新的突破。

概括来说，互联网资源具有成长性、自治性和多样性三个主要的自然特性。成长性是指互联网资源规模不断膨胀，关联关系不断变化的动态特性[11]。随着互联网覆盖地域的扩大，大量分布异构资源的更新与扩展，相应的资源信息也随之不断扩充和变化，从而导致难以获得完整一致的资源信息，资源的组织管理和共享困难；自治性是指互联网资源具有局部自治、自主决策的特性。很多资源面向特定的组织和个人，使得全局化的资源控制不再适用，需要围绕资源共享进行多组织之间的协同。而目前局部自治资源之

间普遍缺乏有效的协同能力，这种相对独立的状态影响了互联网资源综合效用的发挥；多样性是指互联网资源属性存在广泛差异的特性，如互联网上的资源包括计算、数据、信息、应用程序和服务等，甚至还包括各种仪器等物理资源，这些资源在类型、属性上等多方面都存在显著的差异，使得难以给出完整一致的资源描述，难以提供统一的资源访问接口，从而增加了资源建模和管理的困难。

综上所述，由于互联网资源具有成长、自治和多样的自然特性，令传统的全局集中控制的资源管理难以适用。因此，我们希望通过寻求遵循互联网资源规律的资源管理模式，在面向互联网的网络计算核心机理上取得突破。针对上述问题，本课题组在研究过程中，提出了互联网资源管理思路的两个转变，即从试图掌握系统全部资源状况，转向面向需求依靠局部信息动态聚合资源；从传统的集中管理和控制全部系统资源，转向通过分布自治资源间的自主协同来实现资源的共享和综合利用。基于上述思路，提出了虚拟计算环境的概念[7]。

1.1.2 虚拟计算环境的概念

互联网资源的成长、自治和多样性，使得我们难以直接借鉴传统的全局集中控制式的管理。因此，我们提出了面向互联网的虚拟计算环境的概念。所谓虚拟计算环境，是指建立在开放的网络基础设施之上，通过对分布自治资源的集成和综合利用，为终端用户或应用系统提供和谐、安全、透明的一体化服务的环境，其目标是实现资源的有效共享和便捷协作。围绕建立高效、和谐的虚拟计算环境，我们提出了聚合与协同的核心机制[7]。

聚合是指有效获取、汇聚、组织互联网的资源信息，并综合利

用局部的信息，实现资源汇聚、组织和综合利用，形成满足任务需求、相对稳定的资源视图的过程。在传统系统中，由于资源数量较为有限，一般也很少出现动态的加入和退出的状况，因此可以全局地、集中地管理全部资源信息，而互联网资源的多样性使得难以给出资源的统一描述，而成长性使得难以获得传统意义下全面、时空一致的资源信息。因此，资源聚合通常是任务导向的，根据任务需求确定一个局部的资源视图，并在资源发生变化的过程中，确保视图的相对稳定。

资源聚合的难点在于如何适应互联网资源的特点，支持灵活多样的资源共享模式，包括紧耦合、强服务承诺的资源共享模式和多种耦合方式结合、不同承诺强度并存的复杂资源共享模式，资源聚合的范围可能从组织内部、跨组织到互联网范围。

协同是指多个资源为完成共同任务而进行的交互、同步和计算的过程。在传统系统中，资源范围较为有限，也可以接受统一的调度和控制，如集群系统中研究较多的并行计算任务调度问题，多是基于集群节点数量事先预知、各节点完全同构，节点接受集中的调度算法管理这一重要前提进行的。而在互联网这样的开放环境下，资源的自治性导致了资源无法接受全局管理，也不以达到全局最优为目标，而需要各个节点在协商的过程中达到平衡状态；互联网资源的多样性也导致了不同资源的协同模式不同，协同的环境、对象和协议等也往往具有不确定性。这些都对资源自主协同的能力和虚拟计算环境的运行机制提出了更复杂的要求。

资源协同的难点在于如何在开放、动态的环境下，实现从静态的、预设的协同到自主、动态、灵活的多种协同模式，并在此基础上形成虚拟计算环境的核心运行机制。通常一方面需要资源具有一定的自主决策和自适应能力，一方面也需要从运行环境的角度，提供一系列基础服务的支持。

1.1.3 虚拟计算环境的机理

为实现互联网资源的按需聚合与自主协同，本课题组研究了解决问题的途径和相应的机理，并提出了构成虚拟计算环境的三个基本概念[7]，即“自主元素”、“虚拟共同体”和“虚拟执行体”。其中，自主元素借鉴了自主计算的思想[12]，试图对互联网资源实施基本的封装，使资源具有自主行为的能力；虚拟共同体指出了资源的聚合范围和基本机制，从而有效支持资源的按需聚合；虚拟执行体则提供了分布式的执行机制，支持资源之间的自主协同。

下面分别简述这三个基本概念。

1. 自主元素

自主元素是虚拟计算环境中的基本资源管理单位，它能够对互联网资源进行抽象和封装，将被动的、静态的客体资源转变为自主、动态的行为主体。这样，不仅能够有效反映资源的自治性，也在一定程度上有效屏蔽和适应了资源的多样性和动态成长的特点。因此，自主元素从结构上和方法上为实现互联网资源的按需聚合和自主协同奠定了基础。一个自主元素可以对多个资源进行抽象和封装，并赋予了环境动态感知、自主行为决策和协同等能力。自主元素可以相互组合，一个自主元素可以通过组合多个自主元素而得到。此外，自主元素具有持久性和动态性，在其生命周期中可以休眠或唤醒。

由于面向互联网的网络计算支持的应用种类较多，不同的应用也会对资源的共享和协同能力提出不同的需求。有的应用只需要资源具有反应式的能力，如传统的网页浏览中，服务器只需要响应用户当前请求的网页即可；而有的应用可能需要记录并保存一段时间内资源共享的历史信息，在此基础上进行综合分析和判断，