虚拟计算环境中的覆盖网构建技术

● 张一鸣　褚瑞　著 ●

国防科技大学出版社

虚拟计算环境中的覆盖网构建技术

国防科技大学出版社
湖南·长沙

图书在版编目(CIP)数据

虚拟计算环境中的覆盖网构建技术/张一鸣,褚瑞著.—长沙:国防科技大学出版社,2010.12

ISBN 978-7-81099-826-0

Ⅰ.①虚… Ⅱ.①张… ②褚… Ⅲ.①计算机网络-研究 Ⅳ.①TP393

中国版本图书馆 CIP 数据核字(2010)第 233563 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:7 字数:182 千

2010 年 12 月第 1 版第 1 次印刷 印数:1-600 册

ISBN 978-7-81099-826-0

定价:32.00 元

前　言

基于互联网的虚拟计算环境(iVCE)是一种新型网络计算平台。iVCE以互联网资源的自主化为基础,以按需聚合与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。互联网资源的成长性、自治性和多样性等自然特性给iVCE的资源聚合带来巨大的挑战。结构化Peer-to-Peer覆盖网(简称结构化覆盖网)具有可扩展、延迟低、可靠性高等优点,是iVCE应对上述挑战、实现资源按需聚合的重要途径之一。拓扑构建是结构化覆盖网的基础性关键技术,实现了覆盖网的动态维护与消息路由等基本功能。本书面向iVCE资源聚合的需求,对高效的结构化覆盖网构建技术进行研究。

iVCE中的不同应用对下层覆盖网拓扑具有不同的要求,例如路由延迟低、容错特性好或负载平衡等。针对特定要求,现有研究通常选择某种特定的正则图设计专用的覆盖网拓扑构建方法,其设计过程较为复杂,重复工作量大。针对上述问题,本书提出一种适用于任意正则图的通用覆盖网拓扑构建技术:分布式线图(DLG)变换。DLG变换设计了一系列创新性的机制和算法,包括拓扑图统一描述、DL迭代、逻辑点合并与分裂以及高效路由等。理论分析表明,DLG变换具有良好的性能:令 d、N_0、D_0 分别为初始正则图的基、秩和直径,N 为节点个数,则在应用DLG变换构建的覆盖网中节点出度为常数 d,入度在1和 $2d$ 之间,平均入度为

d,网络直径小于 $2(\log_d N - \log_d N_0 + D_0 + 1)$,节点加入/退出维护开销为 $O(\log_d N)$,每次节点加入/退出时最多有 $3d$ 个节点需要更新路由表。

与超立方体、多维花环或 de Bruijn 图等静态拓扑图比较,Kautz 图具有最优直径、最大连通度等良好特性。然而,由于动态维护的复杂性,目前还没有结构化覆盖网能够基于任意 Kautz 图 $K(d,D)$进行构建。本书应用 DLG 变换技术,提出一种基于 Kautz 图 $K(d,D)$的结构化覆盖网:DLG - Kautz(DK)。针对 Kautz 图的特点,DK 设计了高效的资源命名算法、资源 - 节点匹配策略、容错路由算法以及节点动态加入/退出时的资源重分配机制等。理论分析与模拟结果表明,给定平均节点出度($d > 2$),在现有的结构化覆盖网中 DK 的网络直径最小。

DK 的消息路由功能提供了精确匹配的资源查询能力。然而,随着互联网技术的发展,越来越多的上层应用要求下层覆盖网能够提供更加复杂的资源查询能力。高效的分布式索引是实现低延迟、低开销、负载平衡的复杂查询的关键。针对上述需求,本书在 DK 的基础上提出一种支持复杂查询的分布式索引构建技术:平衡 Kautz 树(BK 树)。BK 树通过 Z 曲线实现了资源空间到节点空间的映射,并基于 PHT 技术设计了高效的资源信息索引结构。在 BK 树的基础上,本书进而提出一种支持动态负载平衡并且延迟有界的区间查询算法 ERQ。无论查询区间的大小或资源属性个数的多少,ERQ 都能确保在一定的延迟($\log_d N(2\log_d \log_d N + 1)$)内返回查询结果,从而证明了 BK 树的有效性。本书简要讨论了基于 BK 树实现其它复杂查询(如 Skyline 查询、聚合查询等)的方法。

现有的结构化覆盖网通常采用扁平结构进行组织:所有节点都被理想地认为是同构的,并且所有消息都使用同一种路由算法

进行路由。然而,实际大规模系统中的节点通常是异构的,在计算能力、信誉和稳定性等各方面都存在广泛差异。传统结构化覆盖网的扁平结构难以适应互联网资源的多样性特点。针对上述问题,本书在国际上首次提出一种支持路由控制的覆盖网分组构建技术,允许上层应用根据节点属性的差异对节点进行分组,进而支持在消息路由过程中采用各种灵活的路由控制策略,例如,选择一组计算能力强的节点提供计算服务或者在路由过程中选择一组可信节点作为中间节点等。在分组覆盖网的基础上,本书进而提出一种覆盖网分级构建技术,在多个组之间存在层次关系的情况下,能够以较小的开销在覆盖网中支持分级结构,例如互联网中的管理域结构等。与传统覆盖网比较,分组/分级覆盖网能够使上层应用在性能、可靠性和安全等多方面获益。

作 者

2010.10.10

目 录

第四章 基于 DLG 变换的高性能覆盖网

第五章 支持复杂查询的覆盖网索引构建技术

第一章 绪 论

互联网逐渐成为现代社会的重要信息基础设施。基于互联网的虚拟计算环境(iVCE)是一种新型网络计算平台。iVCE 以互联网资源的自主化为基础,以按需聚合与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。互联网资源的成长性、自治性和多样性等自然特性给 iVCE 的资源聚合带来巨大的挑战。结构化 Peer-to-Peer 覆盖网(简称结构化覆盖网)具有可扩展、延迟低、可靠性高等优点,是 iVCE 应对上述挑战、实现资源按需聚合的重要途径之一。拓扑构建是结构化覆盖网的基础性关键技术,实现了覆盖网的动态维护与消息路由等基本功能。本书面向 iVCE 资源聚合的需求,对高效的结构化覆盖网构建技术进行研究。

1.1 虚拟计算环境概述

随着计算技术与网络技术的广泛应用,互联网逐渐成为现代社会的重要信息基础设施和无处不在的计算平台[1]。目前,互联网上汇聚了大量的计算资源、存储资源、数据资源和应用资源等各类资源。随着国家信息化的推进,经济、行政、科研、教育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。在这种

背景下，基于互联网的虚拟计算环境[2]（Internet-based Virtual Computing Environment，iVCE）应运而生。本节首先简介 iVCE 的基本概念和核心机理，然后概述面向 iVCE 资源聚合的覆盖网技术。

1.1.1 基本概念

互联网资源具有成长性、自治性和多样性等三个自然特性。成长性是指互联网资源规模不断膨胀、关联关系不断变化；自治性是指互联网资源局部自治、自主决策；多样性是指互联网资源的属性存在广泛差异。上述特性使得无法对互联网资源进行全局的集中式控制和管理。针对该问题，文献[2]提出基于互联网的虚拟计算环境（iVCE）。iVCE 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、透明的一体化服务环境，其目标是实现互联网资源的有效共享和便捷协作。

为了实现上述目标，iVCE 提出聚合（aggregation）与协同（collaboration）的核心机制。聚合是指有效获取互联网的资源信息，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；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。为实现互联网资源的按需聚合与自主协同，iVCE 提出三个基本概念，即自主元素、虚拟共同体和虚拟执行体。自主元素是 iVCE 的基本资源管理单位，通过借鉴自主计算的思想对互联网资源进行虚拟化和封装，使资源具有自主行为的能力；虚拟共同体是一组具有共同兴趣和目标、遵从共同原则的自主元素构成的集合，通过提供一定的资源信息管理设施，有效支持自主元素的按需聚合；虚拟执行体是共同承担同一任务的相关自主元素为完成该任务而形成的状态空间的总和，提供了分布的执行机制，支持自主元素之间的协同工作。

iVCE 体系结构主要包括资源虚拟层、聚合层和协同层，如图 1.1[2] 所示。资源虚拟层把互联网资源封装成自主元素，完成对各种资源的虚拟化和自主化；聚合层在虚拟共同体范围内组织和管理 iVCE 中大量的资源信息，并对资源或自主元素进行有效聚合，面向任务形成相对稳定的资源空间和视图；协同层根据任务需求生成相应的虚拟执行体，动态绑定相关的资源或自主元素，进而通过自主协同来完成特定任务。

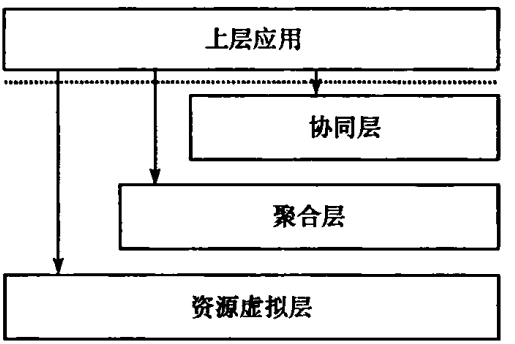

图 1.1 iVCE 体系结构示意图

1.1.2 面向资源聚合的覆盖网技术

互联网是一个不断成长的开放系统，其覆盖地域不断扩大，大量分布异构的资源动态地更新与扩展，资源的规模及其关联关系不断地成长变化，资源管理的范围难以确定。在动态变化的互联网环境下，如何在聚合层支持资源的按需聚合，是 iVCE 面临的重要挑战性问题[2]。

针对上述挑战，iVCE 基于覆盖网技术，在互联网资源之上建立了组织视图和应用视图两个层次的视图，如图 1.2[3] 所示。

(1)组织视图：iVCE 基于覆盖网技术在组织视图中对资源进

行组织与管理，被组织的可以是物理资源，也可以是虚拟化后的自主元素。由于覆盖网技术采用相同的方式对 iVCE 中的物理资源或自主元素进行组织，为行文方便，下文将不再对物理资源和自主元素进行区分，并将它们统称为资源。

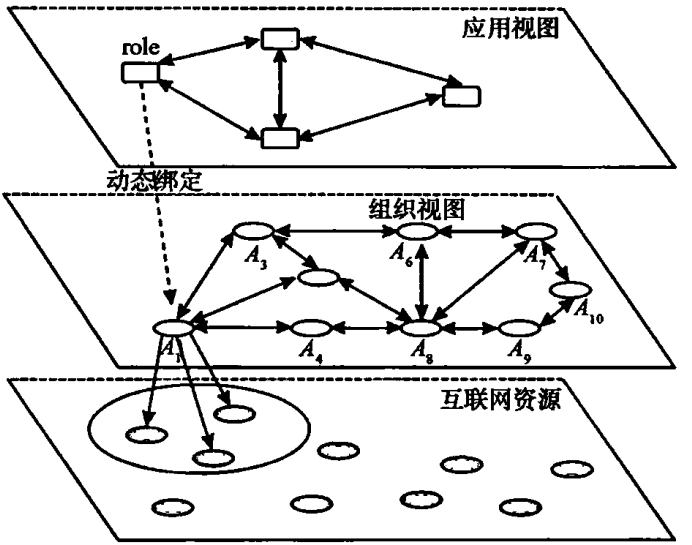

图 1.2　iVCE 的资源视图

(2)应用视图：通过提供聚合模型[4]，iVCE 在组织视图的基础上实现了应用视图与资源的解耦和动态绑定，并在资源发生动态变化时能够维护应用视图的相对稳定。

为了适应互联网资源的成长性、自治性和多样性等特性，iVCE 中的覆盖网需要具有良好的可扩展性、自组织性和适应性。结构化 Peer-to-Peer 覆盖网(简称结构化覆盖网)[5]能够基于各节点的局部决策，适应系统规模的不断成长变化以及自治资源的加入或退出，具有可扩展、延迟低、可靠性高等优点。因此，通过结构化覆盖网动态组织互联网资源并形成相对稳定的资源组织视图，是 iVCE 实现资源按需聚合的重要途径之一[3]。

拓扑构建是结构化覆盖网的关键技术,实现了覆盖网的动态维护与消息路由等基础功能。本书将以在聚合层建立相对稳定的资源组织视图为目标,对 iVCE 中的高效结构化覆盖网构建技术展开深入研究。

1.2 Peer-to-Peer 覆盖网概述

结构化覆盖网是 Peer-to-Peer(P2P)覆盖网的一种。本节首先简介 P2P 覆盖网的基本概念,然后对结构化覆盖网和非结构化覆盖网进行比较。

1.2.1 基本概念

P2P 覆盖网技术是近年来兴起的一种重要网络计算技术。如图 1.3 所示,P2P 覆盖网是一种构建于 IP 网络之上的逻辑网络结构,上层覆盖网中的一跳路由(从节点 *A* 到节点 *B*)可能对应下层 IP 网络中的多跳路由。

随着互联网的飞速发展,P2P 覆盖网已经成为很多互联网应用的基础,例如 Gnutella[6]、OceanStore[7]、PPLive[8]、Skype[9] 等。这些应用的共同特点是无需通过中心服务器的中介,即可直接共享各种资源。据统计,互联网上各类 P2P 应用的流量已占互联网流量的 60%以上[38]。上述构建于 P2P 覆盖网之上的应用系统通常被称为 P2P 系统。近年来,研究者们已经对 P2P 系统提出了多种不同的定义,这些定义的主要区别在于对 P2P 一词所涵盖范围的限定。

文献[10]把 P2P 系统定义为"所有节点在功能和任务上完全对等的全分布系统"。该定义描述了一种理想的纯 P2P 系统,其涵

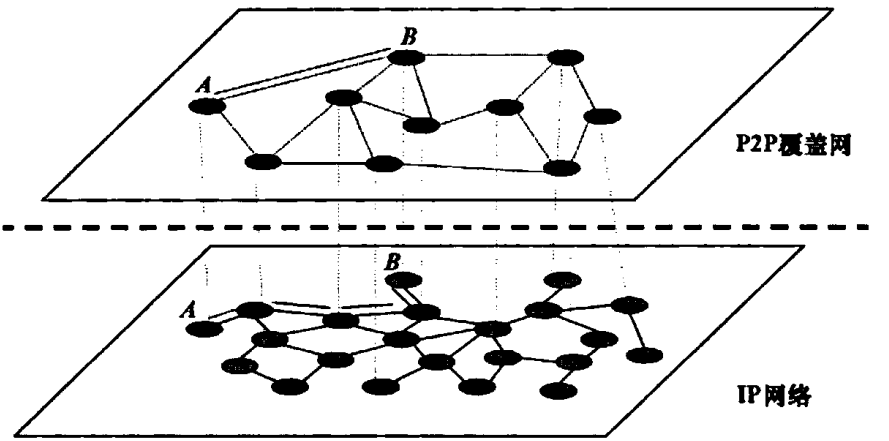

图 1.3　P2P 覆盖网拓扑示例

盖范围过窄。例如，具有超节点结构的 Kazaa[11] 被普遍认为是一个典型的 P2P 系统，但是显然该定义并没有包括 Kazaa。文献[12]把 P2P 系统定义为“能够利用分布在互联网边缘的大量计算、存储、网络带宽、信息和人力等资源的一类系统”。该定义包括了大多数互联网应用，其涵盖范围过宽。例如，依赖于中心服务器的即时通信系统和网格应用系统[13]等通常被认为不属于 P2P 系统，但是却符合上述定义。

目前，在学术界引用较多的是文献[5]给出的定义：“P2P 系统是由多个节点通过自组织形成的、具有特定网络拓扑的、以共享资源(包括文件、CPU、存储和网络带宽等)为目标的分布式系统，其特点是在没有全局集中式服务器或管理者的干涉下能够适应节点或链路失效和瞬时节点数变化，并保证可接受的性能”。该定义较为准确地涵盖了目前被普遍接受的各种 P2P 系统，得到大多数研究者的认同。

目前，P2P 覆盖网技术作为一种通用网络计算技术，已经广泛应用于各种领域，主要包括文件共享、分布式存储、分布式数据库、