

虚拟计算环境实验床平台的设计与实现

张世峰¹ 刘欣然^{1,2} 姚远哲³

(北京邮电大学网络与交换国家重点实验室 北京 100876)¹

(国家计算机网络应急技术处理协调中心(CNCERT/CC) 北京 100029)²

(电子科技大学计算机学院 成都 610054)³

摘要 虚拟计算环境的提出是为了克服互联网资源的“成长性、自治性和多样性”等自然特性与通过全局集中控制达到资源有效共享和综合利用的矛盾。基于虚拟计算环境的体系结构和核心概念,部署了一个真正开放性、分布式的大规模网络环境,在其上设计与实现了一个三层实验床平台系统,包括资源组织、监测、调度、安全等子系统。在平台上进行了网络信息获取实验,验证了该平台设计的科学性与合理性,从实现角度给了虚拟计算环境一个可行的参考。

关键词 自主元素,聚合,协同,虚拟计算,网格监测,作业调度

中图分类号 TP393.02 **文献标识码** A

Design and Implementation of the Experiment Platform of the Virtual Computing Environment

ZHANG Shi-feng¹ LIU Xin-ran^{1,2} YAO Yuan-zhe³

(Network and Switch State Key Lab, Beijing University of Posts and Telecommunications, Beijing 100876, China)¹

(National Computer Network Emergency Response Technical Team/Coordination Center, Beijing 100029, China)²

(Computer School, University of Electronic Science and Technology of China, Chengdu 610054, China)³

Abstract The virtual computing environment is aimed at solving the contradiction between the natural characteristics of internet resources such as “growing, autonomic, diversity” and efficiently sharing and using resources by central controlling. This paper deployed a real large scaled open distributed network based on the architecture and key concepts of the virtual computing environment. The paper designed and implemented a three-tier system of experiment platform including resource organizing, monitoring, scheduling, and security subsystem. The experiment about the national network information acquisition demonstrated the scientific and rational design of the platform. The design offers a feasible reference from the perspective of achieving.

Keywords Autonomic element, Aggregation, Cooperation, Virtual computing, Grid monitoring, Task scheduling

我国已经建立了世界水准的网络基础设施。网络资源的战略价值日益显现,网络资源的有效共享和综合利用能力将直接影响综合国力的提升。世界主要国家正为实现网络资源的有效共享和综合利用^[1]而努力。

近二十年,已经有很多研究成果,比如分布式操作系统 Amoeba, MACH^[2]; 分布计算环境 OSF, DCE 等等。但是类似操作系统的一体化网络资源管理服务仍少有人探索,特别是这方面的基础理论非常薄弱,这导致人们不得不需要针对具体应用开发大量复杂的网络资源管理代码,导致网络应用(例如跨部门一体化网络信息系统)开发成本高,质量低,适应性差。为此,深化对互联网环境本质特征的认识,探索虚拟计算环境的新机理已成为必然的趋势。

虚拟计算环境^[3] iVCE (Virtual Computing Environment) 的目标是能把网络变成一个虚拟计算环境:在网络基础设施上部署相当于操作系统功能的一体化服务环境;以及能够支持用户在开放的网络上有效共享资源,便捷合作与工作。

围绕 iVCE 的体系结构和关键概念,提出了一个比较完善的网络环境下开放的、可扩展的虚拟计算环境实验平台参考模型,模型设计特别针对 iVCE 中资源聚合与协同概念的有效性验证。现在部署的实验床分布于全国多个省市,节点数达到几百个,真正符合了分布式和大规模模拟的要求。在实验床上部署了国家级网络信息资源获取实验,验证了本实验床平台的可用性和合理性。

本文在第1节介绍 iVCE 的体系结构和关键概念,第2节介绍实验床平台的架构,并重点阐述与 iVCE 概念紧密相关的两个子系统,第3节介绍国家级网络信息数据资源获取实验。最后对本文进行总结并提出未来的工作思路与方向。

1 iVCE 简介

iVCE,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,实现有效资源共享和

到稿日期:2008-05-30 本文受国家重点基础研究发展计划 973 项目(2005CB321806)资助。

张世峰(1983-),男,硕士生,主要研究方向为下一代网络等, E-mail: sdu126@126.com; 刘欣然(1971-),男,博士,主要研究方向为网络安全、网格计算等; 姚远哲(1973-),男,博士,主要研究方向为网格计算。

便捷合作工作。聚合是指有效获取、汇聚、组织网上资源特征信息,并综合利用相关信息的过程;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。

iVCE 要解决的 3 个关键科学问题之一是开放环境下的按需聚合问题,如何根据任务需求,运用局部信息,实现资源特征信息的汇聚、组织和综合利用,形成满足任务需求的相对稳定的资源视图,支持任务完成。另一个问题是分布自治资源的自主协同问题,如何支持并实现自治资源的协同,建立可预测、可评估、可调节的协同工作机制和运行环境,达到资源的有效共享和综合利用,完成共同任务。

iVCE 的体系结构是分层结构,包括资源虚拟层、聚合层、自主协同层、编程开发环境、可信保证体系。

iVCE 中资源的抽象模型是 Autonomic Elements 即自主元素。它遵从共同目标、按 Join/Adapt 语义形成自主元素。它拥有传感器和效应器,具有环境动态感知、自主行为决策和协同能力,是能够根据数据与知识自主监测、分析、决策、执行的独立个体。为资源访问、交互提供一致接口。在实验床平台,宏观上它体现为一个节点。

iVCE 的执行抽象模型是 Virtual Executor 即虚拟执行体。它是协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。不同的自主元素分别贡献自己的计算等各种资源,聚合成一个虚拟的作业执行体。体内的自主元素之间协同工作,共同完成同一个任务;示意图如图 1 所示。

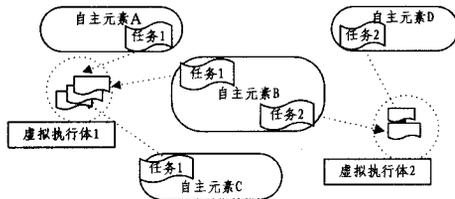


图 1 虚拟执行体

2 实验床平台系统设计

首先部署了一个符合实验床要求的网络环境,具备大规模(已达几百个节点)、广域性(分布于北京、上海、广州等地)、可控性(拥有控制绝大多数资源的能力)的特点。

实验床平台在构建过程中,遵循的第一原则就是必须紧密结合 iVCE 的体系结构和关键概念。研究思路如下,基于主动发现和被动测量的大规模网络测量技术和对真实拓扑与性能数据的网络行为建模,进行大规模虚拟计算环境仿真平台的并行化与实现技术研究,同时结合对虚拟计算环境中聚合、协同机理评价标准研究,最终构建一个实验床平台对按需聚合、自主协同机理进行验证。

2.1 实验床平台结构框架

实验床的分层结构框架如图 2 所示。该框架分为 3 层,5 大功能子系统。包括应用层、中间层、资源层 3 层,应用层的 Portal、中间层的综合调度子系统、资源层的资源组织模块、监测子系统、安全保证模块 5 个功能模块及系统。其中资源层紧扣对应 iVCE 体系中的资源虚拟层;中间层的聚合与协同子层对应 iVCE 的聚合层、自主协同层,另外接口子层负责与

应用层的数据交互。

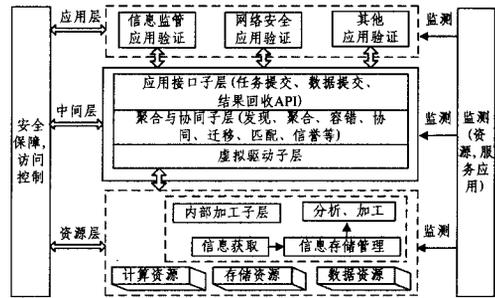


图 2 实验床平台框架

1) 应用层的 Portal 负责各类用户的管理,提供给用户提交作业的接口,与中间层交互以及作业执行结果的展示。

2) 资源层并非一个孤立的模块。这里获取的计算、存储等资源信息都将经过内部分析处理,转变成平台上统一的数据格式(本平台使用 XML 形式的数据存储格式),存储在目录服务器(Directory Server,以下简称 DS)中,供其它模块使用。资源节点(一个节点可以提供不同资源-即自主元素,参与不同任务)可以自主加入或退出实验床平台,与其它节点动态组合成节点组,不同的节点组又可以统一到一个域中,而每个域都有一个统一的 DS,不同域间的交互将通过这些平等、独立的 DS 进行。

3) 监测子系统^[4] RNMS(R-Net Monitoring sub-System)(将与资源层一同叙述,详见 2.2 节)是整个平台系统、网络、应用实时信息数据的来源,这些数据支撑了整个平台的运行。资源层中的自主元素依赖监测子系统,它的传感器和效应器才能与外界环境实时交互,进行实时的数据监测、分析、自主决策等动作,在虚拟执行体中协同工作完成共同任务。同时,实验床上所有节点的自组织、自我管理也必须依赖 DS 中的统计信息才能正常进行。

4) 中间层的应用接口子层负责与 Portal 的交互,接收任务数据和配置文件,将运算结果返回给 Portal,同时这里还是一个 Agent,相当于一个域调度器,从总体上维持整个实验床的负载均衡。虚拟驱动子层是指自主元素在执行任务过程中动态监测自身以及虚拟执行体内的软硬件、应用信息,以做出一系列自主决定。聚合与协同子层中由调度器决定调度策略,然后自主元素接收任务并进行协同执行,必要时进行作业迁移或作业执行出现严重阻碍时通知调度器重新形成虚拟执行体,在作业执行成功后返回计算结果。该层功能由 RNSS(R-Net Scheduling sub-System)来实现。

5) 安全保证模块属于 iVCE 可信保障体系,涉及跨域的授权模型与权限约束方法、资源可信度量与激励机制设计、生存性保障方法和技术等。

鉴于篇幅有限,本文将只详细叙述 RNMS 和资源层、中间层和 RNSS。

2.2 RNMS 和资源组织

RNMS 是建立在实验床平台上的一个分布式监测系统,它跨越多管理域和多种异构资源,是一个符合 GMA 规范、低开销、可扩展、容错的分布式监测系统。用以监测大规模、动态改变的 iVCE 环境,为实验床用户和管理者提供全面、准确

的监测信息。

除了节点的动静态信息、应用信息,在网络参数方面参考了标准组织的一些标准,包括 ITU-T SG13 工作组在建议 Y.1540^[5]中定义的 IP 包传输时延、时延变化、误差率、丢失率、虚假率、吞吐量和可用性参数。IETF 的 IPPM(IP Performance Metrics)工作组定义的连接性测度(RFC2678)、单项延迟测度(RFC2679)、单项分组丢失测度(RFC2680)、往返延迟测度(RFC2681)^[6]等。

如图 3 所示,基于 GMA 规范,整个系统分成 3 个主要部分:生产者、消费者和用于注册的目录服务 DS。生产者从各种传感器(诸如主机资源传感器、网络传感器和应用传感器)收集网络资源的状态信息。目录服务起着连接生产者和消费者的作用。消费者是监测信息的使用者,本系统的监测信息有很多潜在的消费者,如调度器、调试器、数据分析器、预测器等。

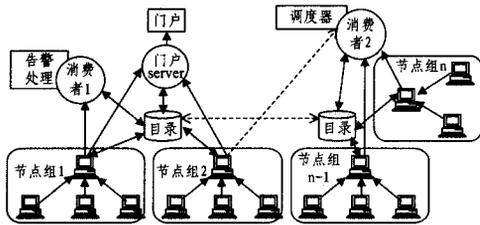


图 3 RNMS 监测子系统结构

在 RNMS 中,生产者映射成一个包括若干节点的节点组,其中有 3 类节点:主节点、备份节点和普通节点。在节点组内部,主节点是节点组的信息收集者和管理者,在外部它具有生产者的特点,执行生产者的功能,负责向目录服务注册,向各种消费者提供监测信息。

节点是 RNMS 中监测和管理的基本单位。RNMS 在每个节点中都置入了传感器和其它必要的管理通信等模块。这些都是节点资源能够形成自主元素的必要条件。它们包括:传感器模块、预处理模块、存档模块、节点内部管理模块、分析模块、通信管理和数据发布模块、数据管理模块和节点组管理模块。后两者只运行在主节点中,它们的主要功能如下:

- 1) 传感器模块。包括主机资源传感器、网络传感器和应用传感器,用于监测硬件的资源状况,软件和应用的状态。
- 2) 预处理模块。将来自传感器的监测数据转换成高效的内部监测信息表示形式。
- 3) 分析模块。对监测数据进行分析,根据预先的设定,产生不同级别和种类的事件。根据功能的不同,可以划分成多个子模块,处理不同的事件,如性能预测子模块、性能分析子模块、告警子模块等。
- 4) 存档模块。从大量的监测信息中过滤出有价值、有意义的信息存储到本地内存或磁盘,并在有需求时,提供这些历史信息。
- 5) 节点内部管理模块。基于管理策略或消费者需求对传感器或者存档模块进行自适应或按需管理。
- 6) 通信管理和数据发布模块。与其它节点或目录服务通信,对监测数据和信息进行发布。
- 7) 数据管理模块。将节点内部监测信息转换为外部的标准格式的信息。
- 8) 节点组管理模块。处理网络节点的动态加入与退出,

协调节点之间、节点组之间的测量。

节点依赖以上的几个模块,拥有了自我感知系统参数、网络状况、作业运行信息,自动做出分析决策的能力。从而能够提供计算或存储等资源,形成一个或多个自主元素。

DS(目录服务,如图 4 所示)提供存储和查询元信息的服务。主要功能就是元信息的发布和发现。由于 iVCE 是一个庞大的计算环境,可能跨越多个管理域,这就需要多个目录服务共存,它们是相互独立的,需要一定的机制对它们进行协调,使其相互协作,通过合作来实现单一的目录服务的映像。

从多样的站点获得复杂的资源信息,再以一个统一一致的映像表达出来是信息服务系统的一个基本要求。在 RNMS 中,全局 DS 由多个相互独立、地位平等的多个局部 DS 组成,它们既相互独立,不能彼此拥有和管理,又密切合作,可以相互查询信息、互通有无。每个 DS 负责在一定的范围内(如一个集群或一组通过局域网连接的多台 PC)生产者与消费者信息的注册与查询,DS 之间通过相互交换信息来了解其它 DS 相联系的虚拟计算环境资源的信息,相互协作提供一个一致的 iVCE 资源的视图。这样,一个 DS 可能直接或间接地和多个生产者和消费者相连,从任意一个 DS 就可了解整个 iVCE 的状态信息。通过多个 DS 的协同工作,生产者能够为其它域中的消费者提供监测信息和事件(如图 3 和图 4 中虚线所示),实现对整个 iVCE(可能跨越多个管理域或虚拟组织)的信息获取和管理。

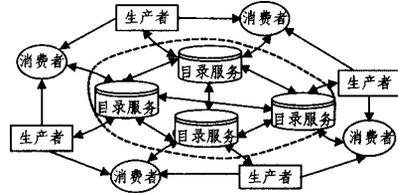


图 4 分布式的目录服务结构

2.3 中间层和 RNSS

实验床平台的中间层的功能包括发现、聚合、容错、协同、迁移、匹配、冗余等。还要负责向上与应用层的接口以及向下对资源层的虚拟驱动。

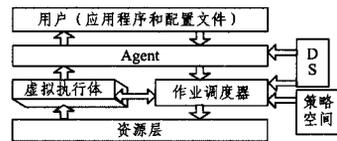


图 5 RNSS 子系统结构

RNSS(R-Net Scheduling sub-System)是根据中间层的功能设计的子系统,其结构如图 5 所示。主要包括 Agent、作业调度器、策略空间 3 部分。RNSS 通过与 RNMS 以及其中 DS 的协作,完成虚拟计算环境中的作业调度。RNSS 各部分功能如下:

- 1) Agent,代理模块。Agent 有两个作用,首先,用户提交的应用程序、作业和配置文件都将在这里进行预处理。作业执行后的结果也会在这里整合后反馈给用户。第二,Agent 相当于整个 iVCE 的域调度器,负责 iVCE 的负载均衡。Agent 会解析配置文件,初步了解该作业所需要的资源种类、数量、处理能力以及对网络带宽、延迟、连接数的要求,同时根据

用户的位置,从 DS 查询相关满足条件的信息,然后将作业配置信息以及作业调度权交给具体某个域作业调度器。Agent 遵循两个原则,原则一是根据作业需求,以尽量满足而不是最大满足的原则向下分派作业;二是根据调度器的地域分布进行分派。这样做既可以满足作业需求,又能尽量节省资源。Agent 不负责安全事宜,认证和授权以及用户管理等将由 Portal 负责。

2)作业调度器。之所以没有将 Agent 的功能放在作业调度器里实现,是为了将纷繁复杂的 RNSS 功能模块化、具体化,从而可以使作业调度器的实现更加精简,只要根据策略空间的策略和知识就能进行作业的有效调配。

作业调度器会根据 Agent 提交的作业以及域信息,与具体域内的 DS 以及某个或某几个节点组的主节点进行交互,结合策略空间的策略决定该作业的调度策略和细节,维持一个可用资源列表,然后将作业分配给具体的节点。

这些接受了任务的节点中的资源(自主元素)即组成了一个虚拟执行体。在这个虚拟执行体内部,自主元素间协同工作,主动地实时监测自身以及环境信息。当自身条件已经不能满足作业运行,它会主动与自己所在节点组的主节点交互,然后由主节点查询本节点组或 DS 域内其它节点组中与其资源能力类似的节点,将作业迁移过去。另外,为了应对自主元素所在节点死机、断网等意外事件,作业调度器也会定期轮询该虚拟执行体,当检测到突发事件时,马上从可用资源队列中选择其他可用资源进行作业迁移。

3)策略空间。在实验床上可以采取多重不同的调度策略。包括基于资源可用门限^[7]的分布式作业调度,基于 FCFS+Backfilling 框架的最优多址协同算法与贪心多址协同算法^[8],以及多 QoS 约束网格作业调度问题的多目标演化算法^[9]等。在此不作详述。

3 实验床的实用性验证

在实验床上部署了基于自主元素结构设计实现的体现 iVCE 特色的 Web 信息采集器(俗称“蜘蛛/Spider”或“爬虫/Crawler”),爬虫具有比较灵活的参数接受能力,能够根据“指令”执行不同的搜索任务。通过实验床的平稳运行,以此验证基于 iVCE 思想及其聚合、协同机制的实验床设计与实现的合理性。

3.1 实验部署

实验床的物理构成是以 CNCERT/CC 北京中心高性能计算集群系统为核心,CNCERT/CC 辐射全国 31 个省份的网络基础设施为纽带,连接各分中心及上海、广州等地外协单位的计算资源,对独立、分布资源进行集成和综合利用,构建起的一个跨地域、跨网、管理域的真正意义上的符合虚拟计算环境要求的网络环境。

根据实验床 RNSS 不同的部署策略,安排的几组实验包括:

1)被访问对象:位于北京、广州、哈尔滨的 3 组 Web 样本网站。访问点:分别从北京、广州、哈尔滨的节点对上述 3 组样本网站进行访问。

2)被访问对象:位于电信、联通、教育网的 3 组 Web 样本网站。访问点:分别从北京电信、北京联通、哈尔滨教育网的节点对上述 3 组样本网站进行访问。

3)被访问对象:均匀分布在电信、联通、教育网的一组 Web 样本网站。访问点:1)从集中部署于北京电信、北京联通、哈尔滨教育网的一组节点分别对上述 Web 样本网站进行访问;2)从分布部署于北京电信、北京联通、哈尔滨教育网的一组节点(性能、数量同上)对上述 Web 样本网站进行访问(同运营商节点访问优先策略)。

3.2 结果分析

实验结果:实验床上的各个模块运行正常且稳定。实验床上所运行的应用(如信息采集器,具有天然的并行性和潜在的非相关性,强弱受搜索策略影响)很容易受以下几个因素的影响:单个节点的处理能力、能够调动的节点数量、网络的实际可利用带宽大小、网络距离远近、集中式部署还是分布式部署等。应用部署与调配方案直接决定应用运行的效率和稳定性。

RNMS 及 RNSS 可以平稳运行,实验床平台的设计,符合、满足 iVCE 体系结构的要求。测试结果和试运行的部分数据表明,基于 iVCE 的节点多、分布广、接入全的实验床平台的设计和实现符合了 iVCE 的要求。它为今后更全面地验证 iVCE“按需聚合”与“自主协同”的思想提供了平台支撑和最大的可能性。

结束语 互联网环境及其资源的自然特性对网络计算环境的构建提出了严峻挑战。本文紧紧围绕 iVCE 体系结构与自主元素、虚拟执行体的核心概念,设计实现了一个虚拟计算环境下的大规模、广域、可控的实验床平台,该平台涉及开放性网络实时监控、数据信息管理、作业调度等一系列功能模块。在该平台上进行了网络试验,验证了该平台设计的合理性以及 iVCE 体系结构的可行性。目前,实验床平台仍有待完善的地方,下一阶段将更深度地进行资源建模、资源聚合、协同计算以及可信保证体系、安全保障的研究。

参 考 文 献

- [1] Gong L. Jxta: A Network Programming Environment [J]. IEEE Internet computing, 2001, 15(3): 88-95
- [2] Andrew S. Tanenbaum, Distributed operating systems. Prentice Hall, 1999
- [3] 卢锡城,王怀民,王戟. 虚拟计算环境 iVCE: 概念与体系结构 [J]. 中国科学(E 辑), 2006, 36(10): 1081-1099
- [4] 姚远哲,杜翠兰,刘欣然,等. R-Net 网络监测与告警系统设计 [J]. 通信学报, 2006, 27(2): 168-177
- [5] IP Performance Metrics (ippm) [EB/OL]. 2004, <http://www.ietf.org/html.charters/ippm-charter.html>
- [6] Beyah R, Sivakumar R, Copeland J. Application layer switching: a deployable technique for providing quality of service [C] // IEEE Global Communication Conference (GLOBECOM). San Francisco: [s. n.], 2003: 3694-3699
- [7] 李慧贤,庞辽军,程春田,等. 基于资源可用门限的分布式作业调度 [J]. 电子科技大学学报, 2007, 36(2): 254-256
- [8] 张伟哲,方滨兴,胡铭曾,等. 计算网络环境下基于多址协同的作业级任务调度算法 [J]. 中国科学(E 辑), 2006, 36(10): 1240-1262
- [9] 张伟哲,胡铭曾,张宏莉,等. 多 QoS 约束网格作业调度问题的多目标演化算法 [J]. 计算机研究与发展, 2006, 43(11)