

文章编号:1001-9081(2006)04-0926-03

## 分布式虚拟环境的可扩充性研究

吴言华<sup>1</sup>, 蔡云泽<sup>2</sup>, 许晓鸣<sup>3</sup>

(1. 上海交通大学 计算机系, 上海 200030; 2. 上海交通大学 自动化系, 上海 200030;

3. 上海理工大学, 上海 200030)

(wyhross@sjtu.edu.cn)

**摘 要:** 讨论了与分布式虚拟环境系统扩充性相关的两类技术, 一类是通过简单的硬件并行来实现系统扩容, 典型的例子是多服务器并行分区技术。这类技术能够支持大规模的分布式虚拟环境(DVE)系统, 缺点是费用太高; 另一类技术试图通过减少服务器通信负载来实现系统容量的增加, 主要包括感兴趣区域(AOI)技术, DR 算法和组播技术。此外, 还讨论了一些复合方法: 如基于 AOI 的组播技术等。网络负载的减少意味着相同的硬件条件可以支持更大规模的客户, 从而有助于解决大规模 DVE 系统的扩充性问题。

**关键词:** 分布式虚拟环境; 可扩充性; 感兴趣区域; DR 算法; 组播

**中图分类号:** TP393 **文献标识码:** A

## Scalability issues of distributed virtual environment

WU Yan-hua<sup>1</sup>, CAI Yun-ze<sup>2</sup>, XU Xiao-ming<sup>3</sup>

(1. Department of Computer Science, Shanghai Jiaotong University, Shanghai 200030, China;

2. Department of Automation, Shanghai Jiaotong University, Shanghai 200030, China;

3. University of Shanghai for Science and Technology, Shanghai 200030, China)

**Abstract:** Two kinds of technologies about the scalability of distributed virtual environment were introduced. One tries to improve system's scalability by paralleling hardware equipments. The example is multi-server parallel partition technology. It can support a large-scale system but with high cost. The other tries to extend DVE system's scale by reducing network traffic, including AOI technology, dead-reckoning algorithm and multicast technology. With the reducing of network load, DVE system can contain more clients, which improves system scalability. As a special refine technology, a combined technology of AOI and multicast was also discussed. With all these techniques discussed in the paper, a system designer can easily construct a large-scale DVE system, which can contain hundreds and thousands users.

**Key words:** Distributed Virtual Environments(DVE); scalability; Area of Interest(AOI); dead-reckon algorithm; multicast

## 0 引言

近年来, 随着网络技术的发展, 基于网络应用的分布式虚拟环境(Distributed Virtual Environments, DVE)在诸多领域取得了快速发展。分布式虚拟环境系统已经在许多领域得到了应用, 包括: 军事仿真系统<sup>[2]</sup>, 在线游戏<sup>[3,4]</sup>, 远程教育<sup>[5]</sup>, 电子商务<sup>[6]</sup>等。

分布式虚拟环境在取得巨大应用发展的同时, 仍然存在一些问题。最主要的问题是当系统用户增加、规模扩大时, 出现网络过载和服务器过载问题, 通称为系统过载。由于 DVE 系统的实时性要求, 任何由系统过载导致的系统延迟都会影响系统的性能。

对于这个问题一般有两类解决方案, 第一类是通过增加硬件的方法实现系统扩容, 主要包括多服务器并行分区方法。这种方法的基本思想是以牺牲硬件费用为代价实现系统扩容。另一类方法是通过减少通信负载来实现系统扩容, 主要包括感兴趣区域(Area Of Interest, AOI)技术、DR 算法和组播技术。AOI 是利用虚拟环境中的参加者一般只跟周围临近的

对象打交道这一特征, 通过只给临近的对象发送状态更新消息而忽视远处的对象的方法来减少通信负载。DR 算法的实质是以部分的牺牲系统一致性的方法来实现通信优化。组播技术是通过将系统用户之间的网络拓扑进行优化, 选取最优的通信渠道和发送方式来实现网络负载的优化。

## 1 多服务器并行分区技术

在 DVE 系统开发中, 客户/服务器结构是应用最广泛的结构。最早的客户/服务器结构一般只有一个服务器, 所有客户端都连接到服务器上并通过服务器来实现通信交互。当系统客户增加到一定程度, 服务器自身的运算能力和所处网络环境的通信带宽变得匮乏, 导致系统出现过载, 延迟增加。为了解决这一问题, 一个容易想到的方法是采用多个服务器并行的方法来分担系统负载。一般的做法是把整个虚拟空间分割成多个分区, 每个分区由一台服务器负责管理<sup>[7~10]</sup>。在 RING 系统<sup>[7]</sup>中, 虚拟环境实现了静态分区, 每个分区由一台服务器负责, 系统拓扑如图 1 所示。通过多个服务器的并行, 系统可以在大负载的情况下保持实时性能。但是这种结构仍

收稿日期: 2005-10-21; 修订日期: 2006-01-09

基金项目: 国家 973 规划资助项目(2002cb312200); 国家自然科学基金资助项目(60575036)支持

作者简介: 吴言华(1976-), 男, 山东新泰人, 博士研究生, 主要研究方向: 分布式虚拟环境系统、应用层组播; 蔡云泽(1975-), 女, 讲师, 博士, 主要研究方向: 时滞系统的鲁棒控制、鲁棒滤波、信息融合; 许晓鸣(1957-), 男, 教授, 博士生导师, 主要研究方向: 智能控制、过程优化控制。

然存在局部过载问题,即在总体系统未饱和的情况下,由于过量用户可能聚集在单个服务器上从而导致单个服务器出现过载,称为局部过载。

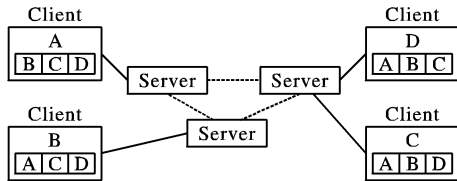


图1 RING 的结构

针对这个问题,提出了多类负载平衡算法。一类方法是在多个服务器之间平衡用户负载<sup>[8,9]</sup>。在 NetEffect 系统<sup>[8]</sup>中,虚拟空间根据虚拟空间中的客户密度分成多个区域,每个区域由一台服务器管理。一台主服务器通过改变服务器的管理区域来维持多个服务器之间的负载平衡,从而降低了过载出现的概率。系统结构如图2所示。其存在的不足主要是分区之间是独立的,用户不能“透明的”跨分区自由移动。这类系统可以看作多个小的 C/S 系统的组合,这打破了 DVE 系统的透明性特征。同时,主服务器的存在难以避免单点错误。

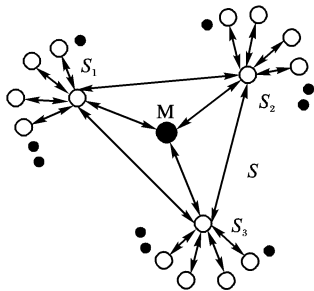


图2 NetEffect 的结构

另外一种方法是通过把每个服务器负载的分区再次划分为小面积的网格,并通过在分区之间移动边界网格实现负载迁移<sup>[10]</sup>。这种移动是在服务器后台实现的,对于用户来讲是透明的。因此保证了 DVE 系统的“用户透明”特征。但随之而来的问题是负载迁移波动问题和负载迁移环形级连问题。

负载迁移波动问题和负载迁移环形级连问题对系统性能影响极大。负载迁移是个费时的过程,这增加了平衡负载所需要的时间。最后迁移波动现象和迁移级连现象增加了系统开销。

针对负载迁移算法存在的问题,一种负载共享算法在文献<sup>[11]</sup>中提了出来。较之负载平衡算法的先过载再平衡的思想,负载共享方法试图通过避免出现过载来解决系统局部过载问题。这种方法的好处是避免了负载迁移带来的庞大系统开销。但是由于服务器之间的松耦合连接现状,处理负载分配异常困难。

文献<sup>[12]</sup>给出了另外一种以时间延迟为指标的启发式算法来进行全局的网络负载分配。虚拟空间被映射为一个图:  $G = (V, E)$ 。V 表示系统内的所有用户, E 表示用户之间的连接。系统在必要的时候通过全局的启发式计算来确定最优的网络拓扑结构并根据计算结果来调整系统连接。其存在的明显缺点是计算过程和调整过程耗时巨大,尤其是当系统规模变大时候更为明显,因此不适合系统的实时性要求。

尽管多服务器并行分区的方法存在许多缺点,但它仍不失为一种解决大规模分布式虚拟环境系统的一个可行的方法。许多大型的商业 DVE 系统采用了这一结构<sup>[3,13,14]</sup>。

## 2 DR 算法

### 2.1 算法介绍

在 DVE 系统中,DR 算法是用来减少网络通信量的基本算法并被大多数 DVE 系统所采用<sup>[15]</sup>。最早的关于 DR 算法的讨论来源于分布式交互仿真 (Distributed Interactive simulation, DIS)<sup>[16]</sup>中关于交通工具的移动仿真的研究。其基本思想是在真实状态与预测状态之间误差超过设定数值之前不进行状态同步从而避免了大量数据包的发送。

DR 算法的前提思想是用户的状态发生突然变化的概率很小。执行过程如下:首先,假定存在三个客户:A, B, C。DR 算法预先设定一个状态预测公式,并定义一个可以接受的最大误差值。其次在仿真过程中,客户 B 和客户 C 不是每个周期都能从客户 A 那里得到 A 的状态更新消息<sup>[17]</sup>。在没有得到更新消息的周期内, B 和 C 利用预定的预测算法和旧的 A' 值来估算新的 A' 的状态值,从而避免了网络通信。最后, A 在调整自身的真实状态的同时也计算预测值并比较两者的差值。当差值大于预先设定的门限时, A 发送状态更新消息给 B 和 C。B 和 C 根据接收到的数据执行一定的收敛算法,并在未来的仿真周期内逐步更新 A 的状态。DR 算法从本质上来讲是以部分的牺牲系统一致性的方法来优化系统通信负载。这种不一致性是被控制在一个可接受的范围内的。

例如在图3所示的例子中,虚线是对象所实际要走的路线,实线是系统通过预测函数计算出来的路线。可以看到,在保证误差维持在一个可接受的范围内的前提下,整个过程中只发生六次网络通信。

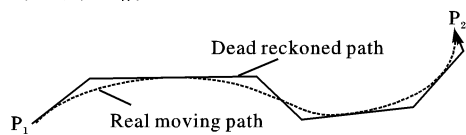


图3 DR 算法曲线

### 2.2 改进 DR 算法

在 DVE 的研究开发过程中,针对不同的系统特点,提出了许多不同的改进 DR 算法。一种典型的方法是根据系统特性变固定的误差门限为可变的误差门限<sup>[17,18]</sup>。一种改变误差门限的方法是其反比于系统刷新周期<sup>[17]</sup>。如果刷新周期长则误差门限小,反之则大。这种方法体现了网络负载与显示效果的一种折中。另外一种思想是利用 AOI 来调整误差门限(AOI 技术将在下一节详细介绍)<sup>[19]</sup>。误差门限正比于对象之间的距离。距离越近,门限越小,负载越高,状态一致性越好。反之,网络负载降低,状态一致性变差。其依据是 DVE 系统中对象以更高概率跟其临近的对象打交道的特征。这种方法仍然体现了网络负载与状态一致性之间的均衡。

## 3 AOI 技术

DVE 系统拥有一个巨大的虚拟空间,但是对于空间中的单个实体对象而言,其一般只对部分虚拟空间感兴趣。也就是说单个实体一般只能感知和影响其周围有限距离内的实体对象。这个有限距离所决定的区域为实体所感兴趣的区域,称为 AOI。在系统运行时,实体主要跟其 AOI 域内的对象进行通信交互并忽视其他的对象。通过只给部分对象发送消息可以有效地减少系统通信量。节省的网络带宽可以用来支持更多的用户,从而提高了系统的规模。图4给出了一个基于 AOI 交互的例子。其中 A 只与 B 交互,而不与 C 发生关系。相对于把状态信息广播给系统所有的用户而言,这种方式可以极大地减少网络通信负载。

AOI 思想在其他技术里面也得到了应用(下一节中将谈到基于 AOI 的多树组播技术),AOI 技术已经成为了 DVE 系统进行负载优化的一项基本技术,并得到了广泛的研究和应用。

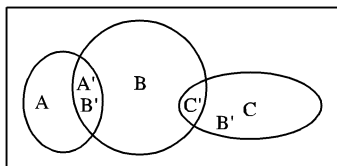


图4 AOI 示意图

## 4 组播

### 4.1 组播结构

组播技术来源于分布式网络应用系统对于一对多和多对多通信模式的需求。典型的应用有视频点播系统、分布式仿真和多人在线游戏等。这些系统的共同特点是系统规模大,网络通信负载高。

组播技术可以描述如下:将相同兴趣的网络用户组成一个组播组,所有的组成员在结构上被连接成为一个斯泰纳树(Steiner Tree),如图5所示。每个组拥有唯一的一个组播地址。任何发送给该组播地址的网络数据包将被发送给组内所有的用户。组播组的建立问题一般可以归结为斯泰纳树的构造问题<sup>[30]</sup>。组播树一般有两类<sup>[19]</sup>:基于源的组播树和基于核的组播树。

图5给出一个基于源的组播树的示例:节点A为组播源,节点组{A,C,E,F,I,G,J}构成了一个组播树。其中{A,F,G,J}是组播组成员,{C,I}为组播前向节点。来自源节点A的数据包沿着被优化了的树路径{A,C,F,I,G}和{A,C,F,I,J}发送给了所有组成员。

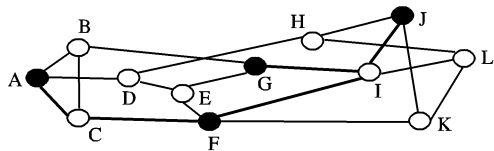


图5 组播树结构

组播树的建立过程虽然消耗一部分系统资源,但是在建立后可以极大地提高网络传输的效率。因此这种方式适合DVE系统的应用。

### 4.2 组播协议

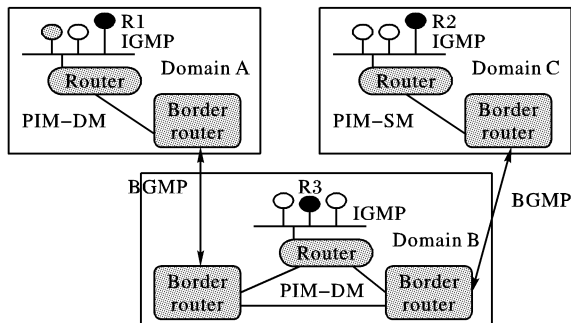


图6 Internet 上的组播协议

组播技术的最初实践体现于 Mbone<sup>[20]</sup> 的建立。Mbone 是一个基于 Internet 的虚拟网络。它提供了一个在传统的 IPv4 网络上实现组播的平台。为了实现组播有三类网络协议被采用:第一类是客户加入和离开组播组的协议,一般采用 IGMP(Internet Group Management Protocol)<sup>[21]</sup> 协议;第二类协议是局域网内的组播实现的协议(Multicast Interior Gateway Protocol, MIGP),包括:距离矢量组播路由协议(Distant Vector Multicast Route Protocol, DVMRP)<sup>[22]</sup>, OSPF 扩展协议(Multicast

Extensions for OSPF, MSPF)<sup>[19,23]</sup>, 核树(Core-Based Tree, CBT)<sup>[19,24]</sup> 和协议独立组播(Protocol-Independent Multicast, PIM)<sup>[25,26]</sup> 等;第三类协议是在广域网边界路由器之间实现组播的协议,如边界网关组播协议(Border Gateway Multicast Protocol, BGMP)<sup>[27]</sup>。这三类协议的内在联系在图6中给出。通过这一系列的网络协议,可以在现有的国际互联网上实现组播。

### 4.3 组播在 DVE 中的应用

组播技术可以减少网络负载,帮助 DVE 系统在相同的带宽下,容纳更多的客户,提高系统可扩展性。因此,组播在 DVE 系统的开发中得到了广泛的应用。

一个应用实例是 DOOM<sup>[28]</sup>。所有的系统客户被连接成一个基于核的树,如图7所示。根节点(核)的选择依据是到各个节点的延迟不会超过平均延迟。消息包通过根节点在全局进行发送。与广播技术相比,该结构可以有效地减少带宽。但是这种结构不能支持规模较大的系统,因为单一树结构决定了当系统规模增大时,根节点到叶子节点的延迟是渐增的;而且当系统用户规模很大时,根节点的全局选举算法将会消耗更多的时间,从而影响系统的实时性能。

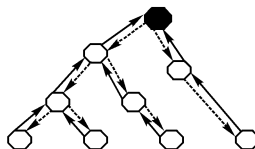


图7 基于核的组播树结构

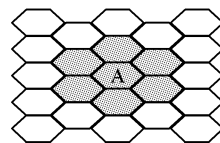


图8 基于 AOI 的组播

另外一种基于 AOI 的多树结构可以避免这个问题。在 NEPSNET<sup>[29]</sup> 系统中,虚拟空间被分成适当大小的六边形,如图8所示。每个六边形所占的区域内的客户构成一个组播组。每个实体的 AOI 内的组播组构成一个组播组集合。发送给该组播组集合的所有消息,都会被实体 A 接收到,称为 A 的“接收集”。而实体 A 的状态更新消息,则只发送给 A 所在的组播组。AOI 随着 A 的移动而移动,当 A 离开一个组播组,加入另一个组播组后,其“接收集”随之发送改变。

尽管通过将 AOI 与组播结合起来,增加了系统处理的复杂性,但是 DVE 系统的网络负载优化效率可以得到极大的提高,从而能够支持更大规模的 DVE 系统。

## 5 结语

规模可扩展性是 DVE 系统的一个重要的主题,本文主要讨论了与之相关的四种技术:多服务器并行分区技术,AOI 技术,DR 技术和组播技术。

多服务器并行分区方法是一种利用简单的硬件并行实现系统扩展的方法。其存在的主要问题是硬件的利用效率低,费用高,以及多个服务器之间负载平衡的问题。

另一类方法试图通过减少网络负载来实现系统的规模扩充,包括 AOI,DR 算法和组播。DR 算法是在可接受的范围内,利用预测状态代替真实状态从而有效的节省网络通信量。AOI 技术则是根据 DVE 系统的一个实体一般只跟其临近的一部分实体打交道的特征而提出的优化系统通信负载的方法。组播技术是通过建立系统客户之间的有序网络拓扑来减少网络负载。其主要问题是并非所有的现有的互联网底层设备都支持组播。

### 参考文献:

- [1] SINGHAL S, ZYDA M. Network Virtual Environments: Design and Implementation[M]. Addison-Wesley, 1999.

虽然图4中显示的两种报文丢失率有一定的差异,但是推测的丢失率还是很接近真实值,反映了真实的报文丢失状况的变化趋势,差异增加是由于误差累积的结果。

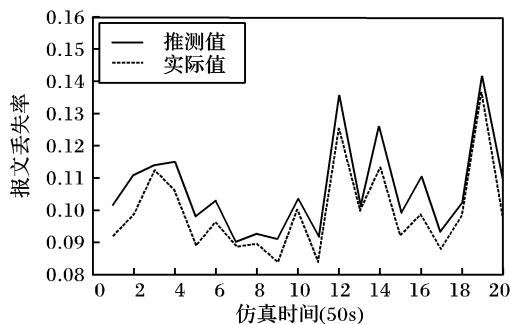


图4 链路1上报文丢失率比较

以图2所示的网络拓扑为例进行误差分析,内部节点2依据本文提出的算法估测的接收到的报文数量与实际接收到的报文数量之差为  $n(X_2 = 1) - \hat{n}(X_2 = 1) = \Delta_{45}$ , 同样节点3估测的误差是  $\Delta_{67}$ 。而节点1的估测误差就成了同时丢失在其左右子树上的报文数量,即  $\Delta_{23} + \Delta_{4567} + \Delta_{267} + \Delta_{345}$ 。从上面误差分析不难看出,越靠近叶节点的节点上的估测误差越小,而越靠近根节点的节点估测误差相对越大。这种误差在实际网络测量中是不可测,因而在本方法中也就不可避免。

## 4 结语

网络断层扫描技术作为一种外部网络测量方法,引起了广泛的关注,主要根据发生在网络边缘的测量数据,推测网络内部的性能指标。目前在推测阶段主要采用的方法是似然估计,这种方法计算复杂,影响在实际网络中的应用。本文提出了一种简单快速的估计方法,通过仿真试验可以看出该方法能够有效地推测出报文丢失率。

### 参考文献:

- [1] MARK C, HERO AO III, ROBERT N, *et al.* Internet Tomography [J]. IEEE Signal Processing Magazine, 2002, 19(3): 47-65.
- [2] ZHU W, GENG Z. A bottom-up inference of loss rate [J]. Computer Communications, 2005, 28(4): 351-365.
- [3] RAMON C, DUFFIELD NG, JOSEPH H, *et al.* Multicast-Based Inference of Network-Internal Loss Characteristics [J]. IEEE Transactions on Information Theory, 1999, 45(7): 2462-2480.
- [4] LIANG G, YU B. Maximum Pseudo Likelihood Estimation in Network Tomography [J]. IEEE Transactions on Signal Processing, 2003, 51(8): 2043-2053.
- [5] Information Sciences Institute of University of Southern California. The Network Simulator 2 [EB/OL]. www.isi.edu/nsnam/ns2, 2005.
- [6] BASSIOUNI MA, CHIU MH. Performance and reliability analysis of relevance filtering for scalable distributed interactive simulation [J]. ACM Transactions on Modeling and Computer Simulation, 1997, 7(3): 293-331.
- [7] IEEE Std 1278.1 - 1995, Standard for Distributed Interactive Simulation-Application Protocols [S].
- [8] SEOK-JONG YU, YOON-CHUL CHOY. A dynamic message filtering technique for 3D cyberspaces [J]. Computer Communications, 2001, 24(18): 1745-1758.
- [9] MOY J. Multicast routing extensions to OSPF, IETF RFC 1584 [S].
- [10] DEERING S. The PIM Architecture for Wide-Area Multicast Routing [J]. IEEE/ACM Transaction on Network, 1996, 4: 53-62.
- [11] BALLARDIE T. Core Based Trees (CBT version 2) Multicast Routing, RFC 2189 [S].
- [12] DEERING SE. Host extensions for IP multicasting, RFC 1112 [S].
- [13] WAITZMAN D, PARTRIDGE C. Distance Vector Multicast Routing Protocol, RFC 1075 [S].
- [14] MOY J. OSPF Version 2, RFC 2328 [S].
- [15] ESTRIN D. Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol specification, RFC 2362 [S].
- [16] THALER D, ESTRIN D, MEYER D. Border Gateway Multicast Protocol (BGMP): Protocol specification [Z]. Internet draft, 1998.
- [17] HOLBROOK H, CHERITON D. IP Multicast Channels: Express Support for large-Scale Single-Source Applications [A]. ACM sigcomm'99 [C]. 1999.
- [18] ERLMAN R. Simple Multicast: A Design for Simple, Low-Overhead Multicast [Z]. Internet draft, 1999.
- [19] RAMAKRISHNA V, ROBINSOM M, EUSTICE K, *et al.* An Active Self-Optimizing Multiplayer Gaming Architecture [A]. Autonomic Computing Workshop [C]. 2003. 32-41.
- [20] MACEDONIA MR, ZYDA MJ. Exploiting Reality with Multicast Groups: A Network Architecture for Large-Scale Virtual Environments [A]. Virtual Reality Annual International Symposium Proceedings [C]. 1995. 11-19.
- [21] SAHASRABUDDHE LH, MUKHERJEE B. Multicast routing algorithms and protocols: A tutorial [J]. IEEE Network, 2000, 14(1): 90-102.

(上接第928页)

- [2] MILLER DC, THORPE JA. SIMNET: The advent of simulator networking [J]. Proceedings of the IEEE, 1995, 83(8): 1114-1123.
- [3] ABRASH M. Quake's game engine: The big picture [J]. Dr. Dobbs' Journal, Spring 1997.
- [4] LEWIS M, JACOBSON J. Game Engines in Scientific Research [J]. Communication of the ACM, 2002, 45(1).
- [5] NITTA T, CONE KF. An Application of Distributed Virtual Environment To Foreign Language [A]. IEEE Education Society [C]. Kansas City, Missouri, 2000.
- [6] ZHANG J, LI FS, LI H. Multi-user Shared Virtual Reality in the Exhibition of Chinese Nationalities-Virtual Museum of Chinese Nationalities [A]. The Sixth International Conference on Computer Supported Cooperative Work in Design [C]. 2001. 83-88.
- [7] FUNHOUSER T. RING: A Client-Server System for Multi-User Virtual Environments [A]. Proceedings of Symposium on Interactive 3D Graphics [C]. 1995. 85-92.
- [8] DAS T, SINGH G, MITCHELL A. NetEffect: A Network Architecture for Large-scale multi-user Virtual World [A]. Proceedings of ACM VRSTI [C]. 1997. 157-163.
- [9] HORI M, ISERI T, FUJIKAWA K, *et al.* Scalability Issues of Dynamic Space Management for Multiple-Server Networked Virtual Environments [A]. Proceedings of IEEE Pacific Rim Conference On Communications, Computers and signal Processing [C]. 2001. 200-203.
- [10] LAU R, NG B, SI A, *et al.* Adaptive Partitioning for Multi-server Distributed Virtual Environments [A]. Proceedings of the ACM International Multimedia Conference and Exhibition [C]. 2002. 271-274.
- [11] NGUYEN T, DUON B, ZHOU S. A Dynamic Load Sharing Algorithm for Massively Multiplayer Online Games [A]. The 11th IEEE International Conference on Networks [C]. 2003. 131-136.
- [12] LUI JCS, CHAN MF. An Efficient Partitioning Algorithm for Distributed Virtual Environment Systems [J]. IEEE Transactions on Parallel and Distributed Systems, 2002, 13(2): 193-211.
- [13] Ultima Online [EB/OL]. www.uo.com, 2005.
- [14] Everquest [EB/OL]. www.everquest.station.sony.com, 2005.