计算机科学 1999Vol. 26№. 6

大型分布式虚拟环境中的消息传递机制的关键问题

The Key Issues of the Messaging Mechanism in Large-Scale Distributed Virtual Environments

郑白玫　施鹏飞

（上海交通大学图像处理与模式识别研究所　上海 200030）

Abstract The messaging mechanism in distributed Virtual Environments(dVEs) deals with the issue of message delivery between participating entities and users. As dVEs are heading toward the large-scale systems, a new messaging mechanism is to be designed in order to enable real-time interactions between entities in large-scale dVEs. This paper discusses some key issues concerning establishing the new messaging mechanism and suggests the directions of future endeavors.

Keywords Virtual Reality(VR), distributed Virtual Environments(dVEs), Messaging, Message packet, Protocol, Entity, Host

1. 引言

目前，分布式虚拟环境(Distributed Virtual Environments，以下简称 dVEs)正向着大规模化的方向发展，旨在同时连接众多的计算资源，以便有意义地重建真实世界的功能。

任何可工作的 dVEs 必须支持系统中用户、信息对象(实体)之间通过消息传递实现的交互，其作用是为 dVEs 中的每位用户(包括原有的用户和新加入的用户)提供关于虚拟世界当前状态的刷新消息。这样，dVEs 的用户便能跟踪虚拟世界的状态的变化，并在此基础上积极地参与到变化中的虚拟世界之中去。为了在 dVEs 中有效地传递消息，必须研究消息传递的策略和方法，也就是有关消息传递机制的问题。

为现有的某些小型 dVEs 系统所采用的一种简单的方法是"广播"方式，即：网络上的所有用户均具有相同的虚拟世界的模型，随着时间的推移和虚拟世界的动作，所有用户均接收关于虚拟世界状态变化的全部刷新消息。

然而，在连接了数量庞大的用户的大型 dVEs 系统中，向网络上的每个用户"广播"虚拟世界状态变化的消息是不大现实也是没有必要的。一种切实可行的方案是：只让每个用户的主机接收和处理"相关的"的消息数据包。大型 dVEs 所必须具备的基本功能是：一种"一般化"的、发布/接收/消化消息数据包的能力；而这些数据包需被"智能地"导向属于某个逻辑组群的站点。因此，必须探索一种新型的消息传送机制，其核心是一种标准的、可扩充的、可实时再设置的虚拟世界的消息协议。

2. 大型 dVEs 消息传递机制的关键问题

大型 dVEs 中的消息传送机制所涉及的问题很多，而且主要是和网件(包括网络软件和硬件)相关的。从 dVEs 和网件发展的情况来看，要解决大型、网连的、多用户的 dVEs 的消息传送问题，网络软件是关键。为了分析 dVEs 系统的消息传送机制的基本构件和动作机理，不妨从分析消息传送过程中需要解决的几个关键问题入手，即：传送什么以及何时传送？以何种形式传送？以何种方式传送？向何处传送？

2.1 传送什么以及何时传送

在 dVEs 系统中，有四种重要的通讯信息，即：轻权交互(Light-weight Interactions)、网络指针、重权对象(Heavy-weight objects)以及实时数据流。我们所说的传送对象是指轻权交互信息，简称为消息。

郑白玫　博士生，研究方向为大型分布式虚拟环境中的消息传递机制的研究。**施鹏飞**　教授，博士导师，上海交大图像处理与模式识别研究所所长，IEEE 高级会员，主要研究领域为模式识别、人工智能、虚拟现实、计算机视觉、图像处理。

消息包括状态、事件以及控制信息。

为了支持 dVEs 系统中实体间的有意义的交互，消息的传送需满足实时性要求。大型的 dVEs 系统包含了数目庞大的实体，每个实体的状态均可能随着时间的推移而发生变化，这便会产生数量巨大的消息；另一方面，dVEs 系统的地理跨度也明显加大。这就产生了一个问题：在消息流量和地理跨度同时剧增以及网络带宽有限的前提下，若不加选择地发送实体的全部刷新消息，必然会造成网络"拥塞"和相当大的延迟，因而必须选择消息发送的时机，以满足实时性的要求。

针对消息发送时机的选择，目前有两种方案，即：分布式交互模拟协议（DISP）所使用的呆推断（Dead Reckoning）方法和 JDS Pugh 提出的复制交互时间偏差（RITW，Replicated Interactive Time Warp）方法。

·呆推断。（DISP）解决带宽和延迟问题的技术称为"呆推断"。其思路很简单：主机发送的一条消息（如 ESPDU）包含了实体的位置、时间戳和速度矢量。使用该信息，网络中的每一主机便可以外推实体的位置，而无需额外的刷新。每个实体运行其自身的模拟程序，同时运行自身简单的呆推断模型，它跟踪由以上两个模型所预测的实体位置，当两个模型得出的位置相差超过某一数值时，它就发送消息刷新所有主机上该实体的模型，使它们与该实体的实际行为相一致。另外，每个实体还周期性地发送"维持生存（keep-alive）"的刷新消息。呆推断取得了令人满意的结果。由于刷新消息只是在需要时发送，通信流量被大大降低。

很明显，呆推断是更加通用的方法的一个特例。也就是说，应当发送关于实体"行为（behavior）"的刷新消息，而不是发送其位置和方向的消息。实体发送的刷新消息将指明某一"行为"和"行为参数"。实体可用的一整套行为应当可以扩充，这样我们就无需事先估计每个可能的行为。需要以标准的方式描述特征，特征描述应足够复杂，这样我们就无需生硬地搬用 PDU 的模板，最终，就使呆推断成为一种"行为库"的操作，因而更一般化，并具有可扩充性，可满足我们的需要。

·复制交互时间偏差 RITW。dVEs 系统中的实体有确定性实体与非确定性实体之分。确定性实体的行为只是时间的函数，其行为在时间上可以被后推（rollback）或前推（rollforward）。非确定性实体的行为具有不可预知的特点，其时间是不可逆的，不能被"倒推（rewind）"至过去的某个时刻。在虚拟世界中，假若所有的实体均为确定性实体，并且各用户的虚拟世界均为完全相同的复制品，那么用户和实体间便无需进行通讯。当非确定性事件介入到确定性的虚拟世界时，如果可以以一种一致的方式插入这一非确定事件，便可以完全消除由确定性实体所产生的消息流，而只需发送关于非确定性实体的消息。这里，"一致的方式"是指：在相同的时刻，将非确定性事件引入到各用户的虚拟世界复制体中。

正是基于这种想法，JDS Pugh 提出了 RITW 方法。RITW 利用确定性对象的行为的可预知性，采用有效的时间后推和前推的方法，维持各虚拟世界拷贝（replicas）的一致性，从而消除由确定性实体产生的消息流，而只是在非确定性事件发生时，才发送刷新消息。其代价是：由 rolling 操作引起的额外的本地计算量以及由于网络延迟和优化的并发控制算法所引起的短暂的用户界面的异常。RITW 的具体实现方法包括：建立输入事件队列，事件存储区，实体状态存储区以及输出事件存储区，撤消事件操作（anti-events）等。

2.2 以何种形式传送

由于大型 dVEs 系统涉及了种类各异的主机和实体，在设计 dVEs 系统时应着重考虑的问题之一便是兼容性。具体到消息传送，也就是消息格式的问题，亦即消息应采取何种形式，才能为不同的对象所理解，即具有可理解性。

DIS 解决消息格式问题的方法是：定义参与模拟各主机信息交换的标准消息格式。这种标准格式称为 PDU（协议数据单元）。有许多种 PDU，其中实体状态 PDU（ESPDU）包含了需要在主机之间交换的类似的信息。ESPDU 封装了一给定的实体在一给定的时刻的位置、姿势（posture）、线速度、角速度、线加速度和角加速度。实体的特殊成分（如活动部件的方向）亦可作为结合参数（articulated parameters），被包括到 PDU 中。一整套的标识性特征（identifying characteristics）唯一地标定了消息源实体（originating entity）。

DIS 是针对一特定的应用而设计的，绝大多数的 PDU 对于通用的虚拟现实系统并不适用。实体状态 PDU 过于特殊化，不适合"发送行为"的一般化概念。另外，它还包括了一大堆多余的信息，这就导致了 PDU 的个体非常庞大，因而大大地甚至没有必要地增加了对带宽的要求。我们所需要的是一种简捷的、开放的消息格式。在最简单的情况，一条

消息必须包括：实体的标识(ID)、时间戳、行为描述。开放格式的消息 PDU 允许将用户定义的扩展加入到 DIS 标准中。这种灵活性，加上 Internet 范围内多信道广播(multicast)消息的高效性，可以支持将面向对象的消息传送范例扩充到规模不限的分布式系统中。使用自由格式的消息 PDU 以及早已存在于 WWW 的网络指针机制，可以向 Internet 上的任一信息站点提供通用的消息传送链接。

2.3 以何种方式传送

Internet 是基于 TCP/IP 协议。其中 IP 是底层协议，它处理寻址和路由。在 IP 的上层有两种协议：TCP(传输控制协议)和 UDP(用户数据报协议)。TCP 提供面向连接的、可靠的位流式通信。TCP 进行复杂的操作，包括：数据流的分割、奇偶校验、数据包的整序，必要时请求重发、流控制等。UDP 不同于 TCP，它提供无连接的、不可靠的数据包传送。"不可靠"是指：不确保一特定数据包的到达。然而，一旦数据包到达，它将是无误的。

TCP 的问题在于其传送效率。TCP 提供的流控制、奇偶校验、整序是以相当大的通讯开销为代价的，后者转化为网络延迟——消息传送的"大敌"之一。与 TCP 相比，UDP 具有一些重要的优点。UDP 数据包比较"轻捷"，尽管它偶尔会丢失，却不会造成网络拥挤和堵塞。另外，不同于 TCP 的是，UDP 数据包可以在子网上广播(因为这里不涉及"逻辑连接")。然而，由于 UDP 是"不可靠"的，它要求使用"无状态"协议；换言之，每条消息必须是完整的和"自主(self-contained)"的，不作关于已到达的前序消息的假设。NFS 和 DIS 便是使用无状态协议的例子。DIS 使用 UDP 广播数据包向其它主机传送它的 PDU。每个 ESPDU 包含了关于实体的完整状态，它们每隔几秒便被重复广播(需要时更加频繁)。

在选择传输协议时，应根据传输的特点和要求，进行权衡。针对大型 dVEs 系统中消息具有"轻权"的特点和消息传送的实时性要求，选择 UDP 将是较合理的。

2.4 向何处传送

广播方式无疑是一种简单的解决方案，即：向模拟世界中的每个主机传送消息。这种方法对于小型的、专用的网络是可以接受，但对于大型的、包含有大量实体的 dVEs 系统是不适用的。随着实体数目的增大，低速链接将趋于饱和。即使不考虑带宽问题，低速主机仍然无法处理大量实体的刷新消息，因为它们必须做输入设备处理等工作。简言之，广播方式无法扩充。

针对以上问题，DIS 的设计者们提出了"刷新过滤"的方法。其思路是：将整个虚拟世界分割成许多"细胞"，这些细胞被用作过滤刷新消息的基地。每个参与模拟的主机确定一个兴趣域(Area Of Interest，AOI)，AOI 包括了一批处于其视野之内的细胞。当参与者四处移动时，细胞将进入或离开其兴趣域。在任一给定时间内，参与者只接收可见细胞的刷新消息，从而使得刷新消息的数量得到降低和控制。具体到"刷新过滤"的实现，DIS 系统使用的是多信道广播(multicasting)方法。其思路是：任一给定的主机除了拥有自己的 Internet 地址之外，还可以属于一些"多信道广播组"。每个多信道广播组具有其特定的 Internet 地址；当任一主机发送消息给一个多信道广播地址时，该消息便被发送到所有属于这个多信道广播组的主机。在效果上，这类似于向一个跨越大陆的子网广播。在 DIS 系统中，每个细胞具有自己的多信道广播组地址，换言之，细胞与多信道广播地址是一对一的关系。当参与者移动时，它们进入或者离开细胞，这对应着进入或者离开多信道广播组。由于参与者只接收他们所在的多信道广播组的消息，多信道广播系统自身便完成了消息过滤的工作。

多信道广播不包含任何高层的过滤功能，其通信开销最小。它存在的主要问题是没有得到全范围的实现。解决途径之一是：将 Internet 上小群的、有多信道广播能力的主机互联，构成多信道广播干线(MBONE)。MBONE 系统使用"隧道法"，它将发往某一多信道广播地址的 IP 数据包打成另一 IP 数据包，后者将经过常规的 Internet 网，从一个 MBONE 子网传向另一个子网。另外，关于兴趣区域亦有许多问题亟待解决。比如：区域的定义、区域的可见性、区域到多信道广播地址的映射等问题。

结语 目前，支持大规模、连网的、多用户的虚拟环境的软件的开发已成为虚拟现实研究工作的一个重点和热点。分布式虚拟环境规模的扩充对消息传递机制提出了新的要求，即：

- 在大型 dVEs 系统中，传送刷新消息应选择时机，以降低消息流量，避免网络拥塞。呆推断和 RITW 方法为此提供了较好的范例；
- 考虑在 dVEs 系统的兼容性，消息应采用标准的格式，如 DIS 所使用的 PDU。大型 dVEs 系统的消息格式应朝着一般化、简捷、开放的方向发展；
- 在选择消息传输协议时，应根据消息传送的

(下转第 40 页)

(4)流水通道技术[5]

传统的MPP系统中互连网络和路由器工作在强同步方式下,需要全局分布时钟且各时钟相位要求严格对准,两相邻结点路由器间消息的传送必须保证在一周期内完成,采用这种强同步方式构造由成百上千个结点组成的MPP系统时,不仅时钟分布和时钟相位对准的难度大,且在不同机柜中的结点间的较长连线限制了网络主频的提高,也限制了系统的可扩展性,而在当今工艺条件下很难进一步减少结点间的互连长度,因此必须研究一种新型的网络互连及路由器技术,以打破这种强同步方式。流水通道即是在这一考虑下提出的一种新的高速互连技术。

在一流水通道互连网络里,路由器间采用源同步技术,采样时钟与被传送数据同时由上一个路由器发出,在一条线上可同时传送多个数据,这使得网络的主频独立于线的长度,与系统中结点间连线长度无关,从而有效地提高了网络传输速度。流水通道的思想在广域网与局域网中早已采用,而在多处理机、多计算机的互连中迄今为止还采用较少,特别是在MPP系统中,目前仅有CRAY T3E等少数几个系统中采用了这类互连技术,在采用流水通道的路由器中必须采用特殊技术,如锁相技术,来解决消息的源同步传送和与源同步接收问题。在MPP系统中实现流水通道是大幅度提高网络的传输速率和系统性能的关键技术。

小结 MPP系统中互连通信性能是影响系统性能的主要因素之一,也是决定其使用效率的关键,要实现高性能的互连通信须从网络接口、通信方式、互连网络技术和线上传输技术等多方面协同考虑,以设计出高效的互连通信系统[6]。此外,在MPP系统的设计中,还可采用隐蔽延迟的技术,将上述多种技术综合应用可更好地提高系统的通信性能。

参考文献

1 Eicken T V, et al. Active messages: a mechanism for integrated communication and computation. In: Proc. 19th Int. Symp. Computer Architecture. 1992. 256～266

2 Blumrich M A, et al. Virtual-memory-mapped network interfaces. IEEE Micro, 1995(Feb): 21～28

3 Pakın S, et al. Fast messages (FM): efficient, portable communication for workstation clusters and massively-parallel processors. available at: achien @ cs. uiuc. edu. 1997

4 Scott S L, Thorson G M. The CRAY T3E network: adaptive routing in a high performance 3D torus. available at: sls@cray. com, 1997

5 刘燕,徐炜遐,杨晓东. 流水通道——一种高速MPP系统互连. 计算机学报,1999

6 刘燕. 大规模并行处理系统高速互连通信技术的研究:[博士论文]. 长沙:国防科技大学,1998

(上接第47页)

特点和要求,综合传输效率和可靠性这两个指标进行考虑;

·考虑到大型dVEs系统规模(包括地理跨度和实体数量)的庞大,已不可能,也没有必要向每个用户发送刷新消息。这里需对消息传送的目的地进行选择和过滤。为此,DIS采用了AOI与多信道广播相结合的方法,并在节省带宽方面取得了显著的成效。

DIS是迄今为止最成功的分布式虚拟现实设计标准,在我们研究大型dVEs系统的消息传送机制时,可从DIS得到许多启发,如:标准的消息格式、全分布式模型、呆推断、使用多信道广播过滤刷新消息的尝试等。然而,这并不意味着我们可以简单地使用DIS。DIS是针对一特定应用领域而设计的,因而存在一些局限性。如:扩充性不好,解释其位模式的计算开销较大,是一种个体较"大"的标准。下一代的DIS应具备:更简捷、开放、可扩充以及可动态修改的特点。这种可动态调整的协议对于测试和评价分布式实体的交互的效率是必需的。

参考文献

1 Stytz M R. Distributed Virtual Environments. IEEE Computer Graphics and Application, 1996(May): 19～31

2 Brutzman D. Graphics Internetworking: Bottlenecks and Breakthroughs. Available at: Http: // www. stl. nps. navy. mil/～ brutzman/VRml/breakthroughs. html

3 Macedonia M R, et al. Exploiting Reality with Multicast Groups: A Network Architecture for Large-Scale Virtual Environments. macedonia@cs. nps. navy

4 JDS Pugh. Eliminating Network Traffic Caused by Deterministic Objects in Multi-user Virtual Worlds. jds @postoffice. utas. edu. cn

5 Macedonia M R, et al. NPSNET: A Network Architechture for Large Scale Virtual Environment. macedonia@cs. nps. navy

6 Macedonia M R, et al. A Taxonomy for Networked Virtual Environments. mmacedon@crcg. edu