doi:10.11959/j.issn.1000-436x.2016058

未来网络 XIA 中的虚拟机跨子网迁移

孟宏伟[1,2,3]，陈钟[1,2,3]，孟子骞[1,2,3]，SONG Chuck[4]

(1. 北京大学信息科学技术学院，北京 100871；2. 北京大学高可信软件技术教育部重点实验室，北京 100871；
3. 北京大学网络和软件安全保障教育部重点实验室，北京 100871；4. 卡耐基梅隆大学计算机学院，匹兹堡 15213)

摘 要：在 IP 网络中，虚拟机跨子网迁移后其网络地址发生了变化，将面临 IP 移动性问题。主要研究如何在未来网络体系结构—XIA（expressive internet architecture）中解决这一问题。利用 XIA 中标识与地址分离、基于 DAG（directed acyclic graphs）的灵活路由等特性，提出了基于集合点代理的虚拟机在线迁移方法，并进行了具体实现和实验验证。结果表明，所提出的方法可满足虚拟机迁移后与通信对端网络连接的快速恢复，并具备控制平面简单和数据平面高效的优点。

关键词：虚拟机在线迁移；未来网络体系结构；可表达网络；集合点代理

中图分类号：TP302 **文献标识码：**A

VM migration across subnets in future internet architecture—XIA

MENG Hong-wei[1,2,3], CHEN Zhong[1,2,3], MENG Zi-qian[1,2,3], SONG Chuck[4]

(1. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China;
2. Key Laboratory of High Confidence Software Technologies (Peking University) Ministry of Education, Beijing 100871, China;
3. MoE Key Laboratory of Network and Software Security Assurance, Peking University, Beijing 100871, China;
4. School of Computer Science, Carnegie Mellon University, Pittsburgh 15213, USA)

Abstract: When a VM migrates among hosts residing in two IP subnets, the network attachment point of VM changed, which causes the IP mobility problem. It was meaningful to design and evaluate VM migration performance under the future internet architecture. XIA (expressive Internet architecture) was focused, a novel future internet architecture, support for VM live migration. Motivated by the natural f res of ID/location decoupling, versatile routing with DAG (directed acyclic graphs) in XIA, The design and implementation of rendezvous agent based migration (RABM) approach was given. In demonstrate that XIA supported VM migration can achieve fast network re-connection between VM and correspondent node, while keeping the network control plane simplicity and data plane efficiency.

Key words: virtual machine live migration, future internet architecture, expressive internet architecture, rendezvous agent

1 引言

由于传统互联网在路由可扩展性、安全性、移动性以及满足用户需求变化等方面的诸多缺陷，未来网络体系结构（FIA，future internet architecture）正逐步成为全球性研究热点[1~4]。未来网络体系结构是区别于现有 IP 网络，采用新的命名与路由规则、网络协议、运行机理以及管理机制设计的网络体系结构。近年来，世界各国针对未来网络研究已经制定了系统性的研究计划，如美国的 FIA 计划、欧盟的 FP7 计划、日本的 AKARI 计划等，这些计划支持了众多新型网络体系结构的研究项目。典型的未来网络体系结构如美国 NSF 在其 FIA 计划[5, 6]中资助的 NDN（named data networking）[7]、MobilityFirst[8]和 XIA[9]。这些新型的网络体系结构，围绕内容高效获取、泛在移动、安全可信等互联网发展趋势与特点进行了全新设计。其中，NDN 注重提升内容的共享和分发效率，MobilityFisrt 主要解

收稿日期：2015-07-25；**修回日期：**2015-11-06
基金项目：国家自然科学基金资助项目（No.61170263）
Foundation Item: The National Natural Science Foundation of China (No.61170263)

决海量设备的泛在移动，XIA 综合考虑了网络演进、内生安全和灵活路由等特性。到目前为止，学术界关于未来网络体系结构没有达成一致的意见，尚处于百家争鸣阶段。但是从研究的角度所达成的共识是：未来网络体系结构需要在具体应用场景的驱动下不断优化完善。目前，内容共享、VoIP、车联网、物联网等应用，已经在这些新的网络体制下开展了实验和验证[10, 11]，并促进了网络体系结构的改进和提升。

本文基于未来网络体系结构 XIA[12~15]，针对虚拟机在线迁移（virtual machine live migration）这一应用场景进行研究。虚拟机在线迁移是指在不影响虚拟机运行的前提下，将虚拟机从一台宿主机移动另外一台宿主机的过程。这一技术可使虚拟机作为独立的计算单元按需动态移动，契合了未来互联网众多应用需求，主要包括如下内容。1) 企业服务向数据中心转移。在商用领域，中小企业为了节约成本，会逐步将自建 IT 环境中的软件服务转移到数据中心，而且在转移过程中不影响业务的正常运行。在军事领域，美国国防部在其云计算战略[16]中也提出了将遗留系统向云环境过渡的目标，并且要求在迁移过程中保证重要军事服务的连续性和服务质量。2)数据中心服务质量保证。为了保证租户的服务质量，数据中心会根据负载均衡、流量控制、灾难恢复等要求，将虚拟机在同一数据中心内部或多个数据中心之间动态迁移，迁移过程中要尽量保证虚拟机所提供服务的连续性，不能造成服务中断或服务降级。3)提供就近服务。服务提供商为了提高服务响应速度，会根据服务/内容访问热度将特定的服务即时迁移到距离用户较近的位置。4)支持“Follow Me Cloud”[17, 18]。为满足时间敏感性高的应用要求，让虚拟机跟随“主人”迁移到本地运行，以降低跨广域网的传输时延。

虚拟机监视器（virtual machine monitor）又称为 hypervisor，如 VMWare、Xen、KVM 等，能够较好地支持虚拟机在局域网内的在线迁移，但对于跨子网迁移的情况不能很好地满足。这是因为虚拟机跨子网迁移后网络接入点地址（IP 地址）发生了变化，需要额外的机制恢复虚拟机与通信对端（CN）的通信连接。针对这一问题，研究人员基于 IP 网络和 ILNP（identifier-locator network protocol）、LISP（locator/identifier separation protocol）等新型网络开展了大量研究[19~28]。但这些解决方案相对复杂，灵活性不强，当大量虚拟机频繁迁移时，面临可扩展性问题。本文基于未来网络体系结构 XIA，提出了基于集合点代理的虚拟机跨子网迁移方法，并进行了实验验证和分析。该方法降低了虚拟机迁移后数据分组重定向的复杂度和开销，避免了三角路由和隧道转发，不仅能够支持虚拟机在线迁移，也可作为一种通用的移动性管理方案，解决 XIA 网络中主机和设备的移动性问题。

2 相关工作

2.1 虚拟机在线迁移及其问题描述

虚拟机在线迁移主要包括虚拟机镜像文件传输和网络连接重同步 2 个过程[19]。

1) 虚拟机镜像传输是指虚拟机文件系统、CPU 和内存状态的拷贝传输过程。源宿主机在保持虚拟机运行的同时，使用预拷贝算法，采取多轮迭代的方式，将虚拟机内存页面拷贝到目的宿主机。当内存脏页数小于一定阈值时，将虚拟机停机。在完成最后的内存脏页传输后，虚拟机在目的宿主机重新启动。虚拟机的停机时间大致相当于最终内存脏页的传输时间。由于最终内存脏页与整个镜像文件相比数据量很小（通常为几十兆字节），在线迁移可极大地缩短虚拟机停机时间。

2) 网络连接重同步是指恢复虚拟机与通信对端之间已经建立的通信连接。网络连接重同步根据虚拟机迁移网络边界范围不同而有所差别，包括局域网内部和跨子网（或跨网络域）2 种情况。虚拟机在局域网内迁移时，IP 地址可以保持不变，迁移后只需要 LAN 交换机根据虚拟机 ARP 广播更新其所对应的 ARP 表和端口，就能将发往原虚拟机的数据分组发送到迁移后位置。但是对于跨子网迁移情况，迁移完成后并不能正常工作。这是因为迁移到其他子网后，虚拟机的网络接入点地址发生了变化，这将带来 2 个问题：1) 发往原地址的数据分组无法路由到虚拟机当前位置；2) 虚拟机中与原 IP 地址关联的 TCP 连接失效。从根本上看，这是由于 IP 网络中传输层连接标识与网络层地址耦合造成的。通信连接中断会严重影响虚拟机所提供服务的连续性和服务质量，对于关键服务和时间敏感性服务尤为重要。

2.2 现有研究工作

2.2.1 IP 网络的解决方案

1) 数据链路层的解决方案

利用 VXLAN[20]、VPLS（virtual private lan service）[21]等技术，在不同子网之间建立二层虚拟

2016058-2

网络，以此屏蔽虚拟机迁移后 IP 地址变化。但是建立二层虚拟网络需要交换设备支持，以及相对复杂的配置和维护，对于有大量虚拟机动态迁移的情况，无疑会增加数据中心网络基础设施管理的成本。另外，通常数据中心出于安全防护、负载均衡等考虑，会将数据中心网络划分为不同的子网以方便管理。但是为了虚拟机迁移目的将不同数据中心中不同的子网划分到同一个二层的网络域，会与数据中心网络的管理策略相冲突，给数据中心网络维护和配置带来了负担。

2) 网络层的解决方案

网络层的解决方案[22, 24]主要使用 IP 隧道和动态 DNS 组合的方法。1)通过在虚拟机原地址和新地址之间建立 IP 隧道，将发往虚拟机原地址的数据分组重新封装并重定向到虚拟机的新地址，解决已建立连接的数据分组转发问题。2)虚拟机迁移后更新 DNS 中虚拟机的 IP 地址，确保新建连接使用的是虚拟机最新的 IP 地址，但是这种方法并不高效。①所有发往虚拟机原地址的数据分组都需要经过 IP 隧道，增加了数据分组经过隧道封装和解封装的处理开销，降低了数据分组转发的效率；②直到没有发往原地址的数据分组后，隧道才能关闭，增加了状态维护的开销；③发往原地址的数据分组都要经过原来虚拟机所在的网络，造成了三角路由问题，数据分组传输路径并不是优化路径；④如果 IP 隧道设置在源宿主机中，会增加源宿主机的资源占用，违背了由于宿主机资源受限而需要将虚拟机迁移到他处运行的初衷；⑤每次虚拟机的迁移都需要单独配置和管理隧道，可扩展性不强，无法应对有大量虚拟机迁移的情况。使用 Mobile IPv6 中的路由优化机制可以消除三角路由问题[25]，在虚拟机迁移后，向已经与其建立连接的所有 CN 发送绑定更新消息（BU, binding update）。CN 在收到 BU 消息后更新虚拟机地址，然后直接向虚拟机的新地址发送数据分组，从而消除了虚拟机原地址和目的地址之间建立的长期隧道和三角路由。但是该方法增加了虚拟机和通信对端的移动性管理开销，重新同步过程需要 CN 的配合，增加了资源受限等便携移动终端设备的负担。

3) 传输层的解决方案

通过引入新的传输层标识（TIFID, transport independent flow identifier），并将 TIFID 与 IP 地址动态绑定，也能够支持虚拟机在线迁移[26]。在虚拟机迁移后，由其内部的同步代理（synchronization agent）向 CN 发送绑定更新。CN 端的同步代理收到绑定更新后，将与 TIFID 对应的 TCP 连接重新绑定。但这种方法不能支持虚拟机和 CN 同时移动的情况，若发生同时移动，虚拟机和 CN 都无法与对方取得联系。

2.2.2 其他网络协议的解决方案

ILNP[27]和 LISP[28]都采用了标识和位置分离的设计，其中，ID 是传输层标识，Locator 是网络层标识。在 ILNP 中，由边界路由器（SBR, site border router）负责虚拟机 ID 和 Locator 的动态绑定。虚拟机迁移之后，向 CN 发送(LU, locator update)消息，将移动后的新 Locator 通知 CN 所在网络的 SBR。CN 所在网络的 SBR 收到 LU 后，更新虚拟机 ID 与 Locator 的映射，进而完成向虚拟机最新地址的转发。在 LISP 中，由映射系统完成 ID 与 Locator 的绑定。虚拟机迁移完成后，一方面向映射系统和源网络路由器发送 Locator 更新消息。另一方面，由源网络路由器根据之前的通信列表向所有 CN 所在网络的路由器发送更新通告，触发其查询映射系统，并获取虚拟机的最新地址。

ILNP 和 LISP 中的方法不仅需要在映射系统中更新 ID 对应的 Locator，还要求所有 CN 所在网络的路由器更新关于虚拟机的绑定缓存。每次迁移会引发全局性的更新，当 CN 数量巨大且在网络中比较分散时，会带来巨大的信令开销，存在可扩展性问题。

2.2.3 虚拟机跨子网在线迁移的要求

通过对虚拟机迁移应用场景和相关研究工作综合分析，虚拟机跨子网在线迁移需满足以下要求。

1) 最小化对虚拟机中服务的影响。用户感受到的服务质量不能因为虚拟机迁移而降级，更不能造成重要的应用或服务中断，要求虚拟机迁移后与其通信对端连接重新同步时间越短越好。2)不增加数据中心网络配置的复杂度。数据中心网络不需要为支持虚拟机迁移而进行特殊的设置，虚拟机的迁移也不会影响到现有数据中心网络配置。3)降低虚拟机迁移后维护网络连接的开销。尽量减少原网络中 hypervisor 或路由器对维护虚拟机迁移后网络连接的资源占用。另外，还要考虑尽量降低整个网络的开销，不能因为虚拟机网络接入点的变化而引发全局性路由表更新。4)对上层应用和通信对端透明。对应用层来说，应用程序应该感知不到虚拟机迁移的发生，无需增加额外的编程来支持虚拟机迁移，最好由网络层来提供对虚拟机迁移的支

持。CN 无需增加特殊的功能配合虚拟机迁移，要尽量减少与 CN 控制消息的交互，降低相关协议的复杂度。

2.2.4 XIA 概述

XIA 是美国国家科学基金（NSF, national science foundation）未来网络体系结构研究计划支持的主要项目之一，由卡耐基梅隆大学的研究团队提出。XIA 早在 2010 年开始的 FIA 计划[5]中就得到了支持，并在新一期的 FIA-NP 计划[6]中得到了继续资助，XIA[12~15]的主要特点包括可演进、可信（trustworthy）和灵活路由（flexible routing）3 个方面。

1) 可演进。网络体系结构的细腰不再局限于某一种特定通信主体，而是可包容多种通信主体，并支持未来可能的新的通信规则定义与加入，从而支持网络长期演进。目前，XIA 中定义了 4 种通信主体，包括网络域（network domain）、主机（host）、服务（service）和内容（content）。这 4 种通信主体标识分别用 NID、HID、SID 和 CID 表示。NID 代表网络域或子网，路由时用来定位网络地址；HID 代表主机标识，支持单播（unicast）路由寻址；SID 代表网络中的服务实例，支持任播（anycast）；CID 是内容标识，便于网络中的内容获取。

2) 可信。NID、HID 和 SID（统称为 XID）都是通过对各自公钥进行散列算法得到的 160 bit 二进制数据。基于 XID，XIA 设计了支持数据分组的源地址验证的机制。首先第一跳路由器（first-hop router）负责验证与其连接的主机是否有源地址欺骗行为（HID verification）。其次，数据分组所经过的所有的边界路由器上，会检查数据分组的上一跳 NID 是否合法（NID verification）。这种认证机制保证了数据分组从主机到边界路由器，以及边界路由器之间的可信转发，是一种内生安全机制。

3) 灵活路由。XIA 使用了有向无环图（DAG, directed acyclic graphs）的地址结构，支持灵活路由。XID 作为 DAG 中的节点（node），通过相互指向表示一个地址结构，这种方法区别于 IP 地址和 NDN 命名采用的字符串方式。

一台主机基本的 DAG 如图 1（a）所示。Ÿ表示概念上的起点，没有具体含义，HID 为终点，是主机标识。终点又称为“Intent”，为想要到达或者获取的对象。如果终点为 HID，表示想要到达的主机；若终点为 CID，表示想要获取的内容；若终点为 SID，表示想要获取的服务。这种地址表示方式，将意图直接表达在 DAG 中，也就是 XIA 中“expressive”的内涵。这种路由方式为网络提供了针对内容或服务的优化机会。由于网络中 HID、SID 和 CID 的数量巨大，为了提高路由效率，图 1（b）给出了支持层次化路由的 DAG。路由器首先要将数据分组发送到以 NID 为标识的子网，当到达 NID 后，再将数据分组转发到标识为 HID 的主机上。这种方式相当于 IP 中按照网络地址和主机地址的寻址路由方式。图 1（c）说明了对于不能直接寻址到 HID 的情况，可通过 fallback 路径（虚线），先转发到备用的重定向节点 *N*，然后再转发到 HID。虚线表示的 fallback 路径在路由时优先级较低，当路由表中不存在到下一个节点直接路径（实线）的情况下使用。这种路由机制为路由可靠性（fallback 指向备用路由节点）和未来网络与现有网络兼容（fallback 指向 IP 节点）提供了很大的灵活性。

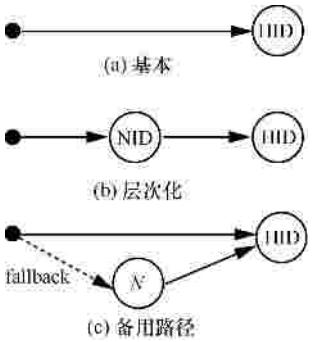

图 1 主机 DAG 举例

3 基于集合点代理的虚拟机在线迁移

3.1 基本机制

在 XIA 中，虚拟机的 DAG 可表示为 NID_1→HID，其中 NID_1 为虚拟机所在的网络标识，HID 为虚拟机标识。虚拟机跨子网迁移到目的地网络 NID_2 后，虚拟机的 DAG 变为 NID_2→HID。迁移后，发往虚拟机的数据分组依然会按照原来的路径发送到 NID_1。由于虚拟机已经迁移离开，所以该数据分组会被丢弃。

为了解决这一问题，利用 DAG 中的 fallback 机制，并引入了集合点代理（RA, rendezvous agent），提出了基于集合点代理的虚拟机迁移方法。按照 XIA 中基于 DAG 的路由原理，当路由器无法直接路由时，会根据 DAG 中的 fallback 路径转发。因此，在虚拟机的 DAG 中增加一个 fallback 路径，并且将 fallback 指向 RA，由 RA 负责数据分组转发到虚拟

机最新位置。采用了 fallback 的虚拟机 DAG 如图 2 所示。

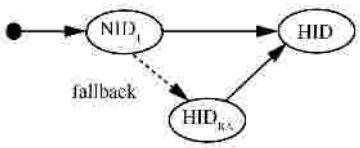

图 2　采用了 fallback 的虚拟机 DAG

基于集合点代理的虚拟机迁移机制如图 3 所示。1)虚拟机迁移后，发往虚拟机的数据分组按照原来的 DAG 发送到虚拟机原来所在网络路由器（源路由器）；2)源路由器中指向虚拟机的路由表项将被超时删除，源路由器接收到 CN 发往虚拟机的数据分组后，无法通过正常路径（DAG 中的实线）路由，会根据 fallback 将数据分组转发到 RA；3)RA 接收到转发来的数据分组后，根据虚拟机迁移后注册的最新 DAG，将数据分组转发到虚拟机的最新位置；4)虚拟机接收到 RA 转发的数据分组后，使用新的 DAG 向 CN 进行回复；5)CN 收到虚拟机迁移后发来的数据分组，根据虚拟机的新地址更新 Xsocket 绑定（类似 socket），之后按照虚拟机最新 DAG 发送数据分组，完成通信连接中同步。

RA 作为虚拟机迁移后的移动锚点，作用相当于 Mobile IP 中的家乡代理（HA, home agent）。RA 维护虚拟机的最新位置信息，并且负责转发数据分组，但是 RA 与 IP 网络中的 HA 有所区别。1) 网络位置不同。RA 不必限定在虚拟机的家乡网络中，而 HA 必须位于移动节点的家乡链路上。RA 可以作为一种网络服务部署在网络中任意位置。移动节点可以自由选择 RA 作为其 fallback 指向的移动锚点，选择的策略能够根据不同的应用场合灵活设置。比如在考虑灾难恢复时，RA 的选择就要避开原来虚拟机所在的网络。2) 数据分组转发方式不同。RA 通过修改数据分组头的 DAG 进行转发，而 HA 则需要通过隧道方式进行转发。采用隧道方式的情况下，HA 端的数据分组封装和虚拟机端的数据分组解封装会带来额外开销和性能损失。从网络层次结构上看，本文提出的方法属于网络层的解决方案，虚拟机迁移后可以自动与通信对端恢复通信连接。对上层应用透明，无需在源宿主机上增加如隧道、重定向、主动更新路由表等额外的操作。

3.2　主要流程

虚拟机从开始准备迁移，到迁移后完成通信连接恢复，主要分为 4 个阶段，如图 4 所示。

1) 迁移前，虚拟机与通信对端正常通信

虚拟机启动后，接收到 XHCP 广播（类似 DHCP），得到当前网络 NID 和 RA 的相关信息。

虚拟机向 RA 发送注册请求消息（registration request message），注册当前 DAG。

RA 收到注册请求消息后，将虚拟机加入绑定缓存（BCE, binding cache entry），并返回注册响应消息（registration reply message）。

虚拟机收到注册响应消息后，将 fallback 指向 RA，完成 DAG 配置，与通信对端建立连接。

2) 迁移开始，虚拟机正常运行

源宿主机输入迁移命令，虚拟机开始迁移。通过不断地迭代拷贝，将虚拟机的镜像文件和内存状态向目的宿主机传输。虚拟机没有停机，与外界通信正常。

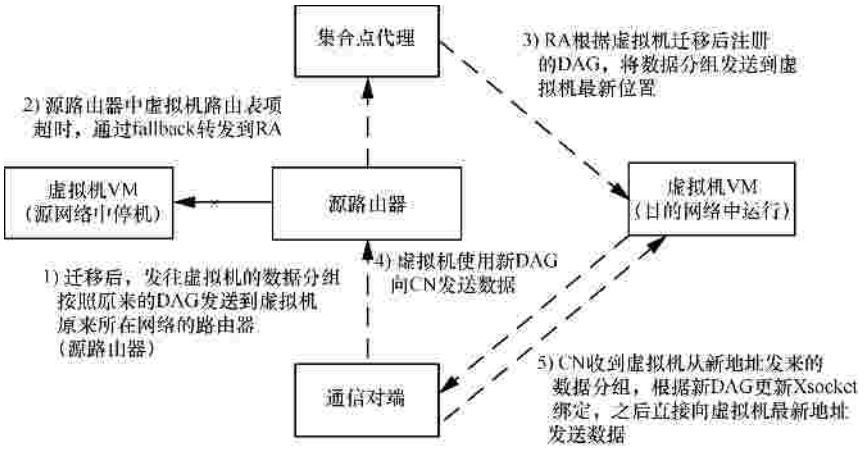

图 3　基于集合点代理的虚拟机迁移机制

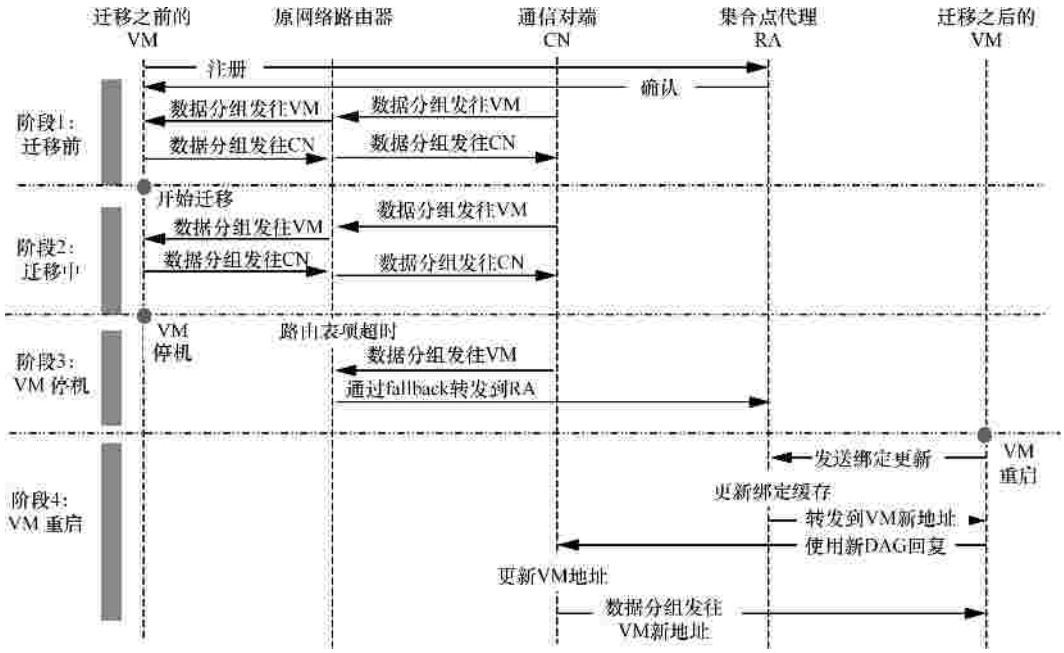

图 4 基于集合点代理的虚拟机迁移流程

3) 虚拟机停机，完成最后的脏页传输

迭代拷贝进入最后阶段，虚拟机停机，等待最后内存脏页数据的传输完成。虚拟机停机后，无法向源路由器发送路由更新消息，导致源路由器虚拟机路由表项超时。源路由器认为虚拟机已经不再网络中，把虚拟机路由项从路由表中删除。

4) 虚拟机重启，与通信对端恢复通信

虚拟机在目的宿主机重新启动后，同样会接收到目的地网络的 XHCP 广播消息。通过对 XHCP 中的 NID 信息与之前 DAG 中的 NID 进行比对，虚拟机发现所在网络的 NID 发生了变化（如 NID_1 变为 NID_2）。虚拟机更新 DAG 中的 NID，将 NID_1 变为 NID_2。

虚拟机发现 NID 变化后，一方面，使用新的 DAG 向 RA 发送绑定更新（BU, binding update），另一方面把 DAG 更新情况通知 Name Service。

RA 收到 BU 消息后，更新绑定缓存。

以虚拟机原来的 DAG 为目的地址的数据分组，被路由到虚拟机原来所在网络的路由器。由于源路由器中不存在虚拟机 HID 对应的路由项（被超时删除），源路由器会使用 fallback 将数据分组发送到 RA。

RA 收到源路由器转发的数据分组后，根据绑定缓存修改目的 DAG 进而转发。

虚拟机收到由 RA 转发的数据分组后，使用新 DAG 直接向 CN 发送数据分组。

在反向路径上，如果 CN 收到由虚拟机最新位置发来的数据分组，CN 会自动更新其 Xsocket 中与虚拟机绑定的 DAG。之后虚拟机通信中所有数据分组都会使用新的 DAG 直接发送到虚拟机最新位置，不需要经过 RA 转发。虚拟机与 CN 之间的通信连接重同步过程完成。

4 设计与实现

按照虚拟机跨子网迁移的应用场景构建的实验环境如图 5 所示。

实验环境分为 NID_1 和 NID_2 这 2 个子网，XIA $Router_1$ 和 XIA $Router_2$ 分别是 NID_1 和 NID_2 的网关路由器。宿主机 $Host_1$ 位于 NID_1，宿主机 $Host_2$ 位于 NID_2，分别连接在 $Router_1$ 和 $Router_2$ 上。RA 位于子网 1 中，虚拟机的初始位置在宿主机 $Host_1$ 上，在实验中动态迁移到宿主机 $Host_2$。通信对端 CN 连接在 $Router_2$ 上，在虚拟机迁移过程中与虚拟机进行通信，用来观察迁移过程对服务的影响。Router、Host 都是运行了 XIA 协议栈（XIA 协议栈通过 Click 路由器实现）的 PC 机，操作系统为 Linux (Ubuntu 12.04) 。PC 机之间通过 100 bit/s 以太网交换机连接。基于 Click 的 XIA 协议栈从 GitHub 获得（http://www.github.com/xia-project/xia-core）。下面介绍每个网络元素的实现细节。

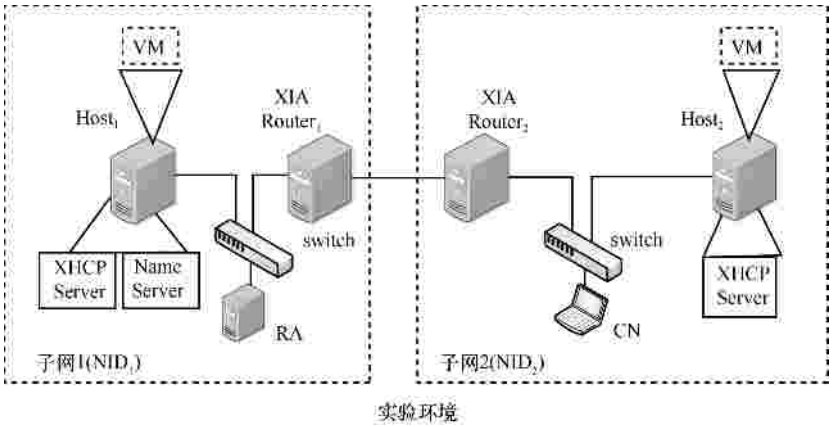

图 5　虚拟机迁移实验环境

4.1　宿主机与虚拟机

宿主机使用的虚拟机监视器为 KVM（kernel-based virtual machine），宿主机和 VM 之间的网络连接选取的是桥接模式（bridging mode）。$Host_1$、$Host_2$ 和 VM 采用了 XIA 原型系统中主机的典型配置。为缩短虚拟机镜像文件的传输时间，源宿主机和目的宿主机之间采用了共享存储的方式，使用网络文件系统（NFS）共享虚拟机镜像文件。这样，虚拟机迁移传输的主要是虚拟机内存和 CPU 状态，以省去镜像文件传输，方便多次实验。

4.2　XIA 路由器

$Router_1$ 和 $Router_2$ 由 2 台配置了双网卡的 PC 机实现，使用的是 XIA 中 Router 的典型配置。本文在 Router 中增加了路由表项超时机制。如果路由器在一定时间内没有收到主机注册信息，定时器发生超时后将删除对应的路由表项。当虚拟机迁移后，源路由器中关于虚拟机的路由表项被超时删除，确保了源路由器将发往虚拟机的数据分组通过 fallback 转发给 RA。

4.3　Name Server

Name Server 在 XIA 中类似于 DNS，提供名称解析服务，完成名称与网络地址 DAG 的绑定。通信发起方可以通过查询 Name Server 获取通信对端的 DAG。当虚拟机的 DAG 改变时，要在 Name Server 中进行更新，以方便后续其他用户查找。在实验环境中，Name Server 运行在 $Host_1$ 上。

4.4　通信对端

在 XIA 传输层协议中，接收方会对数据分组头中的源地址字段进行检查，对比 DAG 是否有变化，若有变化就进行更新，下次回复数据分组将使用更新后的地址。在虚拟机迁移后，CN 会发现虚拟机 DAG 中的 HID 相同但 NID 不同，则判断数据分组来自同一个虚拟机，与之前是同一个连接，将替换并使用最新的 DAG 为目的地址。在虚拟机迁移之后，只要 CN 收到虚拟机发送的数据报，就可以自动完成通信连接的重新同步，直接向虚拟机的新地址发送数据，而无需虚拟机向 CN 发送专门的通告消息，简化了路由优化机制。

4.5　集合点代理

RA 是在 XIA 网络中新增的功能实体，提供对移动节点或者虚拟机的移动性管理功能。在部署上，RA 既可以独立运行，也可以与路由器绑定到一起。RA 作为移动锚点的功能不仅局限在移动节点的家乡网络中，它也可以作为其他与 RA 不在同一个 NID 中移动节点的代理。也就是说，在 XIA 网络中有了 RA，移动节点可以自由选择 RA 作为其 fallback 指向的移动锚点。这样对移动性的支持更加地灵活，能够根据支持移动节点的具体要求（安全、实时性等）灵活设置。RA 的主要功能包括：1）虚拟机位置管理，负责对虚拟机位置进行注册，接收虚拟机迁移后的 BU 消息，更新虚拟机对应的绑定缓存；2）数据分组转发，接收 CN 发往虚拟机的数据分组，并按照虚拟机对应的绑定缓存，修改数据分组的目的地址并转发。

5　定量与定性分析

本节通过实验测试 XIA 中虚拟机迁移给用户使用服务带来的影响，并给出定量和定性分析。

5.1　服务响应时间

服务响应时间是指从用户的角度来看，由于虚拟机迁移引起的服务迟滞时间。为了测定服务响应时间，在虚拟机中运行 plus server 程序，在通信对

端运行 plus client 程序：双方建立连接后，通信对端周期性地向虚拟机发送一个数字，虚拟机将用户端发送的数字加 1 后返回。通信对端记录下每个数字返回的时间，即为服务响应时间。需要指出这是一种简化方法，精确服务响应时间的计算要比本文复杂。虚拟机迁移前后的服务响应时间如图 6 所示。

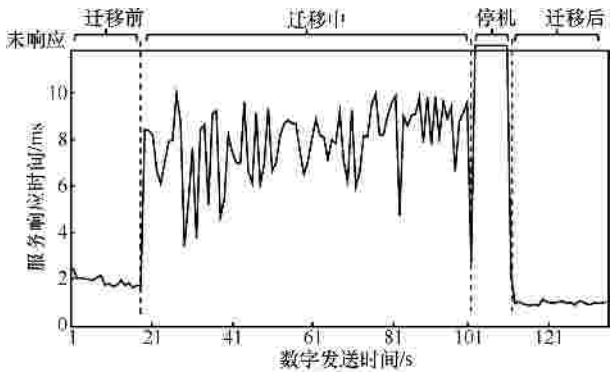

图 6 虚拟机迁移前后的服务响应时间

从图 6 中可以看出，服务响应时间呈现出 4 个阶段的差别，每个阶段服务响应时间的均值和方差如表 1 所示。

表 1 服务响应时间

阶段	均值/ms	方差
迁移前	1.925	0.205
迁移中	8.166	1.55
停机	—	—
迁移后	1.028	0.23

其中，在虚拟机迁移阶段的服务响应时间（均值为 8.166 ms）明显大于迁移前、迁移后的服务响应时间（均值分别为 1.925 ms、1.028 ms），而且服务响应时间的波动也很大。这是因为在迁移过程中要执行拷贝传输，虚拟机内存脏页传输占用了大量网络带宽，增加了虚拟机与通信对端的通信时延，并且由于带宽占用的不规律性，造成了通信时延方差较大。而在虚拟机迁移前和迁移后，并没有其他程序占用带宽，所以这 2 个阶段的服务响应时间相对较小（2 ms 以内），并且波动不大。在虚拟机停机阶段，虚拟机无法响应通信对端，所以此阶段的服务响应时间认为是无穷大。

另外，虚拟机迁移前和虚拟机迁移后的服务响应时间略有不同（均值分别为 1.925 ms 和 1.028 ms），前者大概为后者的 2 倍。这是因为，在迁移之前，虚拟机与通信对端位于 2 个子网中，通信路径上要经过 2 台路由器（如图 5 所示），会引入额外的传输时延和处理时延。而迁移后二者位于同一个局域网中，相比之下时延会显著降低。可以预料，如果 2 台路由器之间的传输时延增加（如广域网），迁移前后的服务响应时间差别也会随之变大。

5.2 服务中断时间

服务中断时间是由于在迁移最后阶段，虚拟机停机拷贝引起的服务“无响应”时间。从过程上看，服务中断时间是从虚拟机停机开始，直到虚拟机重启并完成网络连接恢复所花费的时间，包括虚拟机停机时间、虚拟机网络配置并完成与 RA 绑定更新的时间。在这段时间内，由于虚拟机挂起，网络没有配置完成，虚拟机中的服务无法与外界发生交互。其中，虚拟机停机时间是指虚拟机在源宿主机停机，完成最后的内存脏页传输，以及在目的宿主机重启的时间。虚拟机网络配置时间是指虚拟机在目的地网络获取新的网络接入点地址，以及向新的接入路由器完成路由表注册的时间。虚拟机与 RA 绑定更新的时间，是指虚拟机获得新 DAG，向 RA 发送绑定更新消息并收到绑定更新确认的时间。

在实验中，通信对端若在一定时间内（实验中为 1 s）没有收到虚拟机回复，则会进行重发，直到收到回复为止，这段时间即为服务中断时间。经过多次实验得到的服务中断时间均值为 4 350 ms。应当注意到，这一时间主要被最后阶段的内存脏页传输所占用。实验中，需停机拷贝阶段的脏页内存为 54 MB，网卡速率为 100 Mbit/s，内存脏页传输的时间大致为 4 320 ms。而虚拟机启动、进行网络配置并完成与 RA 绑定更新的时间仅为 50 ms。因此，服务中断时间主要取决于最终脏页内存的传输时间。如果提升网络传输速率到 1 Gbit/s，服务中断时间可以缩小到 1 s 以内。当前数据中心之间多以专线连接，带宽通常达到 1~10 Gbit/s，因此将虚拟机停机时间控制在 1 s 内甚至几百毫秒是可能的，可基本满足网页浏览、视频播放等应用要求。在这种情况下，虚拟机与 RA 的绑定更新时间就会成为影响服务中断时间的关键因素。如果虚拟机与 RA 在广域网中相隔较远，通信传输时延有可能达到秒级。所以，有必要研究 RA 合理部署以缩短绑定更新时间。

5.3　路由优化

三角路由问题存在于 Mobile IPv4 中，这是因为节点移动后数据分组必须通过家乡代理进行转发。Mobile IPv6 中采用了路由优化机制，通过向通信对端发送通告消息的方式避免了三角路由。在 XIA 中，本文提出的方法在避免三角路由的同时，简化了路由优化机制和流程。

为了测试路由优化的效果，比较了 plus server 和 Xping（类似于 IP 中的 ping）的服务响应时间。在 Xping 中，通信对端以一定的时间间隔向虚拟机发送 Xping 命令，通过测量收到 Xping 回复的时间，得出其服务响应时间。虚拟机迁移前和迁移后 2 种应用的服务响应时间如图 7 和图 8 所示。

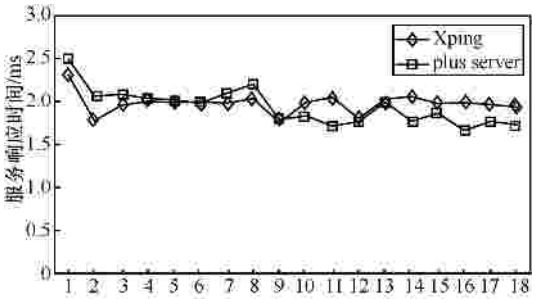

图 7　迁移前服务响应时间对比

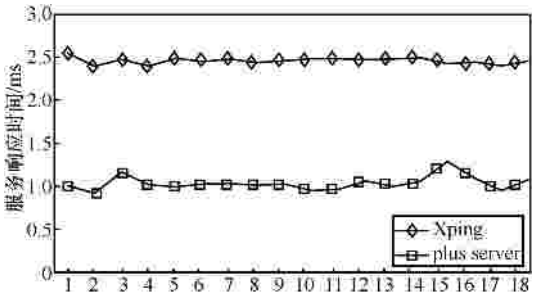

图 8　迁移后服务响应时间对比

从图中可以看出，迁移之前 2 种应用的服务迁移时间大致相同，而迁移之后，Xping 的服务迁移时间明显大于 plus server。这是由于 Xping 基于 UDP，迁移后数据分组发送的目的地址仍然是虚拟机迁移之前的 DAG，导致数据分组首先到达虚拟机源网络路由器，然后再由 RA 转发到目的地网络，形成了三角路由。而 plus server 基于 TCP，虚拟机迁移后，只要通信对端收到虚拟机从新地址发来的数据分组，其传输层协议就会根据新的 DAG 更新目的地址，从而实现路由优化，避免了之后的三角路由。这种方式也无需虚拟机主动向通信对端发送专门的通告消息进行位置更新，实现起来更加简化。因此，对基于 UDP 的应用，需要由上层协议定期查询 DNS（或 name server）获取最新地址，才能避免三角路由。

5.4　关于 RA 的讨论

虚拟机迁移与设备移动的移动特点不同。移动设备的移动性具有一定随意性，而虚拟机迁移是在数据中心预先规划下进行的，有一定的计划性，迁移的目的地是预先确定好的。利用这一特性，可以将虚拟机迁移的目的网络地址作为先验信息在迁移之前写入 RA 中，这样可以省去虚拟机在目的地网络启动后向 RA 发送的绑定更新的时间（通常为一个 RTT），进而缩短服务中断时间。对于网络游戏等交互性强的时敏性应用，这种方法能进一步提高服务质量。

另外，ID 与 Locator 分离的网络体系结构如 ILNP、LISP、MobilityFirst 等，在支持移动性方面都具有一定优势。但是在这些网络体系结构中，ID 与 Locator 的动态绑定是在路由器上完成的（或者依赖于全局的名称解析服务），单次移动事件就会引发全局路由表的更新，当大量设备或虚拟机移动（迁移）时，将导致可扩展性问题。在 XIA 中，通过 RA 支持虚拟机迁移或设备的移动不存在这样的问题。因为 RA 可以根据数据中心或用户的需要按需设置和部署，为网络的移动性支持提供了较大的灵活性。

6　结束语

本文针对虚拟机跨子网在线迁移中的通信连接快速恢复问题，利用未来网络体系结构 XIA 中 ID 与 Locator 分离、基于 DAG 的灵活路由等特点，提出了基于集合点代理的虚拟机迁移方法，并通过实验对服务响应时间、路由优化等进行了验证和分析。结果表明，本文提出的方法具有一定的灵活性，能够满足虚拟机在线迁移的要求。

参考文献：

[1] 黄韬, 刘江, 霍如, 等. 未来网络体系架构研究综述 [J]. 通信学报, 2014, 35(8):184-197 .
HUANG T, LIU J, HUO R, et al. Survey of research on future network architectures [J]. Journal on Communications, 2014, 35(8):184-197.

[2] 任勇, 徐蕾, 叶王毅, 等. 未来网络的研究进展和发展趋势[J]. 中国科技论文在线, 2011, 6(4):247-255.
REN Y, XU L, YE W Y, et al. Research progress on future Internet [J]. Science Paper Online, 2011, 6(4):247-255.

[3] 毕军. SDN 体系结构与未来网络体系结构创新环境[J]. 电信科学,

2013, 1(8): 7-15.
BI J. SDN Architecture and future network innovation environment[J]. Telecommunications Science, 2013, 1(8): 7-15.

[4] **林涛, 李杨, 韩言妮, 等. 融合内容和服务的未来网络体系架构**[J]. **网络新媒体技术**, 2012, 1(6):52-57.
LIN T, LI Y, HAN Y N, et al. A future internet architecture of content and service aware network[J]. Network New Media, 2012, 1(6):52-57.

[5] NSF FIA project[EB/OL]. http://www.nets-fia.net/, 2013.

[6] NSF FIA next phase[EB/OL]. http://www.nsf.gov/news/newssumm.jsp?cntn_id=131248, 2014.

[7] NDN project[EB/OL]. http://www.named-data.net/, 2013.

[8] MobilityFirst project[EB/OL]. http://mobilityfirst.winlab.rutgers.edu/, 2013.

[9] XIA project[EB/OL]. http://www.cs.cmu.edu/~xia/,2013.

[10] ZHANG L X, AFANASYEV A, et al. Named data networking[C]// ACM SIGCOMM Computer Communication Review, c2014.

[11] RAYCHAUDHURI D, NAGARAJA K, VENKATARAMANI A. MobilityFirst: a robust and trustworthy mobility centric architecture for the future Internet[J]. ACM SIGMobile Mobile Computing and Communication Review (MC2R), 2012, 16(4).

[12] ANAND A, DOGAR F, et al. XIA: an architecture for an evolvable and trustworthy internet[R]. Technical Report CMU-CS-11-100, Carnegie Mellon University, 2011.

[13] HAN D S, ANAND A, et al. XIA: efficient support for evolvable internetworking[C]//The 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI'12). San Jose, CA, c2012.

[14] ANAND A, DOGAR F, HAN D S, et al. XIA: an architecture for an evolvable and trustworthy internet [C]//Tenth ACM Workshop on Hot Topics in Networks (HotNets-X). Cambridge, MA, c2011.

[15] NAYLOR D, MUKERJEE M K, et al. XIA: architecting a more trustworthy and evolvable internet[J]. ACM SIGCOMM Computer Communication, 2014, 44(3): 50-57.

[16] Department of Defense of US. Cloud computing strategy[R]. July, 2012.

[17] ANTONESCU A F, GOMES A, et al. Follow-me cloud: an openflow-based implementation[C]// 2013 IEEE International Conference, c2013.

[18] SATYANARAYANAN M, BAHL P, CACERES R, et al. The case for VM-based cloudlets in mobile computing[C]//Pervasive Computing, IEEE, 2009.

[19] FISCHER A, FESSIY A, CARLEY G, et al. Wide-area virtual machine migration as resilience mechanism[C]//30th IEEE Symposium on Reliable Distributed Systems Workshops, c2011.

[20] MAHALINGAM M, DUTT D, DUDA K, et al. VXLAN: a framework for overlaying virtualized layer 2 networks over layer 3 networks[EB/OL]//https://tools.ietf.org/html/draft-mahalingam-dutt-dcops-vxlan-02.

[21] KOMPELLA K and REKHTER Y. Virtual private LAN service (VPLS) using bgp for auto-discovery and signaling[EB/OL]. http://www.ietf.org/rfc/rfc4761.txt.

[22] AJILA A S, IYAMU O. Efficient live wide area vm migration with IP address change using type ii hypervisor[C]//IEEE IRI 2013. San Francisco, California, USA, c2013:14-16,

[23] BRADFORD R, KOTSOVINOS E, et al. Live wide-area migration of virtual machines including local persistent state[C]//VEE'07. San Diego, California, USA, c2007.

[24] LI Q. Hypermip: hypervisor controlled mobile ip for virtual machine live migration across networks[C]//High Assurance Systems Engineering Symposium. c2008: 80-88.

[25] HARNEY E, GOASGUEN S, et al. The efficacy of live virtual machine migrations over the internet[C]//VTDC'07. Reno, NV, USA, c2007.

[26] KALIM U, GARDNER M K, BROWN E J, et al. Seamless migration of virtual machines across networks[C]//Computer Communications and Networks (ICCCN). c2013.

[27] RAAD P, COLOMBO G, et al. Achieving sub-second downtimes in internet-wide virtual machine live migrations in LISP networks[C]// Integrated Network Management. c2013.

[28] BHATTI S N, ATKINSON R. Secure & agile wide-area virtual machine mobility[C]//Military Communications Conference. c2012.

作者简介：

孟宏伟（1983-），男，山西神池人，北京大学博士生，主要研究方向为未来网络体系结构、网络与信息安全。

陈钟（1963-），男，江苏徐州人，北京大学教授、博士生导师，主要研究方向为计算机软件与理论、密码学、网络与信息安全。

孟子骞（1990-），男，北京人，北京大学博士生，主要研究方向为未来网络体系结构、网络与信息安全。

SONG Chuck（1957-），男，卡耐基梅隆大学高级研究员，主要研究方向为计算机网络、未来网络体系结构。