doi:10.3969/j.issn.1000-436x.2014.02.021

现代数据中心网络资源管理技术分析与综述

邓罡，龚正虎，王宏，陈琳，刘志宏
（国防科学技术大学 计算机学院，湖南 长沙 410073）

摘 要：现代数据中心网络结构特征和应用模式的深刻变革，给网络资源管理带来了全新的挑战。地址自动配置技术、传输控制技术、流量管理技术以及虚拟化管理技术等是现代数据中心网络资源管理的重要内容，也是近年来学术界研究的重要方向。结合当前的研究现状，对以上几个方面的最新研究成果进行分析综述，并就网络资源管理未来的发展趋势进行展望。

关键词：数据中心；网络资源管理；地址自动配置；拥塞控制；流量管理；虚拟化管理；云计算
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2014)02-0166-16

Analysis and survey of resource management on modern data center networks

DENG Gang, GONG Zheng-hu, WANG Hong, CHEN Lin, LIU Zhi-hong
(Department of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: Data center is the basic facility of Internet and cloud computing. Since network resource is critical resource in data center thorough study and analysis of its management, which can help to improve data center's performance, save costs and expenses, and is of great significance. However, as the structure characteristic and application mode are profound changing, it brings tremendous challenges to modern data center network resource management. Currently, lots of studies have been made in this field. In order to have an overall perspective of current research, some of the most important aspects were summarized and analyzed, which are automatic address configuration, congestion control, network traffic management and virtualization management. They are also the hotspots in current academe. Based on the comprehensive comparison and analysis, some trends of the future data center network resource management were pointed out in the end.

Key words: data center; network resource management; automatic address configuration; congestion control; traffic management; virtualization management; cloud computing

1 引言

数据中心是互联网和云计算的基础支撑平台，是信息技术发展的重要标志。数据中心虽然由来已久，但围绕数据中心的研究方兴未艾，特别是近年来，随着互联网和云计算的不断发展，数据中心逐渐从后台走向前台，得到了产业界和学术界的高度重视。Google、微软、IBM、SGI、思科、惠普等国际 IT 公司纷纷推出并部署了自己的数据中心，国内 IT 企业也加紧抢占数据中心市场，如中国电信、中国联通、百度、阿里巴巴等都建立了自己的大型数据中心，世纪互联推出了“云立方”，浪潮在 2011 年推出了“Smart Cloud”。学术界也对数据中心给予了很高关注，SIGCOMM、INFOCOM 等

收稿日期：2013-05-05；**修回日期**：2013-11-07
基金项目：长江学者和创新团队发展计划基金资助项目（IRT1012）；湖南省高校科技创新团队支持计划基金资助项目；湖南省自然科学基金资助项目（11JJ7003）；国家高技术研究发展计划（“863”计划）基金资助项目（2011AA01A103）
Foundation Items: Changjiang Scholars and Innovative Research Team in University (IRT1012); Technology Innovative Research Team in Higher Educational Instituions of Hunan Province; The Natural Science Foundation of Hunan Province (11JJ7003);The National High Technology Research and Development Program of China (863 Program)(2011AA01A103)

重要国际会议均开辟了专门的数据中心研讨专题，汇聚相关领域的前沿研究成果。

近年来，围绕数据中心的研究成果大量涌现，为了厘清各类研究的脉络，部分研究者已对相关研究进行过分析综述，文献[1]分析了 Internet 数据中心资源管理面临的挑战，并以挑战为主线对近年来国内外在满足 SLA、降低功耗方面所取得的资源管理研究成果进行了概括总结；文献[2]重点对降低云计算数据中心能耗为目标的资源调度方法，以提高系统资源利用率为目标的资源调度方法，以及基于经济学模型的云资源管理方法进行了分析比较；文献[3]则主要从虚拟资源管理的角度对云数据中心资源调度模型与算法以及基于能量优化和负载均衡的虚拟机迁移技术的研究现状进行了综述。值得注意的是，以上对数据中心资源管理的研究，都主要是针对计算资源进行的。其面临的问题主要是当前计算资源利用率低、费效比高的问题，目标是在满足用户服务等级协议（SLA）的前提下，实现能量和计算负载的优化。然而，由于传输能力的增长往往滞后于计算能力的增长，与计算资源相比，网络资源更为紧缺。最近的研究表明，网络性能往往成为数据中心的性能瓶颈，网络配置错误、网络拥塞、负载不均衡等将导致服务瘫痪、分组丢失、重传、超时等，严重影响数据中心性能，进而影响到服务质量、用户体验和投资回报。网络资源的管理也更为复杂，其原因在于，网络资源往往是分布式的，同一网络资源常常被众多的计算节点所共享，网络资源的管理，不仅牵涉到网络本身拓扑、配置、容量等固有属性，还常常与计算资源、存储资源及应用分布等紧密相关，因此，研究网络资源的管理，将面临更大的困难和挑战，也具有更加的紧迫性。

然而，随着数据中心网络的飞速发展，其组成、结构、功能、规模及应用模式等各方面正发生深刻的变革，传统的资源静态分配、工作负载静态管理、应用与基础设施紧密耦合的网络管理方式已经不能适应现代数据中心网络的新要求，亟待研究新的技术和方法加以解决。深入研究分析现代数据中心网络资源管理的技术和方法，对于揭示数据中心网络的基本工作原理，提高数据中心网络运行效率，节省成本和开销，具有十分重要的理论和现实意义。地址自动配置技术、传输控制技术、流量管理技术以及虚拟化管理技术等是现代数据中心网络资源管理的重要内容，也是近年来学术界研究的重要方向，本文将结合当前的研究现状，对以上几个方面的最新研究成果进行分析综述。据笔者所知，本文尚属首次对数据中心网络资源管理技术进行综述研究，希望本文的工作能对数据中心网络资源管理的研究和系统设计提供抛砖引玉的借鉴作用。

2　现代数据中心网络地址自动配置技术

网络地址配置是数据中心对外提供服务的基础。在数据中心网络对外提供服务之前，需要首先对其节点配置正确的地址。此外，当一个应用从企业数据中心向云端迁移时，为了保持其原有的网络布局不变，需要为其配置相同的网络拓扑和地址。传统的地址自动配置技术主要有 DHCP[4]、Zeroconf[5]等。DHCP 是应用最为广泛的主机地址配置协议，在 DHCP 中，DHCP 服务器保存可用的 IP 地址，当主机加入子网时，通过广播搜寻 DHCP 服务器并获取一个未使用的 IP 地址作为本机地址。为了能够接收广播信息，主机和 DHCP 服务器需在同一子网内。在 Zeroconf 协议中，需要进行地址配置的主机随机产生一个地址，并将该地址广播到网络中，如果没有回复表明该地址被占用，则保留该地址作为本机地址，否则重复以上过程直到找到未被占用的地址，Zeroconf 仍只能对同一子网内的主机进行地址配置。与传统数据中心网络不同的是，为了充分利用网络的结构特性以提供高效的路由，现代许多数据中心网络地址和网络位置常常是相关的，如 Fat-tree[6]、Portland[7]、DCell[8]、BCube[9]等均将位置信息编码到逻辑地址中。此外，现代数据中心网络可达百万节点的规模，传统的地址自动配置协议如 DHCP、Zeroconf 等只能对同一子网内的主机进行地址配置，且需要通过广播进行信息交互，不仅不能适应于大规模的网络，而且随着网络规模的增大，将导致低效。另外，传统的地址通常指的是 IP，而在现代数据中心网络中，地址可能仅仅意味着一个逻辑标识，既可以是传统的 IP 地址，也可能是非 IP 的其他标识，如 Dcell、Bcube 中的 ID 等。因此，传统的地址自动配置技术将不适用于现代数据中心网络。围绕数据中心网络地址自动配置，当前的研究主要可分为 2 类：一类是拓扑相关的地址自动配置，主要是基于图的基本理

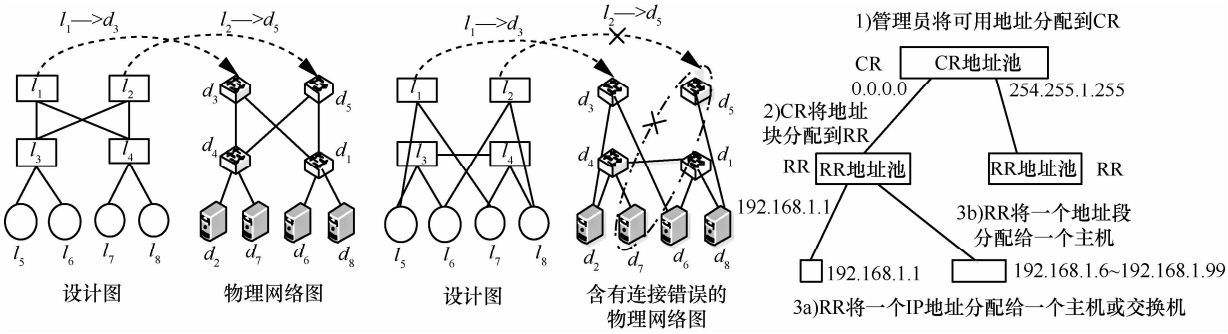
论，将地址自动配置问题转化为图同构问题加以解决，另一类则是拓扑无关的配置技术，典型代表是 DCZeroconf，主要是借鉴 Zeroconf 的思想，实现数据中心网络地址的自动配置。前者主要解决逻辑地址与位置相关的问题，而后者主要是解决大规模的数据中心网络中地址配置的动态适应性问题。

2.1 拓扑相关地址自动配置

如前所述，最近提出的许多数据中心网络结构均是拓扑相关的，在网络提供服务前，需要根据设计图配置网络地址，或者当企业应用向云中迁移时，为了保持原有结构，需要按原有拓扑进行地址配置。一般地，拓扑相关的网络地址自动配置问题可表述为：给定相应的设计图（blueprint）和物理网络图（physical network graph），网络地址自动配置寻找设计图和物理网络图的一种同构映射，从而为每个物理节点分配相应的逻辑 ID，其中，设计图代表了网络的逻辑拓扑，每个节点被赋予一个逻辑 ID，如 IP，而物理网络图则包含了网络的物理连接关系及设备 ID，如 MAC 地址。当前，拓扑相关的地址自动配置算法主要有 DAC 和 ETAC。文献[10]提出了一个通用数据中心网络地址自动配置算法 DAC，DAC 根据设计图及物理网络拓扑实现逻辑 ID 到设备 ID 的映射，如图 1(a)所示。DAC 问题本质上是图同构问题，但是，针对大规模的数据中心网络，图同构算法复杂度极高，为此 DAC 结合数据中心网络结构特性，将问题转化为子图同构问题加以解决。在此过程中，DAC 主要使用了 3 个启发式策略，分别是通过最短路径长度分布选择候选者，通过轨道(orbit)过滤候选者及有选择的撕裂（splitting）。大规模的仿真结果表明，DAC 对地址自动配置问题有较高的效率，对几万节点的 BCube 网络、几十万节点的 Fat-tree 网络及上百万节点的 DCell 网络能在 10 s 内完成配置。DAC 的主要缺点是在网络发生错误时，算法将无法完成配置，需要人工检测并修复，且其错误检测算法需要较长的时间，这极大地降低了网络发生错误时 DAC 的效率。针对这一问题，Ma 等人对 DAC 进行了补充和完善，提出了一种容错的地址自动配置算法 ETAC[11]。ETAC 包含了一个错误检测模块，当无错误时，则对全网进行配置，当发现网络发生错误时，首先在逻辑图和物理网络中逻辑地移除错误节点，然后再对剩余的子图进行同构映射，如图 1(b)所示，从而使得网络在发生错误时，仍能部分地对网络进行配置，提高了地址自动配置的适应性。仿真结果表明，对 10 000 节点的网络规模，ETAC 能在 300 s 内完成逻辑 ID 到物理 ID 的映射。DAC 与 ETAC 的主要缺点是需要首先输入设计图，对百万节点的网络，这是不小的规模。此外，DAC 与 ETAC 均是在设计图与物理网络拓扑同构（可能存在少量故障）的前提下，对网络地址实施同构映射，但某些情况下，可能需要根据设计图在一个庞大的网络中首先找出与之同构的子网（如企业应用向云迁移），且这一子网需要满足某种限制，如占用的带宽最少或相对集中分布等，然后再对该子网进行地址配置。这一过程本身是非常关键和复杂的，但当前尚未见相关研究。最后，对 ETAC 而言，当故障涉及核心节点时，逻辑地移除相关节点将可能使得网络被撕裂为不同的碎片，从而无法正常提供服务，这将严重影响 ETAC 配置的效果。理想的状态是，在节点涉及故障，而非节点本身发生故障时，仍能正常配置并提供服务。

2.2 拓扑无关地址自动配置

DAC 和 ETAC 适用于拓扑与地址严格相关的网络，在进行地址配置时首先需要输入设计图，需要大量的手工输入，当有节点动态加入和退出时，需要对全网进行重新配置，代价开销大，对动态的网络环境如云计算等适应性差。为此，IBM 的研究人员针对拓扑无关的网络提出了一种无需设计图的地址自动配置方法 DCZeroconf[12]。DCZeroconf 的设计目标主要是 3 个方面：①能够对任意网络拓扑的虚拟机（VM, virtual machine）或主机进行 IP 地址配置；②当网络拓扑发生变化时，算法能动态适应这种变化；③能够适应不同的网络规模。DCZeroconf 采用了一种层次式的配置思想，主要包括 3 步，如图 1（c）所示。首先，网络管理员决定可用的 IP 地址范围并将其分配给地址配置集中控制器（CR），这也是 DCZeroconf 唯一需要人工干预的地方，随后，CR 将可用 IP 地址分成不同的段，并告知每一个机架地址控制器（RR）该机架可用的地址池，最后，当机架内有主机、虚拟机或交换机请求地址配置时，RR 即可从地址池中任选一个未使用的地址对其进行配置。为了完成 CR 与 RR 的通信，需要在 CR 与 RR 之间构建专门的配置网络，虽然配置网络规模相对较小且对带宽等要求不高，甚至可以用无线的方式构建，但这也将增加



(a) DAC　　(b) ETAC　　(c) DCZeroconf

图 1　典型数据中心网络地址自动配置算法

DCZeroconf 的部署难度和一定的额外开销。DCZeroconf 的主要优点是实现了完全的地址配置自动化，包括无需输入设计图，且能适应网络拓扑的动态变化，虚拟机的动态加入与退出等，其缺点主要是所配置的网络地址与位置无关，且需要增加专门的地址配置网络。

表 1 对几种地址配置方法进行了对比，由表 1 可以看出，传统的 DHCP、Zeroconf 等配置方法无论是在可配置设备的类型、配置规模和适用的范围上，均不适用于大规模现代数据中心。而 DAC、ETAC 与 DCZeroconf 的主要区别则是在地址与位置是否严格相关上。对网络节点动态加入或退出以及网络故障引起的拓扑动态变化，除 DAC 和 ETAC 外均具有高适应性。就配置效率而言，DHCP、Zeroconf 与 DCZeroconf 均不受故障影响，且只对同一子网（DCZeroconf 为同一机架）进行配置，效率高，但 Zeroconf 需要通过广播信号在所有设备之间进行协商，因而效率较 DHCP 低。DAC 在无故障时具有极高的配置效率，但在发生故障时需人工干预，配置效率低，ETAC 部分克服了 DAC 的局限，但受问题复杂度的影响，在网络规模较大时，配置效率降低。

2.3　小结

当前的数据中心网络的地址自动配置技术主要包括拓扑相关和拓扑无关 2 类，前者主要根据逻辑拓扑和物理网络图完成逻辑 ID 到设备 ID 的映射，后者主要针对地址与拓扑无关的网络实现地址的自动配置。就前者而言，当前的算法仍存在容错性、动态适应性、配置效率等问题，且当节点涉及故障时，DAC 需要人工干预才能完成配置，ETAC 则可能因为移除了关键节点而不符合配置预期。就后者而言，地址与拓扑无关可能导致不能满足应用拓扑相关的特殊要求，也不能有针对性地进行性能和路由等优化。此外，并非所有网络地址均拓扑严格相关或无关的，某些网络的地址与拓扑可能仅是部分相关的，如 VL2[13]，某些网络出于性能或管理的原因，仅仅需要将某一主机放置在某一机架内，对这类地址与拓扑部分相关网络的地址配置问题，

表 1　　**地址自动配置算法对比**

名称	配置类型	配置规模	实现方式	适用范围	容错性	动态适应性	配置效率		适用于数据中心
							无故障	带故障	
DHCP	主机	*K* 级	软件实现	同一子网	高	高	高	高	否
Zeroconf	主机	*K* 级	软件实现	同一子网	高	高	较高	较高	否
DAC	主机及网络中间设备	*M* 级	软件实现	拓扑相关数据中心	低	低	高	低	是
ETAC	主机及网络中间设备	*M* 级	软件实现	拓扑相关数据中心	较高	低	较高	较高	是
DCZeroconf	主机及网络中间设备	*M* 级	需额外硬件设备	拓扑无关数据中心	高	高	高	高	是

注：中间设备指的是交换机或路由器等网络中间连接设备，容错性高指的是在网络发生错误时仍能对地址进行配置，较高指的是仅能对部分设备进行配置，而低则指不能完成配置。

当前未见专门的研究。

3 现代数据中心网络传输控制技术

TCP 协议自诞生以来，取得了巨大的成功。然而，最近的研究表明，传统的 TCP 协议应用于数据中心网络将导致低效，这是因为 TCP 主要是为了满足 Internet 数据传输需要而设计。与数据中心网络相比，Internet 具有分布式、自组织、低带宽、高延迟和低吞吐率的特点，而数据中心网络则表现为集中式、高带宽、低延迟和高吞吐率，网络有集中统一的控制。数据中心运行着各种关键核心业务，如搜索引擎、Web 服务、在线购物、网络游戏等，对网络性能有极高的要求，网络运行效率和性能优劣将直接影响到各种服务的性能进而影响到用户体验。最近的研究指出，许多数据中心应用面临某种软实时限制（soft-real-time constraint），如搜索引擎的响应时间通常应小于 300 ms，响应时间超过这一时限，将导致软超时，影响用户体验进而影响到投资回报。传统基于 TCP 协议的传输控制机制应用于数据中心将极易导致网络性能下降、拥塞、超时、网络利用率低等问题，为此，需要针对数据中心特殊的网络环境，对 TCP 协议进行改进或重新设计，研究新的传输控制机制。当前的研究主要分 2 类：①软超时无关的传输控制；②软超时敏感的传输控制。

3.1 软超时无关的传输控制

标准 TCP 协议当收到网络拥塞通知时，即将发送速率减半，TCP 对拥塞的响应与拥塞程度是不成比例的，在高带宽、低延迟的数据中心网络环境，将导致链路利用率降低和吞吐量下降。软超时无关的传输控制技术主要是通过对传统 TCP 协议的拥塞控制机制进行修改，使之更适应于数据中心传输特性，从而提高数据中心网络吞吐率。文献[14]提出了一种数据中心网络 TCP 协议 DCTCP。DCTCP 主要的设计目标是针对数据中心网络高带宽、低延迟的特点，提供比 TCP 协议更高的吞吐率、更低的延迟，并能够适应网络突发流。DCTCP 利用交换机的显式拥塞通知（ECN, explicit congestion notification），每条流根据拥塞标志 CE 被置位的数据分组所占的比例估计网络的拥塞程度，并据此动态地调整流的发送速率，从而既能降低拥塞，又能减少队列延迟和分组丢失。仿真结果表明，与 TCP 相比，DCTCP 能获得更高的吞吐率和更低的延迟。遗憾的是，DCTCP 是软超时无关的，DCTCP 仅根据感知的拥塞程度，“公平地”调节流的发送速率，而不能使接近超时的流得到优先传输，研究表明，在高扇入、低延迟的应用中，约 25%的流将发生软超时[15]。DCTCP 试图从传输层的角度，设计一种通用高效的拥塞控制机制。而文献[16]则专门针对多对一通信中的拥塞控制开展研究，提出了一种 Incast 拥塞避免机制 ICTCP。多对一通信中接收端容易发生拥塞，引起分组丢失、重传，导致性能下降。ICTCP 在接收端根据剩余可用带宽的大小以及流的期望吞吐量与实际测量值之间的差异率，每条流独立地通过调整接收窗口的大小调节发送速率，达到避免拥塞的目的。简单地说，即当差异率小于某一阈值且网卡有可用带宽时，则增大接收窗口，反之则减小接收窗口，其他则保持不变。

3.2 软超时敏感的传输控制

与软超时无关的方法不同，软超时敏感的传输控制机制的主要目标是最大限度满足流的软超时时限，以提高用户体验，同时兼顾网络吞吐量等其他性能，典型的方法有 D^3、D^2TCP 等。微软研究院的 Wilson 等人首先对软超时问题进行了研究，提出了一种超时敏感的数据中心网络传输协议 D^3[17]。在 D^3 中，应用程序显式地向传输层提供超时的时限和需发送流量的大小，每一个 RTT(round trip time)，发送端主机根据超时时限和传输流量大小向路由器请求发送速率，流的传输路径上的每一个路由器按照先来先服务的原则贪婪地为流分配速率以使尽可能多的流能在软时限内完成并形成速率分配向量，反馈回发送端主机，发送端主机根据速率分配向量选择最小的速率作为下一个 RTT 的发送速率。D^3 开创性地提出了软超时敏感的传输控制协议，但 D^3 的缺陷也是明显的。①需要应用程序显式地提供软超时信息，这可能导致恶性竞争或恶意的带宽占用。②需要修改主机、路由器和应用程序，部署难度大。③与现有 TCP 协议不兼容。④D^3 是不公平的，当不能满足所有流的软超时时限时，D^3 保证了某些流按时传输完成，而另一些流则软超时被丢弃，在需要同步的应用中，可能因为少数流被丢弃而长期等待，从而影响总的响应时间，导致性能不可预期。⑤D^3 按照先来先服务的策略为每条流分配带宽，这可能导致某些稍早到达但离超时尚远的流获得了带宽而某些稍晚到达但接近超时的流无法分配带宽，从而导致优先权

的反转。如前所述，DCTCP 拥塞敏感但却软超时无关，D^3 软超时敏感却拥塞无关，D^2TCP[15]则综合了 DCTCP 和 D^3 的基本思想：发送端根据网络的拥塞程度和流的软超时时限动态地调整发送速率，当检测到拥塞发生时，每条流结合网络拥塞程度和软超时时限动态调节发送窗口，网络越拥塞，离软超时越久，则发送窗口减少越多，从而既能实现拥塞控制，又使得更接近软超时的流能得到优先传输。但是，D^2TCP 仍需应用程序显式地提供软超时时限。D^2TCP 与 D^3 一样不支持抢先调度，这仍然可能导致某些稍晚到达但更紧迫的流发生超时。为此，文献[18]提出了一种分布式的流抢先调度策略 PDQ。PDQ 通过发送端、接收端和交换机的协作，实现了一个分布式超时敏感的抢先调度算法，算法在流有软超时时限时，时限越小的流优先，如果流没有软超时时限，则完成时间小的流优先，但 PDQ 不仅需要上层应用显式提供软超时时限，而且需要对交换机进行修改以支持抢先调度，实现难度大。文献[19]通过理论分析认为导致软超时的原因主要是流完成时间的重尾分布，从概率意义上讲，减少重尾效应就相当于减少了流软超时的概率，而导致重尾的原因主要有 3 个：①分组丢失和重传；②流优先权反转；③负载不均衡。为此，文献[19]提出了一种跨层的传输控制框架 DeTail。图 2 展示了 DeTail 的协议栈结构和跨层的信息交互。在数据链路层，DeTail 通过端口缓冲区占用构造一个无分组丢失网络，无分组丢失网络只会由于硬件错误或失效而导致数据分组丢失，而不会因为突发拥塞而分组丢失，在现代数据中心网络中，由于硬件错误极少发生，这使得分组丢失成为小概率事件，分组丢失的减少也降低了重尾的概率；在网络层，DeTail 通过端口缓冲区的占用信息执行分组级的动态负载均衡，减少了网络拥塞的可能性；在传输层，由于底层网络仅在硬件错误或失效时发生分组丢失，因此，无需对乱序敏感地作出反应，DeTail 通过简单地移除 TCP 的分组丢失重传机制或降低对乱序的敏感度以达到抗乱序的目标；在应用层，DeTail 通过允许应用定义流的优先级实现对延时敏感流的优先传输。DeTail 首次提出了通过跨层的协作机制解决数据中心网络拥塞控制和软超时保证的问题，但是，DeTail 仍需要应用显式定义流的优先级，同时需要交换机及主机协议栈的支持。

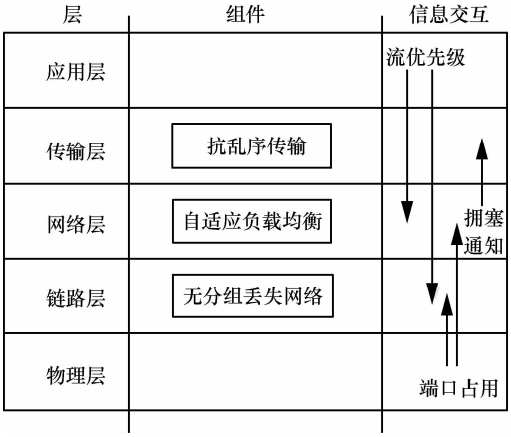


图 2　DeTail 的跨层网络栈结构

3.3　小结

与 Internet 相比，数据中心网络在网络结构、通信模式、流量特征等方面均有着不同的特点，传统的 TCP 协议主要针对 Internet 而设计，不能充分利用数据中心网络特性，直接运用于数据中心网络面临功能和性能等方面的不足，使得新型的数据中心网络传输控制协议成为新的研究热点。当前的研究主要集中在拥塞控制机制和软超时保证 2 个方面，总体而言，单纯的拥塞控制机制虽然能够提高网络吞吐率，但由于对软超时不敏感，可能导致部分流因未得到及时传输而软超时，最终影响到应用的性能。而软超时敏感的传输控制机制虽能够根据超时时限和网络拥塞程度进行流的优先调度，但目前的研究仍都需要上层应用显式地提供超时信息，这容易导致带宽资源被恶意抢占，且许多技术都需要修改网络中间设备，这增加了部署难度。表 2 对几种典型传输控制机制进行了综合比较。

表 2　数据中心网络典型传输控制机制对比

名称	软超时敏感	抢先调度	交换机支持	应用程序支持	兼容 TCP
DCTCP	否	否	ECN	无	是
ICTCP	否	否	无	无	是
D^3	是	否	速率请求应答	提供超时时限	否
D^2TCP	是	否	ECN	提供超时时限	是
PDQ	是	是	流分级、抢先调度	提供超时时限	否
DeTail	是	否	无	提供超时时限	否

4　现代数据中心网络流量管理技术

流量管理是数据中心网络资源管理中最重要

也是动态性最强和最具挑战性的内容。本节将首先对数据中心网络流量特征进行研究分析，然后分别对 2 类不同的流量调度策略进行综述，最后给出各种流量调度方法的小结比较。从网络流级看，现代数据中心网络普遍支持节点之间的多路径连接，多路径提供了更高的传输能力，但现有 TCP 协议的单路径传输特性与网络的多路径支持之间并不相适应，如何根据数据中心网络结构及流的特性，在不同路径间分配和平衡流量，以最大化数据中心网络性能是当前的研究热点之一，可分为 2 种不同的研究思路。①以网络流为中心的调度策略：通过对流的传输路径的分配调度或通过虚拟机的优化布局，实现网络流量平衡或资源利用率的最大化。②以网络结构为中心的调度策略：从网络结构属性的角度研究网络资源高效利用的支持机制，发掘现代数据中心网络的多路径传输能力。

4.1 现代数据中心网络流量特征

网络流量特征是进行流量管理的基本依据。最近的研究表明，数据中心网络流量具有局部性、动态性和不平衡性等特点。文献[20]通过对 1 500 个服务器 2 个月期间的网络流量的测量和分析表明，相同机架内的节点更可能发生通信，一个服务器或者与同一机架内所有节点通信，或者仅与不超过 25%的节点通信；或者不与机架外地其他节点通信，或者仅与 1%~10%的节点通信。数据中心网络流量分布表现出 2 种不同的模型：work-seeks-bandwidth 模型和 scatter-gather 模型。work-seeks-bandwidth 模型表现为相邻或相近的节点之间具有较大的数据通信，如相同机架或相同 VLAN 的节点之间具有更多的通信，这主要是因为设计者通常希望把相同的作业放在同一区域以获得更高的带宽。scatter-gather 模型表现为一个服务器与多个服务器之间的通信，这主要是由于数据中心典型的应用模型如 Map-Reduce 等本质上要求数据有分发和汇聚的过程，需要在一个节点与多个节点之间传递数据。文献[21]通过对 10 个不同类型（大学、企业和云计算）数据中心网络流量的测量也发现，对云计算数据中心而言，80%的数据流量发生在机架之内，这表明对大部分应用而言，流的分布具有某种局部性。文献[21]的研究还表明，随着应用的不同，同一时刻不同数据中心网络中流的数目存在着较大的差异，范围从几条到接近上万条不等，流的到达时间也从几毫秒到上百毫米不等，但是，绝大部分流的到达时间都不会超过 100 ms。在文献[20]的测量结果中，就整个簇而言，流的平均到达速率达 10^5 条/秒，即每毫秒有约 100 条到达，文献[22]的测量结果也表明，数据中心中流的数目巨大，同时存在的跨机架的流平均可达 10^5 的数量级，平均到达时间也达 5×10^5 条/分钟。这意味着数据中心网络流具有极高的突发性和动态性，从资源管理的角度看，集中的流调度算法将很难奏效。最近的研究还表明，数据中心网络流存在不平衡性，这种不平衡表现在 2 个方面：大小不均匀和分布不平衡。在流大小上，文献[21]的测量结果表明，流的大小和长度表现出大象流和老鼠流的特性，80%的流大小不超过 10 KB，10%的流占据了绝大部分的数据流量。无独有偶，文献[8,20]也观察到了类似的现象，文献[8]中存在明显的老鼠流现象，99%的流均小于 100 MB，然而超过 90%的字节却包含于 100 MB 到 1 GB 的流中，这就意味着不到 1%的流包含了超过 90%流量。文献[20]中持续超过 200 s 的流不到 0.1%，80%的流持续时间都比较短，不超过 10 s。在流分布上，文献[20]的分析表明，在采用三层结构（核心层、汇聚层、边缘层）的数据中心网络中，各层之间的流量并不平衡，核心层通常具有较高的链路利用率，而边缘层和汇聚层则相对较低。文献[23]也观察到类似的结论，核心层链路利用率高于汇聚层和边缘层，但分组丢失率却相反，核心层链路利用率是汇聚层的 4 倍，95%的汇聚层链路利用率不超过 10%。这表明流的分布并不均匀，核心层相对集中且稳定，而其他层分布较少但突发性更高，从而导致分组丢失率更高。网络流的这些特性给数据中心网络管理带来了严峻的挑战。

4.2 以网络流为中心的调度策略

实际的测量和理论分析表明，数据中心网络的流量主要是服务器之间的流量，除一对一（1→1）通信外，一对多（1→N）、多对一（N→1）、多对多（N→M）等集群通信模式是数据中心网络的主要通信模式[8,9,20]，集群通信即多节点之间的通信。大量节点之间通信容易引发网络拥塞，导致分组丢失、重传和性能下降。如前 2.3 节所述，数据中心的一些应用还面临某种软实时限制（soft-real-time constraint），响应时间超过一定限度（如 300 ms），将影响到用户体验进而影响数据中心的收益。为此，需要通过网络流的优化调度及服务器的合理布局提高网络性能。当前，针对网络流调度的研究主

要可分为 2 类：一类是仅考虑主机出口带宽的调度方法；另一种是考虑全网容量限制的调度策略。

4.2.1　主机出口带宽限制的调度方法

由于流的突发性和动态性特征，当前，针对主机出口带宽的流调度策略主要是将网络流看作随机变量，采用随机装箱的思想加以解决。文献[24]研究了通过虚拟机的合并提高网络资源利用率的问题。其问题可描述为：给定一序列元素列表 $\boldsymbol{L}=(x_1, x_2, \cdots, x_n)$，其中 x_i 是规格化且互相独立的随机变量，则最少需要多少单位容积的箱子才能装下所有的元素，使得元素的容积之和大于箱子容积的概率不大于某一给定常数 $e\in(0, 1)$。这里的元素就相当于每个虚拟机需要的带宽，而箱子则相当于主机的出口带宽，则网络带宽资源的分配转化为随机装箱问题(SBP, stochastic bin packing)，Wang Meng 等人采用启发式策略，使得对于任意 $\varepsilon>0$，在满足带宽需求的情况下，所需服务器数量不大于最优值的 $(1+\varepsilon)(\sqrt{2}+1)$倍；文献[25]对 Wang Meng 等人的研究工作进行了改进，使得所需服务器数量进一步减少到不大于$(2+\varepsilon)$倍。以上解决方案仅考虑单台服务器的出口流量，并未考虑外部网络的传输能力。而实际上，网络流的优化不仅受到主机出口带宽的制约，同时还受到网络拓扑及路径容量的制约。因此，另一类的流调度策略则结合了网络拓扑进行考虑。又可以进一步分为 3 种策略：①通过调度流的传输路径实现流量优化；②通过虚拟机迁移，优化虚拟机的布局以改变流量的分布；③结合虚拟机迁移和路由共同优化流的传输。

4.2.2　全网带宽限制的调度方法

通过调度流的传输路径实现流量优化。加州大学的 Mohammad Al-Fares 等人提出了一种通过控制流的传输路径实现流量优化的方法 Hedera[26]。Hedera 首先通过网络流量的测量并估算出每条流可能的流量需求，然后采用集中控制的算法，为流预留路径，并更新转发表。为了提高算法的效率，Hedera 仅对大流进行调度，只有当流的持续时间及带宽需求超过某一阈值时，才激活调度算法。Hedera 在路径改变时需要更新转发表，将一定程度上影响流的传输性能。为了支持修改转发表，Hedera 采用 OPenFlow[27]交换机，OPenFlow 交换机虽然已有部分商业产品出现，但其性能并未得到充分检验，当前并未成为数据中心交换机的主流。

通过虚拟机的优化布局改变流量的分布。Hedera 通过改变转发表进而对流实施调度，但如前所述，普通商业交换机并不支持修改转发表。现代数据中心广泛采用虚拟化技术，可以通过对虚拟机的优化布局从而改变流的分布。为此，IBM 的 Meng 等人将流调度问题转化为流量感知的虚拟机的优化放置问题（TVMPP）[28]，问题的输入是虚拟机之间的流量矩阵及主机之间的通信代价矩阵，目标是寻找一种虚拟机的放置方案，使得全网总的流量最小。文献[28]证明了这一问题为 NP 完全问题，并设计了一种启发式策略加以解决，其复杂度为 $O(n^4)$。该方法的主要缺陷是在分配虚拟机时仅考虑了网络出口带宽限制，并未考虑网络容量的限制，因此，分配方案可能不满足 QoS 要求。为使全网总的流量最小，意味着虚拟机应尽量集中放置，这使得配置方案扩展性较差，难以适应带宽需求的变化，此外，算法的复杂度也较高。与此相反，文献[29]将流量优化问题转化为最小割率感知的虚拟机放置问题 MCRVMP，目标是虚拟机的分配不仅要满足当前的需求，还应最大限度适应未来需求的变化，即实现网络流量平衡。MCRVMP 主要针对树型网络，首先根据虚拟机之间的流量矩阵及虚拟机的放置位置找出负载率最高的关键割，再在所有方配方案中选出关键割负载率最低的分配方案作为虚拟机的分配方案。MCRVMP 问题仍为 NP 完全问题。文章提出了 2 种启发式算法。一种采用分治和递归思想的 2PCCRS，首先将 VM 分成不同的簇，再对每一个小簇进行虚拟机分配，从而降低问题的规模。另一种是采用贪婪算法 GH。2 种方法所获得的关键割的负载率均较小（小于 0.5），但是 2 种算法的复杂度均很高，仿真结果表明，对 3 430 个 VM 的网络，在最坏情况下，2PCCRS 需要超过 3 000 s，GH 算法需要超过 8 000 s 才能完成分配。

结合虚拟机迁移与路径选择的调度策略。以上 2 种对网络流优化调度的方法，或者只考虑虚拟机的放置位置，或者只考虑流的路径选择，而实际上，网络的性能跟虚拟机的位置与路由通常是紧密相关的，仅考虑某一维度，优化效果有限。普林斯顿大学的 Jiang 等人结合虚拟机迁移和路由对网络性能优化进行了讨论，将路由选择和虚拟机放置的联合优化问题转化为静态优化问题，并基于马尔可夫模型构造了一个优化算法 VMPR[30]。但是，该模型引入了过多的参数，参数需要人工调节，因此需要专门的经验和知识，且为了降低算法的复杂度，模

型每次只允许调整一个 VM 的位置，在大规模的网络中，这将对算法的优化效率产生重要影响。

4.3 以网络结构为中心的调度策略

现代数据中心网络的新型拓扑结构如 Fat-tree、VL2、DCell 等都提供节点之间的多条路径连接。多路径提供了更高的网络容量，通过在多条不同的路径之间分配和平衡流量，可以减少网络拥塞，提高网络资源利用率，但是，现有 TCP 协议的单路径传输特性和数据中心网络结构的多路径支持之间并不适应。以网络结构为中心的调度策略，通过发掘网络结构固有的传输能力，特别是对多路径传输的支持，并发地传送流量，以到达最大化网络传输性能的目的。针对如何并发利用网络的多路径特性，在不同的路径之间分配和平衡流量，以提高网络吞吐率，当前的解决方案主要有 3 种。

1) 采用固定的转发规则。根据源地址或目标地址将流映射到固定的路径，如 Fat-tree。在 Fat-tree 网络结构中，每个终端主机有$(k/2)^2$条到达核心层的路径，Fat-tree 根据源节点的地址，为不同源地址的流选择不同的路径，从而实现在不同路径间平衡流量。

2) 采用随机负载均衡的办法。在所有可用的等价路径中为每条流随机选择一条路径，如 ECMP[31]、VLB[8]等。ECMP 在多条等价路径中为每一个数据分组随机选择一条路径，而为了防止乱序，VLB 则是为每条流随机选择路径。

3) 使用集中控制策略。根据网络结构属性和流量矩阵，通过全局的控制器进行路径选择。但是，这些方案都存在各自缺陷，固定规则的路径选择策略和随机负载的方法虽然能在一定程度上分散网络流量，但不能根据路径的负载进行动态调度，可能导致网络局部拥塞[32]；而集中调度算法由于需要流的全局信息，运行效率将受限。研究表明，在大规模的数据中心网络中，网络流的数目巨大，且绝大部分的流均为小流，持续时间极短，集中调度的算法将需要频繁地对流进行调度。一种可行的改进是仅对大流进行调度，研究表明，在流的大小服从指数分布，到达时间间隔服从泊松分布到情况下，对大的流进行调度可以获得较高的带宽利用率和较小的时间开销[26]，然而实际的测量和分析表明，数据中心网络流并不服从这样的分布[8]，这种方法仍需要频繁地调度且性能接近于随机调度[32]。基于此，文献[32]通过理论分析认为数据中心网络应由 TCP 自然演进到多路径 TCP(multipath TCP)[33]。但是，多路径 TCP 当前并未得到广泛支持。

4.4 小结

以上分析表明，数据中心网络流数目巨大，且具有极强的突发性和动态性，对网络流量的管理提出了严峻的挑战。网络流的调度问题往往是 NP 完全问题，具有极高的复杂度。以网络流为中心的调度策略通过感知与监测网络的负载情况及流量矩阵，实时地为每条流或主要的流选择传输路径，或通过虚拟机的迁移改变流量分布，达到流量平衡或优化的目的。但是，在流的到达速率达 10^5 条/秒且绝大部分流为短流的情况下，这样的调度算法占用大量的资源，其有效性也难以得到保证，而且，以网络流为中心的调度策略往往只能支持流的单路径传输，不能有效利用现代数据中心网络的多路径特性。以网络结构为中心的调度算法结合网络本身固有的传输特性，按照一定的规则将流分配到不同的路径上，这种方法虽然有利于发挥网络的多路径传输能力，但或者由于缺乏路径负载信息和流量矩阵信息，可能导致局部拥塞，或者仍需要集中的控制，难以适用于大规模数据中心网络。

5 现代数据中心网络虚拟化管理技术

5.1 主机内部网络虚拟化技术

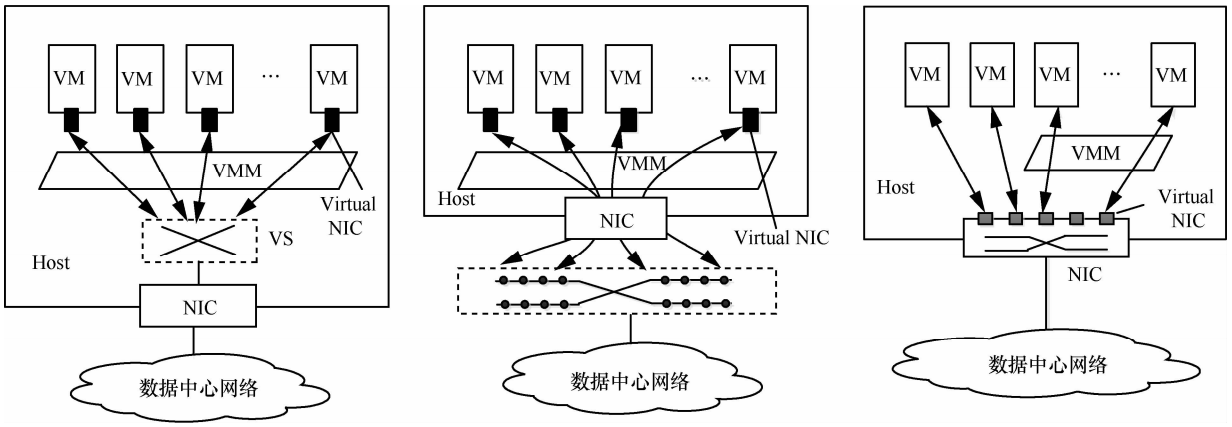
服务器虚拟化技术使得网络进入主机内部，多个虚拟机可以在同一主机内组网共存。虚拟化的这一特点使得主机内部网络的管理与控制成为新的问题。在传统非虚拟化条件下，物理主机作为终端计算节点而存在，内部本身不存在网络功能，物理主机与网络的连接通过链路状态体现，主机的位置相对固定，外部网络可以对其实施访问控制、流量监测等监控功能，主机的状态可以完全被网络感知与控制。服务器虚拟化后，一个物理主机内有多个虚拟机并存，为了支持不同虚拟机之间的通信，并对其实施监控与管理，需要在主机内引入虚拟网络功能，当前主要有 3 种方式。

1) 采用虚拟交换（virtual switch）的方式。由宿主机操作系统或虚拟机管理系统内的软件虚拟交换机完成虚拟机间数据交换，如开源项目 Open vSwitch[34]、VMWare 的 vNetwork Distributed Switch[35]和思科的 Nexus v1000[36]等，如图 3(a)所示。虚拟交换采用软件的方式在主机操作系统或虚拟机管理系统（hypervisor）内部实现一个虚拟交

换机，这种方式下，虚拟交换机、虚拟网卡均需要与 VM 竞争使用主机 CPU 资源，当端口的线速达到 10 Gbit/s 或更高时，将严重影响主机性能，且很难保证 QoS 和 VM 隔离。为此，文献[37]利用多核可编程网卡的支持，提出了一种基于可编程网卡的虚拟交换技术，通过将 VM 的虚拟网卡迁移到可编程网卡上，避免了在高速网络中虚拟网卡竞争使用主机资源，提高了虚拟数据中心资源利用率，同时可编程网卡为每个虚拟机提供单独的域，提高了 VM 间的安全性和隔离性。但该方法仍需要借助虚拟交换机进行数据交换，且同一主机内的虚拟机间的数据需在 I/O 总线上传输 2 次，浪费了主机资源。总体而言，采用虚拟交换的方式存在 2 大问题[38]：①虚拟机之间的流量监控问题，传统的网管系统无法深入服务器内部进行流量监控，造成安全隐患；②性能问题，虚拟机网络流量越大，虚拟交换机就会占用越多的 CPU 资源进行报文转发。

2) 利用外部交换机进行数据交换。如 EVB[39]，VN-tag[40]等，如图 3(b)所示。为了解决服务器虚拟化后网络边界模糊，虚拟机感知与控制困难等问题，IEEE 标准化组织制定了 EVB（edge virtual bridging）标准（即 IEEE 802.1Qbg），EVB 本质上是在 VM 与边缘交换机之间定义一组标准接口，使得 VM 的数据流量完全由外部物理交换机转发，虚拟机接入网络之前首先需要通过虚拟工作站接口发现协议（VDP, virtual station interface discovery protocol）发送关联请求并与边缘交换机建立关联，虚拟机的迁移、销毁也分别需要重新关联和去关联。通过这种方式，使得 VM 可以实时被外部网络感知与控制，物理主机内部的网络功能重新被移回主机外部，网络边界变得更加清晰，可以采用传统的网络管理的策略对 VM 进行流量监控、QoS 监控等高级网络特性的管理。EVB 虽然解决了 VM 的感知与控制的问题，但在 EVB 中，VM 的所有流量都要经外部交换机进行交换，同一主机内 VM 间的流量需要穿越主机网卡 2 次，这不仅占用主机资源、浪费主机 I/O 带宽，同时也增加了外部网络压力。VN-tag 是 Cisco 的私有技术，其实现机理与 EVB 相似，通过在以太网帧中增加 VN-TAG 标记，用以标识 VM 连接的通道和映射到交换机的虚端口号。

3) 利用主机网卡进行数据交换。利用增强功能的主机网卡，实现内部网络的接入与交换，如图 3(c)所示。为了解决虚拟机与外设之间的 I/O 操作引起 hypervisor 陷入带来的平台资源开销，同时保持 I/O 设备在多虚拟机间的共享特性，PCI-SIG 联盟提出了 I/O 虚拟化标准 SR-IOV(single root I/O virtualization and sharing)[41]，SR-IOV 允许多个虚拟机共享一个网卡。一个 SR-IOV 设备具有一个或多个物理功能单元(PF, physical function)，同时可以创建多个虚拟功能单元(VF, virtual function)，每一个 PF 是一个标准的 PCIe 功能部件，每个 PF 可与多个 VF 关联，VF 就像一个轻量级的 PCIe 功能部件，具有唯一的请求标识（RID, requester identifier）及性能攸关的关键资源，并共享大部分的设备资源，提供独立的中断、队列及地址转换机制，从而允许虚拟机直接对与之关联的 VF 进行控制，就仿佛是独立使用专门的网卡，部分网卡也提供板上 VF 间数据交换功能。文献[42]基于 SR-IOV 提出了一种跨平台的主机内网络虚拟化体系结构，该结构主要由 3 部分组成：PF driver、VF driver 和 SR-IOV Manager (IOVM)。PF driver 工作在物理主机操作系统或 XEN 的 Domain0 上，负责直接访问 PF 及配置管理 VF，



(a) 虚拟交换方式　　(b) 外部交换机交换方式　　(c) 网卡交换方式

图 3　3 种主机内网络虚拟化技术对比

VF driver 工作在客户机（VM）操作系统上，直接访问与之对应的专门的 VF，而不需经过虚拟机管理系统 VMM 的干预，IOVM 工作在 VMM 上，为每一个 VF 提供虚拟化的配置服务，使得客户操作系统可以像普通设备一样枚举和配置 VF。该结构使得 VM 可以无需 VMM 干预直接访问 SR-IOV 网卡，从而能在不同的虚拟化平台上实现，在提高 I/O 性能的同时具有良好的可扩展性。但是，网卡的硬件资源总是有限的，随着服务器性能的提高，一台服务器内可以虚拟出几百个虚拟机，如果为每个虚拟机均分配专门的硬件资源，网卡将不堪重负。为此，文献[43]提出了一种折中的方案 Crossbow。Crossbow 采用了硬件和软件相结合的实现方式，当硬件资源充足时，为每一个 VM 分配专门的硬件虚拟网卡，而当硬件资源耗尽时，则采用软件实现虚拟网卡。以上方案虽然无需引起 VMM 陷入，但主机内 VM 间流量仍需经网卡进行交换，将引起 I/O 中断且占用网卡资源，文献[44]提出了一种在直接访问网卡(direct access NIC)上进行 VM 接入和流量交换的体系结构 sNICh。为了减少对主机 I/O 带宽的浪费，sNICh 对主机内 VM 间的数据交换采用了直接内存拷贝的方法，降低数据传输开销。但直接访问网卡仅具有基本的交换机功能，不能实现企业数据中心网络交换机所需的主要特性，也不能支持大量的 VM。为此，sNICh 结合了软硬 2 种实现方式，采用数据平面和控制平面分离的策略，控制平面由软件实现，数据平面采用硬件转发，从而加快处理速度。同时由于控制功能由软件实现，易于实现如分组过滤、安全检查等各种控制策略。

总体而言，基于软件的虚拟交换方式成本低，但对数据分组的处理需要占用大量的 CPU 资源，开销大、性能低。随着对网络性能要求的提高，利用外部硬件支持的虚拟化方式成为必然趋势，但当前的技术发展尚不完善。利用外部交换机的方式能够对虚拟机实施实时的监控，并进行高级的网络管理功能，但是占用了大量的 I/O 带宽资源，同时也增加了外部网络压力；利用主机网卡进行交换的方式对网卡性能要求高，价格昂贵，且网卡对交换机高级特性的支持有限，削弱了对虚拟机的管理功能。

5.2 数据中心骨干网络虚拟化技术

为了叙述方便，本文把数据中心内主机内部网络以外的网络统称为数据中心骨干网，简称骨干网。传统条件下，用户与资源静态耦合，不同用户之间的资源互相隔离，不能共享使用，资源使用率低。现代数据中心逐渐从企业独占向公共云服务转变[45,46]。云计算一方面要求多个租户共享使用数据中心网络资源，运行不同的企业应用，资源以动态、弹性、按需的方式分配和共享使用，并按实际的资源使用情况付费[47]，同时提供各用户间可靠的安全隔离及不同的 QoS 保证[48]。用户可以定制个性化的计算环境，如服务器数量、软件环境、网络配置等。资源可以根据需要动态申请，也可以动态回收。如亚马逊弹性云 EC2[49]和安全存储服务 S3[50]、IBM 蓝云[51]等。另一方面又要求底层数据中心网络基础设施可虚拟化为统一的虚拟资源池，以便能够根据用户需求，灵活地分配资源。按照目标的不同，当前数据中心骨干网虚拟化技术可分为 3 类。

5.2.1 以资源隔离为目标的虚拟化技术

传统的数据中心网络使用划分 VLAN(802.1q)[52]的方式提供对资源隔离的支持，但是这种方法存在缺陷：①为了支持生成树协议，每个 VLAN 只能是底层网络的无环图，从而限制了二分带宽；②将所有的地址空间暴露给交换机，导致转发表过大，不能支持大规模网络，可扩展性差。为了支持大规模的网络虚拟化，业界提出了很多改进方案，如 QinQ(802.1ad)[53]、MinM(802.1ah)[54]等，其基本思想是将网络划分为不同的层次，采用嵌套封装的方法，由每层的边缘交换机进行封包和解包，从而极大地减少了上层网络的地址空间，可以支持更大规模的网络。但目前 QinQ、MinM 等并未得到广泛支持，同时，生成树协议的单路径特性也不能充分利用网络冗余的连通性。为了克服生成树协议的不足，IEEE 802.1aq 工作组提出了最短路径桥接(SPB, shortest path bridging)[55]协议，SPB 支持异构的网络结构，能够并发地利用多条最短路径进行数据传输，具有更快的收敛时间和更好的可扩展性，但是 SPB 仍需使用 QinQ 或 MinM 方式进行虚拟网划分。基于 VLAN 的虚拟网的划分方案均是静态的，每个用户或每个应用运行在单独的虚拟网之中，占用固定的资源，为了保证用户的峰值性能，虚拟网需要预留资源，降低了网络的效能和灵活性。

为了支持现代数据中心网络环境下多租户动态共享的需求，同时提供端到端的带宽保证，文献[56]提出了一种多用户条件下具有带宽保证的数据

中心网络虚拟化体系结构 SecondNet，其中每个用户构成一个虚拟数据中心(VDC)，VDC 中的每一个虚拟机对之间都有明确的带宽需求，SecondNet 将 VDC 的分配问题转化为最小代价网络流问题，采用集中控制的策略和端口交换的源路由算法，提供虚拟机之间的带宽保证及高效的网络资源利用率，但 SecondNet 需要交换机支持端口交换和高优先级流的抢先调度。通常，用户对计算和存储资源的需求可以明确地给出，但是却难以提出明确的网络资源需求，对此，文献[57]提出了一种虚拟网络抽象方法 Oktopus，明确定义了用户—提供者之间的网络资源需求接口，对无阻塞网络和阻塞网络，Oktopus 分别利用一个二元组<*N,B*>和四元组<*N,B,S,O*>代表一个租户的需求，其中，*N* 代表 VM 数，*B* 代表每一个 VM 的带宽需求，对阻塞网络，VM 被划分为不同的组，*S* 代表每组中的 VM 数，*O* 为阻塞因子。Oktopus 根据租户需求为每个租户分配符合要求的虚拟网络，从而使得网络提供者可以像出租计算资源、存储资源一样，对出租的网络资源按带宽进行收费，双方能够在网络性能保障、开销和收益之间作出权衡。但由于 VM 流量需求的突发性和不均衡性，通常难以用统一的标准描述 VM 的流量需求，此外，Oktopus 基于终端主机的速率限制机制，需要对 VM 增加一个增强模块。文献[58]从安全的角度，提出了一种居于身份映射的多租户隔离机制，形式化定义了隔离的概念，在此基础上，以系统开销和资源利用率建立目标函数并给出了解决算法。

为了实施灵活的虚拟化策略，需要对交换机进行精细的控制，但出于商业机密和安全的考虑，普通商业交换机开放程度限制，影响了虚拟化策略的应用。为此，斯坦福大学的研究人员提出了一种 OpenFlow 交换机[27]，通过定义标准接口，用户可对 OpenFlow 交换机的每条流进行精确的控制，包括转发路径、QoS 限速以及分组过滤等，通过 OpenFlow 技术可以很容易地实现虚拟化，并且可以灵活地利用诸如虚拟机的位置等策略进行虚拟网划分。OpenFlow 提出的初衷是为了在校园网上开发大规模的测试床供研究人员进行网络协议测试，但是由于 OpenFlow 的灵活性，它也可以应用于数据中心。但就目前而言，OpenFlow 还存在一些限制。①性能限制：OpenFlow 的控制策略主要由集中控制器负责，在高速网络中，将给集中控制器带来巨大压力，影响数据转发性能，此外，为了实施精细的控制，如入侵检测等，若需要执行逐包处理，将严重影响 OpenFlow 的性能。②安全问题：OpenFlow 虽然增加了灵活性，但也带来了安全隐患，由于采用集中的控制方式且用户可以控制数据转发策略，一旦集中控制器遭到攻击或流表被恶意篡改，将可能导致服务中断或瘫痪。③当前 OpenFlow 尚未得到主流交换机厂商的广泛支持。尽管如此，OpenFlow 仍引起了学术界和产业界的高度关注，当前已经发展成为一个颇具影响的技术流派，OpenFlow 不仅可以用于数据中心网及局域网虚拟化，还可用于广域网虚拟化，有关 OpenFlow 的研究已经超出了本文的范围，在此不再赘述。

5.2.2　以资源整合为目标的虚拟化技术

资源隔离的目的是为了将同一网络逻辑地分成不同的网络，以供不同用户使用。与此相反，资源整合的目标是将逻辑上分离的网络整合为同一网络，以增强应用部署的灵活性，如支持虚拟机的自由迁移等。典型的虚拟化方案有 VL2、PortLand 等。VL2 采用名字和位置分离的思想，应用程序使用应用相关地址(AA, aplication-specific address)发送数据，驻留在服务器上的轻量级 VL2 代理将目标服务器所在 ToR 的位置相关地址(LA, location-specific address)作为目的地址封装到分组内，目的 ToR 解包并将数据转发到目标服务器。为了确定目的 AA 所在 ToR 对应的 LA，VL2 采用一个集中的目录系统维持 AA 到 LA 的映射，通过目录系统可以实现访问控制，如仅允许相同租户的应用之间互相访问。VL2 的主要缺点是需要集中的目录系统，集中控制器可能成为系统性能瓶颈，VL2 可以实现不同租户间的访问控制，但是，不能实施基于租户的性能隔离，租户间竞争共享网络资源，不能提供带宽保证等 QoS 支持，也不支持基于租户权重分配带宽资源等公平性策略。PortLand 的思想与 VL2 类似，仍采用名字和位置分离的思想，每一个终端主机都有唯一的伪 MAC 地址(PMAC，pseudo MAC)，PMAC 是位置相关的，网络上所有的数据转发均基于 PMAC，集中的网络控制器负责完成 IP 地址到 PMAC 的映射，响应 ARP 请求，仅在数据分组到达最后一跳时由边缘交换机完成 PMAC 地址重写，使用真实 MAC 地址(AMAC, actual MAC)将数据最终发送到目标服务器。VL2 与 PortLand 均实现了一

个完全的二层网络抽象，任何应用可以被放在任何服务器上，从而支持服务器池的快速增长和收缩以及任意位置的虚拟机迁移。但是，两者均需要集中控制器进行地址映射，存在单点失效，且两者均不支持基于租户的 QoS 策略、公平性策略等。为此，文献[59]提出了一种数据中心网络虚拟化体系结构 NetLord，NetLord 对 MAC 帧格式和 IP 报文格式进行了重定义，在 IP 报文的头部显式包含了租户标识(Tenant_ID)，并通过一个常驻 hypervisor 的轻量级代理负责封包、解包和路由，虚拟机发出的数据分组经代理计算路径，确定目的边缘交换机地址后，在原报文的头部添加一层 MAC 帧头，将源边缘交换机地址和目的边缘交换机地址封装在 MAC 帧中后在二层网络上传送，当到达边缘交换机时，由边缘交换机去掉帧头，并根据封装在分组中的 IP 地址确定转发端口，将数据分组传送到目标代理，目标代理最后将数据转发到目的虚拟机。NetLord 的虚拟化方法带来了几大好处：首先是通过重新封包屏蔽了虚拟机的 MAC 地址，减少了核心网络地址空间的规模，使得使用商业以太网交换机即可构建大规模多租户网络；其次是提供了一个完全的二层和三层网络抽象，每一个租户都有一个完全的二层和三层地址空间；第三，由于分组头显式地包含租户标识，NetLord 可以对每个租户实施单独的流量管理策略，如增加 QoS 策略、公平性策略等。但是 NetLord 需要修改虚拟机管理系统，且额外的分组封装、解分组可能降低数据处理性能。

5.2.3 以公平共享为目标的虚拟化技术

随着云计算和虚拟化技术的发展，数据中心网络在应用模式和资源配置方式上发生了巨大变化。云计算要求数据中心网络基础设施作为一种服务供用户按需申请共享使用，并按资源使用情况付费。这就要求网络资源应该与用户付费成比例，即资源共享应满足某种公平性。与计算资源和存储资源的使用可以明确的计量不同，网络资源通常是分布式的，如何保证网络资源的公平共享成为虚拟化的研究课题之一。文献[22]提出了一种基于实体（entity）权重公平地共享网络资源的方法 Seawall，每个实体获得的资源与其权重相关，这里的实体可以是进程或虚拟机。通过合理地分配权重，可以实现对网络资源高效利用的同时实现某种公平性，但是实体的资源需求通常是动态的，不合理的权重将导致资源闲置。文献[60]的研究进一步指出，Seawall 的分配策略本质上是一种基于源的分配策略，这种策略对源节点的资源分配公平，但却可能导致对目的节点资源分配的不公平，基于目的节点的分配策略亦如此。为此，文献[60]总结了公平地共享带宽资源所需的 5 个基本原则，如对称性（symmetry）、独立性（independence）等，并提出了一种基于源节点和目标节点双边权重的资源共享策略 PES，通过调节参数，PES 能够实现带宽保证和公平性之间的某种平衡，但需要修改交换机或虚拟机管理程序（hypervisor）。研究资源共享的公平性问题，不仅可以促进资源更加合理的分配，还可以通过定义网络资源使用的标准接口，使得网络资源与计算资源和存储资源一样，可以按照使用量进行收费，使得使用者和提供者之间能够更加清晰地在投入和回报之间做出权衡。但当前对该问题的研究尚少，在实际系统中也未见相关的技术报告，总体而言尚处于起步阶段。几种典型数据中心骨干网虚拟化技术的对比如表 3 所示。

表 3 数据中心骨干网虚拟化技术对比

虚拟化方法	虚拟化方式	虚拟化目标	虚拟机迁移	访问控制	资源利用率	多路径
VLAN	静态	资源隔离	不支持	支持	低	不支持
VL2	动态	资源整合	支持	支持	高	支持
PortLand	动态	资源整合	支持	支持	高	支持
Seawall	动态	公平共享	—	—	高	—
SecondNet	动态	性能隔离	不支持	支持	高	不支持
Oktopus	动态	性能隔离	—	—	高	—
NetLord	动态	资源整合 性能隔离	支持	支持	高	支持
OpenFlow	动态	性能隔离	支持	支持	高	支持

6　结束语

数据中心网络的深刻变化，给网络资源管理带来诸多挑战，传统的资源静态分配、工作负载静态管理，应用与基础设施紧密耦合的网络管理方式已经不能适应现代数据中心网络的新要求，亟待研究新的技术和方法加以解决。本文从网络资源管理的几个重要方面，包括地址自动配置技术、传输控制技术、流量管理技术以及虚拟化管理技术等，对现代数据中心网络资源管理技术的最新研究成果进行了分析综述，以期能为数据中心网络资源管理的研究和系统设计提供借鉴。尽管努力试图发现各种研究、各类技术之间的关联，努力分析其优长与特点，但由于有关现代数据中心网络资源管理的研究正方兴未艾，各种观点、各种技术正处于百家争鸣的态势，以当前的认知，有些相关的研究尚难以用统一的标准加以分析比较，笔者将密切跟踪有关的研究进展，以期能有更加全面深入的了解。就笔者的研究和分析，未来数据中心网络资源管理将呈现以下趋势。

1) 配置自动化。随着数据中心网络规模的不断扩大，传统的人工配置方式不仅效率低，而且容易引发错误，据统计，50%～80%的网络宕机都是由于人工配置错误引发[61,62]。特别是在云计算环境下，应用需要根据资源需求动态申请和释放资源，如创建和销毁虚拟机、配置虚拟网等，人工的配置方式将难以胜任，网络管理系统将需要根据预先设定的配置策略和配置描述文件，自动完成配置操作，如根据逻辑图完成逻辑地址到物理地址的映射。

2) 管理智能化。数据中心是数据计算和存储的中心，运行着各种关键核心业务，如 Web 服务、Map-Reduce 集群计算、在线购物等，其通信模式表现为集群通信，即大规模节点之间的通信，对网络性能要求高，动态性强，传统的通过 VLAN 的资源划分方式导致应用与资源紧密耦合，大量资源被闲置，限制了网络性能的发挥，资源利用率低。为了充分发挥网络性能，提高资源利用率，网络管理系统应具有根据网络负载和资源使用情况，进行动态资源分配调度、流的动态路径规划等能力。

3) 接口标准化。企业应用将逐渐向云服务转变，数据中心网络资源管理的功能和角色正在发生着深刻的变化。网络资源管理的作用在于以服务的方式为数据中心网络应用、数据中心网络路由、数据中心监控等提供支撑。为使上层应用能在运行时动态申请管理服务，同时能在不同的云服务环境中自由迁移，网络管理系统需要提供标准的服务接口，如配置接口、资源申请接口、流量监控接口等。

4) 基础设施虚拟化。虚拟化已经成为数据中心网络的基本特征之一。资源管理系统应提供对虚拟化全方位的支持，如虚拟网的划分与配置、多租户的管理与隔离、虚拟资源的感知与控制、虚拟机创建、回收、迁移等，能够满足现代数据中心网络资源动态共享，按需分配的需求。

参考文献：

[1] ZHANG W, SONG Y, RUAN L, *et al*. Resource management in internet-oriented data centers[J]. Journal of Software, 2012,23(2):179-199.

[2] LIN W, QI D. Survey of resource scheduling in cloud computing[J]. Computer Science, 2012,39(10):1-6.

[3] QIAN Q, LI C, ZHANG X, *et al*. Survey of virtual resource management in cloud data center[J]. Application Research of Computers, 2012,29(7): 2411-2415.

[4] Dynamic Host Configuration Protocol[S]. RFC2131, 1997.

[5] Dynamic Configuration of IPv4 Link-Local Addresses[S]. RFC3927, 2005.

[6] AL-FARES M, LOUKISSAS A, VAHDAT A. A scalable, commodity data center network architecture[A]. Proc of SIGCOMM '08[C]. Seattle, WA, USA, 2008.63-74.

[7] MYSORE R N, PAMBORIS A, FARRINGTON N, *et al*. PortLand: a scalable fault-tolerant layer 2 data center network fabric[A]. Proc of SIGCOMM '09[C]. Barcelona, Spain.2009.

[8] GUO C, WU H, TAN K, *et al*. Dcell: a scalable and fault-tolerant network structure for data centers[A]. Proc of SIGCOMM'08[C]. Seattle, WA, USA, 2008.75-86.

[9] GUO C, LU G, LI D, *et al*. Bcube: a high performance, server-centric network architecture for modular data centers[A]. Proc of SIGCOMM'09[C]. Barcelona, Spain, 2009.

[10] CHEN K, GUO C, WU H, *et al*. Generic and automatic address configuration for data center networks[A]. Proc of SIGCOMM'10[C]. New Delhi, India, 2010.39-50.

[11] MA X, HU C, CHEN K, *et al*. Error tolerant address configuration for data center networks with malfunctioning devices[A]. Proc of INFOCOM'12[C]. Florida, USA, 2012.

[12] HU C, YANG M, ZHENG K, *et al*. Automatically configuring the network layer of data centers for cloud computing[J]. IBM Journal of Research and Development, 2011,55(6):1-10.

[13] GREENBERG A, HAMILTON J R, JAIN N, *et al*. VL2: a scalable and flexible data center network[A]. Proc of SIGCOMM'09[C]. Barcelona, Spain, 2009.51-62.

[14] ALIZADEH M, GREENBERG A, MALTZ D A, *et al*. Data center TCP (DCTCP)[A]. Proc of SIGCOMM'10[C]. New Delhi, India, 2010.

[15] VAMANAN B, HASAN J, VIJAYKUMAR T N. Deadline-aware

datacenter TCP (D2TCP)[A]. Proc SIGCOMM'12[C]. Helsinki, Finland, 2012.

[16] WU H, FENG Z,GUO C, *et al*. ICTCP: incast congestion control for TCP in data center networks[A]. ACM CoNEXT'10[C]. Philadelphia, USA, 2010.

[17] WILSON C, BALLANI H, KARAGIANNIS T, *et al*. Better never than late: meeting deadlines in datacenter networks[A]. Proc SIGCOMM'11[C]. Toronto, Ontario, Canada, 2011.

[18] HONG C, CAESAR M, GODFREY P B. Finishing flows quickly with preemptive scheduling[A]. Proc of SIGCOMM'12[C]. Helsinki, Finland, 2012.

[19] ZATS D, DAS T, MOHAN P, *et al*. DeTail: reducing the flow completion time tail in datacenter networks[A]. Proc of SIGCOMM'12[C]. Helsinki, Finland,2012.

[20] KANDULA S, SENGUPTA S, GREENBERG A, *et al*. The nature of data center traffic: measurements and analysis[A]. Proc of IMC'09[C]. Chicago, Illinois, USA, 2009.202-208.

[21] BENSON T, AKELLA A, MALTZ D A. Network traffic characteristics of data centers in the wild[A]. Proc of IMC'10[C]. Melbourne, Australia, 2010.267-280.

[22] SHIEH A, KANDULA S, GREENBERG A, *et al*. Sharing the data center network[A]. Proc of Usenix NSDI'11[C]. Berkeley, CA,USA, 2011.

[23] BENSON T, ANAND A, AKELLA A, *et al*. Understanding data center traffic characteristics[A]. Proc of SIGCOMM'09[C]. Barcelona, Spain, 2009. 92-99.

[24] WANG M, MENG X, ZHANG L. Consolidating virtual machines with dynamic bandwidth demand in data centers[A]. Proc of INFOCOM'11[C]. Shanghai, China, 2011.

[25] EPSTEIN A, BREITGAND D. Improving consolidation of virtual machines with risk-aware bandwidth oversubscription in compute clouds[A].Proc of INFOCOM'12[C]. Florida, USA, 2012.

[26] AL-FARES M, RADHAKRISHNAN S, RAGHAVAN B, *et al*. Hedera: dynamic flow scheduling for data center networks[A]. Proc of Usenix NSDI'10[C]. California, USA, 2010.

[27] MCKEOWN N, ANDERSON T, BALAKRISHNAN H, *et al*. OpenFlow: enabling innovation in campus networks[J].Computer Communication Review.2008,38(2):69-74.

[28] MENG X, PAPPAS V, ZHANG L. Improving the scalability of data center networks with traffic-aware virtual machine placement[A]. Proc of INFOCOM'10[C]. San, Diego, CA, USA, 2010.

[29] BIRAN O, CORRADI A, FANELLI M, *et al*. A stable network-aware vm placement for cloud systems[A]. Proc of 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing[C]. Ottawa, Canada, 2012.

[30] JIANG J, LAN T, HA S, *et al*. Joint VM placement and routing for data center traffic engineering[A]. Proc of INFOCOM'12[C]. Florida, USA, 2012.

[31] HOPPS C. Analysis of an Equal-Cost Multi-Path Algorithm[S]. RFC 2992, IETF, 2000.

[32] RAICIUC, PLUNTKE C, BARRE S, *et al*. Data center networking with multipath TCP[A]. Proc of HOTNETS '10[C]. Monterey, CA, USA, 2010.

[33] FORD A, RAICIU C, HANDLEY M, *et al*. TCP Extensions for Multipath Operation with Multiple Addresses[S]. Internet-draft, IETF, 2012.

[34] Open vswitch project[EB/OL]. http://www.vswitch.org.

[35] VMWare vSphere: vnetwork distributed switch[EB/OL]. http://www.vmware.com/products/vnetworkdistributedswitch.

[36] Cisco nexus 1000V series switches[EB/OL]. http://www.cisco.com/en/US/products/ps9902.

[37] LUO Y, MURRAY E, FICARRA T L. Accelerated virtual switching with programmable NICs for scalable data center networking[A]. VISA'10[C]. New York, NY, USA, 2010.

[38] XU L, ZHANG Y, WU J, *et al*. Network technology research under cloud computing environment[J]. Journal on Communications, 2012, 33(Z1): 16-221.

[39] 802.1Qbg-edge Virtual Bridging[EB/OL]. http://www.ieee802.org/1/pages/802.1bg.html.

[40] 802.1Qbh-bridge port extension[EB/OL]. http://www.ieee802.org/1/pages/802.1bh.html.

[41] PCI-SIG. Single root I/O virtualization and sharing specication [EB/OL]. http://www.pcisig.com/specifications/iov/single_root/.

[42] DONG Y, YANG X, LI J, *et al*. High performance network virtualization with SR-IOV[J]. Journal of Parallel and Distributed Computing, 2012,(72):1471-1482.

[43] TRIPATHI S, DROUX N, SRINIVASAN T. Crossbow: from hardware virtualized NICs to virtualized networks[A]. Proc of VISA'09[C]. Barcelona, Spain, 2009.

[44] RAM K K, MUDIGONDA J, COX L A. sNICh: efficient last hop networking in the data center[A]. Proc of ANCS '10[C]. San Diego, CA, USA, 2010.

[45] NAHIR A, ORDA A, RAZ D. Workload factoring with the cloud: a game-theoretic perspective[A]. Proc of INFOCOM'12[C]. Florida, USA, 2012.

[46] HAYES B. Cloud computing[J]. Commun, ACM, 2008,51(7):9-11.

[47] LUO J, JIN J, SONG A, *et al*. Cloud computing: architecture and key technologies[J]. Journal on Communications, 2011, 32(7):3-21.

[48] FEHLING C, LEYMANN F, MIETZNER R. A framework for optimized distribution of tenants in cloud applications[A]. Proc of 3rd International Conference on Cloud Computing'10[C]. Miami, FL, USA, 2010.

[49] Amazon elastic compute cloud[EB/OL]. http://aws.amazon.com/ec2.

[50] Amazon simple storage service[EB/OL]. http://aws.amazon.com/s3.

[51] IBM blue cloud project[EB/OL].http://www-03.ibm.com/press/ us/en/pressrelease/ -22613.wss.

[52] 802.1Q-virtual LANs[EB/OL]. http://www.ieee802.org/1/pages/802.1Q.html.

[53] 802.1ad-provider bridges[EB/OL]. http://www.ieee802.org/1/pages/802.1ad.html.

[54] 802.1ah-provider backbone bridges[EB/OL]. http://www.ieee802.org/1/pages/802.1ah.html.

[55] 802.1aq-shortest path bridging[EB/OL]. http://www.ieee802.org/1/ pages/802.1aq.html

[56] GUO C,LU G, WANGH J, *et al*. SecondNet: a data center network virtualization architecture with bandwidth guarantees[A]. Proc of Philadelphia[C]. New York, USA, ACM, 2010.

[57] BALLANI H, COSTA P, KARAGIANNIS T, *et al*. Towards predictable datacenter networks[A]. Proc of SIGCOMM'11[C]. New York, ACM, 2011.242-253.

[58] QIN A, YU H. Research on the multi-tenant isolation mechanism based on indentity mapping[J]. Chinese High Technology Letters, 2011,21(10):1007-1013.

[59] MUDIGONDA J, STIEKES B, YALAGANDULA P, *et al*. NetLord: a scalable multi-tenant network architecture for virtualized datacenters[A]. Proc of SIGCOMM'11[C]. Toronto, Canada, 2011.

[60] POPA L, KRISHNAMURTHY A, RATNASAMY S, *et al*. FairCloud: sharing the network in cloud computing[A]. Proc of HOTNETS'11[C]. Cambridge, MA,USA, 2010.

[61] KERRAVALA Z. As the value of enterprise networks escalates, so does the need for configuration management[EB/OL]. http://www.cs.princeton.edu/courses/archive/spr12/cos461/papers/Yankee04.pdf.

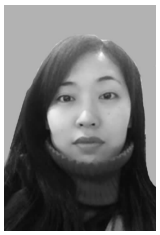
[62] What's behind network downtime?[EB/OL]. http://f.netline. juniper-marketing.com/netline000s/?msg=msg.htm.txt&_m=26.11dk.1.ua06p11znd.3to.

作者简介：

邓罡（1983-），男，贵州晴隆人，国防科学技术大学博士生，主要研究方向为数据中心网络管理、云计算。

龚正虎（1945-），男，湖南长沙人，国防科学技术大学教授、博士生导师，主要研究方向为计算机网络体系结构、高性能计算机网络、数据中心网络、网络管理等。

王宏（1964-），男，博士，湖南益阳人，国防科学技术大学研究员、硕士生导师，主要研究方向为计算机网络协议软件、网络流量测量与分析、数据中心网络、网络管理等。



陈琳（1976-），女，福建陇海人，国防科学技术大学副研究员、硕士生导师，主要研究方向为网络故障诊断、网络配置与管理、数据中心网络、网络管理等。

刘志宏（1986-），男，广东韶关人，国防科学技术大学博士生，主要研究方向为新型网络体系结构。