计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
49(2): 286-293, 2012

面向云计算的数据中心网络体系结构设计

王 聪[1] 王翠荣[2] 王兴伟[1] 蒋定德[1]
[1](东北大学信息科学与工程学院 沈阳 110819)
[2](东北大学秦皇岛分校电子信息系 河北秦皇岛 066004)
(congw1981@gmail.com)

Network Architecture Design for Data Centers Towards Cloud Computing

Wang Cong[1], Wang Cuirong[2], Wang Xingwei[1], and Jiang Dingde[1]
[1](*School of Information Science & Engineering, Northeastern University, Shenyang* 110819)
[2](*Department of Electronic Information, Northeastern University at Qinhuangdao, Qinhuangdao, Heibei* 066004)

Abstract In recent years, the rapid development of cloud computing brings significant innovation in the whole IT industry, which makes the Internet change from translating information data to translating service directly. In cloud computing, enterprises need build their own data centers as privacy clouds or the backbone of public cloud to deploy new Internet information services. Traditional network structure and control plane mechanism in current data centers have their own inherent limitations in network capacity and price quality. The current practicalities also cannot support network multi tenanting which is essential for cloud computing. Therefore, this paper proposes a novel data center architecture which just uses low-cost commercial programmable switches and servers to build the data center networks; and also gives a flexible virtual network bandwidth management mechanism based on convex optimization, which supports the coordinated work between programmable switches and 2.5 layer agents resident in servers. Through a dynamic bandwidth allocation algorithm, it can provide more preferably supports to the resource virtualization which is very common in cloud computing applications. Experimental results show that the proposed data center network architecture can reduce the building-cost significantly as well as improve the network throughput remarkably, and that the virtual network management mechanism provides a flexible manner in bandwidth allocation.

Key words data center network; cloud computing; bisection bandwidth; network throughput; network virtualization

摘 要 近年来,云计算技术的蓬勃发展为整个IT行业带来了巨大变革.传统数据中心网络拓扑构建方式及网络层控制平面的运行机制存在固化性,已经难以满足新形势下日益增长的高性能及高性价比需求,并且无法支持云环境下更加灵活的按带宽租赁数据中心网络的运营方式.因此,提出了一种通过低造价的可编程交换机来构建具有高连通性的非树状数据中心网络的方式,并设计了可编程交换机与服务器2.5层代理协同工作的基于凸优化的虚拟网络带宽控制管理机制,从而提供足够的灵活性以对资源虚拟化技术提供更好的支持.实验表明,新型体系结构在降低构建成本的同时大幅提高了数据中心网络的吞吐量并提供了更加灵活的网络带宽分配机制.

收稿日期:2010-07-06;修回日期:2010-09-25
基金项目:国家自然科学基金项目(61070162,71071028,60802023,70931001);高等学校博士学科点专项科研基金项目(20070145017);中央高校基本科研业务费专项基金项目(N090504003,N090504006)

关键词　数据中心网络;云计算;对剖带宽;网络吞吐量;网络虚拟化

中图法分类号　TP393.02

20 世纪 90 年代,客户端/服务器的计算模式得到了广泛应用,在这种计算模式中,数据中心用来存放服务器并提供服务.近几年,互联网技术的蓬勃发展掀起了建设数据中心的高潮,网上银行、证券和娱乐资讯等网络服务逐渐普及,特别是云计算[1]技术的发展为网络服务形式带来重大变革,使数据中心的发展进入了鼎盛时期.

在云计算环境下,Internet 网络由传送信息数据到直接传送服务.数据中心作为企业构建私有云的硬件平台或者公有云的骨干资源,运行其上的网络服务更加多样化、复杂化,在性能、可靠性和可管理性上的要求越来越细化,这就需要新的设计理念和运行机制的支持,特别是作为信息传输的基础部分——网络层的数据转发和管理机制需要重新设计以满足越来越复杂和多样的数据流传输需求.

随着存储虚拟化[2]等资源虚拟化技术的发展,通常数据中心内的物理主机上会搭载若干独立的虚拟主机,并且虚拟主机可以根据需要在不同的物理主机上迁移.在面向云计算的数据中心内部,资源虚拟化技术的出现使得多个具有独立 IP 地址的虚拟主机公用同一条物理链路,即使目前使用了一些过渡的方式令网络层能够提供一定的支持,但是传统的 TCP/IP 或者 UDP 等协议已经越来越无法为各种服务应用提供足够的性能保障,这点在虚拟机迁移及多 QoS 个性化需求方面显得尤为突出,因此需要将数据中心的网络硬件进行虚拟化以形成多个不同的虚拟网络拓扑从而对资源虚拟化应用提供更好的支持.

为此,本文设计了通过低造价的可编程交换机和商业级服务器来构建具有高连通性的数据中心网络拓扑的方法,并且提出了虚拟网络的控制管理机制,形成了面向云计算的数据中心底层网络体系结构.实验表明,本文提出的网络体系相较于传统树形网络在网络吞吐量及容错方面都有较大提升,并且为虚拟机的迁移提供了更加灵活、高效的支持.

1　相关研究

在数据中心拓扑构建上,传统数据中心所采用的树形分层结构[3](如图 1 所示)通常为包括接入层、汇聚层、核心层的 3 层结构.在树形的末端,一个机架上通常存放数 10 台服务器,这些服务器通过接入层的交换机连接到网络上.在汇聚层和核心层,为了提供尽可能高的性能,采用了造价极高的高端交换机(10 GE 级别)形成高连通网络拓扑.

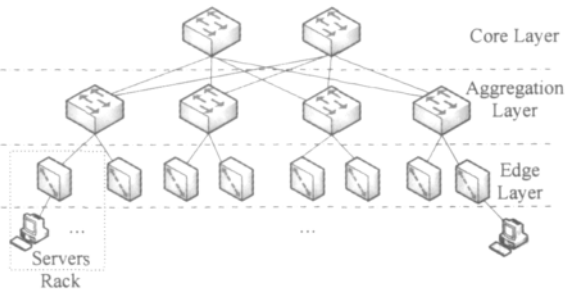

Fig. 1　Traditional tree structure of data centers.

图 1　传统数据中心树形网络结构

传统树形网络构建方式已经难以满足新一代网络服务的需求.首先,这类结构无法提供足够高的对剖带宽、吞吐量及实时通信等性能要求,也无法提供高可扩展性;其次,树形结构在上层存在单点故障,一台核心层或者汇聚层的交换机故障将会导致很大数量的服务器无法进行通信;另外由于传统分层结构在汇聚层与核心层需要部署昂贵的高端高带宽交换机,性价比不高.

关于云计算数据中心非树状网络拓扑构建方面的研究比较具有代表性的有微软的研究团队和美国加州大学的研究团队.加州大学的 Al-Fares 等人提出了 Clos Networks[4],Clos Networks 是一种基于胖树(fat tree)的类树形结构拓扑,其主要目的是在网络端节点处实现更高的聚合带宽.网络分 3 层:核心交换机、聚合交换机及端交换机,通过增加一定的布线复杂度来连接成一个胖树形网络,端交换机用来连接 PC 机.在网络层,Clos Networks 使用两层路由表及多路径的流调度机制,在全负载最坏的情况下可以实现约 87%的聚合带宽.加州大学的 Guo 等人引入了并行计算的一些思想,提出了 DCell[5],DCell 是一种递归构建方式的数据中心,使用的是商业级 PC 和低端交换机,高一级的 DCell 由若干低一级的 DCell 组成.DCell 的扩展性相对于节点的度具有双倍指数增长关系,并且大大降低了数据中心的成本.微软亚洲研究所的 Dan Li 联合加州大学的研究团队提出了一种使用双网卡 PC 机和低端交

换机来构建数据中心的思路，命名为 FiConn[6]，FiConn 同样使用递归构建模式，具有很好的可扩展性和连通性. 在 FiConn 的递归拓扑中，链路被分为若干级别，在网络层中通过一种低开销的流量自适应路由机制来平衡各级别链路中的流量从而达到平衡负载和提高网络吞吐量的目的.

在上述的 3 种新型数据中心网络拓扑中，Clos Networks 网络结构中的主机在网络满负荷情况下仍能够以网卡硬件端口允许的最大带宽进行通信，从而提供了最高的网络对剖带宽，并且由于 Clos Networks 实际上是一种特殊的树形结构变体，因此能够提供最好的兼容性，但是造价要高于 DCell 和 FiConn. DCell 具有最好的可扩展性，但是需要在主机上安装更多的网卡，FiConn 只需在每台主机上安装两块网卡，两者都增加了布线的复杂性. 在容错性上，FiConn 和 DCell 由于采用了递归的拓扑结构，主机需要承担路由功能并且网络内的交换机和主机存在级别差异，因此在交换机及主机故障的情况下将导致网络内数据流分配不平衡，从而导致网络性能的显著下降. 文献[7]比较了这 3 者的容错性，FiConn 和 DCell 的容错性能明显低于 Clos Networks 网络结构. 另外，这两种结构都需要使用全新的路由算法，对于现有应用的支持还有待于进一步解决.

在应对资源虚拟化应用方面的相关研究上，微软研究团队的 Greenberg，Hamilton，Jain 等人提出了 VL2[8]，VL2 主要考虑如何使得虚拟机在服务器上进行灵活的迁移，力求使得虚拟机的迁移对客户及程序设计者透明. VL2 使用了 Clos Networks 的拓扑结构，并在网络层和数据链路层之间加入了相应的路由控制机制. 在 VL2 所提出的体系结构中，应用程序使用服务地址通信而底层网络使用位置信息地址进行转发，这就使得虚拟机能够在网络中任意迁移而不影响服务质量. 加州大学的 Mysore 等人提出了 PortLand[9]，PortLand 使用的同样是 Clos Networks 的拓扑结构，通过在 2.5 层中使用虚拟 MAC 地址来实现虚拟机自由迁移. 在服务器上的某个虚拟机与端交换机第 1 次通信时，端交换机建立该虚拟机的实际 MAC 地址到虚拟 MAC 地址并将其发送给网络底层控制程序，虚拟机进行 ARP 广播时端交换机将广播拦截，然后查询底层控制程序并返回相应的 IP 地址，这样使得虚拟机可以自由迁移不用考虑寻路问题.

文献[8-9]的解决方法在一定程度上提高了虚拟机迁移后的再寻址时间延迟问题，目前通过 DNS 的实现方式需要数分钟甚至更久，采用新方法后可以缩减到数十秒. 但是前者需要在网络内部署专门寻址服务器，并且需要有良好的分布式实时通信机制支持，后者主要在可编程交换机上实现，对交换机性能的要求较高. 另外，这两种方法都无法将隶属于不同应用的虚拟主机进行有效的隔离，在网络通信量较大时，不用服务的数据流由于竞争带宽会相互影响.

2 低成本高连通性的网络拓扑结构

在数据中心网络构建方案设计中，如何在保证足够高性能的前提下尽量减少造价是至关重要的. 本文的目的就是利用低造价且型号单一的可编程交换机及商业级服务器来构建数据中心网络拓扑，新的网络拓扑在对剖带宽及网络聚合吞吐量上要高于传统树形分层结构，并能够对资源虚拟化技术应用特别是虚拟主机在网络中的迁移行为提供更加灵活的支持.

2.1 网络拓扑结构

从降低成本方面考虑，利用低造价交换机及商业级服务器来构建数据中心可以大大减少成本支出. 另外，鉴于现在的 PC 机和服务器都具有至少两个网络端口，充分利用这些端口可以大大提高拓扑内节点的连通性以获得更大的网络吞吐量. 这样的构建方案具有更好的性价比.

图 2 是本文提出的数据中心网络拓扑构图，整个网络由同一型号的可编程交换机组成，中间的服务器将网络分割成两个对称的 Fat-Tree 结构的特殊变体. 每个这样的 Fat-Tree 结构包含核心层、汇聚层和接入层 3 个层次，使用这样结构的好处是可以保证每台服务器的任意网络端口都可以同时以网络硬件接口所允许的最大带宽进行通信而不受网络通信带宽瓶颈的制约.

网络能够容纳的服务器数量取决于构建网络所使用的交换机的端口数 k(图 2 中，$k=2$). 网络中的服务器被分成 k 组，每组包含 $(k/2)^2$ 台服务器. 网络被服务器分成上下两个部分，每部分的接入层和汇聚层对应每组都有 $k/2$ 台交换机，每个接入层的交换机分别连接 $k/2$ 台服务器，剩下的端口分别连接上层的交换机. 核心层有 $(k/2)^2$ 台交换机，每台交换机的第 i 个网络端口连接到第 i 组的汇聚层的某台交换机，这样每个汇聚层的交换机都有 $k/2$ 条链路

与核心层的各个交换机相连接.使用具有 k 个端口的交换机组建的网络可以容纳 $k^3/4$ 台服务器.本文提出的这种结构适用于任意端口数的交换机,如采用常见的 48 口交换机,那么按照本文提出的方法构建的数据中心网络可以包含 27 648 台服务器,足够支持企业构建自己的私有云平台.

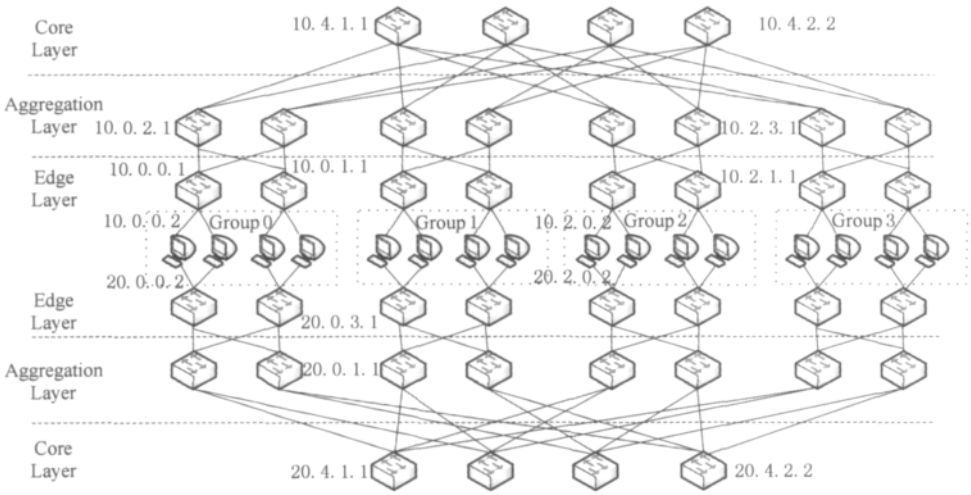

Fig. 2　High connectivity, low-cost network structure.

图 2　高连通性低造价网络拓扑结构图

本文提出的这种网络拓扑构建方式有以下 4 个优点:1)虽然相较于传统树形结构使用了更多的交换机,但是由于无需在核心层和汇聚层采用造价昂贵的高端高性能交换机,因此减少了总体构建成本;2)对于网络中任意的两台服务器之间都存在多条等长度的路径可供选择;3)充分利用了服务器的两个网络端口,提高了网络的连通性和吞吐量;4)在本文提出的网络结构中不存在像传统树形结构中的单点故障,因此容错性得到了加强.

2.2　网络地址分配

在地址分配方式上,为了保证兼容性,采用了与 IP 地址相同的结构.对称的上下两部分网络分别采用 10.0.0.0/8 和 20.0.0.0/8 两个地址段.接入层及汇聚层交换机地址形式为 10.g.s.1 及 20.g.s.1,其中 g 为组号,s 为交换机号(由 0 开始,从左至右,从下到上递增).核心层交换机的地址分别为 10.k.j.i 和 20.k.j.i,其中 k 为组号,j 为汇聚层交换机编号(从 0 开始,由左至右递增),i 为该交换机与核心层交换机连接的端口序号(0~k/2).

主机网络端口的地址与其连接的接入层交换机处于同一网段,形式为 10.g.s.id 和 20.g.s.id,其中 id 为服务器编号,由左至右递增,范围为(2, k/2+1).

这样的地址结构使得后续给出的虚拟网络构建及控制管理机制都能够良好地兼容目前广泛的基于 IP 协议的上层应用.只要 IP 地址结构不变,那么对于网络层作出的相应改动相对于上层应用来说就是透明的,上层应用不必理解地址字段的特殊含义.

3　虚拟网络构建及控制管理机制

云计算环境下,数据中心内服务器上通常运行多个虚拟机来提供不同的服务,这样能够便于应用服务的迅速、灵活的部署.在服务器硬件故障发生时,虚拟机将迁移到另外的服务器上.目前虚拟机迁移后的恢复寻址工作通常由 DNS 服务器来完成,但是由于 DNS 系统被动的工作机制导致效率不高.从底层网络支持情况来讲,运行不同应用服务的虚拟机对于网络的 QoS 具有不同的个性化需求,目前情况下,底层网络普遍采用的基于尽量交付机制的 IP 协议对于个性化的支持显然不够.另外,让隶属于不同应用服务的众多虚拟机同时运行在同一个物理网络上也造成了管理和带宽分配的混乱.

因此,本文提出了一种底层网络控制管理体系结构,通过将硬件网络分割成不同的虚拟网络来实现对上层应用灵活的支持.不同的虚拟网络内运行隶属于不同应用服务的虚拟机,这样可以根据应用服务的 QoS 需求来决定网络所运行的协议及参数也更加有利于虚拟机的控制和管理.虚拟网络的构建也将为数据中心的多用户租赁服务和云计算环境下的资源分配提供配套的灵活管理和带宽控制体系,进一步促进虚拟化技术与数据中心的融合.

3.1 对虚拟机迁移的支持体系

图 3 为虚拟网络划分及控制管理系统示意图. 在开源操作系统的 2.5 层添加一个代理与交换机协同工作是可行的[8-9]. 本文的思路是借鉴应用于 Internet 骨干网络交换机上的网络虚拟化技术[10-11],并使驻留在主机上的代理与可编程交换机的控制软件实时地交换网络运行时参数,以达到动态调整各虚拟网络带宽的分配及控制管理参数的目的. 这样的协同工作机制使得不同的虚拟网络运行不同的网络层路由协议成为可能,这将能够保证运行于不同虚拟网络上的应用服务对于 QoS 的个性化需求.

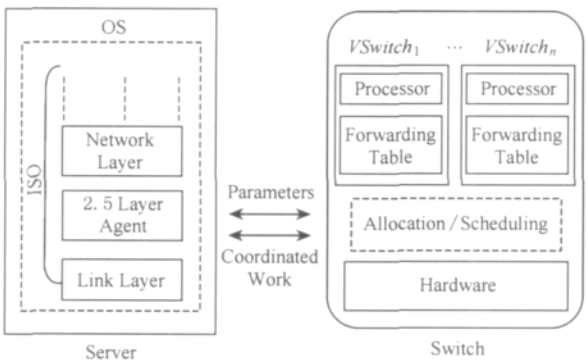

Fig. 3 Virtual network management architecture diagram.

图 3 虚拟网络管理体系结构示意图

2.5 层代理的另一个重要功能就是实现虚拟机的迅速迁移,文献[8-9]都设计了 2.5 层代理的映射功能,通过将虚拟地址和实际地址进行一次映射以隔离上层应用使用的网络地址和底层网络进行交换时使用的物理地址之间的联系. 不同的是文献[8]映射的是服务地址和位置信息地址,而文献[9]映射的是虚拟 MAC 地址和实际 MAC 地址.

本文提出的办法是令每台主机上驻留的代理为每个虚拟网络创建一个映射表(如图 4 所示),用来记录网络内的虚拟机 IP 地址与物理主机 MAC 地址的对应关系. 代理之间可以实时通信,通过类似路由发现的分布式通信机制,周期性更新运行于各台服务器上的虚拟主机与硬件网络地址的对应关系.

在某个服务器上建立新的虚拟主机时,代理会记录新的虚拟 IP 地址与物理主机 MAC 地址的对应关系并在该虚拟网络内进行广播,这样各个代理上对应该虚拟网的映射表都将被更新. 当虚拟机间通信时,请求通信的主机发送的 ARP 探测包将直接被代理捕获,检索本机代理上的映射表,如果有匹配项将直接返回对应的 MAC 地址,如果没有,代理

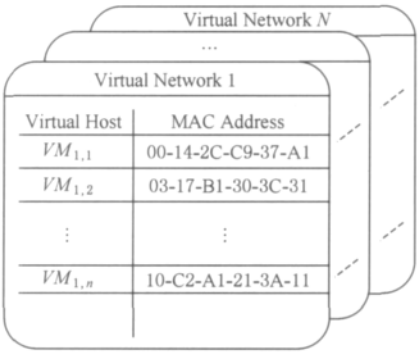

Fig. 4 Mapping table between VMs and MAC address created and managed by the agent.

图 4 代理负责创建和管理的虚拟机与 MAC 地址映射表

将负责进行广播以获得正确的 MAC 地址. 当出现服务器故障时,虚拟主机迁移到其他服务器上之后,服务器上的代理将会在网络中主动广播迁移后虚拟主机地址与服务器端口的对应关系从而加快虚拟机迁移后的恢复时间.

3.2 虚拟网络带宽分配机制

本文提出的带宽分配及控制机制建立在图 3 所示的虚拟网络管理体系之上,可编程交换机负责分配带宽, 2.5 层代理负责参数的协调和反馈. 带宽分配机制示意图如图 5 所示:

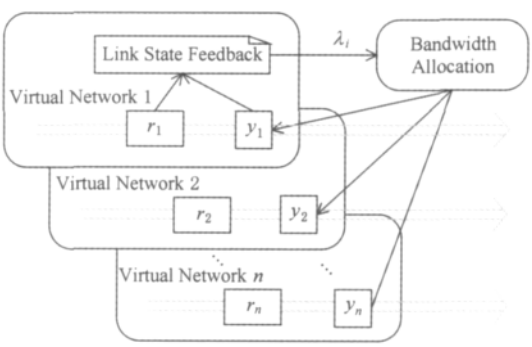

Fig. 5 The structure of bandwidth allocation mechanism.

图 5 带宽分配机制结构图

每台可编程交换机上都部署一个这样的带宽分配逻辑单元,以便对于虚拟网内的各条链路都能够进行分配和控制. 在图 5 中,r_i 是由主机上的 2.5 层代理返回的虚拟机 $VM_{i,n}$ 网络端口的发送速率,以类似于 TCP/IP 拥塞窗口慢启动机制的算法进行增减,y_i 是当前虚拟链路占用的带宽,λ_i 是通过 r_i 和 y_i 计算出的用于带宽分配决策的输入参数.

带宽分配的目的是在保证所有虚拟网络带宽的总和不超过物理链路最大带宽的前提下,使所有虚拟网络的带宽利用率达到最优,这样就把带宽分配

问题抽象成了一个最优化的数学问题从而根据经典数学算法进行求解.本文采用了根据链路的拥塞状态反馈来周期性的分配虚拟网络带宽的方法.拥塞状态 S_l^k 的计算函数如下:

$$S_l^k(T+t)=\left|S_l^k(t)-\beta(y_l^k(t)-r_l^k(t))\right|, \quad (1)$$

其中 t 代表时间,T 是一个时间周期,β 是用于平滑结果的 0,1 之间的步进值.由式(1)可见,虚拟网络 k 在链路 l 上的链路拥塞状态由其得到的带宽 y_l^k 和链路负载 r_l^k 决定.通过式(1),每个虚拟网络可以动态地调整所占用的虚拟链路的网络带宽.

在交换机端,目的是让所有的虚拟链路能够协同工作以获得最大利益,这是一个总体最优化问题,其数学模型如下:

$$\max\sum_k \varphi^k U^k(y^k,r^k),$$
$$\text{s.t.}\ \sum y^k \leqslant C, \quad (2)$$

其中 φ^k 为预设的虚拟网络 k 的优先级别,U^k 为底层网络提供商给虚拟网络 k 分配 y^k 的带宽时获得的利益,即在云计算环境下多租赁机制获得的收益函数,在实际中还可能与租赁时间等参数有关.C 是物理链路的最大带宽.

细化到某台交换机上属于虚拟网络 k 的链路 l 上的带宽分配模型,上述最优化模型可以拆分为在链路 l 上通过图 5 所示的带宽分配逻辑单元上运行的算法:

$$\lambda_l^{*(k)}(t)=S_l^{*(k)}(t)+\frac{\partial U_l^k(y_l^k,r_l^k)}{\partial y_l^k},$$
$$V_l^k(T+t)=\left|y_l^k(t)+\alpha\times(\varphi^k)\times(\lambda_l^{*(k)}(t))\right|,$$
$$y_l^k(T+t)=\arg\min_{y_l^k}\left[y_l^k-V_l^k(T+t)\right],$$
$$\text{s.t.}\ \sum y_l^k \leqslant C_l. \quad (3)$$

式(3)将为某台交换机上属于虚拟网 k 的虚拟链路实时地分配带宽.其中 $S_l^{*(k)}$ 指的是虚拟网 k 通过链路拥塞程度计算式(1)得到的一个收敛值.V_l^k 是带宽分配逻辑单元在链路 l 上应该分配给虚拟网络 k 的带宽,它由链路状态参数 $\lambda_l^{*(k)}$ 和虚拟网的优先级 φ^k 决定,α 是一个类似式(1)中 β 的平滑常数.由于每个 V_l^k 都是单独决定的,因此在链路 l 上所有的 V_l^k 的总和可能超过链路的物理带宽 C_l,因此需要进行一次协调修改,最终的结果 y_l^k 将作为实际的带宽分配给虚拟网络 k.

本文关于带宽分配机制的设计主要目的是给出运行于可编程交换机上的系统逻辑结构,在带宽分配算法上还可以采用其他的数学模型进行求解.

4　实验与讨论

本文的实验软件采用了斯坦福大学研发的 OpenFlow VM[12] 仿真平台.采用这一工具的原因是目前已经有一些厂商(如 NEC 公司)开始生产基于 OpenFlow 的可编程交换机,仿真结果具有较强的实际价值.实验所用硬件平台采用的是 IBM X3650 服务器,4 核 Xeon 3.06 GHz 处理器×2,16 GB 内存,操作系统为 CentOS 5.

仿真实验在 OpenFlow VM 环境内分别构建如图 1 所示的传统树形分层网络结构以及本文提出的如图 2 所示的网络结构.两种网络结构均采用 8 口交换机进行构造,其中树形结构的汇聚层及核心层的带宽为 10 Gbps,其余链路均为 1 Gbps,各网络结构部署的服务器数量均为 128 台.在构建的虚拟网络结构上,通过 D-ITG 模拟出数据中心内部的 all-to-all 通信流量,测试了在交换机故障下的网络聚合瓶颈吞吐量变化趋势.聚合瓶颈吞吐量是网络内各数据流获得的最小带宽与网络内总数据流数量的乘积,它可以反映一个数据中心网络拓扑的性能.

由图 6 可见,本文提出的体系结构在无故障情况聚合瓶颈吞吐量 3 倍于传统的树形结构,并且随着交换机故障率的上升,呈现出更加平滑的下降趋势.导致这样结果的原因主要是由于传统树形结构在高层存在不可避免的对剖带宽限制,尤其是当某台交换机出现故障时将给网络其他交换机带来很大的负载,而本文的结构由于采用了更多的交换设备并充分利用了服务器的两块网卡构建网络,因而网络容量和容错性都得到了很大提升.

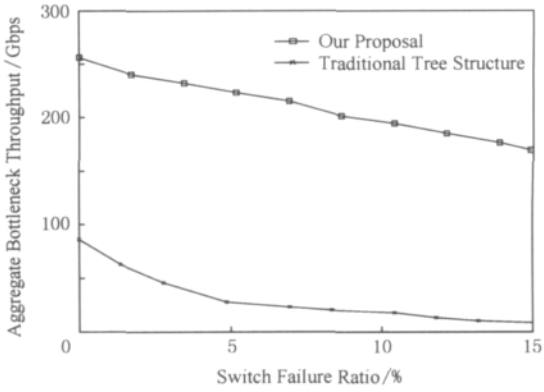

Fig. 6　The aggregate bottleneck throughput under switch failure.

图 6　聚合瓶颈吞吐量相对于交换机故障的变化率

第 2 个实验主要测试本文提出的虚拟网络控制管理机制内的带宽分配机制及算法. 在前面实验的基础上在实际网络拓扑上划分出两个逻辑拓扑,在虚拟 OpenFlow Switch 上编写了上节给出的带宽分配算法. 具体参数设置为: $T=100$ ms,优先级 φ^k 均设置为 1,$\alpha=\beta=0.5$,$U^k=5r^k-2\times|(y^k)^2-r^ky^k|$. 同样在两个虚拟网络内部模拟了 all-to-all 的通信流量,其中虚拟网 1 从一开始便全开流量,虚拟网 2 从 200 ms 后开始产生流量,根据实验数据统计了如图 7 所示的两个虚拟网络的吞吐量变化曲线及收敛情况:

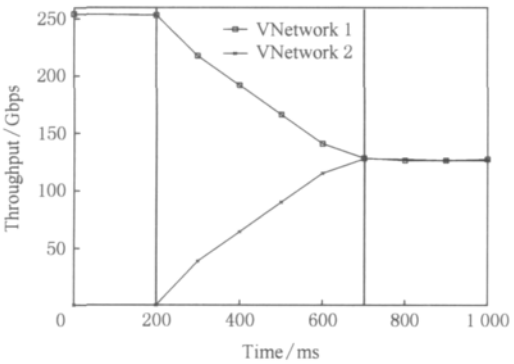

Fig. 7 Throughput of two virtual networks.

图 7 虚拟网络吐量变化曲线

从图 7 可以看出,两个虚拟网络的吞吐量收敛迅速(大概经过 5 个周期)由于优先级相等,两个虚拟网络最终占用的带宽也相等(基本等于实际物理链路带宽的一半). 该实验结果反映了本文提出的虚拟网络带宽分配管理机制对于虚拟网络带宽的分配是符合预先设计目标的.

5 结论和进一步工作

随着数据中心的不断发展,特别是云计算技术应用热潮的到来,传统数据中心的拓扑结构和网络体系结构必将面临变革. 为此,本文提出了一种新型的面向云计算的高性价比数据中心网络体系结构,给出了网络拓扑的构建方式和虚拟网络的划分管理及带宽分配机制,为上层的应用服务及资源虚拟化应用提供了更灵活的支持.

在构建方式上,本文提出的拓扑结构虽然采用了两倍于 Clos Network 结构的交换机数量,但是无需使用造价昂贵的高端高性能交换机并充分利用了服务器的两个网络端口,因此在提高数据中心构建性价比的同时大幅提高了网络的吞吐量. 实验表明,本文提出的新型低成本网络拓扑可以实现更大的网络聚合瓶颈吞吐量和更强的容错性能. 另外,本文提出的虚拟网络带宽分配机制工作良好,在网络流量变化的情况下能够迅速收敛,这将为云计算环境下资源虚拟化技术提供支持,并为网络虚拟化的发展提供一定的技术储备.

为了应对目前虚拟机迁移遇到的困难,解决通过被动更新 DNS 来实现迁移的不足,本文提出的基于 2.5 层代理的方法还需要进一步的研究和实验. 下一步主要工作将论证 2.5 层嵌入程序对于操作系统带来的开销以及代理间通信所造成的网络开销的大小及合理性. 另外,由于新型拓扑仿真困难,所需工作量巨大,而相关研究中介绍的部分网络结构在 OpenFlow VM 环境下的实现存在一些技术问题,下一步将考虑采用合适的仿真工具对这些新提出的技术方案进行比较.

参 考 文 献

[1] Chen Kang, Zheng Weimin. Cloud computing: System instances and current research [J]. Journal of Software, 2009, 20(5): 1337-1348 (in Chinese)
(陈康, 郑纬民. 云计算: 系统实例与研究现状[J]. 软件学报, 2009, 20(5): 1337-1348)

[2] Wang Di, Xue Wei, Shu Jiwu, et al. Fault tolerance with virtual disk replicas in the mass storage network [J]. Journal of Computer Research and Development, 2006, 43(10): 1849-1854 (in Chinese)
(王迪, 薛巍, 舒继武, 等. 海量存储网络中的虚拟盘副本容错技术[J]. 计算机研究与发展, 2006, 43(10): 1849-1854)

[3] Barroso L A, Dean J, Hölzle U. Web search for a planet: The Google cluster architecture [J]. IEEE Micro, 2003, 23(2): 22-28

[4] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture [C] //Proc of ACM SIGCOMM'08. New York: ACM, 2008: 63-74

[5] Guo Chuanxiong, Wu Haitao, Tan Kun, et al. DCell: A scalable and fault-tolerant network structure for data center [C] //Proc of ACM SIGCOMM'08. New York: ACM, 2008: 75-86

[6] Li Dan, Guo Chuanxiong, Wu Haitao, et al. FiConn: Using backup port for server interconnection in data centers [C] //Proc of IEEE INFOCOM 2009. Piscataway, NJ: IEEE, 2009: 2276-2285

[7] Guo Chuanxiong, Lu Guohan, Li Dan, et al. BCube: A high performance, server-centric network architecture for modular data centers [C] //Proc of ACM SIGCOMM'09. New York: ACM, 2009: 63-74

[8] Greenberg A, Hamilton J R, Jain N, et al. VL2: A scalable and flexible data center network [C] //Proc of ACM SIGCOMM'09. New York: ACM, 2009: 51-62

[9] Mysore R N, Pamboris A, Farrington N, et al. PortLand: A scalable fault-tolerant layer 2 data center network fabric [C] //Proc of ACM SIGCOMM'09. New York: ACM, 2009: 39-50

[10] He Jiayue, Zhang Shenrui, Li Ying, et al. DaVinci: Dynamically adaptive virtual networks for a customized Internet [C] //Proc of ACM CoNEXT 2008. New York: ACM, 2008: 556-567

[11] Bhatia S, Motiwala M, Muhlbauer W, et al. Hosting virtual networks on commodity hardware, GT-CS-07-10 [R]. Atlanta, Georgia: Georgia Tech University, 2008

[12] Stanford Clean Slate Program. The OpenFlow switch consortium [EB/OL]. [2010-06-30]. http://www.openflowswitch.org/

Wang Cong, born in 1981. Received his PhD degree in computer application technology from Northeastern University in 2011. His main research interests include routing protocol, network virtualization and data center networks.

Wang Cuirong, born in 1963. Received her PhD degree in software and theory from the Northeastern University, China in 2004. Professor in the Department of Computer Engineering at the Northeastern University at Qinhuangdao, China. Member of China Computer Federation. Her main research interests include routing protocol, network security and sensor networks.

Wang Xingwei, born in 1968. Received his BSc, MSc and PhD degrees from Northeastern University in 1989, 1992 and 1998 respectively. Professor in the College of Information Science and Engineering, the Northeastern University. His main research and teaching interests focus on computer network, distributed computing and information security.

Jiang Dingde, born in 1974. Received his PhD degree in communications and information systems from the School of Communications and Information Engineering, University of Electronic Science and Technology of China, in 2009. Associate professor in the College of Information Science and Engineering, Northeastern University. His main research interests include network measurement, network security, and cognitive networks.