TURING 图灵程序设计丛书

ELSEVIER 爱思唯尔

云数据中心网络技术

由英特尔专家撰写，让读者掌握云网络基本软硬件设施和设计方法，了解云数据中心发展趋势

【美】Gary Lee 著　唐富年 译

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

云数据中心网络技术

【美】Gary Lee 著 唐富年 译

人民邮电出版社
北 京

图书在版编目（CIP）数据

云数据中心网络技术 /（美）李（Lee,G.）著 ；唐富年译. -- 北京 ：人民邮电出版社，2015.12
（图灵程序设计丛书）
ISBN 978-7-115-40518-0

Ⅰ. ①云… Ⅱ. ①李… ②唐… Ⅲ. ①计算机网络－数据处理 Ⅳ. ①TP393

中国版本图书馆CIP数据核字(2015)第230561号

内 容 提 要

本书聚焦于数据中心内部的网络，所讨论的话题集中在大型云数据中心内部组网所需的设备、软件和标准。主要内容包括：云计算和云端网络互连概述，数据中心的演变，交换结构技术，云数据中心网络拓扑结构，网络虚拟化，软件定义网络，等等。

本书适合云计算网络、网络建设、网络管理、系统集成行业的开发人员、技术工程师等阅读。

◆ 著　[美] Gary Lee
译　唐富年
责任编辑　朱　巍
执行编辑　杨　琳
责任印制　杨林杰

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路11号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：12.5
字数：296千字　2015年12月第1版
印数：1-4 000册　2015年12月北京第1次印刷

著作权合同登记号　图字：01-2015-2830号

定价：49.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京崇工商广字第0021号

版权声明

Cloud Networking: Understanding Cloud-based Data Center Networks, 1st Edition

Gary Lee

ISBN: 978-0-12-800728-0 .

译者序

这是一本从网络技术的视角讲述云计算的图书，它的主要特点在于通俗易懂而且没有炒作云计算概念的篇幅。这本书是一部耗费心血完成的作品，对很多概念及技术的理解和解释都非常准确，适合刚刚接触云计算技术的读者阅读。

在信息行业从业多年，我目睹了云计算技术从“云雾”中慢慢落地，同时也亲身感受到数据中心的规模正在变得越来越庞大。我的身边也有不少人正在玩命地开发那些只有在计算集群上才能跑得动的应用程序，但是很少有人能够把云计算讲得透彻明白。曾经听一位计算机行业的老专家笑谈，所谓云计算有三层意思：一是“云里雾里”，二是“众说纷纭（云）”，三是“不知所云”。这基本上反映了相当一部分人对云计算的印象。然而不管怎么说，云计算技术的出现和发展，让客户端越来越轻薄短小，让服务端越来越庞大臃肿；这门技术还充满了商业的味道，骨子里流露出“把方便留给用户，把难题留给自己”的服务运营理念。

本书的作者 Gary Lee 有丰富的电气工程学知识背景和从业经验，这也正是他能够把云网络相关技术的发展讲述得如此清晰的原因。本书的第一个关键词是“云网络”，网络工程技术人员可以从这本书中了解到云网络的基本软硬件设施和设计方法；第二个关键词是“云数据中心”，数据中心建设和管理人员可以从中了解数据中心发展的历史和趋势，将其作为日常工作的参考。

由于工作原因，我也时常与浪潮、华为以及一些科研院所的技术研发人员打交道。在他们的设计方案中，动辄需要十台八台服务器作为各类节点。这让人不禁暗自感叹，不知从什么时候开始，服务器这样的东西居然也跟白菜一样了。看了本书才发现，这算什么，人家大型云数据中心是开着卡车更换模块的。

我一直固执地认为，未来全球最先进的数据中心必然在中国，这是因为中国人面对着世界上最复杂的数据管理难题。在享受“双十一”盛宴的时候，绝大多数人不会想到阿里巴巴强大的服务支持团队其实在那一天如临大敌；在发泄对 12306 网上购票的各种不满时，绝大多数人并不明白 12306 的工程师们心中的苦涩和无奈。这些人的坚守和拼搏才是中国速度的来源和中国数字的支撑。因此，在翻译这本外国人撰写的图书之时，我也想借用译者序的这一点空间表达对他们的敬意。

最后，感谢图灵公司的朱巍和杨琳等诸位编辑在本书翻译和出版过程中给予的帮助和付出的心血，也感谢家人对我的支持和谅解，因为这本书的翻译工作几乎占用了我所有的业余时间。由于水平有限，译文之中难免出现疏漏，请读者海涵之。

唐富年于济南

前　言

在过去的 30 年里，我目睹了半导体行业和网络行业的许多进展。正是由于网络系统依赖于半导体技术的演进，这两方面的进展在许多方面相互交织。鉴于在半导体和网络行业工作多年，我想先介绍一些与此相关的背景，这样你们就会明白我的观点源自何处。

当我以应届大学毕业生的身份加入半导体行业时，研究实验室仍在努力确定可用于高性能逻辑器件的最佳技术。刚开始，我是硅双极芯片设计师，后来很快转到了 GaAs（Gallium Arsenide，砷化镓）芯片的设计上。20 世纪 90 年代，我亲眼见证了 CMOS 成为在行业中占主导地位的半导体技术。我刚从大学毕业的时候，以太网还只是众多建议网络协议中的一个，但是到 20 世纪 90 年代，它已经发展到了开始主导各种网络应用的地步。现在，在局域网络、数据中心网络、运营商网络和模块化系统背板等领域，已经很难找到其他可以与以太网竞争的网络技术了。

1996 年，我在 Vitesse 半导体公司工作。在大约 12 年的 GaAs 芯片设计经历之后，我开始探索将 GaAs 技术用于新的交换结构（switch fabric）架构。当时，硅技术在最大带宽容量上仍然落后于 GaAs 技术，我们今天的交换结构芯片架构尚不存在。我有幸与网络工程顾问 John Mullaney 在同一团队共事，一起设计了一种新的高速串行交换机架构，并且因此获得了两项专利。在这一时期，我们研究关于交换结构架构的论文时，有一个名字频频出现——Nick McKeown。当 Nick McKeown 还是加州大学伯克利分校的博士生时，他和学生们就进行了大量的基础性研究，其中的很多理念被当时设计的新兴交换结构架构所采用，也促进了今天交换结构架构设计的诞生。20 世纪 90 年代末，CMOS 技术在性能水平上很快超过了 GaAs 技术，所以 Vitesse 半导体公司的团队也改弦更张，开始为各种各样的通信市场研发大型 CMOS 交换结构芯片组。我们并不是唯一这样做的公司。

从 1996 年到 21 世纪初电信泡沫结束，人们提出了 20 ～ 30 种崭新而独特的交换结构芯片组设计，主要面向飞速发展的电信行业。这些设计有些来自 IBM 这样的老牌公司，有些则来自从思科和北电网络等公司跳槽出来的设计工程师所成立的创业公司，还有一些来自斯坦福大学和华盛顿大学这样的机构。泡沫最终破裂，资金随之枯竭，这些研发成果绝大多数都已绝迹。现在，这些公司当中仅有很少一部分幸存了下来，被 Broadcom 公司收购的 Sandburst 公司和 Dune Networks 公司就是两个例子。

电信行业的飞速发展期结束之后，在英特尔公司的主导下，交换结构芯片行业仅存的几家

公司联合起来组建了 ASI-SIG（Advanced Switching Interconnect Special Interest Group，高级交换互连特别兴趣小组）。该小组的目标是为遵循 PCI Express 接口规范建立的通信系统创建一个标准的交换结构架构。在 ASI 董事会制定规范期间，我作为 Vitesse 公司的代表参加了 ASI-SIG。人们很快发现，这样的规范显得过于雄心勃勃了。这最终导致英特尔等公司慢慢从中抽身，直到 ASI 走到尽头。但是对我而言，这是一次很好的学习经历，既让我明白了标准机构应该如何运作，也促使我对当今计算机行业广泛应用的 PCI Express 标准有了一些技术上的见解。

在 ASI 完全退出历史舞台之前，我就开始为 Xyratex 公司工作了。这家存储器公司致力于基于 ASI 标准为服务器开发共享 IO 系统，希望以此来扩大自己的市场。其共享 IO 计划最终搁浅，所以我转变方向，开始研究面向存储器应用的 SAS 交换机。虽然只在 Xyratex 工作了 2 年时间，但是我学到了很多关于光纤通道、SAS 和 SATA 存储阵列设计的知识，并且从 Xyratex 公司的工程师和科学家们那里了解到基于闪存的存储器具有何种优势和缺陷，他们甚至在 Xyratex 公司从 IBM 公司分离出来之前就已经花费了多年时间来研究这些技术。

纵观研究专用交换结构架构的那段时光，我那些在 Vitesse 公司以太网部门工作的同行总会批评我们所做的工作，并且说“绝不要跟以太网作对”。如果是在 20 世纪 90 年代末，我可以就为什么不能在电信交换结构设计中使用以太网给出一长串理由。然而这些年以来，以太网的标准在不断演进，以至于现在大多数模块化通信系统都在其背板中使用了以太网。可以说，如果电信泡沫还没有让做交换结构的创业公司全军覆没，那么以太网将会做到这一点。

我职业生涯的下一站是第三家创业公司 Fulcrum Microsystems。在我加入的时候，公司刚刚推出了为数据中心设计的最新 24 端口 10GbE 交换机芯片。虽然我在大部分职业生涯中都在和电信式交换机打交道，但是在过去的几年里也学会了很多关于数据中心网络的知识，最近又学会了如何运营大型云数据中心。我还深入了解了我们一直在交换机芯片产品中支持的以太网和第 3 层网络（Layer 3 Networking）的各种标准。2011 年 9 月，英特尔公司收购了 Fulcrum Microsystems 公司。作为英特尔公司的一分子，我学到了更多关于服务器虚拟化、RSA（Rack Scale Architecture，机架规模架构）、微服务器设计和软件定义网络（Software-Defined Networking）等方面的知识。

人生是一个不断学习的过程，我也一直对技术和技术的演进充满兴趣。我的兴趣有一部分可能是从祖父和父亲那里继承来的：我的祖父在 1920 年前后成为一名电子工程师，而父亲在 1950 年左右成为了一名机械工程师。我学到的很多东西还来自于这些年共事过的同事们。尽管因为人数众多而无法在此列出，但是他们中的每一个人都曾经在某些方面影响和教育过我。

我要特别感谢英特尔公司的同事 David Fair 和 Brian Johnson，他们对本书的一些关键章节提出了很有帮助的意见。我还要感谢我的家人，尤其是妻子 Tracey。当我带着她在全国各地一次次地加入创业公司时，她始终是我最大的支持者。

目　录

第 1 章

欢迎来到云网络

欢迎阅读这本专门讲述云网络的书。不管你是否意识到，“云”都已经对你的日常生活产生了重要影响。每当你在 Facebook 上查看某人的状态时，在亚马逊上购买商品时，从谷歌地图上获取方位时，你都在访问着某个大型云数据中心的计算机资源。这些计算机被称作服务器，它们必须彼此互连，并且通过运营商网络与你相连，你才能访问上述这些信息。在后台，你的每一次点击都可能导致数据中心的服务器之间产生数百个事务。这些事务都必须发生在高效且成本低廉的网络之上，这样的网络也使数据中心充满活力。

本书主要关注数据中心内部的网络，而非那些在数据中心和个人设备之间传递信息的运营商网络。本书的主题聚焦于网络设备、软件以及用于建设大型云数据中心网络的标准，供希望深入了解如何运营大型数据中心网络的人使用。本书并不是一本关于网络的教科书，不包括深奥的协议细节、方程式和性能分析等内容。相反，我们希望这是一本易于阅读的概述性书籍，能够告诉你云数据中心的网络是如何构建和运营的。

1.1 介绍

在全世界，已经部署或正在建设的全新云数据中心都拥有数以万计的服务器，有时候甚至拥有数十万台服务器。这些云数据中心有时被称作超大规模数据中心。你可以把服务器看作去掉图形界面和键盘的台式电脑，不过加上了增强的处理器和网络连接。服务器的用途是为客户端设备（如笔记本电脑、平板电脑或智能手机）提供信息“服务”。在许多情况下，在客户端设备上点击单个网站就能够导致数据中心的服务器之间产生显著的流量。在云数据中心，这些服务器之间以及与相关联存储器之间的高效通信都依赖于先进的数据中心网络技术。

本章为本书的后续内容奠定了基础，为尚不了解本书主题的读者提供了一些基础性的网络背景知识，还对云计算和云网络进行了概述。本章讲述的背景知识能够帮助你更好地理解后续内容所涉及的主题。云数据中心网络的一些关键特性是本书许多章节的基点，将在本章的最后

予以介绍。

1.2 网络基础

本书的目的并非让读者深刻理解网络协议和标准，而是对云数据中心网络内部的技术进行全面概述。为了更好地理解本书介绍的某些主题，最好重温网络原理的基础知识。如果你对网络基础知识已经非常熟悉，那么可以跳过本节的内容。

1.2.1 网络协议栈

几乎每一本教科书都会介绍 7 层 OSI（Open Systems Interconnect，开放系统互连）网络协议栈的相关知识。该模型设计于 20 世纪 70 年代，最初是 OSI 项目的一部分，目标是提供一个支持多厂商互操作的通用网络标准。OSI 并未得到认可，相反，TCP/IP（Transmission Control Protocol/Internet Protocol，传输控制协议 / 网际协议）变成了今天互联网通信的主流标准。不过，OSI 协议栈现在仍然出现在许多技术论文和教科书中。

虽然网络界仍然会提到 OSI 模型，但是当前使用的大多数协议都少于 7 层。在数据中心网络中，虽然以太网包含第 1 层和第 2 层两部分，但我们将其看作一个第 2 层的协议。同样，虽然 TCP/IP 也有第 3 层和第 4 层两部分，但我们仍将其看作第 3 层协议。第 5 ～ 7 层在业界中通常被看作应用层。在本书中，我们将第 2 层视为交换层（也就是以太网），将第 3 层视为路由层（也就是 TCP/IP）。在此之上的其他层被视为应用层。图 1-1 展示了一个简化的模型，以及简单的数据中心事务处理。

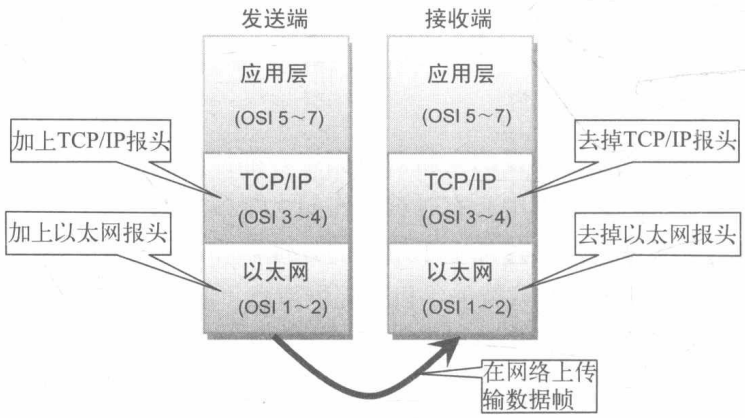

图 1-1
简单的数据中心处理示例

在这个简化了的例子中，发送端应用程序将数据提交给 TCP/IP 层（有时简称为第 3 层）。

这些数据被分割成若干帧（数据包），在将每一帧数据提交给以太网层（有时简称为第 2 层）之前都需要添加 TCP/IP 报头。接下来，为数据帧添加一个以太网报头并将其发送到接收设备。在接收端，接收到的帧首先在以太网层去掉以太网报头，然后在 TCP/IP 层去掉 TCP/IP 报头，之后被重新组装成数据提交给应用层。虽然这是一个非常精简的解释，但是可以让你了解一些有关第 2 层和第 3 层协议的背景知识。本书的后续章节还将更加详细地介绍相关内容。

假设你要从公司的邮件收发室发送一个包裹。你本人就好比应用层，告诉收发室必须把包裹送到公司在另一个城市指定的邮件服务站。收发室就是第 3 层，将包裹放在一个盒子里，查看目的地邮件服务站的编码并附上一个基于该编码的地址，然后将包裹交给运输公司。运输公司拿到包裹后，会查看目的地的地址，然后加上自己特定的条码标签（相当于第 2 层），以便包裹能够到达目的地配送中心。在运输途中，运输公司只需要关注第 2 层的标签。到达目的地配送中心之后，需要再次检查本地地址（第 3 层）以确定最终目的地。这种分层方法简化了第 2 层运输公司的任务。

1.2.2 包与帧

几乎所有的云数据中心网络都使用可变长度帧来传输数据，这些帧也被称为包。这两个术语我们都会在本书中用到。大的数据文件要在通过网络发送之前分割成帧。一个帧格式的示例如图 1-2 所示。

图 1-2
帧格式示例

首先使用第 3 层报头（如 TCP/IP）对数据进行封装，然后再使用第 2 层报头（例如以太网，参见 1.2.1 节示例）对数据进行封装。报头通常包含源地址和目的地址信息，以及帧类型、帧优先级等其他信息。在许多情况下，在帧的末尾使用校验码来验证整个帧的数据完整性。被传输数据有效载荷的大小和帧的大小取决于协议。标准以太网帧的大小范围在 64 ～ 1522 字节。在某些情况下，也可以支持大小超过 16K 字节的巨型帧。

1.2.3 网络设备

云数据中心可以使用多种类型的网络设备。服务器都配有 NIC（Network Interface Card，网络接口卡），这些 NIC 用来为服务器 CPU 提供外部以太网端口。有了它们，就可以通过数据线将服务器连接到网络上的交换机。术语“交换机”通常是指采用第 2 层报头信息来转发数据的设备。有时，以太网交换机也称为以太网桥，而且这两个术语也可以互换使用。术语“路由器”

通常是指采用第3层报头信息转发数据的设备。交换机和路由器都可以在大型云数据中心网络中使用。在有些情况下，以太网交换机也可以支持第3层路由选择。

1.2.4 互连

数据中心的服务器可以彼此互连，也可以连接到存储器，还可以通过交换机和路由器连接到外部网络。实现这些连接需要使用铜缆或光缆。从历史的角度来看，铜缆布线一直是一种成本较低的解决方案，但是当需要较高的带宽和（或）较长的布线距离时，就需要使用光缆。例如，机架内服务器与交换机之间的连接就是布线距离较短的情况，可以使用铜缆；而机架对外的上行链路就需要使用高带宽的光缆，以便传输更长的距离。在本章的后续内容中，还将给出更多有关电缆类型的资料。

1.3 什么是云数据中心

在早期的万维网（还记得这个词吗）中，传输到家庭电脑的数据很有可能来自于某企业级数据中心的一间堆满服务器的机房。此后，互联网的规模开始呈爆炸式增长。访问网络的人数、可用网站的数量以及平均数据下载量都呈指数级增长。像谷歌和亚马逊这样广受欢迎的Web服务公司需要迅速扩大自己的数据中心来满足需求。很快，他们就意识到需要建立大型的专用服务器仓库。这就是今天人们所说的云数据中心。

“云”这个词的出现和无线手持设备开始在市场上流行的时间大体一致。通过无线手持设备访问网络时，你仿佛在从云里“拉出”数据。于是，提供这些信息的数据中心被很自然地称为云数据中心。现在，似乎每个人都在追赶着“云”的潮流，各种各样的云公司、云产品和云服务正在不断地涌入市场。

云数据中心正在被迅速地部署到世界各地。由于这些部署地必须支持多达数十万台服务器，因此数据中心的效率和运营成本成为了关键。也正因为如此，一些云数据中心会建立在靠近廉价电力来源的地区（如水电大坝），或者建立在气候相对寒冷的地区以降低冷却成本。有些公司（例如微软公司）采用箱式结构建立模块化数据中心，也就是自包含的服务器存储和网络模块，大小接近于集装箱。这些模块经过运输、组合、接电、冷却和组网等流程后成为一体。其他数据中心则使用服务器机架作为基本的构建砌块。在整个数据中心里，这样的机架一排接着一排。不管采用何种结构，组网都是这些大型云数据中心网络的重要组成部分。

最近，思科公司发布了一本名为《2012～2017年思科全球云指数：预测和方法》[①]的白皮

① *Cisco Global Cloud Index: Forecast and Methodology, 2012–2017*。

书，对云数据中心给出了一些引人关注的见解。他们预测，从现在到2017年，全球IP数据中心流量将每年至少增长25%。他们还预测，到2017年，数据中心的流量有超过三分之二发生在“云”里，而且这些流量的约76%是在云数据中心内的设备之间发生的。这与当前数据中心里数据需要不断进进出出的情形截然不同。他们还预测，服务器虚拟化（一台物理服务器上运行多个虚拟服务器）将会对云数据中心网络产生巨大影响。他们用服务器工作负载总数除以物理服务器总数，将得到的比值作为评价指标，并且预测：到2017年，这个比值将达到16以上。相比较而言，当前传统数据中心所对应的比值大约是2～3。换句话说，服务器虚拟化（将在本书稍后讨论）将一直是云数据中心的主要特征。所有这些因素都会对如何设计和运行大型云数据中心以及如何实现云数据中心网络产生影响。

1.4 什么是云网络

由于云数据中心使用大量的机架式服务器或箱式数据中心模块，因此将所有这些组件进行组网连接就成了一种挑战。云数据中心管理员希望缩减投资和运行费用，包括网卡、交换机、路由器和电缆等方面的支出。以太网已经成为支持这些大型数据中心的低成本第2层网络，但是这些网络有特殊的要求，不同于传统的企业局域网和企业数据中心网络。在整本书中，我们将这种类型的网络称作“云数据中心网络”，并且将阐述其与传统企业网络之间的关键差异。

1.5 云网络的特征

大多数云数据中心对于在性能最大化的同时实现成本最小化都有特殊需求，这些需求在其网络设计中有所反映。云数据中心通常采用以太网设备建网，这样就可以充分利用以太网的规模效益，同时还可以提供高带宽和数据中心的功能定制。在本节中，我们将讲述这些趋势的背景知识，包括有关以太网布线技术的内容，以及对网络虚拟化、网络融合和可扩展性需求等方面的概述。

1.5.1 以太网的使用

20世纪90年代中期，当我开始从事交换结构芯片设计的时候，人们认为以太网是一种局域网技术；对于重要的电信应用来说，这是一种不可靠的传输机制，会在严重拥塞的环境下丢包。然而，由于广泛的部署使用和大批量制造，以太网始终是成本最低的网络技术。从那时到现在，以太网已经走过了漫长的道路，并且在过去的10年中，对规范加入了很多功能和改进。今天，以太网真正实现了无处不在，既可以用于网络设备电路板之间的互连，又可以用于长距离运营商网络的连接。