> 华为ICT认证系列丛书

HUAWEI

SDN原理解析

——转控分离的SDN架构

闫长江　吴东君　熊怡　著

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

> 华为ICT认证系列丛书

HUAWEI

SDN原理解析

——转控分离的SDN架构

闫长江　吴东君　熊怡　著

人 民 邮 电 出 版 社
北　京

图书在版编目（CIP）数据

SDN原理解析 : 转控分离的SDN架构 / 闫长江, 吴东君, 熊怡著. -- 北京 : 人民邮电出版社, 2016.4
ISBN 978-7-115-40724-5

Ⅰ. ①S… Ⅱ. ①闫… ②吴… ③熊… Ⅲ. ①计算机网络－网络结构 Ⅳ. ①TP393.02

中国版本图书馆CIP数据核字(2015)第262537号

内 容 提 要

SDN 是对现有传统分布式控制网络架构的一次重构。通过 SDN 把网络软件化，可以提升网络可编程能力，并大大简化现有通信网络并提高网络业务创新速度。本书重点介绍了 SDN 的定义、价值和 SDN 架构的基本原理；同时，还介绍了 SDN 中最重要的系统——SDN 控制器实现架构和原理；随后，介绍了 SDN 面临的各种挑战和可能的应对措施；最后，介绍了如何从现有网络架构演进到 SDN 架构。

◆ 著　　闫长江　吴东君　熊　怡
责任编辑　李　静　乔永真
责任印制　彭志环
◆ 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
北京艺辉印刷有限公司印刷
◆ 开本：787×1092　1/16
印张：15　　2016 年 4 月第 1 版
字数：215 千字　　2016 年 4 月北京第 1 次印刷

定价：59.00 元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

序

SDN 作为一种新的网络架构，它实现网络的软件化，试图对传统分布式网络架构进行重构，由传统分布式网络转向集中控制的 SDN 网络，从而给运营商带来巨大价值，包括简化网络、提升网络可编程能力、支持业务快速创新、设备白牌化、业务自动化等。SDN 的理念一经提出，立即受到了产业界的关注。网络架构重构既意味着机会也意味着风险，各个厂家都做出了积极应对。华为的应对策略是积极支持 SDN 网络变革，争取成为 SDN 网络时代的领导者，并积极投入 SDN 控制器的研究、设计开发和试验中。

但是从 SDN 出现至今，争论就没有中断过，一千个人眼里有一千种 SDN 解释和认识。很多时候，大家在谈论一个概念，相互之间却不知道对方心里的概念到底指什么，以致出现各种误解。无论是设备商还是运营商，都有各自的认识。我作为华为 SDN 控制器架构师，在负责华为 SDN 控制器的架构设计过程中，接触到了各种 SDN 信息，既有来自客户的也有来自友商的，既有来自华为高层的也有来自内部相关部门的；再加上华为 Fellow 吴东君的指导，这些都使我对 SDN 有了一个自认为相对全面的认识，我眼中也有了一个自己的“哈姆雷特”。今天我把对 SDN 的各种认识总结并分享出来，与广大读者交流，希望读者能从中有所收获。

为了方便理解，本书有些章节以华为的实现作为示例，这些示例是当时的快照，并不代表华为正式产品的架构和方案。关于华为控制器产品架构、支持的功能和解决方案等，可参考华为官方正式的产品手册、配置指南、命令参考等资料。本书并不代表任何华为产品路标和承诺，只是华为 SDN 架构团队工作过程中的总结和认识，也包含了个人的观点，错漏不可避免，请广大读者谅解，并欢迎批评指正。

本书介绍的各种思想和观点主要来自华为 SDN 架构师核心团队的吴东君、闫长江和熊怡，当然也有来自华为参与 SDN 架构设计的小组，包括白涛、倪辉、李艳民、黄铁英、曾玮等同事。我们在一起设计，一起讨论，对 SDN 逐渐形成了基本概念。为此，感谢华为 SDN 设计团队，他们拥有丰富的设计经验，在整个架构和特性设计过程中提出了大量的细节问题，使得我们架构师团队能够更加准确地把握细节。非常感谢一起走过来的华为的大师们，包括涂伯颜、陈杰、杨宏杰、郭锋、施震宇、饶远、吕鑫等人。同时更加感谢潘梅芳、郝冠普，他们不辞辛苦对本书进行评审、校对和插图整理。没有他们的辛苦付出，这本书不知道是否能够完成。还要感谢李怿敏、陈双龙、阴元斌、倪辉、黄铁英、董林等人的评审。

最后，感谢我的夫人刘远碧、女儿闫叶赫和闫娜拉，她们这些年一直在背后默默地支持我，从来没有抱怨过我没有时间陪她们，使我能全身心地投入我热爱的工作中，并能够利用各种节假日完成这本书的写作。

前　言

一次，我驾车回家，途中使用了导航软件导航。导航软件有个功能叫作“避开拥堵”，由于高速拥堵，它建议我提前从某个出口下高速，但等我到了那个出口前 2 公里处时就发现车道开始塞车，我估计一定有不少人使用了导航相同的功能，从而导致这里出口也出现拥挤。于是，我决定继续沿着高速前行，选择其他出口。出人意料的是，我发现前面的拥堵其实已经解除了。我在想，如果导航软件能够选择性地通告一些人走某些出口，让另外一些人走其他出口，进行流量分担，是不是会更好一些呢？如此一来，导航软件对交通网络的状态更新的实时性将变得非常重要。假定导航软件把拥堵信息通告给驾驶员，由驾驶员自行决定如何绕开拥堵时，我认为大部分驾驶员会选择最短的不拥堵路径行进，而这种决策很可能会导致新的拥堵。如果导航软件能够实时了解交通状态，并对每个驾驶员规划出不同的路径，总体上使得所有人都绕开拥堵，而不会产生新的拥堵，这是一个最为理想的结果。这个导航软件实际上对所有的车辆实施了集中控制，使得整体交通状态变得更加通畅。如果只是通告交通状态信息给驾驶员，由驾驶员自行决定行程，这种分布式路径决策，极有可能导致新的拥堵，而这恰恰是我们最不期望的结果。

通信网络和交通网络是一样的，几乎面临相同的问题。作为现代通信网络的基础——IP 分组交换网络而言，其基本原理就是全分布式路径决策。这种分布式控制网络的选路过程和上面的交通案例中的驾驶员自主根据道路拥堵状态进行选择绕路是一样的原理，其结果会产生新的拥堵。网络设备可以通过一些路由协议收集网络拓扑，然后选择最短路径转发报文，每个网络设备都自主选路并遵从该原则，这就是分布式自主选路过程。这种网络架构具有极高的网络生存能力，使得网络本身能够全自治完成业务，但是这种网络的可控性就受到极大限制。一方面，在商用通信网络中，网络设备供应商不只是一家，经常是多厂家设备共同组网，就要求这些不同厂家的设备需要协议互通，而这个互通要求导致网络支持一些新业务的时间非常漫长，通常数年之久。另一方面，这种分布式自主网络，在部署新业务时，需要对网络设备进行软件升级。然而，这些设备数量众多，同时还承载着业务，如何不间断业务升级如此众多的设备，对网络运维人员而言是一个巨大的挑战。再者，在这样分布式控制的网络中要部署业务，需要逐个操作网络设备，并需要保持这些网络设备配置的一致性，以便它们能协同完成网络业务。这个过程也经常会出现问题，部署业务人员必须学习大量的分布式控制协议的技术细节，才能很好地进行业务部署和运维，这种把内部技术细节暴露出来的做法增加了网络运维难度，提升了对维护人员的技能要求，从而增加了运维成本。

而 SDN（Software Defined Network，软件定义网络）则是试图重构传统分布式控制的

网络，实现集中控制的网络。通过集中控制把网络进行软件化，这样可以更好支持网络业务自动化和自治，并简化了网络的复杂度，向运维人员屏蔽了网络技术细节，降低对网络运维人员的要求，降低了运维成本。同时 SDN 可以支持业务快速创新，增加网络的盈利能力。SDN 控制器对网络的集中控制和交通拥堵时导航软件根据交通道路状态为不同的汽车安排不同的路径是一样的道理。通过在网络中部署一个集中控制的 SDN 控制器，当我们需要调整网络的行为，不再需要去修改网络设备本身，而是只要调整 SDN 控制器内部的软件就可以了。由于采用了集中控制，原来很多的网络分布式控制协议就不再需要了，网络得到极大的简化，进一步降低了对人员技能要求，提高了网络可运维能力，总体降低了运维成本。由于集中的 SDN 控制器可以提供网络端到端的业务，并提供这些业务的完全自治能力，这样使得网络的业务自动发放能力得到加强，能更快地部署网络业务。这种做法使得网络运维者可以很简单地部署业务，甚至可以把这些业务接口直接开放给他们的客户，从而实现无人工干预的业务销售能力。SDN 这样的能力就满足现在互联网公司对网络快速部署业务的需求，提升运营商的业务创新发展速度。

SDN 概念一出来，立即引起了业界的广泛关注，从运营商、网络设备供应商、技术研究人员、初创公司都积极参与讨论 SDN 的价值、需求、实现、标准定义等问题。原因很简单，SDN 是对网络架构的重构，网络架构的重构通常意味着产业链的重构，也意味着产业分工的重构。那么，在这次 SDN 重构中，哪些公司能够最后胜出，哪些公司最后被大潮抛弃，都取决于每个公司是否找对了自己的位置，在产业链占据自己的一席之地。运营商积极响应，探讨需求和价值，积极推动部署，进行 SDN 实验。传统供应商在传统网络有领先优势，很难放弃目前的利益。所以这些具有优势的供应商是比较矛盾的。而一些初创公司则积极投入 SDN 的研发中，试图在 SDN 时代成为新的主流供应商。还有其他相关企业也积极投入 SDN 的探讨中，包括芯片供应商、OSS 供应商等。于是有的供应商积极拥抱，有的供应商则半推半就，有的供应商则一边想办法拖延一边积极做好各种准备。

由于有众多的企业参与 SDN 的讨论，于是在 SDN 概念认识上也就产生了各种思路。

1．传统派，认为 SDN 就是把原来的 OSS 做的更加实时，支持网络业务自动发放就是一种 SDN。这种思路主要解决了目前网络普遍存在的业务自治能力差的问题。通过这种 SDN，可以实现网络业务快速部署，提供网络对客户需求的响应速度，满足现在各种互联网对网络的快速业务部署需求。

2．演进派，SDN 需要把网络功能集中控制，比如可以把网络内部的交换路径进行集中控制，也可以把网络的边缘接入业务进行集中控制，或者把两者都集中控制。通过这种灵活的集中控制，使得这种 SDN 可以很好的解决从现有的传统分布式网络向 SDN 演进的问题。而凡是被 SDN 集中控制的功能，都可以简化网络，降低运维成本，加速业务创新的进程。

3．创新派，SDN 需要支持完全转控分离，集中控制，实现 OpenFlow 技术，把设备完全白牌化，设备控制面全部在 SDN 控制器。这种思路是一种彻底的革命，能够构建全

新的网络架构，实现所有 SDN 的价值。ONF 标准组织推动这种技术的进步和成熟，初创公司也积极支持这种思路。创新派面临的主要问题是如何能够兼容当前已经广泛部署的分布式基础网络问题。

上述各种 SDN 思路，其实也反映了产业界对 SDN 网络的各种期望和诉求。创新派可能是 SDN 网络的终极目标，为了达到终极目标，要走的路可能会很长。因为目前现网已经有海量的传统分布式网络存在，如何把海量现网逐渐迁移到 SDN 网络，必然经过一个较长演进阶段，逐渐对网络进行一些集中控制，逐渐对网络进行改造，这样也就是上面说的演进派思路。而在为了能快速解决网络现在的业务自动化能力不足的问题，传统派的思路就能够快速的满足这种诉求。所以这些 SDN 思潮不是对立矛盾的，而是 SDN 的不同发展阶段的必然产物，他们是对立统一的。本书主要目标向读者介绍演进派和创新派的 SDN 基本原理和概念，较少介绍传统派思路是因为这个阶段相对业内人员已经比较熟悉了。

本书作者作为华为 SDN 设计研究人员，把自己对 SDN 的理解分享出来，主要向读者阐述 SDN 的基本原理，什么是 SDN，SDN 给网络带来什么变化，SDN 能够产生什么价值等，希望通过对 SDN 的解析，让读者了解 SDN 的精髓，把握 SDN 的发展趋势。

本书主要内容

第 1 章，SDN 概述

本章介绍了 SDN 的产生历史，以及相比传统分布式控制网络，SDN 网络带来了哪些价值。

第 2 章，SDN 网络的工作原理

本章介绍了 SDN 网络的基本工作原理，并用实例方式介绍 SDN 网络如何实现其价值的。

第 3 章，SDN 控制器实现原理

本章给出了 SDN 网络中的核心部件 SDN 控制器的需求、基本实现框架架构和范例。

第 4 章，SDN 网络的可靠性

本章给出了 SDN 网络面临的可靠性挑战，分析了 SDN 网络的故障模式，并给出了可能的解决建议。

第 5 章，SDN 网络收敛问题

本章介绍了 SDN 网络收敛和传统网络收敛时间对比，并给出了可能提升 SDN 网络收敛时间的建议方案。

第 6 章，SDN 的开放性

本章介绍了 SDN 网络的南北向开放的定义以及如何通过开放南向接口来保证控制器兼容多厂家转发器，介绍 SDN 应该开放哪些北向接口，给出应用程序如何使用这些接口的建议。

第 7 章，SDN 网络的安全性

本章介绍 SDN 网络面临的主要安全威胁，以及如何应对这些网络安全威胁，确保 SDN 网络的安全可用。

第 8 章，从现网演进到 SDN 网络

本章以多个实例详细介绍了现有网向 SDN 网络演进的途径，通过这些演进技术，可以解决现有网络如何平滑迁移到 SDN 网络的问题。

第 9 章，SDN 控制器实现架构实例分析

本章介绍了华为 SDN 控制器、ONOS 开源控制器和 ODL 开源控制器的实现架构分析。

关于本书读者

本书适合有一定的 IP 网络知识背景的学生、院校研究人员、企业人员阅读使用。如果是行业内人士，比如电信运营人员、企业网络管理维护人员、电信设备供应商、网络研究人员，他们对现在的网络理解深刻，如果希望进一步了解 SDN 的相关概念和知识，本书则可以较好地满足他们的需求。

关于本书图标

SDN 控制器，负责网络的实时控制。

传统路由器，通常运行了分布式控制协议。

传统路由器，通常运行了分布式控制协议。

传统路由器，通常运行了分布式控制协议。

转发器，是和控制器配套运行的数据转发设备，通常不运行分布式控制业务路由协议，其业务控制面通常运行在控制器内。

转发器，是和控制器配套运行的数据转发设备，通常不运行分布式控制业务路由协议，其业务控制面通常运行在控制器内。

通用服务器或者虚拟机。

目　录

第1章 SDN概述

1.1 传统网络

1.1.1 传统网络架构采用分布式控制

IP 通信技术已经成为今天通信网络的核心技术。今天的通信网络，从庞大的全球互联网到大小不一的企业网、私有网络，全部都是基于 IP 构建的。这些 IP 网络中承载着各种各样的业务，包括数据业务、视频业务、传统的语音业务，人们在互联网上进行购物、社交、娱乐、金融等相关的活动。

IP 技术之所以能够成为通信网络的核心技术，首先是因为其具有简单性。通过全球统一的 IP 地址编址，任何两台主机就可以进行通信，而通信的主机之间不用关心对方的具体位置，也不用关心对方具体的网络细节，这种简单性使得构建全球范围的大规模互联网成为可能。IP 技术的另外一个重要基因是采用分布式控制架构。

传统的 IP 网络的自治系统内的基本通信模型如图 1-1 所示。

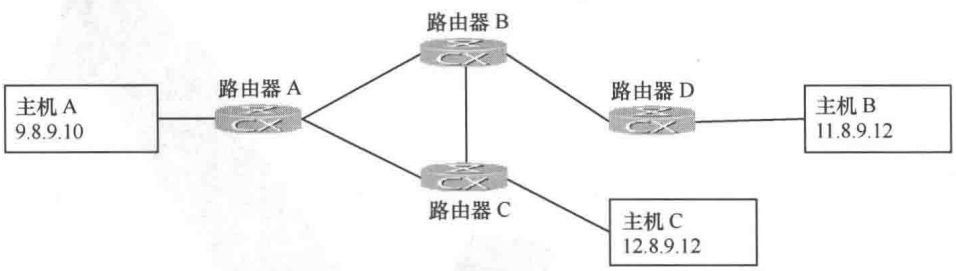

图 1-1 传统 IP 网络的基本通信模型

在同一个自治系统内，当主机 A 希望和主机 B 发起通信时，首先主机 A 需要知道主机 B 的 IP 地址为 11.8.9.12，主机 A 会把发给 B 的 IP 报文先发送给网关路由器 A，接下来网关路由器 A 必须决定到底这个 IP 报文是发送给路由器 B 还是路由器 C。路由器 A 在做决策时要把报文发送给路由器 B 或者 C 所依赖的数据，称作路由表。路由器 A 会根据 IP 报文里面携带的目的地址 11.8.9.12，在路由表中查找最长匹配的路由表项，在这个表项中会获得一个下一跳路由器和出接口的信息：

```
IP prefix=11.8.9.0/24, Nexthop = 路由器 B, OutgoingInterface=AB
```

路由器 A 根据这些信息知道应该把目的地址为 11.8.9.12 的报文转发给路由器 B。

同样道理，路由器 B 必须决定是把这个 IP 报文发送给路由器 C 还是路由器 D。这里，路由器 B 会根据其本地的路由表获得如下的转发信息：

```
IP prefix=11.8.9.0/24, Nexthop = 路由器 D, OutgoingInterface=BD
```

路由器 B 根据这个信息，会把该报文转发给路由器 D。当路由器 D 收到报文时，也一样会查询本地的路由表，获得的路由表信息会发现这台主机和路由器 D 是直连网络，所以直接投递该报文到路由器和主机 B 连接的接口，这样最终这个报文被逐跳转发后，递交给了目的主机 B。

上述过程是 IP 报文逐跳转发的基本原理。在这个过程中，每个路由器始终依赖其本地的路由表数据来进行寻路，决定到底应该把报文发送给哪个下一跳路由器。这个路由表数据是如何生成的呢？一个简单的想法可能是网络管理人员自己去配置，比如给路由器 A 配置静态路由：

```
IP route 11.8.9.0/24 nexthop 路由器 B interface AB
```

给路由器 B 配置静态路由：

```
IP route 11.8.9.0/24 nexthop 路由器 D interface BD
```

通过这些配置，每台路由器都会在本地生成路由表。当目的地址为 11.8.9.12 的报文进入路由器时，这些路由器就能够根据这些数据进行报文转发。可是，上面的做法对于一个小的网络是可以工作的，但当网络规模很大，路由数量可能达到几十万的时候，这种手工静态配置的方法就不能工作了。另外，还有一个原因也使得这种静态配置方式无法工作，那就是当网络的某些链路发生故障，比如路由器 A 和路由器 B 的连接接口 AB 中断时，报文就不能再正确转发到目的地了，从而将导致通信中断。直到人工介入再次给路由器 A 和路由器 C 配置静态路由，才能恢复主机 A 和主机 B 的通信。

为了解决此类问题，引入了动态路由协议方式来学习这些路由信息，而不是通过手工静态配置。现在常见的域内路由协议（IGP）主要是 OSPF 和 ISIS 协议，这些协议能够学习完整的网络拓扑，然后根据拓扑计算出任何两点之间的最短路径，并自动生成路由信息。通过这种 IGP 的自动学习，生成路由器转发所需的路由表的方法，解决了上面的两个问题，从而不用人工给每个网段配置静态路由。在网络拓扑变化时，也不用人工干预，路由协议会重新计算出一条新的转发路由。当上面说到的接口 AB 中断时，路由协议会在 1s 时间内自动重新学习网络拓扑并计算出可用路由：路由器 A 会计算出走 AC 口，把报文送交给路由器 C；而路由器 C 会计算出需要把这个报文送交给 B；路由器 B 则会计算出把报文送交给 D。如此会使得主机 A 和 B 的通信可以在故障后 1s 内恢复。这些 IGP 需要在每台路由器上运行，并且这些路由器之间会通过 IGP 路由协议交互拓扑信息，每台路由器的 IGP 通过交互都拿到同样的全部网络拓扑数据，然后每台路由器的 IGP 分别独立计算出转发报文所需的路由表数据。这个过程是完全分布式计算的，没有集中点，网络中任何路由器出现故障，其他路由器都会重新计算路由，保持网络的最大通信连接能力。这种在路由计算和拓扑变化后全分布式地重新进行路由计算的过程，称为分布式控制过程 。传统的网络被认为是全分布式控制的。

为了能够大规模组网，IP 网络架构的设计者对网络进行了区域划分，每个区域是一个自治系统，自治系统内部运行 IGP 来完成路由计算，域间则采用另外一种路由协议来传递和扩散路由信息，其基本组网架构模型如图 1-2 所示。

在图 1-2 中，主机 A 希望和主机 B 通信，主机 A 发送报文给主机 B，当这个报文进入到路由器 A 时，路由器 A 必须做出决定，是把报文送给路由器 B 还是路由器 C。此时，这个主机 B 的网段路由不能通过 IGP 学习，因为该主机 B 不在自治系统 118 网络内，而是在另外一个自治系统 1098 网络内。自治系统 118 网络内的路由器如何能够学习到网段 11.8.9.0/24 的路由呢？为了解决域间路由学习问题，IETF 标准工作组定义了域间路由协议——BGP。通过 BGP，把这种不在同一个域内的路由前缀进行扩散，以便所有的网