

# 第 1 章 云数据中心网络演进 1

- 1.1 传统的 3-Tier 架构 1
- 1.2 设备“多虚一”——虚拟机框 2
  - 1.2.1 Cisco VSS 2
  - 1.2.2 Juniper VC 与 H3C IRF 4
- 1.3 高级 STP 欺骗——跨设备链路聚合 4
  - 1.3.1 Cisco vPC 4
  - 1.3.2 Juniper MC-LAG 和 Arista M-LAG 6
- 1.4 变革 3-Tier——向 Leaf-Spine 演进 6
- 1.5 初识大二层 9
- 1.6 插叙——虚拟机的接入 10
  - 1.6.1 VEB 10
  - 1.6.2 Cisco VN-TAG 11
  - 1.6.3 VEPA 12
  - 1.6.4 VEB 性能优化 13
- 1.7 消除 STP——Underlay L2MP 14
  - 1.7.1 TRILL 15
  - 1.7.2 SPB 17
- 1.8 Cisco 私有的大二层——FabricPath 19
  - 1.8.1 整体设计 19
  - 1.8.2 控制与转发过程分析 21
  - 1.8.3 其他技术细节 25
- 1.9 Juniper 私有的大二层——QFabric 25
  - 1.9.1 整体设计 26
  - 1.9.2 集中式的控制机制 29
  - 1.9.3 控制与转发过程分析 30
- 1.10 Brocade 私有的大二层——VCS 32
  - 1.10.1 整体设计 33
  - 1.10.2 控制与转发过程分析 33
  - 1.10.3 其他技术细节 35
- 1.11 跨越数据中心的二层——DCI 优化 36
  - 1.11.1 Cisco OTV 36
  - 1.11.2 HUAWEI EVN 与 H3C EVI 38
- 1.12 端到端的二层——NVo3 的崛起 39
  - 1.12.1 VxLAN 39
  - 1.12.2 NvGRE 41
  - 1.12.3 STT 42
  - 1.12.4 Geneve 43
- 1.13 新时代的开启——SDN 入场 45
- 1.14 Overlay 最新技术——EVPN 46
  - 1.14.1 传统网络对 SDN 的反击 46
  - 1.14.2 组网与数据模型 47
  - 1.14.3 控制信令的设计 48

- 1.15 Underlay 最新技术——Segment Routing 55
  - 1.15.1 SID 与 Label 56
  - 1.15.2 控制与转发机制 57
  - 1.15.3 SDN 2.0? 60
- 1.16 本章小结 62

## 第 2 章 杂谈 SDN 63

- 2.1 SDN 与传统网络——新概念下的老问题 63
- 2.2 转控分离——白盒的曙光 66
  - 2.2.1 芯片级开放 68
  - 2.2.2 操作系统级开放 71
  - 2.2.3 应用级开放 75
  - 2.2.4 机箱级开放 76
  - 2.2.5 白盒的“通”与“痛” 77
- 2.3 网络可编程——百家争鸣 78
  - 2.3.1 芯片可编程 78
  - 2.3.2 FIB 可编程 80
  - 2.3.3 RIB 可编程 83
  - 2.3.4 设备配置可编程 85
  - 2.3.5 设备 OS 和控制器可编程 88
  - 2.3.6 业务可编程 88
- 2.4 集中式控制——与分布式的哲学之争 89
  - 2.4.1 在功能上找到平衡点 90
  - 2.4.2 在扩展性和可用性上找到平衡点 91
- 2.5 回归软件本源——从 N 到 D 再到 S 94
  - 2.5.1 模块管理 94
  - 2.5.2 模块间通信 95
  - 2.5.3 接口协议适配 96
  - 2.5.4 数据库 97
  - 2.5.5 集群与分布式 98
  - 2.5.6 容器与微服务 99
- 2.6 本章小结 100

## 第 3 章 SDDCN 概述 101

- 3.1 需求 101
  - 3.1.1 自动化与集中式控制 101
  - 3.1.2 应用感知 103
- 3.2 整体架构 105
  - 3.2.1 实现形态 105
  - 3.2.2 功能设计 107
- 3.3 关键技术 107
  - 3.3.1 网络边缘 107

3.3.2	网络传输	110
3.3.3	服务链	112
3.3.4	可视化	115
3.3.5	安全	117
3.3.6	高可用	120
3.4	本章小结	122

## **第 4 章 商用 SDDCN 解决方案 123**

4.1	VMware NSX	123
4.1.1	从 NVP 到 NSX	124
4.1.2	NVP 控制平面设计	125
4.1.3	NVP 数据平面设计	125
4.1.4	NVP 转发过程分析	126
4.1.5	NSX-V 整体架构	128
4.1.6	NSX-V 管理平面设计	129
4.1.7	NSX-V 控制平面设计	130
4.1.8	NSX-V 数据平面设计	132
4.1.9	NSX-V 转发过程分析	132
4.1.10	NSX-MH 与 NSX-T	139
4.2	Cisco ACI	140
4.2.1	整体架构	141
4.2.2	管理与控制平面设计	142
4.2.3	数据平面设计	145
4.2.4	转发过程分析	152
4.2.5	议 ACI 与 SDN	154
4.3	Cisco VTS	155
4.3.1	整体架构	156
4.3.2	管理与控制平面设计	158
4.3.3	数据平面设计	159
4.4	Juniper Contrail	162
4.4.1	整体架构	164
4.4.2	管理与控制平面设计	167
4.4.3	数据平面设计	173
4.4.4	转发过程分析	175
4.5	Nuage VCS	176
4.5.1	整体架构	178
4.5.2	管理平面设计	179
4.5.3	控制平面设计	179
4.5.4	数据平面设计	180
4.6	Arista EOS 与 CloudVison	181
4.6.1	整体架构	183
4.6.2	管理与控制平面设计	185
4.6.3	数据平面设计	187

4.7	HUAWEI AC-DCN	187
4.7.1	整体架构	187
4.7.2	管理平面设计	189
4.7.3	控制平面设计	189
4.7.4	数据平面设计	193
4.8	Bigswitch BCF 与 BMF	194
4.8.1	整体架构	195
4.8.2	BCF 控制平面设计	196
4.8.3	BMF 控制平面设计	201
4.8.4	数据平面设计	205
4.9	Midokura Midonet	207
4.9.1	整体架构	207
4.9.2	控制平面设计	210
4.9.3	数据平面设计	213
4.10	PLUMgrid ONS	217
4.10.1	整体架构	217
4.10.2	数据平面设计	219
4.10.3	控制平面设计	221
4.10.4	转发过程分析	222
4.11	Plexxi Switch 与 Control	225
4.11.1	整体架构	225
4.11.2	数据平面设计	227
4.11.3	控制平面设计	229
4.12	Pluribus	230
4.12.1	Server Switch 设计	231
4.12.2	Netvisor 设计	232
4.12.3	再议数据中心 SDN	235
4.13	本章小结	236

## **第 5 章 开源 SDDCN: OpenStack Neutron 的设计与实现 237**

5.1	网络基础	237
5.1.1	网络结构与网络类型	238
5.1.2	VLAN 网络类型中流量的处理	239
5.2	软件架构	242
5.2.1	分布式组件	242
5.2.2	Core Plugin 与 Service Plugin	243
5.3	WSGI 与 RPC 的实现	245
5.3.1	Neutron Server 的 WSGI	245
5.3.2	Neutron Plugin 与 Neutron Agent 间的 RPC	247
5.4	虚拟机启动过程中网络的相关实现	248
5.4.1	虚拟机的启动流程	248
5.4.2	Nova 请求 Port 资源	250
5.4.3	Neutron 生成 Port 资源	250

5.4.4	Neutron 将 Port 相关信息通知给 DHCP Agent	252
5.4.5	DHCP Agent 将 Port 相关信息通知给 DHCP Server	252
5.4.6	Nova 拉起虚拟机并通过相应的 Port 接入网络	252
5.5	OVS Agent 的实现	253
5.5.1	网桥的初始化	253
5.5.2	使能 RPC	255
5.6	OVS Agent 对 Overlay L2 的处理	256
5.6.1	标准转发机制	256
5.6.2	arp_responder	258
5.6.3	l2_population	260
5.7	OVS Agent 对 Overlay L3 的处理	261
5.7.1	标准转发机制	261
5.7.2	DVR 对东西向流量的处理	262
5.7.3	DVR 对南北向流量的处理	267
5.8	Security-Group 与 FWaaS	268
5.8.1	Neutron-Security-Group	268
5.8.2	FWaaS v1	269
5.8.3	FWaaS v2	269
5.9	LBaaS	270
5.9.1	LBaaS v1	270
5.9.2	LBaaS v2	271
5.9.3	Octavia	271
5.10	TaaS	272
5.11	SFC	274
5.12	L2-Gateway	275
5.13	Dynamic Routing	277
5.14	VPNaaS	279
5.15	Networking-BGPVPN 与 BagPipe	280
5.15.1	Networking-BGPVPN	280
5.15.2	BagPipe	280
5.16	DragonFlow	282
5.17	OVN	287
5.18	本章小结	290

## **第 6 章 开源 SDDCN: OpenDaylight 相关项目的设计与实现 291**

6.1	架构分析	291
6.1.1	AD-SAL 架构	292
6.1.2	MD-SAL 架构	293
6.1.3	YANG 和 YANG-Tools	294
6.1.4	MD-SAL 的内部设计	294
6.1.5	MD-SAL 的集群机制	296
6.1.6	其他	298
6.2	OpenFlow 的示例实现	298

6.2.1	OF 交换机的上线	299
6.2.2	l2switch 获得 PacketIn	301
6.2.3	l2switch 下发 PacketOut 和 FlowMod	302
6.3	OpenStack Networking-ODL	303
6.3.1	v1	303
6.3.2	v2	304
6.4	Neutron-Northbound 的实现	306
6.4.1	对接 Networking-ODL	306
6.4.2	RESTful 请求的处理示例	306
6.5	Netvirt 简介	307
6.5.1	OVSDB-Netvirt 和 VPNService 的合并	307
6.5.2	Genius	309
6.6	Netvirt-OVSDB-Neutron 的实现	311
6.6.1	net-virt 分支	311
6.6.2	net-virt-providers 分支	317
6.7	Netvirt-VPNService 的实现	321
6.7.1	elanmanager	323
6.7.2	vpnmanager	326
6.8	SFC 的实现	328
6.8.1	sfc-openflow-renderer 分支	328
6.8.2	sfc-scf-openflow 分支	335
6.9	VTN Manager 的实现	336
6.9.1	neutron 分支	337
6.9.2	implementation 分支	339
6.10	本章小结	342

## **第 7 章 开源 SDDCN: ONOS 相关项目的设计与实现 343**

7.1	架构分析	343
7.1.1	分层架构	344
7.1.2	分层架构的实现	345
7.1.3	模块的开发	347
7.1.4	分层架构存在的问题	347
7.1.5	数据存储与集群	348
7.1.6	其他	349
7.2	OpenFlow 的示例实现	349
7.2.1	OF 交换机的上线	350
7.2.2	fwd 获得 PacketIn	352
7.2.3	fwd 下发 PacketOut 和 FlowMod	356
7.3	ONOSFW 的实现	359
7.3.1	vtnmgr 分支	359
7.3.2	sfcmgr 分支	363
7.4	SONA 的实现	365
7.4.1	openstacknode 分支	366

7.4.2	openstacknetworking 分支	368
7.5	CORD 简介	371
7.5.1	R-CORD 的架构	372
7.5.2	R-CORD 的控制与转发机制	373
7.6	本章小结	376

## **第 8 章 学术界相关研究 377**

8.1	拓扑	377
8.1.1	FatTree	377
8.1.2	VL2	379
8.1.3	DCell	380
8.1.4	FiConn	382
8.1.5	BCube	384
8.1.6	MDCube	385
8.1.7	CamCube	387
8.2	路由	388
8.2.1	Seattle	388
8.2.2	FatTree	391
8.2.3	VL2	393
8.2.4	PortLand	396
8.2.5	SecondNet	400
8.2.6	SiBF	401
8.2.7	SPAIN	402
8.2.8	WCMP	404
8.2.9	OF-based DLB	406
8.2.10	Flowlet 与 CONGA	406
8.2.11	Hedera	408
8.2.12	DevoFlow	409
8.2.13	MicroTE	409
8.2.14	Mahout	410
8.2.15	F10	410
8.2.16	DDC	411
8.2.17	SlickFlow	412
8.2.18	COXCast	413
8.2.19	Avalanche	415
8.3	虚拟化	416
8.3.1	NetLord	416
8.3.2	FlowN	418
8.3.3	FlowVisor	420
8.3.4	ADVisor	421
8.3.5	VeRTIGO	423
8.3.6	OpenVirteX	424
8.3.7	CoVisor	426

8.4	服务链	427
8.4.1	pSwitch	427
8.4.2	FlowTags	428
8.4.3	Simple	430
8.4.4	StEERING	432
8.4.5	OpenSCaaS	434
8.4.6	SPFRI	435
8.5	服务质量	437
8.5.1	NetShare	437
8.5.2	Seawall	438
8.5.3	GateKeeper	439
8.5.4	ElasticSwitch	440
8.5.5	SecondNet	441
8.5.6	Oktopus	441
8.6	传输层优化	443
8.6.1	MPTCP	443
8.6.2	DCTCP	446
8.6.3	D3	447
8.6.4	pFabric	449
8.6.5	Fastpass	450
8.6.6	OpenTCP	451
8.6.7	vCC	452
8.7	测量与分析	453
8.7.1	Pingmesh	453
8.7.2	OpenNetMon	454
8.7.3	FlowSense	455
8.7.4	Dream	455
8.7.5	OpenSample	457
8.7.6	Planck	458
8.7.7	OpenSketch	458
8.8	安全	460
8.8.1	SOM	460
8.8.2	FloodGuard	462
8.8.3	TopoGuard	463
8.8.4	FortNox	464
8.8.5	AVANT GUARD	466
8.8.6	OF-RHM	468
8.8.7	Fresco	470
8.9	高可用	471
8.9.1	ElastiCon	471
8.9.2	Ravana	473
8.9.3	BFD for OpenFlow	474
8.9.4	In-Band Control Recovery	476
8.9.5	OF-based SLB	477

- 8.9.6 Anata 478
- 8.9.7 Duet 480
- 8.10 大数据优化 482
  - 8.10.1 BASS 482
  - 8.10.2 OFScheduler 482
  - 8.10.3 Phurti 483
  - 8.10.4 Application-Aware Networking 484
  - 8.10.5 CoFlow 485
- 8.11 本章小结 486

## **第 9 章 番外——容器网络 487**

- 9.1 容器网络概述 487
- 9.2 容器网络模型 488
  - 9.2.1 接入方式 488
  - 9.2.2 跨主机通信 491
  - 9.2.3 通用数据模型 492
- 9.3 Docker 网络 494
  - 9.3.1 docker0 495
  - 9.3.2 pipework 496
  - 9.3.3 libnetwork 496
- 9.4 Kubernetes 网络 498
  - 9.4.1 基于 POD 的组网模型 498
  - 9.4.2 Service VIP 机制 499
- 9.5 第三方组网方案 501
  - 9.5.1 Flannel 501
  - 9.5.2 Weave 502
  - 9.5.3 Calico 504
  - 9.5.4 Romana 506
  - 9.5.5 Contiv 507
- 9.6 Neutron 网络与容器的对接 508
- 9.7 本章小结 510

## **第 10 章 番外——异构网络与融合 511**

- 10.1 融合以太网基础 511
  - 10.1.1 PFC 512
  - 10.1.2 ETS 513
  - 10.1.3 QCN 513
  - 10.1.4 DCBX 514
- 10.2 存储网络及其融合 514
  - 10.2.1 FC 的协议栈 515
  - 10.2.2 FC 的控制与转发机制 516
  - 10.2.3 FCoE 的控制与转发机制 517

10.2.4	昙花一现的 SDSAN	520
10.3	高性能计算网络及其融合	524
10.3.1	InfiniBand 的协议栈	525
10.3.2	InfiniBand 的控制与转发机制	526
10.3.3	RoCE 与 RoCEv2	528
10.4	本章小结	530