



教育部—神州数码网络教学改革合作项目成果教材
神州数码网络认证教材

DCN



神州数码
Digital China

创建高级路由型互联网

全国职业技能大赛推荐参考书

神州数码网络认证指定教材

校企合作新课改教材

程庆梅 主编



机械工业出版社
CHINA MACHINE PRESS



赠电子教学资源

教育部—神州数码网络教学改革合作项目成果教材

神州数码网络认证教材

创建高级路由型互联网

主 编 程庆梅

参 编 徐雪鹏 杜婉琛 岳大安 赵 飞



机械工业出版社

本书首先回顾了路由技术基础知识,深入浅出地探讨了实际项目中常见路由选择协议 RIP、OSPF 以及 BGP 协议的原理和应用,以及路由优化技术中的路由再发布、路由过滤和策略路由等项目应用,并对其他路由选择协议和 MPLS 进行了概念上的介绍,使读者对路由技术的认识既有重点又较为全面。通过本书的学习,读者可以达到神州数码 DCNP 认证中对路由技术的水平要求。

本书读者对象为:本科、职业院校开设计算机网络技术专业相关课程的师生;参加各省市及全国职业技能大赛的师生;考取神州数码网络认证的人士(与工业和信息化部、人力资源和社会保障部联合认证事宜请参考神州数码网络有限公司网站 www.dcnetworks.com.cn);从事大中型局域网工程技术工作的人士以及热爱计算机网络技术的各界人士。

本书配套丰富的电子教学资源,包括网络课程、电子教案、演示文稿等,方便教师教学,需要者可联系责任编辑免费索取(010-88379934),或联系神州数码网络有限公司相关业务人员索取。

图书在版编目(CIP)数据

创建高级路由型互联网/程庆梅主编. —北京:机械工业出版社, 2009.10

教育部—神州数码网络教学改革合作项目成果教材

神州数码网络认证教材

ISBN 978-7-111-28679-0

I. 创... II. 程... III. 计算机网络—路由选择—教材 IV. TN915.05

中国版本图书馆 CIP 数据核字(2009)第 196841 号

机械工业出版社(北京市百万庄大街 22 号 邮政编码 100037)

策划编辑:孔熹峻

责任编辑:蔡 岩

封面设计:鞠 杨

责任印制:洪汉军

三河市国英印务有限公司印刷

2010 年 1 月第 1 版第 1 次印刷

184mm×260mm·9.75 印张·2 插页·237 千字

0001—3000 册

标准书号:ISBN 978-7-111-28679-0

定价:22.00 元

凡购本书,如有缺页、倒页、脱页,由本社发行部调换

电话服务

网络服务

社服务中心:(010) 88361066

门户网:<http://www.cmpbook.com>

销售一部:(010) 68326294

教材网:<http://www.cmpedu.com>

销售二部:(010) 88379649

读者服务部:(010) 68993821

封面无防伪标均为盗版

前言

本书根据神州数码多年积累的项目实际应用编写,以信息产业人才需求为基本依据,以提高学生的职业能力和职业素养为宗旨,坚持以能力为本位的课程设计原则,明确计算机网络人才的培养定位。

“创建高级路由型互联网”是一门理论与实践紧密结合的实操类课程。在机械工业出版社联合国内知名计算机网络厂商“神州数码”于2009年开发之前,本书的主体内容已在实践教学中得到了很好的使用。为了更好地融合网络最新技术与职业教育最新理念,我们重新编写了本书,以便适应教学改革的需要,更好地为一线教学服务。

本书编写体现了以下特点:

- 以就业为导向,以企业需求为依据。全书按照实际应用特点编写,全部案例来自企业一线,使学生在学校就可以接受岗位训练。
- 更加体现以综合职业素质为基础,以能力为本位。书中内容以由浅入深的方式循序渐进地展开复杂的理论阐述。
- 书中涉及的计算机网络技术先进,设备型号领先,全部设备均为目前国内主流设备且为神州数码主打设备。
- 本书可以作为神州数码网络有限公司 DCNP-IRT 网上认证考试配套教材(相关事宜可登录公司网站 www.dcnetworks.com.cn 查询,与工业和信息化部、人力资源和社会保障部联合认证事宜也可在此网站查询),为学生获得企业资格认证和将来就业创造有利条件。
- 本书配有丰富的电子教学资源,包括网络课程、电子教案、演示文稿等,方便教师教学,需要者可联系责任编辑免费索取(010-88379934),或联系神州数码网络有限公司相关业务人员索取。

本书的编写力求达到以下目标:

- 体现“授之以渔”的方法论理念。考虑到网络工程项目与传统行业项目的技术含量差异,本书在编写过程中增加了关键原理和常识的讨论与介绍,力求在了解和掌握配置方法的同时“知其然,更知其所以然”。
- 集中、专业的交换技术讲解。与以往的网络教材不同,本书集中讨论路由技术,而没有将路由器中可以实现的其他技术一股脑放在教材中加以讨论,重点突出,目标鲜明。
- 配套详尽细致的实训手册。实训手册中设计的实训内容,不仅包含实现目标的具体步骤,更进一步讨论为什么要这样做,以及常见的问题解决方法思路,并且在必要的时候增加 debug 信息的解释,对网络协议过程作了尽可能充分的讲解。

如何使用本书:

1. 关于图标

本书图标采用神州数码图标库标准图标,除真实设备外,所有逻辑示意均使用如下图标。



2. 关于课程组织

- 本书为学生用书,可在课堂上用于理解教师授课使用。
- 本书配套的实训手册,作为学生完成理论学习后实验过程的指导用书。
- 建议理论与实践结合,本书适合在实训室完成教学过程。

3. 关于配套设备

本书配套实训设备的购置可参考神州数码网站,或联系神州数码网络有限公司业务接口人进行办理。

4. 关于小贴士

本书使用 4 种贴士,如下所示:

-  **小学堂:** 一般为非正式的知识讲解。
-  **提醒:** 实际项目中容易出现问题和理解偏差的过程。
-  **小常识:** 非正式的常识,需要了解但不是重点。
-  **提问:** 帮助深入思考的,通常结合上下文可自行解决,认证考试中包含。

本书由程庆梅主编,参与编写的主要人员有:神州数码网络大学教材编委会成员——

徐雪鹏、杜婉琛、岳大安、赵飞、郭薇、黄天生、张向东、李亚峰、王吉忠、王永才、朱建英、吕凯。

同时，在编写和审校的过程中，也得到了许多来自合作伙伴院校一线老师的意见和建议。他们是北京市供销学校赵鹏老师、北京市金驼技术学校葛久平老师、天津中华职专阴海涛老师、贵州电子信息职业技术学院曹炯清老师、福建信息职业技术学院李宏达和詹可强老师、漳州职业技术学院章忠宪和郑东升老师、广州市番禺区工贸职业技术学校赵宏胜老师等。

在此特向参与编审的各界人士提供的大力支持表示衷心的感谢！

本书的主体内容已经在实践中使用多次，但由于编者水平所限，疏漏之处在所难免，敬请广大读者批评指正。编者邮箱：dcnu_2007@163.com。

编 者

目 录

4.4 策略路由与路由映射.....	86
总结与习题.....	90
第 5 章 BGP 协议.....	91
5.1 概述.....	92
5.2 何时使用 BGP.....	95
5.3 BGP 的术语和属性.....	96
5.4 BGP 同步概念.....	115
5.5 BGP 路由选择进程.....	116
5.6 BGP 联盟与反射.....	119
5.7 BGP 基本配置.....	121
总结与习题.....	123
第 6 章 其他路由协议介绍.....	124
6.1 IGRP 协议介绍.....	124
6.2 EIGRP 路由协议介绍.....	125
6.3 集成 IS-IS 路由协议介绍.....	129
第 7 章 MPLS 技术基础.....	133
7.1 MPLS 介绍.....	133
7.2 MPLS 包头.....	133
7.3 标签交换路由器.....	134
7.4 标签交换路径.....	134
7.5 转发等价类.....	135
7.6 标签分发协议 LDP.....	135
7.7 MPLS VPN.....	137
7.8 MPLS VPN 访问公网.....	143

第1章 路由选择原理



内容提要

- 能够阅读路由表
- 掌握路由表条目的优先级比较顺序
- 掌握 Route-cache 的作用及其影响
- 能够灵活应用静态路由实现网络互通
- 了解距离矢量路由协议的工作原理
- 了解链路状态路由协议的工作原理

1.1 路由概述

1.1.1 什么是路由

Routing 是一种动作,在计算机网络中,一个逻辑网段到另一个逻辑网段的数据需要经过路由才可以到达。

“路由”,是指把数据从一个地方传送到另一个地方的行为和动作。而路由器,正是执行这种行为动作的机器,它的英文名称为 Router,即:选择路径的人。

在 TCP/IP 网络中,当子网中的一台主机发送 IP 数据包给同一子网的另一台主机时,它将直接把 IP 数据包送到网络上,对方就能收到。而要送给不同子网上的主机时,它要选择一个能到达目的子网的路由器,把 IP 数据包送给该路由器,由路由器负责把 IP 数据包送到目的地。如果没有找到这样的路由器,主机就把 IP 数据包送给一个称为“默认网关(default gateway)”的路由器上。“默认网关”是每台主机上的一个配置参数,它是接在同一个网络上的某个路由器端口的 IP 地址。



值得注意的是:为路由器添加默认路由和主机的默认网关是不同的概念。通常在设备上可以添加 8 个 0 的路由,从而能够为到达设备的所有数据寻址发送,在这种情况下,没有在路由表中寻找到明确匹配项的数据将遵从这 8 个 0 的路由发送出去。在某些设备中(如二层交换机),可以使用“default-gateway”来为设备添加默认网关,这种做法多见于没有开启 ip routing 的设备如主机和二层交换机中。在路由器中如果开启了 ip routing,则可以使用 8 个 0 来设置默认路由。

路由器转发 IP 数据包时, 只根据目的 IP 地址的网络号部分, 选择合适的端口, 把 IP 数据包送出去。同主机一样, 路由器也要判定端口所接的是是否是目的子网, 如果是, 就直接把数据通过端口送到网络上, 否则, 就要选择下一个路由器来传送数据。路由器也有它的默认路由/默认网关, 用来传送不知道往哪里送的 IP 数据包。这样, 通过路由器把知道如何传送的 IP 数据包正确转发出去, 把不知道的 IP 数据包送给默认路由器, 这样一级级地传送, IP 数据包最终将送到目的地, 送不到目的地的 IP 数据包则被网络丢弃。注意: ICMP 协议可以协助数据发送端得知这一情况。

路由器中的路由表就是它进行选路的依据所在。

1.1.2 路由的两个基本阶段

将数据从一个网络转发给另一个网络需要经过两个阶段的转换, 它们是: 路由和交换。其中“路由”这个过程是三层的内容, 而“交换”过程是二层的内容。

(1) 路由过程 (见图 1-1) 主要完成的工作是

- ① 去掉收到帧的头, 得到一个 IP 数据包。
- ② 读取其目的 IP 地址。

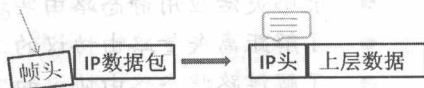


图 1-1 路由的过程

(2) Switching 过程的主要工作是

- ① 如图 1-2 所示, 查询路由表信息、与之前得到的目的 IP 地址比较, 得到下一跳端口或下一跳站点地址, 如图 1-3 所示。
- ② 重新进行二层的帧头封装, 如图 1-3 所示。
- ③ 转发。

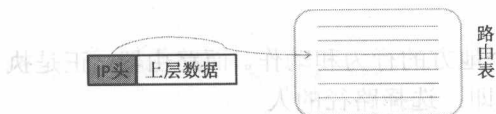


图 1-2 查询路由表

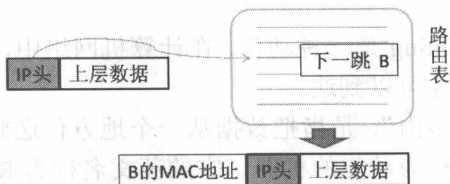


图 1-3 路由重打包的过程

1.1.3 路由的分类

1. 根据路由表生成的方式来划分, 可以分为如下两种

(1) 静态路由

由系统管理员事先设置好固定的路由表称之为静态 (static) 路由表, 一般是在系统安装时就根据网络的配置情况预先设定的, 它不会随网络结构的改变而改变。其优点是几乎不消耗路由器的资源。缺点是不随着网络拓扑结构的改变而改变。

(2) 动态路由

动态路由是网络中的路由器之间相互通信, 传递路由信息, 利用收到的路由信息来更新路由表的过程。它能实时地适应网络结构的变化。如果路由更新报文中表明发生了网络变化, 路由协议就会重新计算路由, 并发出新的路由更新报文。这些报文通过各个网络,

引起各路由器重新启动其路由算法,并更新各自的路由表以动态地反映网络拓扑变化。动态路由适用于网络规模大、网络拓扑复杂的网络。当然,各种动态路由协议会不同程度地占用网络带宽和 CPU 资源,如果网络规划不当,一些低端路由器则根本无法承受大量的动态路由更新的信息。常见的动态路由协议有 RIP、OSPF、BGP 等。

动态 (Dynamic) 路由表是路由器根据网络系统的运行情况而自动调整的路由表。路由器根据路由协议 (Routing Protocol) 提供的功能,自动学习和记忆网络运行情况,在需要时自动计算数据传输的最佳路径。

2. 内部网关协议和外部网关协议

(1) AS—Autonomy System

所谓自治系统就是处于一个管理机构控制之下的路由器和网络群组。一个 AS 内的所有路由器必须相互连接,且运行相同的路由协议,同时分配一个 AS 编号。

(2) IGP

应用于同一个自治系统内部的动态路由协议,常见的有 RIP、OSPF 等。

(3) EGP

应用于多个自治系统之间的动态路由协议,现在常用的有 BGP 等。

3. 其他分类

(1) 静态与动态

静态路由算法很难算得上是算法,只不过是开始路由前由网管建立的表映射。这些映射自身并不改变,除非网管去改动。使用静态路由的算法较容易设计,在网络通信可预测及简单的网络中工作得很好。

由于静态路由系统不能对网络改变做出反映,因此通常被认为不适用于现在的大型、易变的网络。20 世纪 90 年代主要的路由算法都是动态路由算法,通过分析收到的路由更新信息来适应网络环境的改变。如果信息表示网络发生了变化,则路由软件需重新计算路由并发出新的路由更新信息。这些信息渗入网络,促使路由器重新计算并对路由表做相应的改变。

静态路由和动态路由有各自的特点和适用范围,因此在网络中动态路由通常作为静态路由的补充。通常,路由器首先查找静态路由,如果查到则根据相应的静态路由转发数据;否则再查找动态路由。

(2) 单路径与多路径

一些复杂的路由协议支持到同一目的的多条路径。与单路径算法不同,这些多路径算法允许数据在多条线路上复用。多路径算法的优点很明显,即可以提供更好的吞吐量和可靠性。

(3) 平面化与分层

一些路由协议在平面化的空间里运作,其他的则有路由的层次。在平面化的路由系统中,每台路由器与其他所有路由器是对等的;在分层次的路由系统中,一些路由器构成了路由主干,数据从非主干路由器流向主干路由器,然后在主干上传输直到它们到达目标所在区域,在这里,它们从最后的主干路由器通过一个或多个非主干路由器到达终点。

路由系统通常设计有逻辑节点组,称为域、自治系统或区间。在分层的系统中,一些路由器可以与其他域中的路由器通信,其他的则只能与域内的路由器通信。在很大的网络中,可能还存在其他级别,最高级的路由器构成了路由主干。

分层路由的主要优点是它模拟了多数公司的结构，从而能很好地支持其通信。多数的网络通信发生在小组中（域）。因为域内路由器只需要知道本域内的其他路由器，它们的路由算法可以简化，根据所使用的路由算法，路由更新在逻辑节点组之间的通信量可以相应地减少。

（4）主机智能与路由器智能

一些路由算法假定源结点来决定整个路径，这通常称为源路由。在源路由系统中，路由器只作为存贮转发设备，无意识地把数据发向下一跳。其他路由算法假定主机对路径一无所知，在这些算法中，路由器基于自己的计算决定通过网络的路径。前一种系统中，主机具有决定路由的智能，后者则为路由器具有此能力。

主机智能和路由器智能的折衷实际是最佳路由与额外开销的平衡。主机智能系统通常能选择更佳的路径，因为它们发送数据前探索了所有可能的路径，然后基于特定系统对“优化”的定义来选择最佳路径。然而确定所有路径的行为通常需要很多的探索通信量和很长的时间。

（5）域内与域间

一些路由算法只在域内工作，其他的则既在域内也在域间工作。这两种算法的本质是不同的，优化的域内路由算法没有必要也成为优化的域间路由算法。

（6）链路状态与距离向量

链路状态算法（也叫做最短路径优先算法）把链路的状态信息散布到网络的每个节点，也就是说每台路由器只发送路由表中描述其自己链路状态的部分。距离向量算法（也叫做 Bellman-Ford 算法）中每台路由器发送路由表的全部或部分，但只发给其邻居。也就是说，链路状态算法到处发送较少的链路状态，而距离向量算法则向相邻的路由器发送较多的路由表内容。

链路状态算法收敛较快，而且它们相对于距离算法产生路由环的倾向较小。但另一方面，链路状态算法需要更多的 CPU 和内存资源，因此链路状态算法的实现和支持较昂贵。虽然有差异，但并不能笼统地判断哪种协议就是绝对地优于另一种，事实上，针对不同的网络环境，它们都可以工作得非常出色。



所谓收敛一般针对动态路由环境，当网络结构发生变化时，路由器在路由协议的帮助下完成对变化后的拓扑的重新计算和统一的过程就是收敛。

1.2 路由查询

1.2.1 什么是路由表

路由的两个过程中，更加关键的是路由表查询过程，它为数据包的转发指明了方向，它是路由器赖以工作的前提。

路由器的主要工作就是为经过路由器的每个数据包寻找一条最佳传输路径，并将该数

据有效地传送到目的站点。由此可见,选择最佳路径的策略即路由算法是路由器的关键所在。为了完成这项工作,在路由器中保存着各种传输路径的相关数据——路由表(Routing Table),供路由选择时使用。

路由表就像我们平时使用的地图一样,标识着各种路线,路由表中保存着子网的标志信息、网上路由器的个数和下一个路由器的名字等内容。路由表可以由系统管理员固定设置好(静态路由),也可以由系统动态修改(直连路由),可以由路由器自动调整(路由协议),也可以由主机控制。

整个路由表,分成两个部分:

1. Codes 部分

这部分是对路由表中条目类型的说明,描述了各种路由表条目的类型的缩写,其中

- 1) C 表示连接路由,路由器的某个接口设置/连接了某个网段之后,就会自动生成。
- 2) S 静态路由,系统管理员通过手工设置之后生成。
- 3) R RIP 协议协商生成的路由。
- 4) B BGP 协议协商生成的路由。
- 5) BC BGP 的连接路由。
- 6) D BEIGRP 协议生成的路由,兼容 CISCO 的 EIGRP。
- 7) DEX BEIGRP 的外部路由。

8) DHCP 当路由器的某个端口设置为由 DHCP 分配地址时,系统在收到“默认网关”属性之后自动生成的路由,实际上是一条默认路由。

对于 OSPF 协议生成的路由,又有如下的路由表项:

- 1) OIA OSPF 的区域之间路由。
- 2) ON1 OSPF NSSA 路由(类型 1)。
- 3) ON2 OSPF NSSA 路由(类型 2)。
- 4) OE1 OSPF 外部注入路由(类型 1)。
- 5) OE2 OSPF 外部注入路由(类型 2)。

以上这些 codec 信息对于路由表的工作不产生任何影响,但对于管理维护人员的阅读却提供了便利。

2. 路由表的实体

对实体中的每一行,我们都可以发现从左到右有如下几个内容:路由的类型(Codec 表示)、目的网段(网络地址)、优先级(由[AD(Administrative Distance),度量值(Metric)]组成)、下一跳 IP 地址(next-hops)等。

(1) 目标网段 就是网络号,它描述了一类 IP 包(目的地址)的集合。

(2) 优先级描述 由管理距离 AD 和度量值 Metric 组成,通常值越小,优先级越高,更可信。



管理距离 AD 表明了路由学习方法(静态,各种动态协议)的优先级(后面还会有详细介绍);度量值则表示不同的下一跳代表的路径的优先级。

(3) 下一跳网关 被匹配的数据包从哪个端口被转发，有本地端口名字和下一跳 IP 可选。



出端口的二层协议将决定数据包的二次封装类型。

路由器在查询时，就是通过对数据包的目的 IP 地址比较表中的目的网段，如果可以匹配，则得出一个下一跳 (next-hops) 作为出口，以便进行转发；如果有多个条目可以匹配，则根据优先级信息比较并选出一个最优的条目。

如图 1-4 所示的列表中第一行内容：S 代表这是一条由管理员手工添加的静态路由；0.0.0.0/0 代表所有目的网络，表明去往所有网络的数据包都按照后面的方式发送给下一跳 222.66.38.211。

```
Codes: C - connected, S - static, R - RIP, B - BGP, BC - BGP connected
        D - BEIGRP, DEX - external BEIGRP, O - OSPF, OIA - OSPF inter area
        ON1 - OSPF NSSA external type 1, ON2 - OSPF NSSA external type 2
        OE1 - OSPF external type 1, OE2 - OSPF external type 2
        DHCP - DHCP type

VRF ID: 0

S      0.0.0.0/0          [1,0] via 222.66.38.211(on FastEthernet0/0)
C      192.168.0.0/16     is directly connected, FastEthernet0/2
C      222.66.38.208/28   is directly connected, FastEthernet0/0
```

图 1-4 路由表实体

1.2.2 如何查询匹配路由表

路由器查询路由表不仅是找到一个下一跳，更重要的是选出一个最优的下一跳。我们来看一下它是如何选择的。

(1) 掩码最长匹配

路由器收到一个数据包，在查询路由表时，首先是查询目标网段，如果有多个条目同时匹配时，则掩码长者优先，如图 1-5 所示。如果路由器收到一个目的地址为 10.10.10.5 的数据包，它将被转发到哪个端口？

```
S      10.10.0.0/16[0]     is directly connected, VLAN2[0]
S      10.10.10.0/24[0]    is directly connected, VLAN1[0]
C      172.16.19.0/24[0]   is directly connected, VLAN2[0]
C      192.168.0.0/16[2]   is directly connected, VLAN1[2]
```

图 1-5 路由表中的掩码

两个路由项分别是 10.10.0.0/16 和 10.10.10.0/24；显然两条 S 静态路由都可以匹配上述目的地址。

但因为第一跳的掩码长度为 16，而第二跳的掩码长度为 24，所以后者更加优先。所以目的地址为 10.10.10.5 的数据包在转发时，下一跳地址为 VLAN1 这个接口，而不是 VLAN2。可以看出，路由表优先级与其在表中的位置无关。



默认路由（其目的网段为 0.0.0.0/0）是一种特殊的静态路由，虽然 0.0.0.0 是一个可以匹配任何 IP 的网段（任何 IP 地址与它“与运算”之后还是 0.0.0.0），但掩码长度为 0，决定了它的优先级永远比别的路由要低。

以上的路由表来自一台具备三层转发功能的交换机，其出端口以 VLAN1/2 来表示。但实际情况中，多数会使用下一跳路由的 IP 地址。

（2）管理距离

管理距离，是指一种路由协议的可信度，也可以说是协议优先级。该值在 1~255 之间取值，值越小，优先级/可信度越高。

每一种路由协议都有自己的默认管理距离，据此，不同的路由协议可以按可靠性从高到低排除一个优先级。该值可以根据需要进行调整。

对于到达同一目的地的多条不同协议学习到的路由项，路由器会首先根据管理距离决定相信哪一个协议。

路由协议	管理距离
C 直连路由	0
S 静态路由	1
R 路由信息协议 RIP	120
O 域内 OSPF	110
B BGP	20

上述表格中列举了 DCR 路由器中常见路由协议的默认管理距离值。



注意：并不是所有厂家的默认值都是完全一致的，这些默认值都是可以修改的，使用时不要一概而论。

（3）Metric

Metric，路由度量值，也是表示路由项优先级、可信度的重要参数之一，如图 1-6 所示。

不同的路由协议的 AD 值不同，而同种类型的路由项 AD 值相等。如果同种路由协议生成了多个路由表项，目的网段相同，且 AD 值也相同，则可以根据度量值这个信息来判断它们之间的优先级别。

不同路由协议计算度量值的方法是不同的。

（1）RIP

RIP 协议是以跳数作度量值的，两个相邻的路由设备和之间的链路为一跳。跳数越多，优先级越低。

RIP 以跳数为衡量： Σ (Hops)

上虚线：metric=2，下虚线：metric=1



图 1-6 度量值与路径选择

很显然，距离矢量路由协议会选择度量值小的下虚线所示路径。

(2) OSPF

OSPF 的 metric 值计算方法比较复杂，有一个公式 $\sum(10^8)/BW$ ，就是把每一段网络的带宽倒数累加起来乘以 100000000（即 100M）。简单说，OSPF 的度量值与带宽相关。

OSPF 以带宽因子为衡量： $\sum(100M)/BW$

上虚线：metric=10+1=11，下虚线：metric=1562

此时，OSPF 根据其度量值选择了带宽较大的路径传输数据。

(3) BGP

BGP 是一种 AS 之间的外部网关路由协议，它的 Metric 值通常是从内部网关协议中继承的。

BGP 的 metric 值从 IGP 中继承。



注意：

- 1) 只有动态路由才会有 metric 参数，静态路由是没有的。
- 2) 不同类型路由协议的 metric 值没有可比性。

1.3 路由转发

1.3.1 Load-Balance（负载均衡）

如果说路由表里面存在多个路由项，它们的目的网段相同、AD 值相同、metric 值也相同，显然它们的优先级完全相等，那么路由器在转发相关数据时该怎样处理呢？

路由器为了公平起见，只能将用户数据“平分”处理，也就是这里我们要介绍的负载均衡机制。要注意的是：负载均衡并不是说把传输的数据流量等分成 N 份，然后平均放在 N 条线路上传输。

首先需要清楚的是，网络上的数据流在传输时是以 IP 数据包为基本单位的，同时，数据包的大小也是不尽相同的，小到 64bytes，大到 1518bytes。即便每条线路上的 IP 数据包数量相等，最后计算流量的时候也是不同的。

另外，负载均衡还分为两种方式：每数据包和每目标，这两种方式在进行流量均衡时，对于数据包的流量会产生非常大的影响，所以我们要重点介绍一下。

首先需要了解路由器的基本工作原理：

路由器在转发 IP 数据包时，为了提高转发效率，尤其是同类型 IP 数据包的速率，引入了路由缓冲（ip route-cache）的机制。

一个 IP 包经过路由器的 CPU 处理（查询路由表并进行转发）之后，会在入口的缓冲中生成一条 cache，如果后续收到同类型的数据包需要转发，那么 CPU 只要直接查 cache，不需要检查路由表，这样效率就提高了。

同时，由于 cache 缓冲的容量总是有限的，不可能无限地存储，所以给每个 cache 条目增加一个 timeout 时间是非常必要和实际的做法。这样，长期不用的 cache 就会自动被删除。如图 1-7 所示。



图 1-7 ip-cache 示意

图 1-8 所示就是一个 ip route-cache 的实例，在路由器中使用命令 Router#show ip cache 即可得到。

Current ip flow cache in used : 2				
Max available ip flow caches: 12288				
Source	Destination	DstIf	Next Hop	Ref
192.168.19.180	222.66.38.211	f0/0	222.66.38.211	0
222.66.38.211	192.168.19.180	f0/2	192.168.19.180	0

图 1-8 ip-cache 表项

可以看到，其中包括了源 IP 地址、目的 IP 地址、出口及其对应的下一跳地址等。

查 ip cache 的方法和查路由表的方法是有差别的，前者必须匹配源和目的地址，而后者只需匹配目的 IP 地址即可。

默认情况下，出于提高转发效率的目的，route-cache 总是打开的。

在图 1-8 中，192.168.19.180 这个 IP 地址去 ping 222.66.38.211，第一次 ping 的数据包会在 cache 表中留下 2 个条目（包括 echo request 和 reply 两个方向的 IP 报文）。当第二个 ping/IP（跟第一个的 IP 头相同）包过来时 timeout 还未超时（默认为 80s），所以 CPU 直接查询 cache 进行转发，显然其出口还是 F0/0。

依此类推，如果有一系列的数据包，它们具有相同的源 IP 地址和相同的目的 IP 地址，且相互之间的时间间隔都不超过 cache 的 timeout 时间，则它们在转发时的匹配是一致的。这样的数据包序列我们称之为“流”。



注意：当一个“流”的第一个数据包经过 CPU 查询路由表时，如果有多条路由可选，那么路由器选中的出接口是随机的。该流后续的数据包转发时，是根据 cache 中的目标接口，而不是路由表。即便是路由表中的信息存在多个可选出口也不会有影响。

如果 cache 关闭，那么所有的数据包在转发时，都没有 cache 可以参考，它们都要经过 CPU 去查询路由表。那么此时就是以 IP 数据包为单位轮询，在多条线路上进行转发。

关闭 route cache 的命令为在接口模式中 no ip route-cache；全局中 Ip route-cache-timeout *** 可以调整缓冲默认时间；该值建议不要设置得太小，否则在多线路环境中容易导致 TCP 应用频繁的拆链建链。

综上所述，每目标的方式也叫基于流的负载均衡方式；而每数据包的方式则是基于包的负载均衡方式。



思考

路由器某接口上每隔 8s 收到一个数据包，分为红色和黑色两种，如图 1-9 所示。路