第37卷 第16期
Vol.37 No.16

计 算 机 工 程
Computer Engineering

2011年8月
August 2011

·网络与通信·

文章编号：1000—3428(2011)16—0104—04 文献标识码：A 中图分类号：TP393

二层以太网中最优组播树的构建方法

张梦竹[1]，马红兵[2]

(1. 东南大学计算机科学与工程学院，南京 210096；2. 南京贝伦思网络科技有限公司，南京 210017)

摘 要：针对二层以太网难以构建最优组播树且收敛速度慢的问题，提出一种新的最优组播树构建方法。该方法将IS-IS协议用于二层以太网络最短路径树的计算，通过扩展链路状态包的CLV改进其通告机制，以便构建最优组播树，且网络拓扑发生变化时可快速重构此最优组播树。测试结果表明，该方法可将组播树的收敛时间从现有方法的10 s~20 s缩短到50 ms，并能确保该组播树为最优组播树。

关键词：组播树；多生成树协议；IGMP侦听；IS-IS路由协议

Construction Method of Optimal Multicast Tree in Layer 2 Ethernet

ZHANG Meng-zhu[1], MA Hong-bing[2]

(1. School of Computer Science and Engineering, Southeast University, Nanjing 210096, China;
2. Nanjing Balance Network Technology Co., Ltd., Nanjing 210017, China)

【Abstract】Aiming at the problems that it is difficult to build an optimal multicast tree and its convergence is slowly in layer 2 Ethernet, a new optimal multicast tree construction method is proposed in this paper. This method put IS-IS protocol to compute the shortest path tree for layer 2 Ethernet. By means of the Code-Length-Value(CLV) of expanded link-state bag to improve its notice mechanism, an optimal multicast tree can be easily construct. The optimal multicast tree can be reconstructed fast when network topology changed. Test results show that the method can make the time of multicast tree convergence be reduced from 10 s~20 s to 50 ms and ensure that the tree is an optimal multicast tree.

【Key words】multicast tree; Multiple Spanning Tree Protocol(MSTP); Internet Group Management Protocol(IGMP) snooping; IS-IS routing protocol

DOI: 10.3969/j.issn.1000-3428.2011.16.035

1 概述

开展IPTV、视频会议等组播业务时，对于网络而言，需要解决以下3个关键技术[1-2]：

(1)构建一棵无环组播树，从组播业务系统沿着组播树的各节点，将组播业务流逐点复制到各用户，以防止组播业务流在网络中构成环路。

(2)确保组播树是最优树，即组播树的路径最短、链路带宽最大。

(3)组播树可快速收敛与重构，即在网络拓扑发生变化或用户对组播业务的需求状态发生变化时，可快速生成一棵新的组播树，使组播业务的中断时间最短。

在由路由器构建的三层IP网络中，通过IS-IS三层路由协议并结合PIM(Protocol Independent Multicast)组播协议的方法能较好解决以上3个技术问题[1-3]。但在由以太交换机或EPON等设备构建的二层以太网中，目前通常用的MSTP生成树协议结合IGMP(Internet Group Management Protocol)侦听协议的方法[4-5]，可解决如何构建一棵无环组播树和向最优组播树的逼近问题[6]，而组播树的最优化、快速收敛及网络拓扑变化时的快速重构问题还没有有效的解决方案。为解决前文提到的3个关键技术问题，本文提出一种新的最优组播树构建方法。

2 现有的技术方案及存在的问题

在二层以太网中构建组播树时，通常是通过多生成树协议(Multiple Spanning Tree Protocol, MSTP)和IGMP侦听协议的组合和交互来实现，具体实现方法如下：

(1)各互联端口上配置一个专用的虚拟局域网(Virtual Local Area Network, VLAN)，构成一个广播域，用于承载组播业务，且该VLAN称为Multicast VLAN，简称MVLAN。

(2)各节点部署MSTP生成树协议将MVLAN广播域破环，为MVLAN构建一棵无环的生成树，即阻断MVLAN的部分端口的广播或组播的转发行为，使MVLAN不再存在环路。

(3)各节点部署IGMP侦听协议，在MVLAN生成树的基础上构建组播树。

用上述方法构建组播树，对于组播业务存在以下问题：

(1)当网络故障时，组播树重构时间长，将导致业务中断时间也长，通常要中断15 s~20 s。

(2)为最大程度地防止网路环路，MSTP协议将所有端口先设置为阻断状态，只有在检测到某节点或某链路故障并确认本端口置为放开状态并且不构成环路后，才把本端口重置为放开状态。此方法生成树的构建速度比较缓慢，如当网络节点数量越多时，需要阻断与打开的端口数量越多，则生成树的构建时间就越长。一般情况而言，需要10 s以上才能构建或重构好生成树[7]。对重构后的生成树，某端口由阻断状态若切换为非阻断状态，则会在此端口上广播IGMP Query报文到各用户，并侦听来自用户的IGMP Report报文，以完成组播树的重构，这个过程也较长，通常需要1 s~3 s。上述

作者简介：张梦竹(1990－)，女，本科生，主研方向：路由计算；马红兵，高级工程师

收稿日期：2011-02-24 **E-mail：**213082891@seu.edu.cn

2 个因素将导致组播树的重构时间和业务中断时间长达 15 s~20 s。

(3)多生成树协议由于各端口的阻断和放开状态无法完全同步，生成树是在某些端口被设置为阻断状态下计算出来的，因此无法保证生成树最优，即无法保证从根节点到叶子节点的路径最短、链路带宽最大。故基于生成树所构建的组播树自然也无法保证是最优的。

3 二层以太网最优组播树的实现方法

3.1 IS-IS 路由协议简介

IS-IS 是一个无连接网络协议(Connectionless Network Protocol, CLNP)而设计的动态路由协议，已在 IETF RFC1195 中标准化并广泛应用于 IP 网络的路由计算[8]。在 IS-IS 协议中，任何一个中间系统(网络节点)用一个取值唯一的 System ID 标识，并通过直接封装于数据链路层帧结构中的协议数据包(Protocol Data Unit, PDU)进行信息交互。PDU 包括：PDU 通用报文头，PDU 专用报文头，CLV(Code-Length-Value)这 3 个部分。PDU 的通用报文头用于说明报文类型，专用报文头用于节点间相互通告一些网络信息或协商协议参数。对于每一种 PDU 报文有相应的 CLV，并且该报文可用于进一步描述网络链路相关信息。在 ISO10589 和 RFC1195 中定义了相关的 CLV。通过定义新的 CLV，进一步扩展所需要的网络信息。CLV 是一个包括编码(Code)、CLV 报文长度(Length)、CLV 的取值(Value)的三元组。

PDU 包括 Hello、链路状态数据报(Link State Packets, LSP)、完整的序列号包(Complete Sequence Number Packets, CSNP)、部分序列号包(Partial Sequence Number Packets, PSNP)等类型的报文。通过节点间定期发送 Hello 报文，可发现、建立和维系邻接关系。在建立邻接关系后，通过链路状态数据报，相互交换链路状态、链路带宽等网络链路信息。CSNP 用于发布完整的链路数据信息，实现链路状态数据库(Link State Data Base, LSDB)的定时同步。PSNP 用于确认和请求链路数据信息。

用 IS-IS 计算网络路由时，主要包括链路状态数据库(LSDB)的生成和最短路径树(SPT)的计算，具体如下：

(1)链路状态数据库的生成

每个节点在检测到其与邻接节点的状态(Up/down)、相关接口状态(up/down)、接口带宽的 metric 值发生变化时，将产生新的 LSP 通告给其邻居。当接收到来自邻居的 LSP 时，会和 LSDB 中已存在的 LSP 进行比较，如果是新的 LSP，则将其刷新到 LSDB 中，并通过 PSNP 报文确认此 LSP，并继续扩散到其他邻居，最终使所有网络节点拥有相同的链路状态数据库。该链路状态数据库是一个反映整网拓扑结构的“带权有向图”，并在各个节点上完全相同和同步。

(2)最短路径树的计算

每个节点在获取 LSDB 后，使用最短路径优先(Shortest Path First, SPF)算法，以本节点为根节点，其他节点为叶子节点，生成一棵最短路径树(Shortest Path Tree, SPT)，即经过的节点数最少，节点间链路带宽最大。在计算 SPT 时，对 SPF 算法做优化和改进，通过增量路由计算 I-SPF(Incremental SPF)和部分路由计算(Partial Route Calculation, PRC)，加快最短路径树的计算速度[8]。最短路径树描述网络中任何一个节点到其他节点的最优路径树，是 IS-IS 路由计算的核心。

3.2 二层以太网的 SPT 最短路径树计算

要在二层以太网中构建从组播源节点到叶子节点间的最优组播树，需要先计算网络中的任何一个节点到其他节点的最短路径树，故将三层网络中已广泛使用的 IS-IS 路由协议引入二层以太网中，对各节点作如下设置：

(1)在各个节点上设定本节点的 System ID。

(2)根据节点间互联接口的带宽，设置接口的 Metric 值。

(3)在各互联接口上运行 IS-IS 协议。

通过 IS-IS 协议，每个节点都能得到一个反映节点间连接关系、节点间链路带宽等信息的 LSDB。基于 LSDB 即可计算以某个节点为根到其他所有节点的最短路径树。定义最短路径树包括：根节点 ID，入端口，出端口列表。其中，出端口列表形如：<出端口 1，出端口 2，…，出端口 *n*>。以节点 *B* 为例，假定根节点为 *A*，入端口为 *a*，出端口列表为：<端口 1，端口 2，端口 3>，则最短路径树表示对来自根节点 *A* 的广播报文，从端口 *a* 流入，从端口 1、端口 2、端口 3 组播出去。为能够支撑任何一个节点可能会成为组播源节点，全网各节点都需要计算出以任何一个节点为根，其他节点为叶子节点的 SPT。如果整个网络有 *N* 个节点，则每个节点都需要计算出 *N* 棵 SPT 的广播树。

最短路径树是实现最优组播树和组播树快速收敛的技术基础，其用途为：

(1)在获知哪个节点是组播源节点，哪些路由是 DR/BDR (Designated Router and Backup Designated Router)和哪些节点是叶子节点时，则可从以组播源节点为根，其他节点为叶子节点的最短路径树，剪枝出一棵最优组播树。

(2)如果本节点需要向其他节点快速广播一个消息，则可通过本节点为根的最短路径树向全网其他节点快速广播。

3.3 基于 SPT 的最优组播树构建

在计算出任何一个节点为根其他节点为叶子节点的 SPT 树后，需要从中查找出以 DR 为根的最短路径树，并根据各叶子节点的用户对组播组 G 的需求状态，对该 SPT 树进行剪枝后获得组播树。故需要有一种技术手段，让全网所有节点能够快速感知哪个节点是 DR 节点，哪些叶子节点需要组播组 G。IS-IS 协议的 LSP CLV 可以承载链路状态信息并通过通告机制向全网所有节点通告。如果将组播组 G 视为附属在组播源节点和叶子节点上的一种链路信息，则可通过 LSP CLV 及其通告机制来解决上述问题。

对提供组播业务的 DR/BDR 节点，其属性为 G/Tx，表示该节点将发送该组播业务 G，同时还声明本节点是 DR 还是 BDR。对接入用户的叶子节点，其属性为 G/Rx，表示该节点需要接收组播业务 G，同时需声明是 Join 还是 Leave。组播业务的提供节点和接收节点，都需要将组播组 G 这一附属链路信息向全网通告。故对 CLV 扩展，定义一个新的 Code 编码(如 Code=200)，用作对组播这个附属链路信息的通告。扩展 CLV 的 Value 取值包括组播地址 G(48 bit)、接收组播还是发送组播标识位 R/T(1 bit)、加入组播还是离开组播标识位 Join/Leave(1 bit)、组播源主节点还是备用节点标识位 DR/BDR(1 bit)以及保留位 Reserved(5 bit)。

扩展 CLV 用于组播源节点和叶子节点对组播组 G 的链路信息声明。R/T 用于表明是组播接收者还是发送者：如果是接收者，则由 Join/Leave 表明是申请加入组播组 G 还是离开组播组 G；如果是发送者，则由 DR/BDR 表明是组播源主用节点还是备用节点。该 CLV 作为本节点的附属链路信息，通过 LSP 通告机制向全网其他节点通告。

叶子节点和用户之间通过 IGMP Join/Leave/Query/Report

等协议报文感知用户是否需要接收组播业务 G。当检测到节点的首个用户需要加入组播组 G 时，表示该节点需要从上游节点引入组播 G 的业务流；当检测到所有用户离开该组播组 G 时，表示节点不再需要从上游节点引入组播 G 的业务流。当叶子节点的首个用户需要组播组 G 和最后一个用户离开组播组 G 时，表明该节点要通过上述 CLV 的 LSP 报文向全网通告本节点是申请加入组播 G(CLV 属性为 R 和 Join)还是申请离开组播 G(CLV 属性为 R 和 Leave)。网络边缘的组播源节点 DR 和 BDR，在经过主备协商确定为 DR 还是 BDR 身份后，需要通过上述 CLV 的 LSP 报文向全网通告，以确认节点对组播 G 是 DR(CLV 属性为 T 和 DR)还是 BDR(CLV 属性为 T 和 BDR)。通过上述扩展 CLV 以及 LSP 通告机制，各节点的 LSDB 能获知哪个节点是组播组 G 的发送节点，哪些节点是组播组 G 的接收节点。首先在 LSDB 中查找组播组 G 的发送节点(CLV 属性为 T 和 DR)，再在已经计算好的全部 SPT 中查找以组播发送节点为根节点的 SPT，并根据 LSDB 中 CLV 属性为 R 和 Join 的节点的情况，对这棵最短路径树进行剪枝操作，即在最短路径树的出端口列表中，裁剪从本节点开始往下的所有叶子节点中不需要组播组 G 的所有出端口，最终生成组播组 G 的组播树。由于最短路径树的路径最短、链路带宽最大，因此基于最短路径树剪枝而成的组播树也就是一棵最优组播树。为在组播源节点 DR/BDR 发生主备切换时方便组播树的快速重构，除了要确定以 DR 为组播源节点构建的主用组播树之外，还要确定以 BDR 为组播源备份节点构建一棵备用组播树。

3.4 最优组播树的快速重构

3.4.1 引起组播树重构的网络变化事件

当网络拓扑发生以下变化时，需要对已经构建好的最优组播树进行重构。

(1)网络的某个节点或某条链路的状态发生变化

当检测到与邻接节点的状态或某端口的链路状态发生变化时，整个网络拓扑将发生变化，需要基于新的网络拓扑对原先已构建好的最优组播树进行重构。

(2)组播源节点的 DR/BDR 主备关系发生变化

组播源节点在基于 DR/BDR 协议选举后，以其中一个节点为主节点(处于 DR 状态)，提供组播业务，以另一节点为备用节点(处于 BDR 状态)。当 DR/BDR 的主备关系发生变化时，组播源节点将发生变化，需要基于新的组播源节点构建新的最优组播树。

(3)用户对组播业务的需求状态发生变化

当接入用户的叶子节点检测到该节点首个用户需要请求组播组和最后一个用户离开组播组时，叶子节点对组播流的需求状态将发生变化，故需要对原先已经构建好的最优组播树重新进行剪枝操作，以重构组播树。

当检测到以上某个变化时，首先需要向全网快速发出通告，使全网所有节点能快速感知到这个变化，然后每个节点基于这个变化重新构建组播树。

3.4.2 全网快速通告网络变化的方法

当链路状态发生变化时，通过链路状态数据报的 Flooding 机制向全网快速扩散这一变化，并刷新 LSDB 数据库，实现最短路径树和最优组播树的快速收敛(收敛时间通常在 50 ms~200 ms)[8]，为进一步加快最短路径树以及最优组播树的收敛速度，本文提出一种向全网快速广播该链路状态数据报的链路状态变化方法。

二层以太网向全网广播一个消息时，需要将消息封装在目的地址是组播地址的二层组播报文中。为使这个报文以组播方式发送给全网所有节点，需对这个组播地址生成一棵无环组播树，该组播树在每个节点中以组播地址、出端口列表(<出端口 1，出端口 2，…，出端口 *n*>)的组播转发表形式表达。

在 3.2 节中，各节点已经计算出 *N* 棵最短路径树，每棵树对应以某个节点为根，其他节点为叶子节点的最短路径树，该树可用于根节点向全网广播某个消息。为将每棵最短路径树表达为组播地址、出端口列表(<出端口 1，出端口 2，…，出端口 *n*>)的组播转发表，需要以节点 ID 自动映射出一个组播地址，以确保全网的所有节点能够统一标识以该节点为根其他节点为叶子节点的最短路径树，并生成相应的组播转发表。

以节点 ID 映射为 48 bit 的二层组播地址为例，高 25 bit 是固定的组播 MAC 地址标识，低 23 bit 以节点 ID 标识，其相应的组播地址为 01:00:5e:ID。

通过以上映射，全网所有的节点都能在已计算出的最短路径树中为每个节点 ID 生成一个组播树，并表达组播转发表，该表格式如下：组播地址 01:00:5e:ID，出端口列表(<出端口 1，出端口 2，…，出端口 *n*>)。当节点 ID 检测到某个网络变化事件时，可将 LSP 报文直接封装在目的地址为 01:00:5e:ID 的二层组播报文中，并向全网广播该 LSP，各节点在收到该组播报文后，按照组播转发表复制给其他节点。

为能使每个节点的控制平面接收该 LSP 报文，并交由 IS-IS 协议模块处理该链路状态数据报，在组播转发表中，将本节点的 CPU 控制平面作为需要接收此组播报文的一个节点，将数据平面通往控制平面 CPU 的通道加入到组播转发表的出接口列表中，则此时的组播转发表为：组播地址 01:00:5e:ID，出端口列表(<本节点 CPU 通道，出端口 1，出端口 2，…，出端口 *n*>)。

通过以上方法，某节点在检测到 3.4.1 节所阐述的某个网络变化事件时，可将这个事件快速广播给其他节点，使全网的节点能够在同个时刻快速刷新链路状态数据库，并通过 I-SPF 和 PRC 算法，快速重构最优组播树[8]。

3.4.3 最优组播树的快速重构

收到网络变化事件后，在刷新链路状态数据库的同时，重新构建最优组播树。对于不同的网络变化事件，其组播树的重构方法不同。

(1)网络中某个节点或某个链路的状态发生变化

每个节点在收到来自其他节点的拓扑变化消息的链路状态数据报后，通过 I-SPF 算法和 PRC 算法对原先已经计算好的各最短路径树的变化部分进行增量计算，以修正和重构原先的最短路径树。为进一步加速组播树的重构速度，可优先计算以组播源节点为根的最短路径树。根据链路状态数据库已经保存的有关各叶子节点的 G/Rx 的信息，对这棵新的最短路径树进行剪枝操作，可快速完成最优组播树的重构。

(2)组播源节点的 DR/BDR 主备关系发生变化

在 3.3 节中，已对组播源节点生成最优组播树，并确定一棵为主用，另一棵为备用。当收到 DR/BDR 状态发生变化的消息后，直接对组播树进行主备快速切换即可。

(3)用户对组播业务的需求状态发生变化

当每个节点接收到来自某个叶子节点对组播组 G/Rx 的状态变化消息后，如果发现在以组播源节点为根的最短路径

树有一个新的叶子节点需要加入组播组，则对原先的最优组播树进行增枝操作，即将最短路径树从本节点到这个叶子节点之间有上下游树枝关系的端口，增加到组播表的出接口列表中；如果发现某个叶子节点不再需要组播组，则对原先的最优组播树进行剪枝操作，并在组播表出接口列表中删除相关端口数据直到该叶子节点的出端口数据。

4　仿真测试与对比

为比较本文方法和传统基于“MSTP 协议+IGMP 侦听协议”的方案在组播树构建速度、组播路径优化度、网络发生变化时组播树的重构速度这 3 个关键技术指标，仿真测试网络模型如图 1、图 2 所示。

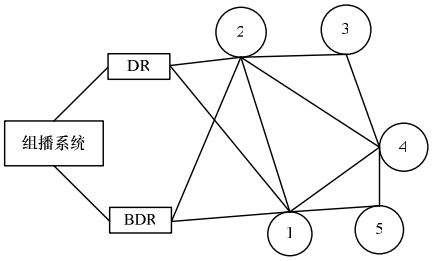

图 1　5 个叶子节点的仿真网络

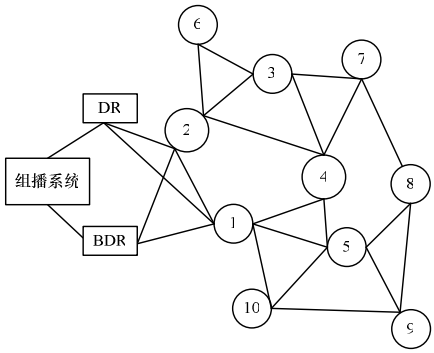

图 2　10 个叶子节点的仿真网络

由图 1、图 2 可知，在仿真网络模型中，DR、BDR 这 2 个节点作为组播源的主备节点，其余节点作为叶子节点，相邻节点间均为两两互联的三角形拓扑结构。

为验证上述 2 种方案对不同网络规模的适应能力和网络构建性能的影响，将仿真网络模型从 5 个叶子节点扩展到 10 个节点、20 个节点、30 个节点、50 个节点，并对这 5 种规模的网络模型进行逐一测试和对比。

传统方法测试时，将各节点 MSTP 协议的相关参数按照缺省的参数进行配置：Forward Delay=10 s，Hello Time=2 s，Max Age=20 s。而在测试本文方法时，只需配置各节点间 IS-IS 协议的 Hello Time=2 s。

4.1　组播树生成速度对比

每种模型的测试时，每个叶子节点都包括用户需要接收的组播业务。2 种方法组播树生成速度对比如表 1 所示。

表 1　组播树生成速度对比

叶子节点数	传统方法结果/ms	本文方法结果/ms
5	11 560	29.5
10	13 580	32.5
20	15 740	36.5
30	17 920	41.5
50	20 320	46.8

由表 1 可知，传统方法中组播树的生成时间在 11 s~20 s 之间，当网络规模越大，组播树生成时间越长。而采用本文方法，组播树的生成时间平均在 30 ms~50 ms 之间，且组播树生成时间对网络规模不敏感。

4.2　组播树优化度对比

在每种模型的测试时，将网络节点间互联端口分别仿真为 FE、GE、10GE 这 3 种速率，且每种速率各占 1/3 的端口。在每种模型的测试统计中，各叶子节点对应着该叶子节点的组播路径带宽，将本文方法构建的组播树的叶子节点所对应的组播路径带宽与传统方法确定的组播路径带宽进行比较，本文方法结果要大于或者等于传统方法的结果。组播树优化度表示：当叶子节点数一定时，本文方法确定的组播路径带宽大于传统方法所确定的结果，所对应的叶子节点数。组播树优化度比较如表 2 所示。

表 2　组播树优化度比较

叶子节点数	组播树优化度
5	1
10	2
20	5
30	8
50	14

由表 2 可知，本文方法对任何规模的模型，总有若干个叶子节点的组播路径的链路带宽要高于传统方法，且网络规模越大，这种叶子节点数量越多，说明本文方法所构建的组播树在链路带宽上优于传统方法。

4.3　组播树重构速度对比

比较 2 种方法在不同网络规模下，当网络拓扑发生变化时组播树的重构速度，该比较结果如表 3 所示。其中，情况 1 表示：当一个网络节点和一条链路发生故障时组播树重构的时间；情况 2 表示：当 DR/BDR 主备关系发生变化时组播树重构的时间；情况 3 表示：当一个节点用户加入组播或另一个节点用户离开组播时组播树重构的时间。

表 3　组播树重构时间比较

叶子节点数	情况 1		情况 2		情况 3	
	传统方法/ms	本文方法/ms	传统方法/ms	本文方法/ms	传统方法/ms	本文方法/ms
5	12 670	22.5	2 650	5.5	2 430	7.3
10	13 740	29.2	2 770	7.2	3 120	7.8
20	16 630	33.3	3 620	7.8	3 970	8.6
30	19 830	37.6	4 320	8.6	4 610	9.5
50	21 020	40.7	5 020	9.7	5 830	10.8

由表 3 可知，在上述 3 种情况下，传统方法的收敛时间基本在 2 s~21 s 之间，而本文方法的收敛时间基本上在 5 ms~50 ms 之间。

由表 1~表 3 可知，本文方法在组播树的构建速度、组播路径的优化度、网络拓扑发生变化时组播树的重构速度等方面优于 MSTP 生成树协议+IGMP 侦听协议的传统方法。

5　结束语

本文提出二层以太网中基于 IS-IS 协议及其扩展，在组播源节点到组播叶子节点之间构建一棵最优组播树的方法。实验结果表明，该方法能解决在二层以太网中开展组播业务时，最优组播树的快速构建以及当网络拓扑发生变化时该最优组播树可快速重构的技术问题。　　(下转第 110 页)