基于Linux的一种MPLS故障恢复的设计与实现

朱瑞新[1],王 芳[1],李 珂[2],秦 臻[3]
(1. 北京邮电大学,北京 100876;
2. 广东电信网络操作维护中心,广东 广州 510083;
3. 中国电子设备系统工程公司研究所,北京 100039)

摘 要:目前主要存在Haskin方案和Makam方案两种不同MPLS故障恢复机制。根据对二者进行分析比较,采用Makam方案,提出在静态LSP下,检测到MPLS故障时可采用的两种实现方法:一种是通过修改IP路由表来实现保护倒换;另一种是通过修改MPLS转发表实现保护倒换。依据实际的网络环境选择后者进行设计与实现,给出了具体的处理流程、相应的数据结构和修改MPLS转发表的方法。最后,对拓扑图配置Lsp进行试验,试验显示故障发生后成功保护倒换到备份Lsp,从而证明本方法的可行性和快速性。

关键词:MPLS;故障恢复;保护倒换;转发;标签交换;边缘路由器
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2009)02-0220-04

Design and Implementation of MPLS Fault Recovery Methods Based on Linux

ZHU Rui-xin[1], WANG Fang[1], LI Ke[2], QIN Zhen[3]
(1. Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Guangdong Telecom Network Operation Center, Guangzhou 510083, China;
3. Chinese Institute of Electronics Systems Engineering Company, Beijing 100039, China)

Abstract: Haskin scheme and Makam scheme are two different way's on MPLS fault recovery. According to the analysis and comparison these two way's, takes Makam scheme, and introduces the methods update IP router table or update MPLS forward table to implement the Makam scheme on static Lsp. Offers handling process, data structure and update MPLS forward table method. At last, config and exam the topology which pictured in article. The experiment shows that the main Lsp is switched to the backup Lsp successfully, and verify that this method is practicability and quickly.

Key words: MPLS; fault recovery; protecting rotate; transmitting; tag exchanging; edge router

0 引 言

MPLS(Multi-Protocol Label Switching)网络已成为运行在IP骨干网的核心技术,它也是下一代网络的核心技术。它通过定长标签进行数据转发,能够在无连接网络中引入面向连接网络的特性,并可以较好地支持QoS保障、流量工程、VPN等增值业务。

MPLS作为骨干网技术,需要对网络故障作出快速响应,避免在路由器中数据包的拥塞,导致网络性能下降及QoS降低。因此,关于MPLS的故障恢复技术具有很大的应用价值,国内外对此展开了广泛的研究,网络故障恢复与网络自身的拓扑结构是密切相关的,由于以往网络规模较小,相关的网络生存性技术大多基于集中式的控制策略,网络向更大规模发展的必然趋势,使得集中式的生存性技术有些力不从心。基于MPLS故障恢复技术针对故障的局部性。在局部范围内实施网络资源的优化调配。能实现快速的故障诊断和恢复,从而成为大型复杂网络生存性的主流技术。

在大型复杂网络中实施故障恢复主要有两方面的工作:一是利用网络探测检测故障。目前常用的故障探测有BFD快速故障检测,MPLS Ping等故障探测方法[1];二是在对故障链路实施保护倒换,将网络数据流

收稿日期:2008-06-16
基金项目:国家863高科技计划(2007 AA01Z2A1,2006AA012Z229);国家自然科学研究基金(60672086)
作者简介:朱瑞新(1979-),男,山东菏泽人,硕士研究生,研究方向为计算机宽带路由及网络仿真;导师:陈山枝,教授级高工,博士,教授,博士生导师,研究方向为宽带交换技术、宽带接入技术、移动互联网、网络发展与演进战略等。

切换到备份路径上。在探测到故障之后,对采取保护倒换的方案进行了分析比较,然后选择其中最优方案进行了设计和实现。

1　故障恢复理论研究

目前针对事先配好的静态 LSP 进行保护倒换,主要有以下两种方案:一种是 Haskin 方案[2],它是在建立备份路径时构成回路,故障后实现快速重路由;另一种是 Makam 方案[3],故障后向上游节点发送故障指示信号(FIS)将流量切换至备份路径;Makam 方案是一种全局方案。它的恢复路径建立在 PSL 与 PML 之间。如图 1 所示,Ler0 和 Ler1 是边缘路由器,其中 Ler0 中运行故障探测程序,主 Lsp 路径是 Ler0 - Lsr0 - Ler0,Makam 方案是建立备份路径 Ler0 - Lsr1 - Ler1,Ler0 检测到故障后会触发保护倒换;由保护倒换程序负责将流量切换至备份路径,故障恢复后同样触发保护倒换程序,以实现 Ler0 将流量切换回原工作路径。

Makam 方案的主要优点是失序分组较少,且恢复路径较短。但必须在故障探测后才会实施切换,因而会造成较大的丢包。

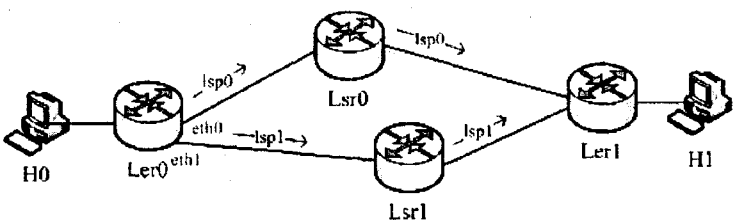

图 1　试验拓扑图

LER:标签边缘路由器。作用是分析 IP 包头,用于决定相应的传送级别和标签交换路径(LSP)。

LSR:标签交换路由器。

故障恢复:在网络发生故障时,及时进行故障切换,保障网络应用不受影响。

备份路径:在建立标记交换路径时,指定其备份路径,在主路径发生故障,通知入口路由器把流量切换到备份路径;它的主要优点是,恢复时间比较快,主要缺点是需要占用额外的资源。

Mpls 入口路由器在 IP 分组中根据不同的转发方式嵌入不同的标签[4],各个中间路由器即根据标签转发数据,而不需要像传统的路由器那样根据路由表深入分析 IP 分组的头部。在出口路由器标签被移去。在入口路由器和出口路由器之间通过标签指定了一条独立与第三层路由的固定传输路径——标签交换路径(LSP)。当网络发生故障时有些可以通过底层修复,但通常不能满足 MPLS 的要求,如底层协议只有针对链路的保护而没有针对节点的保护。通过第三层的路由协议也可以重新计算出新的路径进行数据传输,但其时间较长(几秒到几分钟),这对于服务质量要求较高的服务如语音传输等会带来很大的影响[5]。MPLS 故障恢复机制保证在发生故障时能在期望的时间内恢复数据传输,包括链路保护、节点保护、路径保护和网段保护等。

2　故障恢复方案选择

在实现保护倒换时有两种方案可供选择:一种是通过修改 IP 路由表来实现保护切换,见图 1,在保护倒换之前配置主 LSP 和备份 LSP,在故障检测程序检测到故障后,触发保护倒换程序,保护倒换程序导入配置信息,查找主 LSP 对应的备份 LSP,保护倒换的命令是通过 Netlink 的形式与内核交互的,所以在保护倒换之前要建立 Netlink 连接,以便于与内核交互,根据当前 IP 下一跳查找路由表,从中找到对应的 FEC,修改改 FEC 为备份的 FEC。

另一种是通过修改 MPLS 转发表实现保护倒换,该方案与前一方案在建立 Netlink 连接之前是相同的,连接之后查找主 LSP 对应的 NHLFE,然后修改 NHLFE 对应的 FEC 为备份 LSP 的 FEC。如图 2 所示。

以上两种方案均能实现 LSP 的保护倒换,但是前者需要修改路由表,考虑到在实际的运行网络中,如果修改路由表会影响到其它的路由软件的运行,所以本实现采用了修改 MPLS 转发表的方案。如图 3 所示。

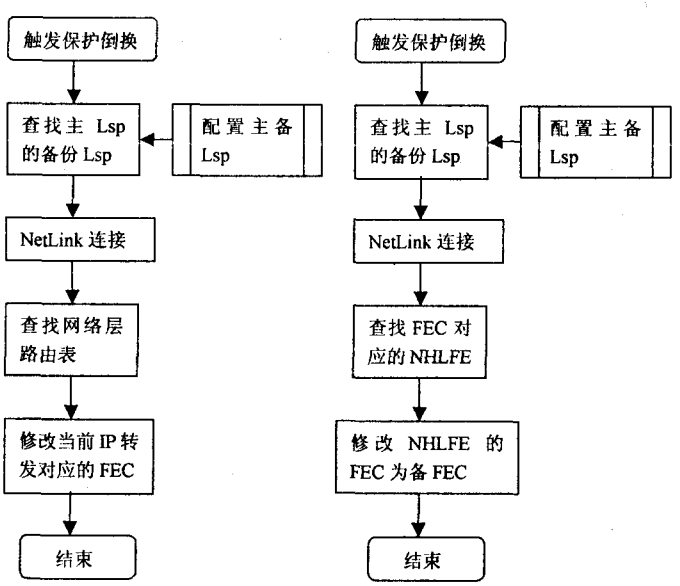

图 2　修改路由表　　图 3　修改 MPLS 转发表

3 故障恢复设计与实现

在 MPLS 固定网络中，通过各种检测手段检测到网络故障，然后进行保护倒换，保护倒换分为动态保护倒换和静态保护倒换。文中针对的是静态保护倒换，通过静态配置多条 LSP 为路由提供保护倒换。

在配置的保护倒换 Lsp 中，1 条备 LSP 为 1 条主 LSP 提供保护。在 Ingress 建立好一条主 LSP 和备 LSP，在链路状态正常的情况下通过 Ingress 的 Selector 将流量分发到主 LSP 上，此时备 LSP 上没有主 LSP 上的流量。当发生故障时启动保护倒换，将流量切换到备 LSP 上，接收方也从备 LSP 上接收数据流。

本系统的实现是基于 Linux 2.6.8.1 内核并且打上 MPLS 1.950 补丁，采用拓扑见图 1。

1) netlink 处理流程。

保护倒换启动时通过 netlink 与内核交互，mpls 协议实现提供对 netlink 的支持。采用 Netlink 的好处是：netlink 是一种异步通信机制，在内核与保护倒换应用之间传递的消息保存在 socket 缓存队列中，发送消息只是把消息保存在系统的 socket 的接收队列，而不需要等待接收者收到消息。

实现步骤如下：

(1)建立 netlink 链接；

(2)获取当前的 key 值；

(3)删除当前 key 值指向的 nhlfe；

(4)用函数 mpls_nhlfe_modify()，发送修改 key 值指向的指令；

(5)接收判断返回信息。

在系统进行保护倒换前存在主 LSP(LSP0)和从 LSP(LSP1)，系统通过从 LSP 对主 LSP 提供备份。当主 LSP 发生故障时，启动保护倒换。保护倒换程序根据主 LSP 的信息，查找其 key 值，然后将这个 key 值指定到从 LSP 的 nhlfe。

2) 数据结构。

·netlink socket 的地址结构：

```
struct sockaddr_nl
{
sa_family_t nl_family;
unsigned short nl_pad;
_u32 nl_pid;
_u32 nl_groups;
};
```

nl_family：表示协议族，当前必须设置为 AF_NETLINK 或者 PF_NETLINK。

nl_pad：当前没有使用，因此要总是设置为 0。

nl_pid：为接收或发送消息的进程的 ID，如果希望内核处理消息或多播消息，就把该字段设置为 0，否则设置为处理消息的进程 ID。

nl_groups：用于指定多播组。

·netlink 消息头：

```
struct nlmsghdr
{ _u32 nlmsg_len;
_u16 nlmsg_type;
_u16 nlmsg_flags;
_u32 nlmsg_seq;
_u32 nlmsg_pid;
};
```

nlmsg_len：指定消息的总长度，包括紧跟该结构的数据部分长度以及该结构的大小。

nlmsg_type：用于应用内部定义消息的类型，它对 netlink 内核实现是透明的，因此大部分情况下设置为 0。

nlmsg_flags：用于设置消息标志。

nlmsg_seq：用于应用追踪消息，表示顺序号。

nlmsg_pid：用于应用追踪消息，表示消息来源进程 ID。

3) 修改 MPLS 转发表。

针对图 1 拓扑中 LSP0 发生故障后，进行路径切换示意图如图 4 所示。

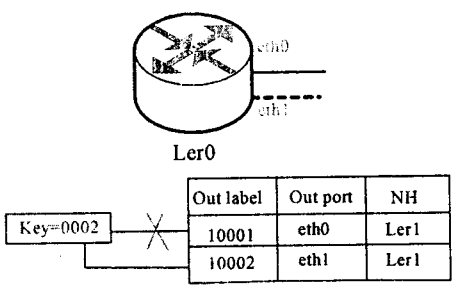

图 4 边缘路由器故障处理

系统正常情况下运行时，LSP1 的 key 值为 0002，标签是 1001，当发生故障执行倒换，首先删除 key 与当前 nhlfe 的关联，然后将 LSP2 的 key 值设为 0002。

4 故障恢复的实现验证

1) 静态 Lsp 配置。

Ler0 上的配置信息如下：

```
ifconfig eth0 192.168.1.81/24
ifconfig eth1 192.168.1.82/24
mpls nhlfe add key 0
mpls nhlfe add key 0
mpls labelspace add dev eth0 labelspace 0
mpls labelspace add dev eth1 labelspace 1
mpls ilm add label gen 20001 labelspace 0
```

```
mpls ilm add label gen 20002 labelspace 1
mpls nhlfe change key 0x02 instructions push
gen 10001 nexthop eth0 ipv4 192.168.1.91
mpls nhlfe change key 0x03 instructions push
gen 10002 nexthop eth1 ipv4 192.168.1.92
ip route add 192.168.1.91/32 via 192.168.1.91 spec_nh
0x8847 0x02
```

2) 在进行保护倒换之前检测 LSP 配置，如图 5。

```
[root@localhost 1]#mpls nhlfe show
MHLFE entry key 0x00000003 mtu 1496 propagate_ttl
    push gen 10002 set eth1 iqv4 192.168.1.92(0 bytes,0 pkts,0
dropped)
MHLFE entry key 0x00000002 mtu 1496 propagate_ttl
    push gen 10001 set eth0 iqv4 192.168.1.91(0 bytes,0 pkts,0
dropped)
[root@localhost 1]#
```

图 5　保护倒换之前的 LSP 静态配置

3) 保护倒换之后，验证转发表被正确修改。从图 6 中可以看出在保护倒换之后，LSP 路径切换到备份路径上去。

```
[root@localhost ProtectSwitch]#mpls nhlfe show
MHLFE entry key 0x00000002 mtu 1496 propagate_ttl
    push gen 10002 set eth1 iqv4 192.168.1.92(0 bytes,0 pkts,0
dropped)
[root@localhost ProtectSwitch]#
```

图 6　保护倒换后验证

5　结束语

通过对 MPLS 故障恢复机制的研究，比较了在全局路径保护的条件下，对两种保护倒换机制进行了比较，并选取其中一个有利于扩展和移植的方案进行了设计和实现，文中给出了主要的数据结构及实现办法，最后对方案的实现进行了测试并验证通过。

参考文献：

[1] Kompella K, Swallow G. Detecting MPLS Data Plane Failures [S]. IETF RFC4379, 2006.

[2] Haskin D, Krishnan R. A Method for Setting an Alternative Label Switched Paths to Handle Fast Reroute [S]. IETF Draft, 2001.

[3] Huang Changeheng, Sharma V, Makam S. A Path Protection/Restoration Mechanism for MPKS Networks [S]. IETF Draft, 2000.

[4] Sharma V, Hellstrand F. Framework for Multi-Protocol Label Switching (MPLS) - based Recovery [S]. IETF RFC3469, 2003.

[5] 王卫民，张　辉，刘　畅，等. 一种支持多种保护类型的 MPLS 故障保护机制[J]. 计算机工程与应用，2004(22)：132-133.

(上接第 219 页)

加急剧，这些都对目前的指挥决策提出了严峻的挑战。而群体决策支持系统正是决策者的有力工具，它可以帮助进行群体决策，使群体迅速达成一致，做出满意、高效的决策。文中针对异步指挥控制环境中的群体决策存在的诸多困难，提出了信息组件的概念，并利用信息组件进行了异步指控 GDSS 的结构设计，采用此结构能较好解决目前决策者之间进行信息共享和交流时的效率低下、决策者存在认知负担和决策者的偏好不能迅速集结等问题，从而提高群体决策的效率和质量。

参考文献：

[1] Fleming R. Information Exchange and Display in Asynchronous C2 Group Decision Making [C]//8th International CCRT Symposium, held at the National Defense University. Washington DC: [s.n.], 2003: 17-19.

[2] 刘　翔，李明星，胡运权. 群决策支持系统研究的现状值得注意的问题及将来的研究方向[J]. 高技术通讯，2000(4)：107-110.

[3] 陈少辉. 群体决策支持系统中知识库的研究[D]. 武汉：武汉理工大学，2006.

[4] 侯引茹. 群体决策支持系统中的信息共享技术研究[D]. 西安：西北工业大学，2001.

[5] 阎礼祥，覃　征，韩　毅. 指挥自动化系统中的群决策支持模式研究[J]. 西安电子科技大学学报：自然科学版，2003，30(6)：839-843.

[6] 叶　丹，陈禹六. 面向问题的动态群体决策支持系统框架研究[J]. 计算机工程与应用，2003(14)：210-213.

[7] 吴健中，周　泓. 群体决策支持系统的理论与应用述评[J]. 系统工程学报，1994，9(2)：119-130.

[8] 颜浩然，袁　捷，陈毛狗. Internet 环境下 GDSS 框架的研究与实现[J]. 计算机工程，2002，28(2)：66-68.

[9] 郑全全，郑　波，郑锡宁，等. 多决策方法多交流方式的群体决策比较[J]. 心理学报，2005，37(2)：246-252.

[10] Fleming R, Kaiwi J. The Problem of Unshared Information in Group Decision-Making [R]. San Diego: SPAWAR Systems Center, 2002.

[11] 蒋　丽，于广涛，李永娟. 团队决策及其影响因素[J]. 心理科学进展，2007，15(2)：358-365.

[12] 杨　雷. 群体决策理论与应用：群体决策中的个体偏好集结方法研究[M]. 北京：经济科学出版社，2004.

[13] Fleming R, Cowen M. Quantification of Subjective Information Assessments in C2 Decision Option Selection [C]//11th International CCRT Symposium, held at the De Vere University Arms. Cambridge, UK: [s.n.], 2006: 26-28.