ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.3, March 2006, pp.445−453
DOI: 10.1360/jos170445

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种集群路由器转发表同步框架及关键算法*

张晓哲+, 卢锡城, 朱培栋, 彭 伟

(国防科学技术大学 计算机学院,湖南 长沙 410073)

A Synchronization Framework and Critical Algorithm Maintaining Single Image of IP Forwarding Tables Among Cluster Router's Nodes

ZHANG Xiao-Zhe+, LU Xi-Cheng, ZHU Pei-Dong, PENG Wei

(School of Computer, National University of Defense Technology, Changsha 410073, China)

+ Corresponding author: E-mail: nudtzhangxz@263.net, http://www.nudt.edu.cn

Zhang XZ, Lu XC, Zhu PD, Peng W. A synchronization framework and critical algorithm maintaining single image of IP forwarding tables among cluster router's nodes. *Journal of Software*, 2006,17(3):445–453. http://www.jos.org.cn/1000-9825/17/445.htm

Abstract: With the rapid development of Internet, traditional centralized routers can not meet the requirements of next generation Internet on reliability, performance scalability and service scalability. Cluster routers will be the most important components of future Internet. It is very important for cluster routers to maintain the same forwarding table images among cluster router nodes. Different synchronization mechanisms have variant performance to control plane and packet forwarding plane. After the analysis of two typical synchronization mechanisms, this paper proposes an asymmetrical forwarding table synchronization framework — AREF (asymmetrical routes electing framework) synchronization framework. It fits the requirements of massively parallel cluster routers architecture perfectly. Continuous route flapping of the backbone network burdens the synchronization mechanisms of cluster routers. AREF synchronization algorithm is proposed to decrease the synchronization costs of AREF synchronization framework during route flapping. It uses route cache to predict a new best route when the original best route is deleted, and reduces the synchronization cost of AREF synchronization framework greatly. AREF synchronization framework and algorithm requires diverse abilities for different routing node types and can be used in heterogeneous cluster router widely.

Key words: cluster router; single image; route synchronization; route prefix; forwarding table; route cache

摘 要: 随着传统体系结构路由器在可靠性和多维可扩展性等方面不能满足下一代 Internet 发展的需要,集群结构的路由器将成为未来骨干网络的核心.如何保证集群路由器各个路由节点转发表的单映像性,对控制平面及转发平面的性能至关重要,是值得研究的重要问题.在分析现有的各种转发表同步机制特点的基础上,提出一

* Supported by the National Natural Science Foundation of China under Grant Nos.90104001, 90204005, 90412011 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.2003CB3148020 (国家重点基础研究发展规划(973)); the National High-Tech Research and Development Plan of China under Grant No.2005AA121570 (国家高技术研究发展计划(863))

Received 2005-03-16; Accepted 2005-08-25

种非对称的路由同步框架——AREF(asymmetrical routes electing framework)路由同步框架,更适合于大规模异构的集群路由器系统的特点.在 AREF 路由同步框架上,进一步提出了 AREF 路由同步算法.算法针对每个路由前缀使用路由 Cache 来缓存次优路由,在全局最优路由被删除时,通过预测次优路由来减少同步开销.模拟实验表明,AREF 同步框架与算法的性能远远优于其他路由同步机制,与理论最优值比较接近.

关键词: 集群路由器;单映像;路由同步;路由项;转发表;路由缓存
中图法分类号: TP393 文献标识码: A

Internet 的飞速发展对网络设备的计算能力、转发能力和端口密度都提出了更高的要求.传统的单节点路由器在可靠性、性能可扩展性、规模可扩展性和服务可扩展性等方面有其难以逾越的障碍,已经不能满足下一代 Internet 发展需要.集群结构的路由器由于其自身的分布式特点,存在着非常广阔的发展空间.G 比特及 T 比特商用路由器都支持多机柜互连的集群技术,如:Cisco 的 CRS 路由器[1]、juniper T640[2],AVICI 公司的 TSR 路由器[3].为了降低高性能路由器越来越高昂的开发成本,可以使用廉价个人工作站或者低端商用路由器通过高性能交换网络连接起来,组成大规模并行转发路由器,如:Pluris 大规模并行路由系统[4]、纽约州立大学的 Suez[5]等.

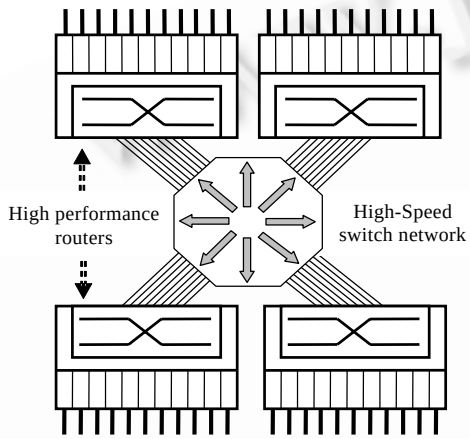

Fig.1 The architecture of cluster router

图 1 集群路由器参考体系结构

集群路由器可以定义为“将常规路由器或通用 PC 计算机连接起来组成的单映像路由系统[6]”.比较有代表性的集群路由器体系结构如图 1 所示,它具有如下几方面的特点:1) 单映像,从组网的角度来看,一个集群路由器是一台路由器,而不是一个网;2) 传统路由器的报文转发工作由集群路由器的各个节点分别承担;3) 路由协议及路由计算功能可以集中在特定的路由节点上,也可以分布到不同路由节点;4) 路由节点之间通过高速网络连接,绝大多数网络支持单播、组播等灵活的传输机制.

集群路由器面临的核心问题之一是“单映像”问题.从报文转发层面上看,由多个路由节点组成的集群路由器在外部行为上能够被称为是一台路由器而不是一个网络的前提条件是:“每个节点对到达相同地址的 IP 报文按照相同的路由策略进行转发”.这要求集群路由器的每个路由节点具有完全相同的转发表映像.在集群路由器运行过程中,路由协议与邻接路由器交换路由信息,不断地更新本地转发表信息.因此如何保证各个路由节点转发表的单映像性,是亟待解决的重要问题.

转发表是在路由协议接收的路由信息基础上,通过执行路由计算过程产生的.因此,路由协议在集群路由器各个节点上的分布情况,直接决定了使用何种算法来保证转发表的单映像性.图 2 给出了集群路由器一种可能的协议分布方式:节点 1~节点 3 除了承担报文转发任务之外,还负责路由协议的执行,而节点 4 只承担报文转发任务.

目前,很多路由协议都不支持分布式实现方式,因此协议在节点上的分布通常是不均匀的.如图 2 所示,BGP 协议分布在节点 1、节点 2 和节点

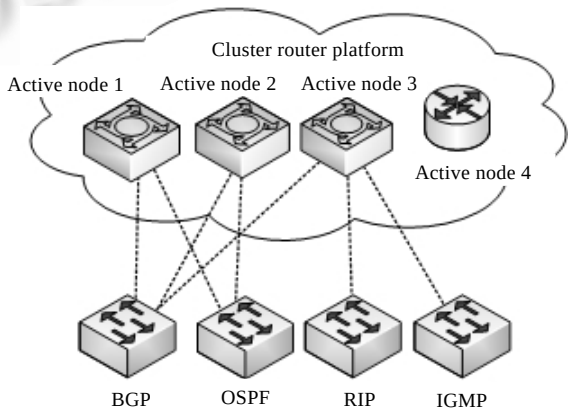

Fig.2 The asymmetric distribution of routing protocols

图 2 路由协议在集群路由器中非均匀分布

3 上;OSPF 协议分布在节点 1 和节点 2;RIP 协议和 IGMP 协议分布在节点 3 上.路由协议的分布式实现问题,不在本文的讨论范围内,请参阅文献[7–9].我们将运行路由协议进程、能够主动变更转发表路由的节点称为主动节点,如节点 1.将节点 4 这种不运行任何路由协议进程、不会引发转发表变化的节点称为被动节点.节点 1 运行 BGP 协议、OSPF 协议,会通过邻接路由器学习大量的路由信息并保存在本地的路由表中.将这种由起源于本节点或者通过本地路由协议获取的路由信息构成的路由表称为本地协议路由表.将根据本地协议路由表执行路由计算过程产生的本地最优路由集合称为本地转发表.将各个节点的本地转发表通过一定的路由同步算法后,获得的全局最优路由集合称为全局转发表.

在全局转发表发生变化时,集群路由器要通过内部互连网络在各个路由节点之间进行路由同步.由于比传统的集中结构的路由器增加了额外的内部通信开销,其对全局转发表的频繁变化更加敏感.Internet 规模的飞速扩展以及支持网络应用类型的快速增长,使得核心路由器控制平面面临一些严峻问题:网络规模的扩展,要求控制平面支持巨大的路由表,对设备的存储能力提出极高要求.高层路由协议的路由更新比较频繁;骨干网络故障频率高,故障间隔及恢复时间非常短[10],关键链路的故障会造成协议路由表剧烈震荡;作为全局转发表主要来源的 BGP 协议,其路由更新变化有很强的周期性[11],40%~55%[12]的路由更新存在周期摆动.由协议路由表产生的全局转发表必然会继承这些特点.因此,如何降低节点之间路由同步开销,提高算法效率对集群路由器转发平面、控制平面及高层路由协议的性能至关重要.

面向由不同计算能力、存储能力与转发能力的路由节点构成的异构集群路由器系统,本文提出一种维持节点转发表单映像性的路由同步框架——AREF(asymmetrical routes electing framework)同步框架.AREF 同步框架要求每个主动节点先执行路由计算过程,生成本节点的本地转发表,再在本地转发表的基础上执行路由同步算法.减少了参与路由同步过程的路由数量.在 AREF 同步框架中,只有全局转发表被复制在各个节点上.协议路由表及本地转发表中冗余路由分布在各个主动节点上,而被动节点只需保存全局转发表副本,降低了对每个路由节点存储能力的需求.为了解决路由抖动造成的同步开销过大的问题,提出了基于本框架的 AREF 路由同步算法.通过在软件转发表的叶节点数据结构中增加一个路由缓存指针,缓存一条来自于其他节点的次优路由.当全局转发表中路由被删除时,使用路由缓存预测新的全局最优路由,极大地降低了路由抖动时的同步开销.

1 相关工作

分布式结构下路由器的转发表同步问题关系到路由器转发平面和控制平面的性能,因此受到研究机构和厂商的广泛关注.其中比较有代表性的工作有:清华大学关于分布式路由器路由管理模型[13]的研究、国防科学技术大学的银河玉衡高性能核心路由器的转发表主从分发机制、Pluris 大规模并行路由器中路由冗余分发机制[4]以及 Cisco 的 CRS 路由器中使用的转发表全冗余备份机制[1].尽管实现细节有所不同,这些已有路由同步机制可以归纳为两类:广播更新方式、全冗余备份方式.早期的集群路由器由于使用中央集中的协议分布方式,在路由同步机制上使用简单的广播更新方式.目前的商用集群路由器通常使用全冗余备份的路由同步机制.

1.1 广播更新方式

广播更新方式主要应用于中央集中的协议分布方式.指定一个具有较强计算能力、存储能力的节点或者机柜作为路由服务器,运行 CLI、网络管理模块以及 OSPF,BGP,RIP 等全部路由协议,负责与网络上其他邻接路由器交换路由信息,并根据接收到的路由信息形成全局转发表.在全局转发表发生变化时,将更新消息通过交换网络广播给集群路由器的所有被动节点.广播更新方式不能充分利用集群路由器的各种分布式资源,使控制平面成为整个系统的性能瓶颈,并且存在单点失效等可靠性问题,应用价值较低.但这种同步方式只在全局转发表发生变化时,才需要与被动节点进行路由同步,是一种理论上最优的路由同步算法,在算法评价中可以作为衡量各种同步算法的标准.

1.2 全冗余备份方式

对如图 2 所示的路由协议非均匀分布方式,由于主动节点之间处于对等地位,不存在一个集中的路由服务

器节点,因此不能直接使用广播更新方式.目前,很多商用集群路由器使用全冗余备份方式,其工作过程是:

1. 每个主动节点在逻辑上包含 3 个路由表:本地协议路由表、全局协议路由表和全局转发表,被动节点只包括全局协议路由表和全局转发表;
2. 主动节点运行路由协议,将协议接收的路由信息保存在全局协议路由表中,并同时通过交换网络将路由更新广播给其他节点.其他节点在接收到广播信息时,相应地更新本节点的全局协议路由表.当所有主动节点都完成更新广播后,每个节点都具有完全相同的全局协议路由表;
3. 每个节点独立地运行路由计算进程,通过本节点的全局协议路由表产生全局转发表.

全冗余备份方式可以较好地解决集群路由器的转发表单映像维护问题,但仍然存在以下不足:由于每个路由节点需要容纳庞大的全局协议路由表,要求每个节点具有同等强大的计算能力和存储能力.全局协议路由表的同步通信开销比较大,进一步加重了内部交换网络的负载;频繁的路由同步通信可能恶化上层协议的路由更新处理,延长路由收敛时间.

2 AREF 同步框架

针对集群路由器节点的异构性以及全冗余备份方式存在的不足,本文在全冗余备份方式的基础上提出了一种非对称的路由同步框架——AREF 同步框架,如图 3 所示.由于只有主动节点产生路由变更事件,AREF 路由同步框架约束参与同步过程的节点集合,只有主动节点能够参与路由同步过程,被动节点通过监听交换网络上的同步消息来完成全局转发表的更新.

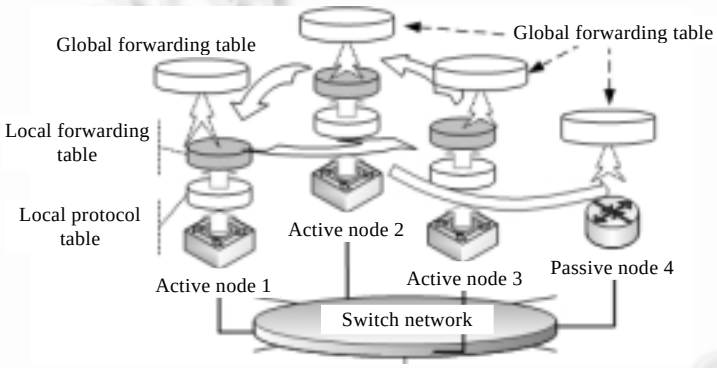

Fig.3 AREF framework maintaining routes synchronization among nodes' forwarding tables

图 3 AREF 转发表单映像同步框架

每个主动节点在逻辑上包含 3 个路由表:本地协议路由表、本地转发表及全局转发表,而被动节点只包括全局转发表.本地转发表由主动节点在本地协议路由表的基础上执行路由计算过程产生,而全局转发表则包含在各个主动节点的本地转发表中选择出的全局最优路由.为了减少了对集群路由器各个节点存储能力的需求,AREF 路由同步框架只将全局转发表复制到各个路由节点上,协议路由表以及转发表中的冗余路由信息都以分布式的方式存储在各个主动节点上.

为了减少路由同步过程中交换的路由信息数量,每个主动节点需要对本节点的协议路由表进行预处理:先执行路由计算过程,在本地协议路由表的基础上生成本地转发表,再在本地转发表的基础上执行 AREF 路由同步算法,这样只有主动节点本地转发表中的路由才可能参与路由同步过程.AREF 同步框架工作过程如下:

1. 运行于主动节点上的路由协议,将接收的路由信息保存在本地协议路由表中;
2. 路由计算过程判断本地协议路由表中的路由变化是否会引发本地转发表的变化.如果需要则调用全局转发表访问接口 API,修改本地转发表;
3. 在全局转发表访问接口中,AREF 同步算法判断对本地转发表的修改是否会引发路由同步过程.如果需要同步,通过交换网络向其他主动节点广播路由同步消息;
4. 其他节点在接收到路由同步消息时,执行 AREF 同步算法的同步消息处理过程,更新本节点的全局转发表;
5. 在 AREF 同步算法结束后,可以确保每个主动节点和被动节点都使用相同全局转发表进行 IP 报文转发.

3 AREF 同步算法

针对 AREF 路由同步框架的特点,对全局转发表中任意一条路由,每个主动节点的本地转发表中或者不包含与其具有相同网络地址的路由(将其标记为空路由),或者存在唯一的一条本地最优路由.如果将为空的路由作为所有有效路由的最小值,路由同步实际上是以分布的方式,在 n 个主动节点的具有相同网络地址的本地最优路由集合中选择一个全局最优路由,并通告给集群路由器所有节点.集群路由器运行过程中,高层路由协议会造成主动节点本地转发表的变化.由于所有节点都保存全局最优路由的副本,处理增加新路由的过程比较简单.当原有的全局最优路由被删除时,由于冗余路由被分布式地保存在各个主动节点上,主动节点之间无法预知剩余的本地最优路由大小,必须通过复杂的路由广播过程才能完成同步.为此,本文在 AREF 同步框架的基础上进一步提出了 AREF 路由同步算法以减少全局路由被删除时的同步开销.

3.1 符号定义

为方便讨论,将路由 r 简化为三元组$\langle dest,metric,nodeid\rangle$.其中 *dest* 指出路由覆盖的目的网络地址,*metric* 表示路由的度量值,*nodeid* 表示路由来源节点的节点号.符号 $r_1>r_2$ 表示路由 r_1 优于 r_2,符号 $r_1\equiv r_2$ 表示路由 r_1 和 r_2 各个域的取值都完全相等.表 1 给出了 AREF 算法描述中要用到的函数声明及对应的功能说明.

Table 1 Functions used in AREF algorithm description

表 **1** AREF 算法描述中使用的函数说明

Function name	Description
Dest(r_1)	Return the *dest* field of triple r_1
ID(r_1)	Returns the *nodeid* field of triple r_1
myid	Macro myid gets the identifier of local routing node
Broadcast(*opt*,r_1)	Broadcast route r_1 with the *opt* code on inner network $opt=\begin{cases}ADD, & \text{announce newroute } r_1\\ DEL, & \text{withdraw route } r_1\end{cases}$
GetLeafNode(*dest*)	Match a leaf node with network address *dest* in forwarding table and return it
Best(R_{List})	Select the best route in set R_{List} and return it

传统路由器 IP 转发表通常使用树形结构,网络地址相同的路由项保存在同一个叶节点上.下面给出叶节点数据结构的定义.由于一个叶节点保存到达某个网络地址的路由集合并标记出该集合中的最优路由,为此将到达网络地址 d 的叶节点 L_d 用二元组表示如下:

$$\begin{aligned}&L_d=\langle r_b,R_{List}\rangle,R_{List}=\{r_i\mid i\ge 1\}\text{ and satisfy}\\&r_b\in R_{List}\bigcap(\forall r_i(r_i\in R_{List}\bigcap(r_i\equiv r_b\bigcup r_i<r_b)))\end{aligned}\tag{1}$$

AREF 算法对 L_d 进行扩充,增加了全局最优路由指针 r_g 和路由 Cache 指针 r_c,扩充后的叶节点 L'_d 表示如下:

$$\begin{aligned}&L'_d=\langle r_g,r_b,r_c,R_{List}\rangle,R_{List}=\{r_i\mid i\ge 1\}\text{ and satisfy}\\&\begin{cases}r_b\in R_{List}\bigcap(\forall r_i(r_i\in R_{List}\bigcap(r_i\equiv r_b\bigcup r_i<r_b)))\\(r_g>r_b\bigcap r_g\notin R_{List})\bigcup r_g\equiv r_b\\r_c=null\bigcup(r_c\notin R_{List}\bigcap r_c<r_b\bigcap(\forall r,r\in R_{List}\bigcap(r=r_b\bigcup r_c>r)))\end{cases}\end{aligned}\tag{2}$$

3.2 算法描述

AREF 同步算法由两部分组成:一是 IP 层向路由协议层提供的全局转发表访问接口 API(application program interface);另一个是节点之间路由同步消息的处理例程.全局转发表访问接口为上层路由协议提供访问、修改全局转发表的手段,与路由同步相关的处理例程包括:

- 全局转发表中增加路由的处理过程 *Proto_Addroute*;
- 全局转发表中删除路由的处理过程 *Proto_Delroute*.

图 4 给出了 *Proto_Addroute* 过程和 *Proto_Delroute* 过程的具体算法描述.在增加路由的处理过程中,r_i 为来自于本节点高层路由协议的一条路由.路由协议调用该接口函数将 r_i 加入到全局转发表中,实现对转发平面的

控制.加入 r_i 时,首先要定位到与 r_i 相对应的叶节点 L'_d,然后再判断 r_i 是否影响 L'_d 的全局最优路由.如果 r_i 优于全局最优路由 $L'_d.r_g$,则将 r_i 赋值给 $L'_d.r_g$,并向其他节点通告该路由.在更新 $L'_d.r_g$ 时,如果原有的全局最优路由是来自于其他节点,并且满足路由 Cache 的更新条件,则将该路由赋值给路由 Cache 指针 $L'_d.r_c$,作为以后进行全局最优路由预测的依据.

Proc *Proto_Addroute*(r_i)	**Proc** *Proto_Delroute*(r_i)

d:=***Dest***(r_i)
L'_d:=***GetLeafNode***(d)
$L'_d.R_{List}$:=$L'_d.R_{List}\cup r_i$
if (r_i>$L'_d.r_g$)
  if ($L'_d.r_g$>$L'_d.r_b$)
  $L'_d.r_c$>$L'_d.r_g$
  $L'_d.r_g$=r_i
  Broadcast(*ADD*,r_i)
if (r_i>$L'_d.r_b$)
  if ($L'_d.r_b$>$L'_d.r_c$)
  $L'_d.r_c$:=*null*
  $L'_d.r_b$=r_i
else
  if (r_i>$L'_d.r_c$)
  $L'_d.r_c$:=*null*
	d:=*Dest*(r_i)
	L'_d:=*GetLeafNode*(d)
	if ($r_i\equiv L'_d.r_g$)
	  $L'_d.R_{List}$:=$L'_d.R_{List}-r_i$
	  if ($L'_d.r_c \neq null$)
	  Broadcast(*ADD*,$L'_d.r_c$)
	  $L'_d.r_g$=$L'_d.r_c$
	  $L'_d.r_c$:=*null*
	  else
	  $L'_d.r_b$:=***Best***($L'_d.R_{List}$)
	  if ($L'_d.r_b \neq null$)
	  Broadcast(*ADD*,$L'_d.r_b$)
	  else
	  Broadcast(*DEL*,r_i)
	  $L'_d.r_g$=$L'_d.r_b$
	else
	  if ($r_i\equiv L'_d.r_b$)
	  $L'_d.R_{List}$:=$L'_d.R_{List}-\{r_i\}$
	  $L'_d.r_b$:=***Best***($L'_d.R_{List}$)
	  $L'_d.r_c$:=*null*
	  else
	  $L'_d.R_{List}$:=$L'_d.R_{List}-\{r_i\}$

Fig.4 API implementation algorithm of global forwarding table

图 4 协议路由更新处理过程

在 *Proto_Delroute* 删除路由过程中,r_i 为高层协议要从全局转发表中删除的路由.如果待删除的路由等于全局最优路由 $L'_d.r_g$,则在删除后 r_i 需要当前节点对全局最优路由进行预测,预测规则如下(按照序号顺序):

1. 如果路由 Cache 域 $L'_d.r_c$ 不为空,向其他节点广播 $L'_d.r_c$;
2. 如果本地最优路由域 $L'_d.r_b$ 不为空,向其他节点广播 $L'_d.r_b$;
3. 否则,向其他节点广播信息删除 r_i.

如果待删除的路由不是全局最优路由 $L'_d.r_g$,当前节点仅仅需要将路由 r_i 从叶节点的路由集合 $L'_d.R_{List}$ 中删除,必要时重新计算本地最优路由 $L'_d.r_b$,不需要向其他节点发送路由同步消息.

节点需要处理的路由同步消息只有两种:路由更新消息和路由撤销消息.图 5 给出了路由更新消息处理过程 *Node_Addroute* 和路由撤销消息处理过程 *Node_Delroute* 的算法伪码.其中 *srcid* 为发送路由同步消息的源节点 *ID*,r_i 为需要同步的路由信息.

当节点通过内部交换网络接收到更新路由 r_i 时,需要判断 r_i 是否会引起全局最优路由 $L'_d.r_g$ 的变化.如果 r_i 优于全局最优路由 $L'_d.r_g$,则将 $L'_d.r_g$ 更新为 r_i.如果发送路由同步消息的节点为当前全局最优路由 $L'_d.r_g$ 的来源节点,表明原有的全局最优路由发生了变化.这时路由 r_i 存在两种情况:

1. r_i 为源节点 *srcid* 根据路由 Cache 预测产生的候选路由.这时需要判断 r_i 是否为本节点的路由,如果 r_i 起源于本节点并且等于本地最优路由 $L'_d.r_b$,表明 r_i 仍然有效.如果 r_i 起源于本节点并且不等于本地最优路由 $L'_d.r_b$,则需要广播消息进行修正.
2. r_i 为源节点 *srcid* 新的本地最优路由;这时判断本地最优路由 $L'_d.r_b$ 是否优于 r_i,如果 $L'_d.r_b$>r_i,则广播 $L'_d.r_b$.

路由撤销过程 *Node_Delroute* 的算法比较简单,只需要判断 r_i 是否为全局最优路由.如果全局最优路由被删除,则广播本地最优路由 $L'_d.r_b$ 给其他节点.

```
Proc Node_Addroute(srcid,r_i)
  d:=Dest(r_i)
  L'_d:=GetLeafNode(d)
  if ((r_i>L'_d.r_g)OR(srcid≡ID(L'_d.r_g)))
    if (ID(r_i) ≠myid)
      if (r_i>L'_d.r_b)
        L'_d.r_g:=r_i
      else
        Broadcast(ADD,L'_d.r_b)
        L'_d.r_g:=L'_d.r_b
        if ((ID(r_i)≡srcid)AND(r_i>L'_d.r_c)AND
          (r_i>BEST(L'_d.R_List−{L'_d.r_b})))
          L'_d.r_c:=r_i
    else
      if (L'_d.r_b≠r_i)
        if (L'_d.r_b≡null)
          Broadcast(DEL,r_i)
        else
          Broadcast(ADD,L'_d.r_b)
        L'_d.r_g:=L'_d.r_b
  else
    if ((ID(r_i)≡srcid)AND(r_i>L'_d.r_c)
      AND(r_i>BEST(L'_d.R_List−{L'_d.r_b})))
      L'_d.r_c:=r_i
```

```
Proc Node_Delroute(srcid,r_i)
  d:=Dest(r_i)
  L'_d:=GetLeafNode(d)
  if (r_i≡L'_d.r_g)
    if (L'_d.r_b≠null)
      Broadcast(ADD,L'_d.r_b)
      L'_d.r_g:=L'_d.r_b
  else
    if (r_i≡L'_d.r_c)
      L'_d.r_c:=null
```

Fig.5 The procedures of route synchronization messages

图 5 路由同步消息处理过程

4 性能评价

为了模拟骨干网络环境中经常遇到的链路故障、协议路由抖动、用户更改配置造成的 IP 转发表路由剧烈变化的问题,本文构造了 3 种典型案例进行性能比较.将链路故障或者用户修改配置造成的转发表路由短时间内一次大范围抖动称为案例 1;将 BGP 协议的持续路由抖动造成 IP 转发表路由频繁切换的情况称为案例 2;最后将路由协议的猝发更新,使得转发表在短时间内注入大量路由称为案例 3.

针对案例 1 与案例 2,本文比较了广播更新方式、全冗余备份方式和 AREF 同步框架 3 种路由同步机制的性能差异;对于案例 3,由于广播更新方式在增加路由的处理上与 AREF 同步框架完全相同,为此只比较了全冗余备份方式和 AREF 同步框架在任意一个主动节点的本地转发表中注入大量路由时的性能差异.

性能模拟时使用网络模拟器 NS2,在 Linux RedHat 9.0 上编译运行.模拟使用与图 3 相同的节点连接方式,即所有的路由节点通过高速交换网络连接起来,任意两个节点都直接可达,并且假设内部交换网络支持单播和广播传输机制.为了消除路由项打包、定时重传、不同的路由项数据结构等具体实现方式对同步机制的性能影响,本文使用节点之间必须同步的路由项数量作为衡量各种算法的性能标准.

对于案例 1,由于链路故障或者用户修改配置通常只影响某个路由节点的本地转发表.图 6 给出了在集群路由器中随机选择一个主动节点,使其本地转发表发生一次 100K 路由抖动后产生的平均同步路由数.横坐标为构成集群路由器的主动节点数,纵坐标为以 100K 为单位的同步路由数.全冗余算法由于要通知所有节点本地转发表的变化,因此一次抖动需要同步的路由数恒定为 200K;AREF 算法与广播更新方式只有在全局最优路由发生变化时才需要路由同步,并且 AREF 算法在单个节点路由抖动时路由 Cache 的命中率非常高,因此在图中二者的性能曲线非常接近,AREF 算法的平均同步路由数只比广播更新方式多几十个.

BGP 协议的路由抖动问题会造成 IP 转发表路由的频繁切换,为此,本文比较了多个主动节点同时出现 IP 转发表抖动情况下的性能差异.图 7 给出了每个主动节点的 IP 转发表抖动概率为 0.5,主动节点数从 4 增长到 128 时,3 种同步机制的性能曲线.

图 8 比较了 128 个主动节点构成的集群路由器在不同抖动概率下的性能曲线.AREF 算法在抖动概率为 0.5

时,仍然与最优的广播更新方式比较接近.即使在 128 个主动节点并且抖动概率升高到 0.8 这种极端情况下,AREF 算法的路由同步开销也只占全冗余方式的 25%左右.

图 9 比较了案例 3 中全冗余备份方式与 AREF 算法的性能差异.随机选择一个节点,在极短时间内向其本地转发表中注入 100K 新路由.AREF 只在全局最优路由发生变化时,才发送路由同步消息.因此,随着主动节点数的增加,AREF 算法产生的同步路由数迅速下降,远远优于全冗余算法.

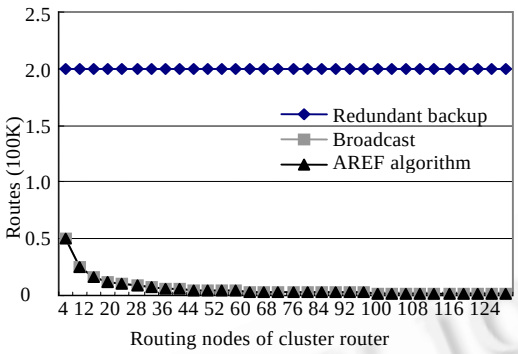

Fig.6 Route flapping of an active node

图 6 单个节点路由波动时的性能差异

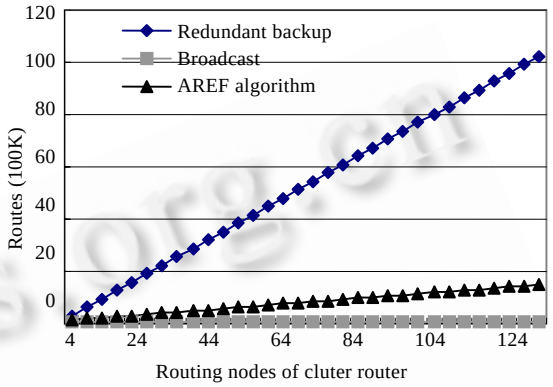

Fig.7 All active nodes flapping in probability 0.5

图 7 每个节点路由抖动概率为 0.5 时的性能差异

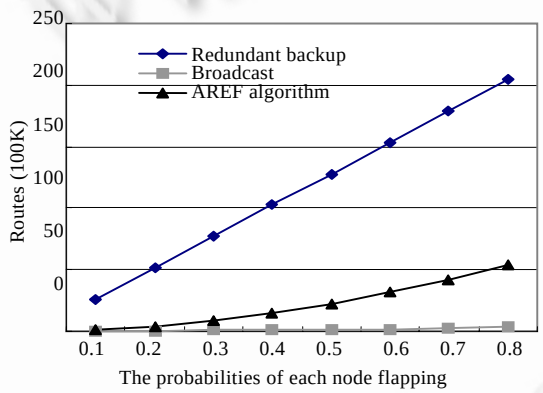

Fig.8 128 active nodes in different flapping probabilities

图 8 128 个节点在不同抖动概率条件下的性能差异

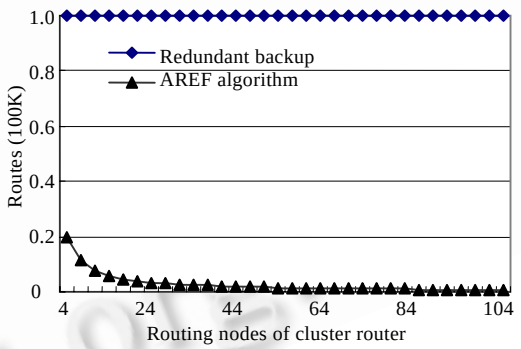

Fig.9 Burst insertion of 100K routes

图 9 注入 100K 路由时的同步路由数

在国防科学技术大学研制的高性能核心路由器项目中,一个 IP 转发表路由项数据结构占用 13 个长整数,共 52 个字节.假设图 9 中注入的 100K 路由是通过 BGP 协议加入到集群路由器某个路由节点的 IP 转发表中,核心路由器的 BGP 协议处理路由更新的能力为 5300(个路由/秒),则全冗余方式占用的内部交换网络的带宽为 2.2M.AREF 算法只在全局最优路由发生变化时才发送路由同步消息.在图 9 中,最坏情况下占用的网络带宽为 0.44M.最好情况下占用的带宽只有 0.017M.可以看出,AREF 同步框架可以有效地降低路由同步的通信开销.

5 结论及下一步工作

转发表单映像同步机制对集群路由器控制平面及报文转发平面的性能至关重要.本文在分析现有的各种转发表同步机制特点的基础上,提出一种非对称的 AREF 路由同步框架.与其他同步机制相比,具有如下一些特点:(1) 支持异构的集群路由器硬件平台;(2) 支持灵活的协议分布方式;(3) 强大的可扩展性;(4) 较低的硬件资源需求;(5) 同步开销比较低.在 AREF 路由同步框架的基础上,为了降低路由同步开销,本文针对骨干网络经常遇到的路由抖动问题,提出了 AREF 路由同步算法.算法针对每个路由前缀使用路由 Cache 来缓存次优路由,在

全局最优路由被删除时,通过快速切换到路由 Cache 缓存的路由以减少路由抖动时的转发表同步开销.模拟实验表明,AREF 路由同步算法可以有效地降低增加、删除路由时的同步开销,在多个节点同时出现路由抖动的极端情况下,AREF 路由同步算法的性能远远优于全冗余路由同步机制,与最优的广播更新方式比较接近,非常适合于异构的大规模并行集群路由器系统.

已有的集群路由器项目由于节点规模有限,所有路由节点都处于同一个连接平面上.随着集群规模的扩大,这种平面结构的连接关系很难容纳数量庞大的路由节点,需要使用具有层次结构的多连接平面.在这种应用背景下,路由节点之间转发表单映像维护问题与平面结构下有很大不同.本文下一步的工作要继续深入研究具有层次结构连接关系的集群路由器的转发表单映像维护问题.

致谢 在此,我们向审稿老师严谨的评审以及对本文提出很多建设性的意见表示诚挚的感谢.对学报编辑老师的辛勤工作表示由衷的感谢.

References:

[1] Cisco Networks. 2004. http://www.cisco.com

[2] Juniper Networks. 2004. http://www.juniper.net

[3] Avici Systems Technology. 2003. http://www.avici.com

[4] Halabi S. Pluris massively parallel routing. White Paper, Pluris Inc., 1999.

[5] Tzi-cker C, Prashant P. Suez: A cluster-based scalable real-time packet router. In: Wen-Tsuen C, eds. Proc. of the the 20th Int'l Conf. on Distributed Computing Systems (ICDCS 2000).Washington: IEEE Computer Society, 2000. 136−145.

[6] Gong ZH,Sun ZG. CRA: Cluster router architecture. Technical Report, Changsha: National University of Defense Technology, 2004 (in Chinese with English abstract).

[7] Maruyama M, Takahashi N, Miei T. CORErouter-1: An experiental parallel IP router using a cluster of workstations. IEICE Trans. on Commun., 1997,E80-B(10):1407−1414.

[8] Xiao XP, Ni LM. Parallel routing table computation for scalable IP routers. In: Panda DK, Stunkel CB, eds. Proc. of the IEEE Int'l Workshop on CANPC. LasVegas: Springer-Verlag, 1998. 144−158.

[9] Zhang XZ, Zhu PD, Lu XC. Fully-Distributed and highly-parallelized implementation model of BGP4 based on clustered routers. In: Lorenz P, Dini P, eds. Proc. of the 4th Int'l Conf. on Networking. Springer-Verlag, 2005. 433−441.

[10] Iannaccone G, Chuah CN, Mortier R, Bhattacharyya S, Diot C. Analysis of link failures in an IP backbone. In: Diot C, eds. Proc. of the 2nd ACM SIGCOMM Workshop on Internet Measurment Table of Contents. Marseille: ACM Press, 2002. 237−242.

[11] Labovitz C, Malan GR, Jahanian F. Internet routing instability. IEEE/ACM Trans. on Networking, 1998,6(5):515−527.

[12] Labovitz C, Malan GR, Jahanian F. Origins of Internet routing instability. In: Proc. of the IEEE INFOCOM'99. New York: IEEE, 1999. 218−226.

[13] Liang ZY, Xu K, Wu JP, Xu MW. Routing management model in distributed routers. Journal of Tsinghua University (Sci & Tech), 2003,43(4):503−506 (in Chinese with English abstract).

附中文参考文献:

[6] 龚正虎,孙志刚.机群路由器体系结构.研究报告,长沙:国防科学技术大学,2004.

[13] 梁志勇,徐恪,吴建平,徐明伟.分布式路由器中的路由管理模型.清华大学学报(自然科学版),2003,43(4):503−506.

张晓哲(1976－),男,博士生,主要研究领域为路由协议,路由器软件系统,路由查找算法.

朱培栋(1971－),男,博士,副教授,主要研究领域为路由技术,移动网络,网络安全.

卢锡城(1946－),男,教授,博士生导师,中国工程院院士,CCF 高级会员,主要研究领域为高性能计算,并行与分布处理,先进网络技术.

彭伟(1972－),男,博士,副教授,主要研究领域为路由协议,移动网络.