路由器分布式控制研究综述

徐明伟[1],江学智[1,2],陈文龙[3]
(1.清华大学计算机科学与技术系,北京 100084;2.石家庄机械化步兵学院,河北石家庄 050083;
3.北京科技大学信息工程学院,北京 100083)

摘 要: 光传输技术飞速发展和互联网流量快速增长对路由器性能提出了更高的需求.路由器经常因控制平面过载导致网络振荡,甚至路由器失效.为了克服路由器集中控制存在的问题,研究人员提出了路由器分布式控制方案.本文深入分析了路由器集中控制面临的问题,围绕路由器实现分布式控制需要解决分布式控制平面、分布式控制平面内部通信和分布式路由协议和算法这三个关键问题,综述和比较了现有的路由器分布式控制方案.最后对下一步工作进行了展望.

关键词: 路由器;分布式控制;内部通信;分布式路由协议和算法
中图分类号: TP393.05 **文献标识码:** A **文章编号:** 0372-2112 (2010) 08-1892-08

Survey on Distributed Control in a Router

XU Ming-Wei[1], JIANG Xue-Zhi[1,2], CHEN Wen-Long[3]
(1. *Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*;
2. *Shijiazhuang Mechanized Infantry Institute, Shijiazhuang, Hebei* 050083, *China*;
3. *School of Information Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*)

Abstract: The rapid development of optical transmission and fast growth of Internet traffics require higher-performance routers. The overload of control plane often causes the oscillation of network. In a serious case, it results in router crashing. To overcome the problems of the monolithic control plane, some schemes of distributed control in a router have been proposed. In this paper, we thoroughly analyzed the problems of centralized control plane. Based on the critical obstacles in realizing router's distributed control, we made summary and comparison of existing schemes from three aspects: distributed control plane, internal communication of distributed control plane and distributed implementation of routing protocol and algorithm. Finally, the future trend is discussed.

Key words: router; distributed control plane; internal communication; distributed routing protocol and algorithm

1 引言

光传输技术飞速发展和核心路由表快速增长对路由器性能提出了更高的需求.互联网快速发展要求路由器随着网络规模和流量增长不断扩展自身性能.虽然通过硬件升级在短期内能够提高路由器的性能,但受硬件性能限制,仅依靠硬件升级无法满足互联网快速发展的需要.为了克服硬件的性能束缚,一些路由器[1~3]在数据平面采用多机柜分布式互连的集群体系结构提高转发性能.但目前路由器控制平面只有一个控制单元处理控制任务,数据平面规模扩展将增加控制平面的负载,容易造成控制单元过载[4].

目前对路由器的研究大部分集中在数据平面[5],对于逐渐成为路由器性能瓶颈的控制平面缺乏成体系的研究.为了解决现有路由器控制平面基于单一控制单元的集中式控制所面临的问题,研究人员提出了路由器分布式控制方案.

为了更好地了解路由器集中控制与分布式控制的特点,我们从可靠性、可扩展性和部署代价等方面对它们进行了比较,如表1所示.

路由器控制平面分布式互连和多实例并行可有效地避免单一硬件或软件失效导致的网络振荡,提高网络稳定性和路由器容错能力.分布式控制能够支持性能和功能的灵活扩展,提高路由器的可扩展性.控制单元之间和控制单元与转发单元之间分担负载,可克服单一硬件的性能瓶颈,减少控制单元过载,提高路由器的可

收稿日期:2009-07-14;修回日期:2010-03-05
基金项目:国家重点基础研究发展计划(973)(No.2009CB320502);国家高技术研究发展计划(863)(No.2007AA01Z2A2,No.2009AA01Z251);国家科技支撑计划(No.2008BAH37B03)

靠性.硬件分布式互连和软件功能分布式、模块化设计和实现可实现不中断服务升级,提高路由器的可用性.但与集中式控制相比,分布式控制存在内部通信开销大和能耗高、管理和维护复杂等不足.

表 1　路由器集中式控制与分布式控制比较

内容 \ 控制类型	集中式	分布式
内部通信协议	专用、私有	开放、标准
可扩展性	软、硬件升级	分布式、模块化,可灵活地扩展控制平面功能和性能
可靠性	单点失效影响可靠性	分布式多实例并行和冗余备份提高了可靠性
并行性	无	不同任务并行;同一任务不同功能并行
可用性	不支持动态升级,影响可用性	可灵活对软、硬件升级,支持不中断服务升级
失效恢复	硬件修复或软件重启时间相对较长	分布式多实例并行,单一硬件和软件模块失效对系统功能影响小,可实现无缝恢复
网络变化感知速度	信令消息传输到控制平面和单进程增加了排队时间	信令功能分布在转发单元和并行处理减小了信令消息排队时间
网络稳定性	单一控制单元过载和故障容易引起网络振荡	分布式并行减少了硬件或软件故障引发的网络振荡
部署代价	低	高
内部通信开销	小	较大
维护管理	简单	较复杂
能量消耗	低	高

文献[5]虽然按照分层模型综述了可扩展路由器目前的研究进展,但它重点分析和比较了数据平面的扩展方案,而对控制平面—这个制约路由器可扩展的瓶颈和关键问题缺少系统和针对性分析.

本文深入剖析了路由器集中控制存在的局限性,总结出路由器控制平面从集中式向分布式发展需要解决三个关键问题:(1)分布式控制平面.由多个控制单元分布式互连而成的分布式控制平面可有效克服单一控制单元的性能瓶颈和可扩展性差等不足,为控制平面的性能和功能灵活扩展提供支持;(2)分布式控制平面内部通信.物理上分布的控制单元和软件功能模块要形成一个整体,需要分布式控制平面内部通信协议实现硬件和软件的透明通信;(3)分布式路由协议.为适应分布式控制平面体系结构,路由协议和算法应分布式实现,充分利用系统的计算和存储资源,提高路由器的性能.本文重点围绕这三个关键问题综述了这一领域的最新研究进展,并对各种方案进行了分析和比较.

2　路由器集中控制面临的主要问题

2.1　性能瓶颈

目前路由器控制平面只有一个控制单元处理协议分组.Iannaccone.G 等人[6]通过网络测量得出,50%的网络故障可能因路由器控制平面过载丢失"心跳"消息引起.根据目前互联网的发展速度和硬件技术发展速度,基于单一控制单元的集中式控制平面很难满足互联网快速增长的需求.

2.2　单点失效

现有路由协议大部分集中在主控制单元上运行,很容易因硬件或软件局部功能失效或代码错误导致整个协议失效,例如:邻居建立与维护功能失效将导致整个协议失效.虽然现有路由器控制平面采用主、从备份方式,但是主、从备份的失效恢复速度相对较慢,影响了网络的可用性.为提高网络可用性,目前通过向网络中增加路由器和运行虚拟路由器冗余协议(Virtual Router Redundancy Protocol,简称 VRRP)实现冗余备份.这提高了网络的运营成本,增加了网络连接的复杂度和网络管理的难度.

2.3　可扩展性差

各路由器厂商都采用私有技术,分别设计各自专用的部件、接口和通信协议.不同生产厂商的路由器部件之间不能互换和通信.因此,在网络中,这些路由器只能作为独立的网络设备互连,而不能通过互连扩展为一台更高性能和更多功能的路由器.技术封闭私有和集中控制严重影响了路由器的可扩展性.

基于路由器集中控制所面临的问题,研究人员提出了路由器分布式控制方案.通过分布式互连、并行处理和冗余备份等技术提高路由器的性能、可靠性和可扩展性.

虽然路由器分布控制是一种发展趋势,但是随着硬件处理能力的不断提高,集中式还将长期存在.Ballani,Hitesh 等人[7]研究表明两种控制方式相结合将有效延长现有路由器的生存周期,改进网络性能.

3　分布式控制平面

路由器分布式控制平面主要分为集群路由器(Cluster Router,简称 CR)[8~11]和转发与控制分离(Forwarding and Control Elements Separation,简称 ForCES)[12,13]两种结构.

3.1　CR

CR 是多个可独立运行的、具有路由功能的节点通过某种互连结构(例如:高速以太网)连接成性能、功能可扩展的单映像路由器.集群路由器根据内部节点的类型可分为软件集群路由器和集群路由器.

软件集群路由器由一台路由器与多台具有路由处理能力的 PC 机相连而成.软件集群路由器通过向集群中增加路由处理节点扩展路由器的性能和功能.目前比较典型的软件集群路由器有 CLARA[8]、Suez[9]和 VERA[10].软件集群路由器借助多个路由处理节点并行,实现分布式控制单元间的负载分担,提高路由处理性能.多台 PC 机和一台路由器组成的分布式控制平面可实现功能和性能灵活扩展,部署代价比较低.但软件集群路由器不支持数据平面的扩展,无法提高数据平面性能.在软件集群路由器中,路由器节点失效会导致整个系统不可用,降低了系统的可靠性.

与软件集群路由器相比,集群路由器 HCR[11]的内部节点都是具有路由和转发功能的路由器.在集群路由器内部,节点之间分担负载,并行处理不同协议的分组.分布式控制单元之间功能可相互冗余备份.能够支持数据平面的灵活扩展.

由于集群路由器的最小组成单元是路由器,控制单元与转发单元之间采用私有协议和专用接口通信,因此,内部节点的互连需要对路由器进行相应的修改,增加了部署代价.

3.2 ForCES

ForCES[12]体系结构如图 1 所示.ForCES 体系结构允许一台路由器的控制平面有多个控制单元.它们之间通过内部高速网络互连.控制单元(CE)与转发单元(FE)可采取预先配置或动态联合的方式实现路由器相应的功能.CE 和 FE 间的自由联合可灵活扩展路由器的性能和功能.控制单元之间可彼此分担负载、冗余备份和实现分布式控制.ForCES 体系结构为一台路由器内部多个控制单元分布式互连和规模扩展提供了一种灵活的机制.FE/CE、FE/FE、CE/CE 间的标准接口和 FE/CE 间的 ForCES 通信协议为控制平面的性能和功能灵活扩展提供了支持.通过标准化的机制,CE 和 FE 变成相互分离的标准组件,克服了集群路由器控制单元和转发单元不能分离的不足.可对 CE 和 FE 数量灵活扩展,有效克服单个部件性能的束缚,延长路由器生存周期,保护投资.

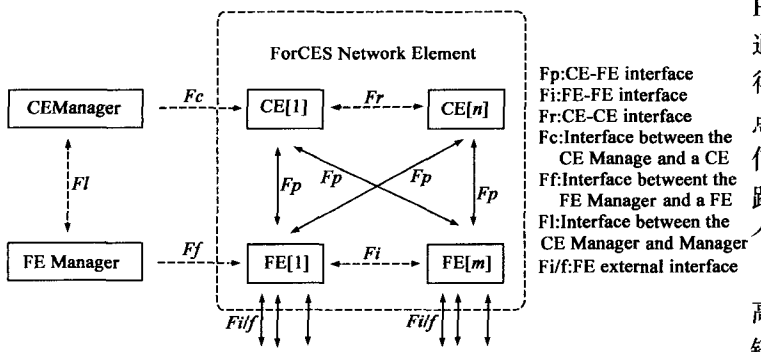


图1 ForCES体系结构

ForCES 方案虽然为路由器实现控制单元、数据单元的规模扩展和扩展路由器的功能提供了一种灵活机制,但是,它只提出了分布式控制平面体系结构和 CE 与 FE 之间的 ForCES 通信协议,没有对路由软件进行分布式、模块化设计.

DMR (Decentralized Modular Router,简称 DMR)[13]是基于 ForCES 框架实现的分布式模块化路由器.在 DMR 中,CE 和 FE 是独立的标准组件,设计有标准的 ForCES 接口,相互通过内部高速以太网互连.DMR 路由器支持控制单元数量和控制平面功能的灵活扩展.通过内部通信协议 Forz(ForCES on zebra,简称 Forz)[14],多个分布式互连的控制单元聚合为一个整体,彼此分担负载和冗余备份.DMR 对路由协议进行了功能分解和模块化设计,将路由协议的邻居建立与维护功能迁移到转发单元上实现,利用转发单元的处理资源分担控制平面的负载,实现了路由协议分组的并行处理,提高了路由器对网络变化的感知能力和路由协议的可用性.多个控制单元分布式互连和路由协议的功能分布可提高路由器的可靠性.

由于 CE 与 FE 的标准化需要一段较长的时间,在短时期内,这种方案还很难体现它的优势.但是,DMR 实现了基于 FE 与 CE 分离的多个控制单元间的分布式互连的原型系统.为路由器分布式控制平面的设计提供了很好的参考和借鉴作用.

4 分布式控制平面内部通信

随着路由器控制平面由集中式向分布式发展,为了使物理上分布式互连的控制单元和转发单元组合成一台完整的路由器,需要设计和实现分布式控制平面内部通信协议.目前分布式控制平面内部通信方案主要有集群路由器内部通信协议(Router Cluster Protocol,简称 RCP)[15]和 Forz 两种.

4.1 RCP

J.B-Guan 等人设计了集群路由器内部通信协议 RCP,其内部接口和协议框架模型如图 2 所示.通过 RCP 协议,集群路由器内部各节点都可获得组成该集群路由器的节点数量、编号、邻居节点的性能和类型、内部端口号和外部端口号等信息.集群路由器内部各节点形成对整个集群路由器一致的内部拓扑视图.在转发平面上,多个路由器节点通过标准互连卡互连.

集群路由器内部互连接口的物理层采用高速光互连技术(Very Short Reach,简称 VSR);链路层使用常用的高级数据链路控制(High Level Data Link Control,简称 HDLC)帧封装格

式.通过制定多种速率等级备选,满足了集群路由器内部不同速率互连的需要,提高了互连的灵活性.传输的数据帧中包含符合 NPSI[16]规范的数据信元和流控信元两种格式.数据信元交换数据报文,流控信元交换集群路由器内部各个交换网络的流量控制信息.数据信元头部附加一个包含信元全局目的端口信息的标签,在它所经过的每一个交换网络节点的内部互连卡上,互连卡根据全局端口设备视图与本地交换网络端口的映射关系确定本地交换网络的目的端口,将信元送往本地交换网络进行交换.

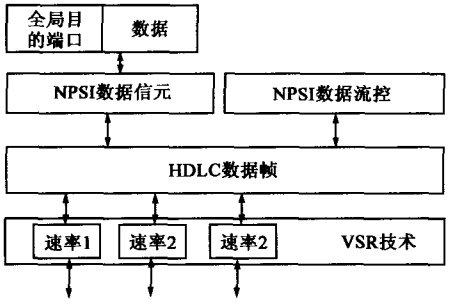


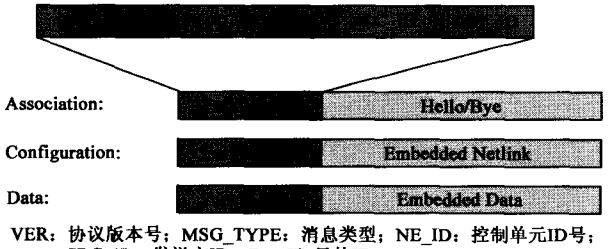
图2 RCP协议内部接口和协议框架模型

在控制平面,多个控制单元通过 RCP 协议交换拓扑信息,形成统一的集群路由器管理视图和设备视图,相互间协同路由计算和协议处理,同步转发表等功能.

由于 RCP 方案只设计了集群路由器内部通信协议的框架,没有可参考的协议的具体工作机制,而且它只适用于集群路由器内部节点间的通信和数据交换,具有一定的局限性.

4.2 Forz

O. Hagsand 等人基于开源路由软件 Zebra[17] 和 ForCES 体系结构设计了路由器分布式控制平面内部的通信协议 Forz. Forz 协议在 ForCES 协议框架的基础上,对其内部通信机制进行了扩展,可实现 CE/FE, CE/CE, FE/FE 间的通信,其消息格式如图 3 所示. Forz 协议的通信机制分为:联合、配置和数据传输三个阶段.



VER: 协议版本号; MSG_TYPE: 消息类型; NE_ID: 控制单元ID号; SRC_ID: 发送方ID; Dsr_ID:目的ID

图3 Forz协议消息格式

在联合阶段,每个 CE 或 FE 通过 IP 可靠组播的方式向同组内的其它成员发送 Hello 消息,报告自身的信息. Hello 机制管理成员的加入和离开,通告内部控制和数据流的组播地址、成员的信息和心跳检测.

当一个成员刚加入时,它的初始 Forz 数据库为空.首先,它向内部控制流的组播地址发送一个 Hello 消息,Hello 消息中包括自身的性能、端口数量、类型等信息.收到 Hello 消息后,组内的其它成员将新成员的信息加入到自己的数据库.同时,组内现有成员向新加入的成员发送 Hello 消息通告自己的功能、资源和端口地址等信息.收到消息后,新加入的成员向自己本地的 Forz 数据库中添加相应成员的信息.当一个成员离开时,它发送 Bye 消息,组内其它成员收到这个消息后,将它的信息从数据库中删除.当一个成员失效时,由于缺少心跳消息,其它的成员将其从数据库中删除.

在配置阶段,Forz 协议用于创建、删除和获取网络接口、IP 地址、路由信息和邻居的信息等. Forz 协议通过可靠组播对配置信息进行分发.建立本地端口与全局端口的映射关系,形成内部一致的路由表.

在数据传输阶段,Forz 协议对本地数据进行封装,然后通过内部网络交换到相应的输出端口,在输出端口解封装后发送到外部网络.

Forz 实现了分布式控制平面内部通信的工作机制和协议消息格式;对分布式控制平面内部路由和通信机制的设计具有很好的参考和借鉴作用.

4.3 小结

现有的分布式控制平面方案及其内部通信协议如表 2 所示.

集中式控制平面路由器内部通信采用私有协议,技术成熟,部署代价相对较低,但性能和功能相对固定,可扩展性差.集群路由器分布式控制单元间采用 RCP 协议通信.但它需要对部分路由器进行修改,增加了部署代价. ForCES 和 DMR 内部采用标准的接口和协议进行通信.基于标准化组件容易实现对控制单元的功能定制,支持性能和功能的灵活扩展.但标准化设计是一个长期的过程,因此部署代价目前相对较高.

表 2 路由器控制平面体系结构与内部通信

内容 / 方案	体系结构	控制单元数量	内部通信协议	部署代价	可扩展性
集中式路由器	集中式	2 个(主、从)	私有	低	低
CR	分布式	多个	RCP	中	中
ForCES	分布式	多个	ForCES	高	高
DMR	分布式	多个	Forz	高	高

5 分布式路由协议和算法

现有路由协议的功能大都集中在控制平面.为充分利用分布式控制单元和转发单元的资源,需要对路由协进行功能分解和分布式设计,将不同的功能分布

在相应的控制单元或转发单元上实现.设计分布式路由算法,充分利用多个控制单元的计算资源,提高路由计算性能.

5.1 分布式路由协议

5.1.1 DCP

M.Deval 等人[18]提出了分布式控制方案(Distributed Control Plane,简称 DCP).将路由协议的功能分为三类:(1)链路相关功能.主要包括分组转发和邻居状态维护.这些功能可分布在转发单元上实现,利用转发单元的处理资源分担控制单元的负载.(2)协议处理功能.例如:路由计算,协议状态机的维护.这些功能需要在多个控制单元间实现分布.(3)更新控制信息功能.例如,更新路由表.这些功能很难进行分布式设计,应在控制单元实现.DCP 方案将 OSPF[19]的 Hello 机制迁移在转发单元实现,有效地利用了转发单元的处理资源处理协议的信令分组,分担了控制单元的负载.减少了 Hello 分组到控制平面的传输时间,缩短了等待控制单元处理的排队时间,加快了路由器对网络故障的感知和响应.信令功能分布在转发单元实现可避免转发单元规模扩展而导致的内部通信开销,缓解了数据平面扩展对控制平面处理能力的需求.协议信令功能的分布和并行,提高了协议的可靠性和容错性.

DCP 虽然提出了路由协议功能分布的原则,但缺乏对路由协议功能分布和模块化设计的细节.

5.1.2 MCPB

模块化 BGP(Modular Control Plane for BGP,简称 MCPB)[20]按照功能将 BGP[21]协议划分为信令模块(BGP Session Manager)和路由处理模块(BGP Processing).信令模块分布在转发单元,用于维护邻居关系、接收邻居路由通告和处理"心跳"消息、分配路由计算任务.路由处理模块分布在控制单元,进行路由计算.通过对 BGP 协议功能分解和模块化设计,MCPB 方案可实现对 BGP 协议分组并行处理,能够利用转发单元的处理资源分担控制单元的负载.BGP 相应的功能模块分布在不同的控制单元或转发单元上并行运行提高了可靠性和可扩展性.

5.1.3 DCPA

K-K.Nguyen 等人[22]设计了路由软件的分布式控制平面体系结构(Distributed Control Plane Architecture,简称 DCPA).分别对 OSPF/ISIS 和 BGP 等路由协议进行了功能分解和模块化设计.例如:将 OSPF 协议分解为 OCC(OSPF Control Component)模块和 OSC(OSPF Signaling Component)模块.OSC 模块分布在转发单元,OCC 模块分布在控制单元.每个控制单元上运行的路由协议模块在其它控制单元上都相应地备份,提高了可靠性.DCPA 将 BGP 协议分解为 BGP 邻居建立与维护模块、本地路由管理(L-RTM)模块和全局路由管理(G-RTM)模块.邻居建立与维护模块分布在转发单元,L-RTM 和 G-RTM 分布在不同的控制单元.每个控制单元的 BGP 邻居建立与维护模块接收来自邻居的路由信息,L-RTM 先决策出本地最优路由,然后发送给 G-RTM 模块,G-RTM 模块从各个 L-RTM 发送的本地最优路由中计算出全局最优路由.BGP 路由协议功能的分布和模块化设计提高了控制平面的可扩展性.

DCPA 方案为路由协议的功能分布和模块化设计提供了很好的参考.由于 DCPA 方案只是从功能上考虑了路由协议的功能分布,没有考虑路由协议模块相互之间的耦合度和通信开销,因此,在实际的路由协议功能分布和模块化设计中需要考虑这些因素.

5.1.4 DRTM

K.Khoa 等人[23]对路由表管理进行了分布式、模块化设计(Distributed RTM,简称 DRTM).DRTM 方案将路由表管理分解为线卡路由表管理模块(LC-RTM)和全局路由表管理模块(G-RTM).LC-RTM 分布在线卡.线卡维护它所在区域的链路状态数据库.当多个线卡具有相同的拓扑信息时,这些线卡组成一个集群,在集群中选举超级节点负责计算重叠区域的路由表.G-RTM 分布在控制单元.G-RTM 模块接收各 LC-RTM 模块发送的路由信息最终计算出全局路由表.

DRTM 方案充分利用数据平面的计算和存储资源来提高路由表的计算效率.但目前路由器的线卡不具备路由计算能力,因此,该方案不适合于常规路由器,实现代价比较高.另外,它在一台集中式控制路由器的控制单元与转发单元之间实现路由表计算的分布式,无法避免控制单元或 G-RTM 模块导致的单点失效,降低了可靠性.

5.1.5 小结

路由协议的分布式和模块化设计方案如表 3 所示.

表 3 路由协议功能分布式设计

方案 / 内容	现有协议	DCP	MCPB	DCPA	DRTM
体系结构	集中式	分布式	分布式	分布式	分布式
内部通信协议	私有	标准化	标准化	标准化	私有
路由协议	-	OSPF	BGP	OSPF,BGP 等	RTM
可靠性	低	中	高	高	中
可扩展性	差	好	较好	好	较好
并行性	无	中	好	好	中

目前路由协议大都集中在主控制单元,影响了可靠性,不能实现负载分担和并行处理.DCP 将路由协议信令功能迁移到转发单元,与集中式控制相比,在一定

程度上提高了系统的可靠性和并行处理能力.MCPB 和 DCPA 对现有协议功能进行分布式模块化设计,提高了并行性、可靠性和可扩展性.DRTM 基于高端路由器的线卡具备计算能力设计了路由表分布式管理方案,对线卡的性能要求较高.由于它在同一路由器的控制单元与线卡之间实现分布式路由表管理,影响了系统的可靠性.但是,利用转发单元计算路由较好地分担了控制单元的负载,提高了路由器的并行处理性能.

5.2 分布式路由算法

为了提高路由计算性能,应设计分布式路由算法,充分利用分布式控制平面多个控制单元的资源,提高路由计算性能.目前的分布式路由算法主要分为:(1)分布式并行 OSPF 路由算法,包括 PDSPT[24]、BPA[25]和 PRTC[26];(2)分布式并行 BGP 路由算法,主要有 FDHP[27]和 ITBGP[28].

5.2.1 分布式并行 OSPF 路由算法

(1)PDSPT

Zhu-Y.B 等人根据最短路径树(Shortest Path Tree,简称 SPT)增量更新的特点,利用原有的 SPT 树设计了并行动态 SPT 算法(Parallel Dynamic SPT,简称 PDSPT).每次网络拓扑发生变化,将原 SPT 树上受链路变化影响的节点加入队列,并分配给相应的控制单元.每个控制单元每次从本地选出距离增量最小的节点.控制单元之间利用广播方式选出全局距离增量最小的节点.每次选出全局距离增量最小的节点后,根据原 SPT 树上的父子关系,相应更新全局最小节点原 SPT 树上子孙的距离,并将它们从受影响节点的队列中删除.然后将新受影响的节点加入队列.反复迭代,直到受影响节点的队列为空.

PDSPT 算法利用分布式控制平面多个处理器资源分担计算负载,实现了 OSPF 路由表的并行计算.由于利用了 SPT 增量更新,减少了整个算法的迭代次数.

当网络拓扑规模比较大,并且链路故障对原 SPT 树上的节点影响比较多时,这个算法的并行性能较好.当网络拓扑中每个受影响节点的入度和出度比较接近时,控制单元的负载比较均衡.由于每次迭代控制单元间需要相互通信,因此,路由器内部通信开销比较大.受网络拓扑结构和迭代算法的影响,系统的负载均衡性差,影响了并行性能.

(2)PSPT

Zhang-X.P 等人针对 OSPF 协议设计了并行 SPT 算法(Parallel SPT Algorithm,简称 PSPT).PSPT 根据路由器中控制单元的数量 p 利用图分割理论将网络拓扑分割成几个区域,每个控制单元负责相应区域的路由计算.首先,每个控制单元并行计算出各自区域内所有边界节点到这些区域每个节点的最短路径.对于包含根节点的区域,计算根节点到这些区域中每个节点的最短路径.然后,将每个区域的边界节点和根节点组成新的拓扑图,再计算根节点到这个区域边界节点的最短路径;最后并行地将根节点到每个区域边界节点的路径和区域边界节点到区域内每个节点的最短路径合并,生成最终的 SPT 树.

每次拓扑变化,PSPT 算法需要重新分割网络拓扑,算法复杂度高.因此,PSPT 算法不适用于网络拓扑频繁变化时的 SPT 计算.基于图分割能够较好地实现负载均衡,并行性能较好.PSPT 算法在网络拓扑规模比较大时能够获得较好的性能.

(3)PRTC

Xiao-X.P 等人根据集群路由器的分布式控制平面提出了并行 OSPF 路由算法(Parallel Routing Table Computation,简称 PRTC).利用集群路由器多个路由节点将 OSPF 区域按照收集的拓扑信息进行分割,每个节点负责维护自己所在区域的拓扑,各自计算这个区域的路由表.当多个路由节点的路由区域重叠时,选举指派节点计算重叠区域的路由表.每个区域的指派节点向所有节点通告自己的路由表.每个节点选择性地将接收的路由合并生成最终的路由表.当自己所在区域的路由变化,每个指派路由节点向其它路由节点广播变化的路由.

PRTC 算法可实现 OSPF 路由表的并行计算,但它对网络拓扑依赖比较大.由于每个节点在网络中所处的位置不同,它们各自的计算负载不同,不能很好地实现节点间的负载均衡,并行性较低.每个节点不维护全局的链路状态数据库,一定程度上降低了可靠性.

5.2.2 分布式并行 BGP 路由算法

(1)FDHP

Zhang-X.Zh 等人设计了 BGP 并行路由算法(Full Distributed High Parallelized BGP,简称 FDHP).集群路由器的每个路由节点分别充当与它相连的 BGP 邻居的代理,负责与其相连的 BGP 邻居交换路由信息.每个节点根据本地路由信息计算本地最优路由.每个节点将本地最优路由广播给集群中其它节点,从而保证每个路由节点都维护一致的全局路由表.当集群中某个节点失效时,将它代理的 BGP 邻居会话和路由计算任务重新分配给其它节点.

FDHP 方案有效地利用了集群内部各个路由节点的计算资源和存储资源.每个路由节点只计算和存储一部分 BGP 候选路由,多个路由节点分担负载提高了路由计算性能.但节点间以广播的方式同步路由信息增加了内部的通信开销.

(2)ITBGP

Wu-Kun 等人设计了“迭代树”BGP 并行路由算法(Iterative Tree BGP,ITBGP).ITBGP 方案根据 BGP 邻居的数量 n 和每个控制单元的 BGP 邻居数量,采用广度优先算法构建一棵 k 阶迭代树.每当路由发生变化,树中相应的叶子节点首先计算出本地最优路由,然后发送给自己的父节点,这样沿着叶子向根的方向反复迭代,最后在根节点计算出全局最优路由.

当负载分布比较均衡时,ITBGP 算法的性能最优.当负载分布不均衡时,例如,到某一目的地的路由只存储在一个内部节点时,基于树形结构需要从叶子到根的方向多次迭代计算,降低了路由计算效率.

5.2.3 小结

PDSPT 将计算 SPT 树步骤中计算邻居节点的距离和搜索全局距离最小节点实现了并行处理,但算法复杂度较高,不能实现较好的负载均衡.PSPT 将拓扑进行分割,能够较好实现路由计算的并行和负载均衡,但算法复杂度较高.PRTC 对网络拓扑进行分割,算法性能受网络拓扑结构影响比较大,负载均衡性能差,但算法复杂度低.

FDHP 按邻居会话将 BGP 路由计算任务进行划分,能够较好地实现 BGP 路由计算并行,算法复杂度低.采用广播的方式进行路由信息同步,减少了路由计算的迭代次数.多个节点并行计算,提高了系统的并行性能.但是,在互联网中,由于多个邻居会通告同一故障,基于广播的方式进行路由信息同步将导致内部大量的通信开销.ITBGP 虽然按邻居会话划分负载并行计算 BGP 路由,但需要经过多次迭代才能计算出全局最优路由,随着树高度增加,内部通信开销和延时不断增大,影响了系统的并行性能.

6 结论和研究展望

路由器控制平面的分布式实现是一种发展趋势,也是路由器体系结构及软件设计必须解决的一个关键性技术问题.通过以上分析和比较,本文认为实现分布式控制平面应该从以下几个方面入手:

(1)分布式控制平面体系结构是实现路由器分布式控制的基础,它为标准化接口和内部通信设计提供依据,为控制平面性能和功能的灵活扩展提供支持.

(2)标准化接口和通信机制是实现与互连技术无关的关键技术,是控制单元与转发单元数量和功能灵活扩展的基本保证.

(3)软件分布式、模块化设计是提高路由器性能,支持路由器功能和性能动态扩展的主要途径.合理设计将有利于提高路由器的可用性和可扩展性.

虽然目前提出了一些路由器分布式控制方案,但是路由器实现分布式控制仍然面临一些关键问题亟待解决,需要进一步深入研究:(1)任务分配.如何将原来并行运行在一个控制单元上的多个路由协议任务合理地分布到多个分布式的控制单元,这需要研究分布式控制平面的任务分配方案.即要考虑每个任务对 CPU 的占用时间,又要考虑不同路由协议模块之间的通信开销,实现负载均衡,使内部通信开销最小,从而优化整个系统的性能.(2)分布式路由算法.应设计高效的分布式路由算法,充分利用各个节点的计算和存储资源,提高系统的性能.

参考文献:

[1] T640 routing node and TX Matrix platform: Architecture[OL]. http://www.juniper.net.

[2] Getting started guide on the Cisco CRS-1 series carrier routing system[OL]. http://www.cisco.com.

[3] The Avici TSR: the world's first scalable router[OL]. http://www.avici.com.

[4] 徐恪,吴鲲,王青青.可扩展路由器控制平面的高性能通信模型[J].软件学报,2007,18(9):2205 - 2215.
Xu Ke, Wu Kun, Wang Qing-qing. The high performance communication model of the scalable router control plane[J]. Journal of Software, 2007, 18(9): 2205 - 2215. (in Chinese)

[5] 张小平,刘振华,赵有健,关洪涛.可扩展路由器[J].软件学报,2008,19(6):1452 - 1464.
Zhang Xiao-ping, Liu Zhen-hua, Zhao You-jian, Guan Hong-tao. The scalable router[J]. Journal of Software, 2008, 19(6): 1452 - 1464. (in Chinese)

[6] Iannaccone. G et al. Analysis of link failures in an IP backbone [A]. In Proceedings of the ACM IMW'02[C]. NewYork: ACM, 2002. 237 - 242.

[7] Ballani Hitesh et al. Making routers last longer with ViAggre [A]. In Proceedings of NSDI'09[C]. Boston, Massachusetts: USENIX Association, 2009. 453 - 466.

[8] Welling G, Ott M, Mathur S. CLARA: a cluster-based active router architecture[J]. IEEE Micro, 2001, 21(1): 16 - 25.

[9] Chiueh T, Pradhan P. Suez: a Cluster-based scalable real-time packet router[A]. In Proceedings of ICDCS'00[C]. Washington, DC: IEEE Computer Society, 2000. 136 - 150.

[10] Karlin S, Peterson L. VERA: An extensible router architecture [J]. Computer Networks, 2002, 38 (3): 227 - 293.

[11] 管剑波.集群路由器体系结构及其关键技术研究[D].国防科学技术大学,长沙,2005.
Guan Jian-bo. Research on the Architecture and Key Technologies of Cluster-Based Routers[D]. National University of Defense Technology, Changsha. 2005. (in Chinese)

[12] L Yang, R Dantu, R Gopal. Forwarding and control element separation (ForCES) framework[OL]. IETF RFC 3746, http://www.ietf.org/rfc/rfc3746.txt, 2004.

[13] Markus Hidell. Decentralized modular router architecture[D]. KTH-Royal Institute of Technology, 2006.

[14] O. Hagsand et al. Design and implementation of a distributed router[A]. In Proceedings of IEEE ISSPIT[C]. Athens, Greece: IEEE Computer Society, 2005. 227 - 232.

[15] 管剑波;胡晓峰. 基于开放接口的异构路由器集群[J]. 信息工程大学学报, 2009, 10(1): 77 - 84.
Guan Jian-bo, Hu Xiao-feng. Heterogeneous cluster router based on open interface[J]. Journal of Information Engineering University, 2009, 10(1): 77 - 84. (in Chinese)

[16] Streaming interface (NPSI) implementation agreement[R]. Network Processing Forum Hardware Working Group, http://www. oiforum. com/public/documents/HWStreamingIA. pdf, 2002.

[17] Zebra[CP]. http://www. zebra. org/.

[18] M. Deval, H. Khosravi et al. Distributed control plane architecture for network elements[J]. Intel® Technology Journal, 2003, 7(4): 51 - 63.

[19] J. Moy, OSPF version 2[OL], IETF RFC 2328, http://www. faqs. org/rfcs/rfc2328. txt. 1998.

[20] Markus Hidell et al. A modularized control plane for BGP [A]. In Proceedings of the 19th IASTED[C]. Cambridge, Massachusetts: ACTA Press, 2007. 168 - 175.

[21] Y Rekhter, T Li, S Hares: A border gateway protocol 4(BGP-4)[OL]. IETF RFC 4271, http://www. faqs. org/rfcs/rfc4271. txt, 2006.

[22] K-K. Nguyen et al. Toward a distributed control plane architecture for next generation routers[A]. In Proceedings of E-CUMN '07[C]. Washington, DC: IEEE Computer Society, 2007. 173 - 182.

[23] Kim-Khoa, Nguyen Jaumard, B. Agarwal, A. A distributed and scalable routing table manager for the next generation of IP routers[J]. IEEE Network, 2008, 22(2): 6 - 14.

[24] Yuanbo Zhu et al. Parallel dynamic SPT update algorithm in OSPF[A]. In Proceedings of PaCT'07[C]. Berlin: Springer, 2007. 346 - 359.

[25] 张小平,吴建平等. 可扩展路由器中 SPT 并行计算的实现[J]. 电子学报, 2007, 35(11): 2129 - 2134.
Zhang Xiao-ping Wu Jian-ping et al. An implementation for parallel computing SPT in cluster router[J]. Acta Electronica Sinica, 2007, 35(11): 2129 - 2134. (in Chinese)

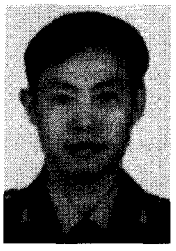
[26] Xi-peng Xiao, Lionel M. Ni. Parallel routing table computation for scalable IP routers[A]. In Proceedings of CANPC'98 [C]. London, UK: Springer-Verlag, 1998. 144 - 158.

[27] Xiao-zhe Zhang, Pei-dong Zhu, Xi-cheng Lu. Fully-distributed and highly-parallelized implementation model of BGP4 based on clustered routers[A]. In Proceedings of ICN'05[C]. New York: Springer, 2005. 433 - 441.

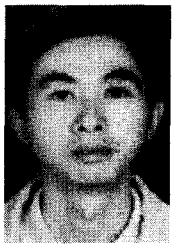
[28] Wu Kun, Wu Jian-ping, Xu Ke. A tree-based distributed model for BGP route processing[A]. In Proceedings of HPCC [C]. Berlin: Springer, 2006. 119 - 128.

作者简介:

徐明伟　男, 1971 年生于辽宁朝阳. 博士. 教授. 博士生导师. 研究方向为计算机网络体系结构、高速路由器体系结构、互联网路由.
E-mail: xmw@csnet1. cs. tsinghua. edu. cn



江学智　男, 1975 年生于江西湖口. 石家庄机械化步兵学院教研部讲师. 清华大学博士生, 研究方向为高速路由器体系结构、路由协议和算法.
E-mail: jxz@csnet1. cs. tsinghua. edu. cn



陈文龙　男, 1976 年出生于江西吉安, 博士生. 主要研究领域为网络体系结构和网络通信协议.
E-mail: wenlongchen@sina. com