文章编号:1001 - 2486(2006)03 - 0040 - 04

软件集群路由器体系结构的研究*

龚正虎,傅 彬,卢泽新
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:分析了集群路由器的研究现状,而后提出了软件集群路由器的两个参考模型:SCR-RM(software-based cluster router reference model)和 PCR-RM(parallel cluster routing reference model),给出两个参考模型的具体描述。这两个参考模型将为集群路由器的后续研究提供参照系统、实验系统和 IP 流处理场景。

关键词:软件集群路由器;软件集群路由器参考模型;并行集群路由器参考模型
中图分类号:TP393.4 文献标识码:A

Research on the Architecture of Software-based Cluster Routers

GONG Zheng-hu, FU Bin, LU Ze-xin
(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: After analyzing the background of cluster routers, this paper proposes two cluster routing reference models: Software-based Cluster Router Reference Model (SCR-RM) and Parallel Cluster Routing Reference Model (PCR-RM). The detailed description of the two models is also made. These models will provide reference systems, experiment systems and IP flow processing scenarios for future studies.

Key words: SCR(Software-based Cluster Router); SCR-RM(Software-based Cluster Router Reference Model); PCR-RM(Parallel Cluster Routing Reference Model)

互联网发展的基本特征是:(1)网络的规模迅速扩大,速度快速提升;(2)新的网络与通信技术(如光通信,移动通信)不断融入;(3)人们日益关心网络安全与可信问题;(4)用户越来越需要更丰富、更方便、更及时的网络应用和服务;(5)网络营运商需要了解网络的动态形式,实施更有效的网络管理和控制;(6)网络营运商需要不断地根据用户需求开拓新的增值业务。互联网的这些特征导致新一代互联网(例如 NGI)概念的产生。什么是新一代互联网还没有严格的定义,但是"更大、更快、更及时、更安全、更方便、更可知、更可管理和更有效益"自然是它的基本特征。为了实现新一代互联网的这些特征,973 项目"新一代互联网体系结构理论研究"提出了规模、性能、安全、服务可扩展体系结构的概念。

从转发平面来看,路由器的体系结构可分为集中式、分布式、并行式和集群式四类[1-2]。

将常规路由器或计算机连接起来组成的单映像路由器叫集群路由器。单映像是指,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。本文将这类路由器叫做软件集群路由器(SCR, software-based cluster router)。SCR 有如下几个方面的发展优势:性能扩展性好;安全服务可扩展性好;开发成本低,开放性好;并发度高,加速比大。

SCR 分为两大类:基于主动网络技术的软件集群路由器和基于软件模块化技术的软件集群路由器。前者的典型代表有普林斯顿大学的 VERA[3] 和 NEC(USA) C&C 实验室的 CLARA[4];后者的典型代表有纽约州立大学的 Suez[5]。软件集群路由器的前期研究实际是软件可扩展路由器(单机环境),它们可成为集群路由器结点操作系统。典型的软件可扩展路由器的代表有亚利桑那大学和普林斯顿大学的 Scout[6],MIT 的 Click[7] 和华盛顿大学的 Plugins[8](三者的分析比较见文献[9])。软件可扩展路由器的研究不但可在计算机集群上展开,还可在网络处理机上展开[10]。

* 收稿日期:2005 - 12 - 01
基金项目:国家 973"新一代互联网路由与交换理论"资助项目(2003CB3148020);国家自然科学基金资助项目(90104001)
作者简介:龚正虎(1945—),男,博士生导师,教授。

本文提出了软件集群路由器的两个参考模型:SCR-RM 和 PCR-RM。所谓"参考模型"有几重含义:(1)它不是唯一的模型,他人可提出更好的模型;(2)它不是集群路由器的具体实现,未来商用的 SCR 可能是另外一种结构;(3)它可描述软件集群路由器 IP 流处理场景;(4)它可作为新一代路由器体系结构和 NPC 的研究/实验平台。

1 软件集群路由器参考模型(SCR-RM)

物理上,一个 SCR 由多个结点组成(图 1)。这些结点分为两大类:高速链路结点(HLN:high-speed link node)和低速链路结点(LLN:low-speed link node)。这里,高速链路和低速链路是相对于集群结点的处理能力而言的。如果一个集群结点的所有链路的入/出网络流量的处理和转发工作量小于该结点的处理能力,这种结点为低速链路结点(LLN);反之为高速链路结点(HLN)。由于 HLN 不能以软件方式线速处理和转发它的所有链路的入/出网络流量,那么,它必须将部分甚至大部分的流量处理工作转移到 LLN 去做,而自己只做分流/合流(multiplexor/demultiplexor)及其链路接口收发工作。HLN 可以是一种特殊的高速链路装置(例如,可采用核心路由器的高速线卡改造成 HLN),其主要功能是分流/合流。LLN 可以是有多个 Ethernet 接口的常规计算机(主板)或刀片服务器,它们运行常规操作系统(如 Linux)。

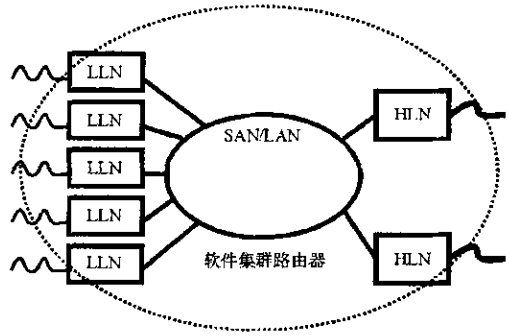

图 1 SCR-RM 物理结构
Fig.1 SCR-RM physical structure

多个 LLN 和多个 HLN 通过 SAN 或 LAN 连成一个集群路由器。它们之间通过有效的转发模型和控制模型形成一个"单映像"的集群路由器。所谓"单映像"是指,从网络工程师的角度看,这样的集群路由器的组网特性和常规路由器完全一样。

当 SCR 作边缘路由器使用时,它通常只有一条高速链路与核心网络连接,但有多条低速链路与接入网连接。此时,SCR 配置一个 HLN,而 LLN 可能多达数十个。

按照 ForCES 术语[11],逻辑上,一个 SCR 包含多个 CE(control element)和 FE(forwarding element)(图 2)。

每个 CE 可包括多个功能模块,如 BGP 协议模块、OSPF 协议模块、LDP 协议模块、QoS 协议模块等。所有 CE 的集合形成 SCR-RM 的控制平面。在一个 CE 内部,多个功能模块是并发的。在 SCR-RM 控制平面内,同类功能模块是并行的(如多个 BGP 协议模块并行执行 BGP 协议)。关于分布在多个 CE 内的 BGP、OSPF、LDP 等协议模块怎样协同工作形成"单映像"路由器的问题属协议并行模型研究范畴。

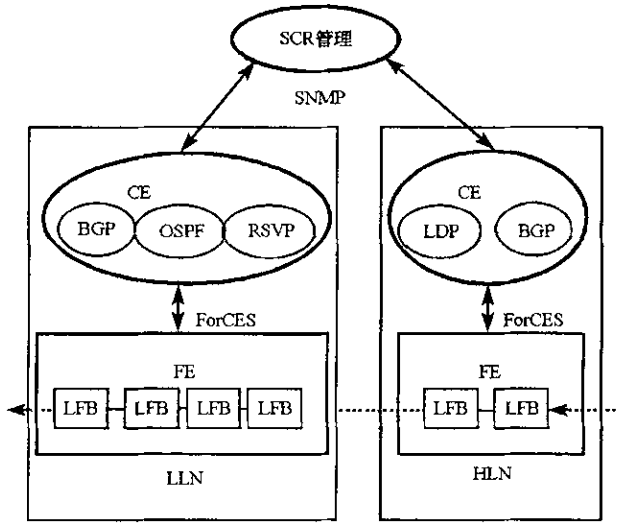

图 2 SCR-RM 逻辑模型
Fig.2 SCR-RM logical structure

按照 ForCES 的 FE 模型,每个 FE 由多个 LFB(logical function block)按照某种方式组成。一个 LFB 执行一种 IP 流量转发功能,如分类、测度、排队、调度、IP 头重写、地址解析等。SCR-RM 的安全、服务、功能可扩展通过在 FE 中插入、删除、修改 LFB 来实现。在 SCR-RM 中,所有 FE 的集合构成它的转发平面。这些 FE 分布在 LLN 和 HLN 中,它们分布/并行工作而获得较高的整机转发能力。关于按照什么方式组织 FE

和 FE 内的 LFB 而获得较好的转发性能属于转发模型的研究范畴。

图 2 所示的 SCR-RM 逻辑模型中有一个位于控制平面之上的管理平面，这一点不同于 ForCES 框架。在 ForCES 框架中，管理属于控制平面的一个功能。考虑到控制平面是分布的，为了提高系统的可管理性，SCR-RM 采用集中式管理方案，其协议为 SNMP。

为了更好地理解 SCR 的工作过程，下面描述 SCR 流量场景。假定 SCR-RM 作边缘（接入）路由器使用 ;其高速链路和低速链路都是以太网 ;每个 LLN 只有一条低速链路，连接一个接入网。IP 流量从多个接入网进入 LLN，经 HLN 汇聚后流向核心网的流量称作入口流量（ingress traffic），此时的 SCR-RM 起入口路由器（ingress router）的作用。IP 流量从核心网进入 HLN，分流后经 LLN 送往接入网方向的流量称作出口流量（egress traffic），此时的 SCR-RM 起出口路由器（egress router）的作用。IP 流量从一个接入网进入 LLN，经另外一个 LLN 又流向接入网的流量称作中转流量（transit traffic）。

下面仅仅讨论入口流量场景和出口流量场景。

图 3 描述入口流量在 SCR-RM 的处理场景之一，图 4 描述出口流量在 SCR-RM 的处理场景之一。图中的小方块表示 LFB，NI 为对外网络接口，SI 为内部 SAN 接口，CL 为分类器，SC 为调度器，FL 为过滤器，RW 为重写 IP 分组头，DX 为分流器。

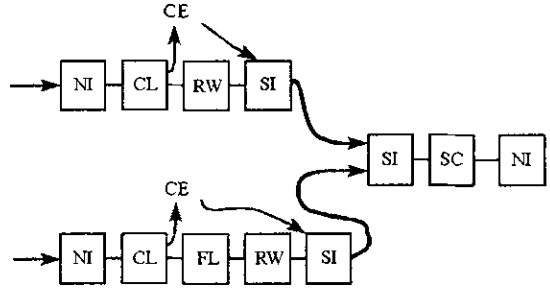

图 3 SCR-RM 入口流量场景
Fig.3 SCR-RM input flow scenario

图 3 中，LLN 处理低速链路的入口流量的过程可以不相同。CE 可以根据接入用户的特点和要求动态插入、删除、修改 LFB。第一个 LLN 的入口流量仅作分类处理（确定从哪个 HLN 的哪个链路转发，或送往本地 CE 处理）。第二个 LLN 的入口流量处理中增加了过滤功能。根据接入用户需求，FL 可以放在 CL 之前或 RW 之后。HLN 汇聚（合流）后各个 LLN 的流量经 SC 调度送往高速链路，调度算法的改变可以通过替换 SC 或改变 SC 的控制参数来实现。

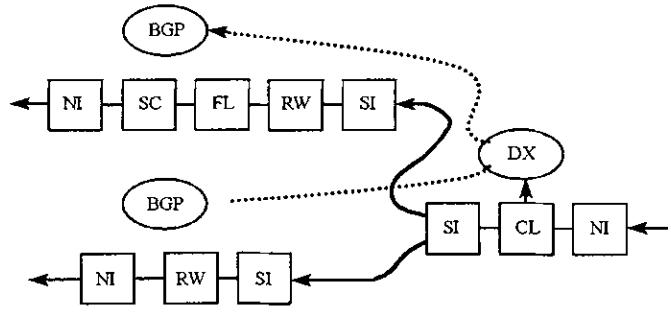

图 4 SCR-RM 出口流量场景
Fig.4 SCR-RM output flow scenario

图 4 中，高速链路进来的出口流量经 CL 分类后送往对应的 LLN（此时的 CL 有分流的功能）。为了减轻 HLN 的负担，流量的其它处理工作（RW，FL 等）放在 LLN 中进行。高速链路进来的路由、信令、控制等报文经 DX 分派到 LLN 处理。例如，图中的 BGP 报文分派到两个 LLN 处理。DX 可认为是 HLN 的控制平面的一个进程。

2 并行集群路由器参考模型（PCR-RM）

图 5 描述 PCR-RM 的逻辑结构。图中，BGP-1、BGP-2 和 BGP-3 代表 BGP 协议引擎（控制平面的软件实体，它们运行在集群路由器多个结点上），RIB（routing information base）为 BGP 协议引擎操作的路由信息库（它分布在集群路由器多个结点上，或集中放在某个结点上），FIB（forwarding information base）为转发信息库（转发平面的 FE 工作的基础，它是 FIB 的映射），e-BGP（external BGP）为管理域外部 BGP 协议，i-BGP（internal BGP）为管理域内部 BGP 协议，c-BGP（clustering BGP）为集群路由器内部 BGP 协议。

BGP-1、BGP-2 等协议引擎并行工作，它们操作一个统一的 RIB（不管 RIB 是分布的还是集中的）。多个 BGP 协议引擎和一个统一的 RIB 构成单映像的 BGP 协议系统，该系统与其它路由器之间的交互仍然是 e-BGP 和 i-BGP。BGP 协议引擎之间的交互按 c-BGP 协议进行。

以 PCR-RM 方式运行的 BGP 协议的工作过程是相当复杂的，本文还无法描述一个完整的场景。下

面给出它的基本工作模式。

BGP 协议引擎分布在集群路由器的各个结点中,它们可以是同构的(即它们的功能和特性完全相同),也可以是异构的(即它们的功能和特性相互不同)。当集群路由器的某个结点的某条链路接收到一个 BGP 报文或检测到网络发生变化,该结点的 BGP 协议引擎被触发而进行路由计算。在某些情况下,被触发的 BGP 协议可能在其它结点中。依据 PCR-RM 的并行模式,一条链路接收到的 BGP 报文可以分派到不同结点处理,一条 BGP 报文可以分派到多个结点处理。BGP 协议引擎在路由计算中要访问、修改的 RIB(还有 PIB:policy information base)可能在本地结点,也可能在其它结点。BGP 更新(UPDATE)报文可能从本地结点的某条链路发出,也可能从其它结点的某条链路发出。集群路由器与其它域内或域外路由器所建立的 BGP 连接(TCP 连接)应该终止于集群路由器的各个结点。

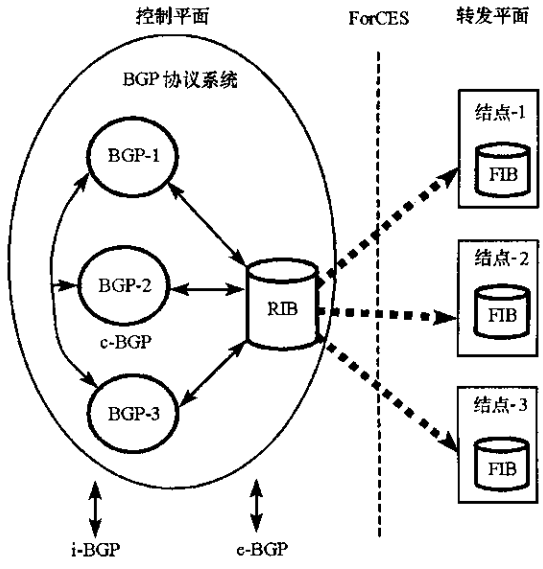

图 5 PCR-RM 逻辑结构

Fig.5 PCR-RM logical structure

3 结束语

本文为软件集群路由器的后续研究提出两个参考模型,一个是转发平面参考模型,另一个是控制平面参考模型。为了验证模型的可行性,项目组建立一个实验系统。该实验系统由 LAN 和 PC 机组成。PC 机运行 Linux 操作系统以及 Click 和 ForCES 软件平台。Click 用于 FE 和 LFB 的管理,ForCES 用于 CE 和 FE 之间的通信。

SCR 研究涉及体系结构,IP 转发模型,协议并行化,性能优化,动态可配置能力,单映像保守性以及路由稳定性等问题。上述研究正在按计划展开,并在实验系统进行验证。

参 考 文 献:

[1] Chan H C B,et al. A Framework for Optimizing the Cost and Performance of Next-generation IP Routers[J]. IEEE J. on Selected Area in Communication, 1999,17(6).

[2] Aweya J. IP Router Architectures: An Overview[EB/OL]. http://www.cs.virginia.edu/~cs757/papers.

[3] Karlin S,Peterson L. VERA: An Extensible Router Architecture[J]. Computer Networks, 2002,38(3).

[4] Welling G,et al. CLARA: A Cluster-based Active Router Architecture[A]. Hot Interconnect '00[C],2000.

[5] Pradhan P,Chiueh T. A Cluster-based Scalable and Extensible Edge Router Architecture[A]. In the Proceedings of MUG-2000[C],2000.

[6] Mosberger D,Peterson L. Making Paths Explicit in the Scout Operating System[A]. In Proceedings of the Second USENIX (1996)[C],1996.

[7] Kohler E,et al. The Click Modular Router[J]. ACM Transactions on Computer Systems, 2000,18(3).

[8] Decasper D,et al. Router Plugins: A Software Architecture for Next Generation Routers[J]. IEEE/ACM Transaction on Networking, 2000,8(1).

[9] Gottlib Y,Peterson L. A Comparative Study of Extensible Routers[A]. 2002 Open Architecture and Network Programming Proceedings[C],2002.

[10] Spalink T,et al. Building a Robust Software-based Router Using Network Processors[A]. In Proceedings of the 18th ACM Symposium on Operating Systems Principles[C],2001.

[11] Yang L,et al. Forwarding and Control Element Separation (ForCES) Framework[R]. RFC 3746, April 2004.