ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.9, September 2007, pp.2205−2215
DOI: 10.1360/jos182205

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

可扩展路由器控制平面的高性能通信模型*

徐 恪[1+], 吴 鲲[1], 王青青[2]

[1](清华大学 计算机科学与技术系,北京 100084)

[2](兰州大学 信息科学与工程学院,甘肃 兰州 730000)

High Performance Control-Plane Communication Model for Scalable Routers

XU Ke[1+], WU Kun[1], WANG Qing-Qing[2]

[1](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[2](School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China)

+ Corresponding author: Phn: +86-10-62795818 ext 6856, E-mail: xuke@tsinghua.edu.cn, http://network.cs.tsinghua.edu.cn/teacher/xuke/

Xu K, Wu K, Wang QQ. High performance control-plane communication model for scalable routers. *Journal of Software*, 2007,18(9):2205−2215. http://www.jos.org.cn/1000-9825/18/2205.htm

Abstract: The inter-nodes communication bottleneck has been a key factor that restricts the large-scale expanding of scalable router's software architecture. To solve the problem, this paper introduces a transmission adapted sub-layer in the supporting model of traditional software architecture. Through feature extracting of the up-going data stream and pattern matching with registered task, data stream can then be classified and divided based on the content of information to increase the effective communication rate. The paper further analyzes the model's performance from task's three characteristics: Distribution rate, spread number and traffic rate, and provides an optimized task dispatching reference. It shows that the introduced transmission adapted sub-layer can reduce inter-layer redundant flow and extensibility bottleneck of communication. Finally, the presented experiment verifies the theoretical analysis.

Key words: router; router software architecture; scalability; transmission adapted sub-layer

摘 要: 可扩展路由器控制平面节点间通信的瓶颈问题是制约软件体系结构大规模扩展的关键因素.针对此问题,在传统的软件体系结构的支撑模型中引入了传输适配子层的结构,上行的数据流经特征抽取与已注册的任务进行模式匹配,从而完成了对控制信息流基于内容的分类与分流,提高了其有效通信率.进一步根据任务的分布率、分散数和流量率这 3 个特征对模型进行了性能分析,表明了适配层的引入可以消除面间冗余流量和通信的可扩展瓶颈.最后通过实验验证了理论分析的正确性.

关键词: 路由器;路由器软件体系结构;可扩展性;传输适配子层

中图法分类号: TP393 **文献标识码**: A

路由器作为现代计算机网络的核心设备,其体系结构的发展经历了集中式到分布式再到可扩展的结构变

* Supported by the National Natural Science Foundation of China under Grant No.60473082 (国家自然科学基金); the National Basic Research Program of China under Grant No.2003CB314801 (国家重点基础研究发展计划(973))

Received 2006-08-25; Accepted 2006-11-24

迁,目前,可扩展的体系结构已成为下一代路由器的主要发展方向.

路由器可以划分为两个主要的功能平面:一是集中于流量转发处理的数据平面;另一个是完成控制与路由交互的控制平面.目前,大量的研究成果集中于可扩展路由器的硬件平面的可扩展性方面[1−7],而对于逐渐成为新瓶颈的可扩展路由器软件体系结构尚缺乏成体系的理论框架和参考模型,已有的模型也仅集中在功能的伸缩与动态调整方面[8−11].随着数据平面和控制平面分别进行的规模扩展,两个平面间的通信量大幅度增加,连接拓扑更为复杂,传统信息流的传输模式会成为限制系统进一步扩展的潜在瓶颈.这个问题在通用分布式系统中未曾遇到,从公开的论文和其他资料中也未看到解决方案.

数据平面到控制平面的信息流是控制平面的主要信息来源,直接影响软件体系结构的运行状况.在传统的路由器软件体系结构中,来自底层数据平面的控制信息,例如路由交互报文等,都是原始的信息,信息在到达控制平面之前不作处理.如果系统中有多个控制平面的节点,数据不作区分地提交到控制平面的每个节点.每个控制平面节点在收到原始控制信息后,首先对信息进行过滤,丢弃不属于自己的信息,选出对本节点有用的信息再次提交到各个相应的协议模块进行处理.这种通信模式在控制平面节点单一或者数目很少的情况下可以使用,但在可扩展路由器软件体系结构中,由于体系结构对可扩展性的提升与全面支持,控制平面和数据平面都包含有多个节点,而且实现高度的系统可扩展能力,在理论上对两个平面的节点数目不再有约束.如果仍然沿用这种基于洪泛的通信模式,会随着控制节点数目的增多带来过大的流量,容易造成内部通信带宽的浪费,从而影响了软件体系结构的扩展能力.

本文研究的主要贡献是,在传统的分布式路由器操作系统中引入了传输适配子层的结构,对通信流的分发进行控制,以消除面间的冗余流量和通信的可扩展性瓶颈,使其适应可扩展路由器体系结构的需求.

本文第 1 节详细分析传统洪泛模式下的性能.第 2 节引入传输适配子层,介绍传输适配子层的交互流程和模式匹配方法.第 3 节在模型性能理论分析的基础上给出性能的评价依据和优化参考方案.第 4 节通过实验验证理论的正确性.最后总结全文.

1 传统通信模式的性能分析

在传统的数据平面与控制平面的通信模式中,来自数据平面的信息流在到达控制平面的预处理部分(例如传输协议层)之前,不对其作关于内容的分类处理.这种不对信息进行关于内容的分类与分流的通信模式称为洪泛模式.

我们对洪泛模式下数据平面到控制平面的控制信息流进行简单的分析,以便为通信结构的改进和改进后的性能比较提供依据.首先对环境进行简单的描述,为方便分析,假定环境中的任务彼此不相关,其中:

- n 表示数据平面的节点个数;
- m 表示控制平面的节点个数;
- t 表示控制平面的任务总数;
- P_i表示控制节点i上分配的任务集合;
- $Q_i(k)$表示数据节点j到任务k的控制信息流量.

洪泛模式的数据平面到控制平面控制信息流通信量为

$$\mathrm{II}_{flood}=\sum_{j=1}^{n}\sum_{k=1}^{t}(mQ_j(k)) \tag{1}$$

在这部分流量中,实际的有效流量为

$$\mathrm{II}_{act}=\sum_{i=1}^{m}\sum_{k\in P_i}\sum_{j=1}^{n}Q_j(k)=\mathrm{II}_{act}=\sum_{j=1}^{n}\sum_{i=1}^{m}\sum_{k\in P_i}Q_j(k) \tag{2}$$

假设所有t个任务在控制平面所有m个节点上的分布矩阵为$W=[w(i,k)]_{t\times m}$,其中,

$$w(i,k)=\begin{cases}1, & \text{任务}i\text{在控制节点}k\text{上}\\ 0, & \text{任务}i\text{不在控制节点}k\text{上}\end{cases}.$$

类似地,n个节点到t个任务的控制信息流流量矩阵为$F=[f(j,k)]_{n\times t}$,为了形式上的统一,如果数据平面节点j_a不包含到达任务k_b的信息流,则$f(j_a,k_b)=0$,即

$$f(j,k)=\begin{cases}Q_j(k), & \text{数据平面节点}j\text{有到任务}k\text{的流量}\\ 0, & \text{数据平面节点}j\text{没有到任务}k\text{的流量}\end{cases} \tag{3}$$

不失一般性,从控制信息流流量矩阵F中取出任意一行$F[j]_{1\times t}=[f_{j1},\ldots,f_{jt}]$,由流量分布模式定义的物理意义可知,从数据平面节点j发出的所有有效控制数据流量为

$$Flow_j=F[j]_{1\times t}\cdot[1,\ldots,1]_{1\times t}=\sum_{k=1}^{t}Q_j(k) \tag{4}$$

另一方面,由矩阵运算有

$$\sum_{i=1}^{m}\sum_{k\in P_i}Q_j(k)=[Q_j(P_1),\ldots,Q_j(P_m)]_{1\times m}\cdot[1,\ldots,1]_{1\times m}$$

由于$P_1,\ldots,P_m$彼此不相交,所以上式右端得到的也是数据平面节点j发出的所有信息流流量之和.再由式(4)可得

$$\sum_{k=1}^{t}Q_j(k)=\sum_{i=1}^{m}\sum_{k\in P_i}Q_j(k) \tag{5}$$

综合式(1)、式(2)可以得到:洪泛模式在可扩展的软件体系结构下,数据平面到控制平面的有效通信率为

$$R=\frac{\Pi_{act}}{\Pi_{flood}}=\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}\sum_{k\in P_i}Q_j(k)}{m\sum_{j=1}^{n}\sum_{k=1}^{t}Q_j(k)}=\frac{1}{m} \tag{6}$$

分析结果说明,在控制平面也采用可扩展的分布式结构时,如果沿用传统路由器系统中的洪泛模式进行控制流传输,其有效传输率仅为实际传输量的 $1/m$,这也与我们直观的理解相一致.从另一个角度来讲,如果要每个控制节点的负载尽可能地增大,到单节点的信息流也会增加,则总的传输量为实际需求量的 m 倍.这是一个与控制平面规模相关的数量,因此,它带来的平面间信息流爆炸会成为严重的系统扩展能力瓶颈.

在分析过程中,我们假定了 t 个任务是不相关的,也就是说,不考虑同一任务的分布式实现情况.在实际的分布式软件体系结构中存在同一任务的分布式实现,所以,实际系统的传输利用率情况比上述的分析结果要好.但是,这一因素并不会从根本上解决通信的可扩展瓶颈问题.

2 改进后的通信模型

2.1 传输适配子层的引入

可扩展的结构带来了控制平面与数据平面之间的通信爆炸问题,成为可扩展结构的一个瓶颈.因此,对这个问题的解决也要从通信结构入手进行分析.

图 1 是可扩展路由器的数据平面到控制平面通信通路的典型结构图.这里描述的是逻辑的通信通路,而忽略了底层的物理连接.这是因为在当前的主流实现方案中,这部分通信都是由交换式的内部网络来实现的,在可扩展的结构下,还可能用到多级的交换网络来进行扩展.这种基于交换的物理通信结构虽然对上层的通信有一定的影响,但是这种影响并非决定性因素.因此在本文的分析中,我们仍然沿用这种假设,而这种假设下可以将平面间的通信抽象为更一般的 mesh 结构.

传统洪泛模式的根本问题在于不对数据流进行基于内容的区分,因此,平面间的逻辑通信结构与实际的数据流量完全重合.也就是说,实际的数据流通路也是图 1 中的 mesh 结构,因此,产生了巨大的无效流量,浪费了带宽.解决此问题的根本方法是根据信息的内容进行分类与分流,只生成有效的流量,才有可能使控制信息流的有效通信率达到最大.

图 2 是改进的通信结构,其中来自数据平面的控制流根据内容进行了分类和分流,信息只发送到目标任务所在的节点或者节点组.这种通信结构下需要有一定的规则和机制来对数据平面发出的控制流作区分,并且由

于控制平面的任务在各个控制节点间的分配是动态的,因此这个过程需要具有动态的可维护特性.为此,本文提出一个“传输适配子层”,插入控制平面和数据平面之间,通过传输适配子层来统一和协调与层间控制流相关的所有操作.在可扩展路由器的系统功能划分中,这个子层从属于操作系统的一部分,它一方面向上提供了数据平面的信息流传输接口与信息流控制接口,另一方面在层间通信的框架下完成数据平面对控制平面的抽象屏蔽.

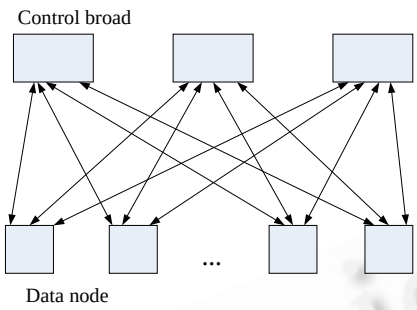


Fig.1 Inter-Layer communication architecture

图 1 平面间通信结构

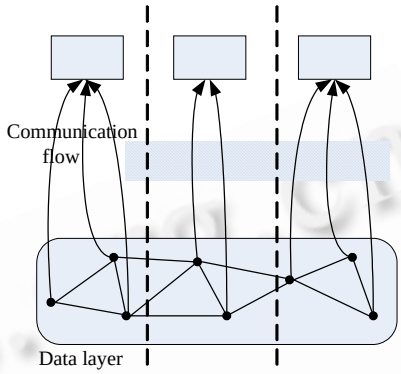


Fig.2 Improved logical communication architecture

图 2 改进的逻辑通信结构

2.2 传输适配子层的结构

如图 3 所示是传输适配子层的控制结构.传输适配子层是跨接在数据平面和控制平面之间的一个功能模块.它在每个数据平面节点和控制平面节点上都有相应的实现.这些模块实现共同完成信息流的分类操作.传输适配子层的主体功能在数据平面上进行,控制平面只提供控制接口的部分.图 3 中的结构为数据平面上的结构,其中的功能部件参见表 1.

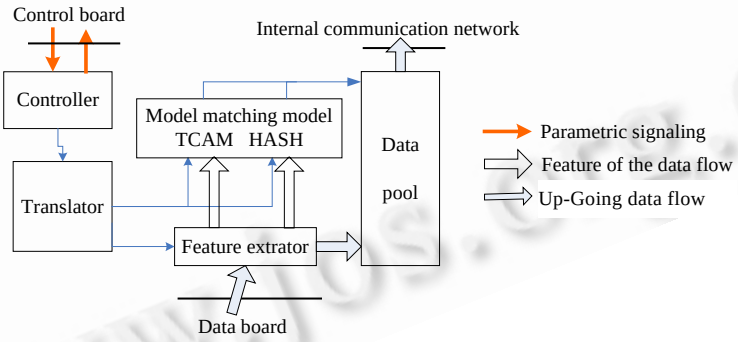


Fig.3 The control structure of transmission adapted sub-layer

图 3 传输适配子层的控制结构

Table 1 Main components of transmission adapted sub-layer

表 1 传输适配子层的主要功能部件

Functional component	A brief description
Data pool	Used to temporarily store up-going data flows
Feature extractor	To extract data flow’s feature
Pattern matching model	To do pattern matching and classifying of the data flow’s feature
Controller	To perform interacting with nodes in control plane
Translator	Translate the commands sent by nodes in control plane into patterns for feature matching

- 数据池:来自底层的控制流数据首先经过节点本地的网络层处理,判断不是转发数据,进而提交为本地的上行控制数据,进入数据池.数据池是一个共享内存的缓冲池,提交给上层控制节点的信息首先存入

数据池中.在实际的系统实现中,数据池不是底层数据的二次拷贝,而是直接利用底层的原始存储.例如,在以嵌入式 Linux 为基础操作系统的数据节点上,数据池中不对 sk-buffer 的内容进行复制,而是在数据池中维护进行缓冲的 sk-buffer 的索引标识.从这个意义上讲,数据池是一个可随机访问的数据缓冲区.

- 特征抽取器:底层提交的上行数据在到达传输适配子层时,数据主体进入数据池进行缓冲,数据的特征由特征抽取器进行抽取.特征抽取器根据来自控制器的指令确定进行抽取的目标数据段.将目标数据段送入模式匹配模块进行分类.目标数据段拥有与放入数据池数据相同的索引标识.这个功能模块的引入主要是为了对模式匹配进行预处理,减少进行模式匹配的数据量,减小开销.
- 模式匹配模块:主要任务是根据来自控制器和翻译器的指令,形成模式匹配的规则,对来自特征提取器的目标数据段进行模式匹配,进而对数据进行分类,选择数据的目标控制平面节点.传输适配子层需要分类的数据是多特征多模式的信息,模式匹配模块包含两个模块,对信息进行不同层次的分类.其中,TCAM 是以硬件为主的分类模块,另一个是基于软件的分类模块,通过软件的哈希表来实现.对拥有简单结构、固定结构、静态结构的特征,采用 TCAM 来完成快速的模式匹配.对于复杂的、动态的结构特征,采用哈希表来完成深度模式匹配.
- 控制器:传输适配子层的数据平面部分通过控制器向控制平面部分提供交互操作.控制器维护与控制平面节点任务之间的关系,处理来自控制平面任务的信息流分配申请,并自动监测任务的离开与迁移.
- 翻译器:传输适配子层提供给控制平面的接口是面向命令的抽象接口,这些命令需要转换成相应的模式匹配规则.此外,不同的任务会有不同的信息流分配,这些分配可能存在着交叉与重叠.在翻译器中解决多个模式之间的交叉、重叠与共享.

在传输适配子层中,存在如下 3 条主要的信息通路,数据流通路如图 3 所示,其中,细线标识了传输适配子层内部的控制操作.

(1) 参数化命令通路:用于传输适配子层在控制平面与数据平面之间的接口交互.

(2) 上行信息流通路:是上行数据的主要通路,由底层提交的上行数据,经过特征提取器取出用于模式匹配的特征,同时进入数据池进行缓冲.

(3) 特征信息流通路:特征信息由特征提取器生成,然后刷新到 TCAM 或者对应的哈希表.

2.3 交互流程

传输适配子层的交互是指数据平面和控制平面在传输适配子层上的控制信息传递,包括两个部分:一是控制平面的任务向数据平面提出的上行信息流申请,也称为注册过程;另一个是反向的从数据平面向控制平面节点的信息流传递,也称为信息发射.因此,传输适配子层的交互过程称作“注册-发射”的流程.

任务的注册通过在传输适配子层控制平面端的注册接口来实现,注册接口在形式上是一个四元组⟨*TgtNodeID*,*TgtTaskID*,*RuleList*,*TimeOut*⟩,其中各字段的含义见表 2.

Table 2 The registration interface field

表 2 注册接口字段

Field	Meaning
TgtNodeID	Identification of the target node that sends registration request
TgtTaskID	Identification of the task that sends registration request
RuleList	Rule list of pattern matching
TimeOut	Overtime period of heartbeats process

传输适配子层在控制平面端维护一个进程,负责接收注册请求,然后转发到数据平面节点,同时将请求按照任务标识保存在一个本地的数据库中.当数据平面有新节点加入时,由这个进程来自动将申请重新向新加入节点再次注册一遍.为了防止任务在节点上退出后还有不必要的流量,任务可以主动地注销注册过的上行信息流.除此之外,传输适配子层与操作系统配合,提供对任务运行状态的监测,在任务发生故障或者任务没有实现注销机制的情况下,对相应的信息流进行自动的注销.

数据的发射相当于一个由数据内容来进行控制的组播过程,由前端的数据分类和后端的数据组播构成.数

据分类的结果是一个与数据平面节点相对应的位图(bitmap),组播部分根据位图的结果从数据缓冲池中将对应的数据发送到目标节点.组播部分的实现机制可与传统模式中的组播相同,不是所有的信息流都会流向所有的控制平面节点,因此,传统模式中的组播是本模型中组播过程的速度下界.

2.4 模式匹配

经过传输适配子层的数据流主要是上行的协议数据,从功能上划分有:各个路由协议的交互信息、网络管理协议的交互信息、其他控制协议(例如流量工程协议)的控制信息等信息流.这些信息流的分布与实现模块的分布息息相关,有些是在协议的级别进行分布,有些由于对应的协议本身就采用了分布式实现,所以在同一个协议内的不同功能数据流也会选择不同的目的节点.因此,传输适配子层的分类是一个多模式多字段的复杂匹配.

根据不同的功能需求,传输适配子层的分类模式有:固定字段模式和非固定字段模式.前者是指在通信报文的基础格式中占据固定位置的模式,这些模式包括网络层源地址与目的地址、传输层协议目的端口号.后者指无法通过通信报文的固定格式来进行匹配的模式,主要指基于内容来进行区分的数据,例如,网络管理协议报文根据管理对象来进行数据流的分布,那么就要由管理协议报文的内容来判断目标节点.在 SNMP 协议框架中,这个模式是非固定结构的.

固定字段模式的匹配和现有的路由器过滤功能前端实现基本相同,因此,可以利用路由器线卡上的过滤功能来辅助实现.这也是为什么要引入传输适配子层,在数据进入控制节点之前对其进行分类的主要原因,就是可以在数据平面节点上通过硬件进行快速的分类与分流.固定字段模式通过 TCAM 来进行匹配,匹配的内容是网络层源/目的地址和传输层目的端口号,这个匹配可以满足大部分的信息流匹配.

匹配规则由若干预定义的特征域与特征域上相关的带参数二值函数共同构成.特征域的集合记作

$$F=(f_1,f_2,\ldots,f_n),$$

其中,f_i是编号为i的预定义特征域.对每一个预定义特征域,都有若干带参数的二值函数与之绑定.特征域f_i绑定的函数列表为

$$G_i=(g_{i1},g_{i2},\ldots,g_{iq}).$$

函数的输入是若干个参数,输出是一个二值逻辑,0 表示不选中流,1 表示选中流.函数g_{ij}的参数列表记作$P_{ij}=(p_{ij1},p_{ij2},...,p_{ijr_{ij}})$.所有规则的特征域、函数及其参数通过一个树形的目录结构组织在一起:

$$ROOT\begin{cases} f_1 \quad \{\ \ldots \\ \vdots \\ f_i \begin{cases} g_{i1} \quad \{\ \ldots \\ \vdots \\ g_{ij} \begin{cases} p_{ij1} \\ \vdots \\ p_{ijr_{ij}} \end{cases} \\ \vdots \\ g_{iq} \quad \{\ \ldots \end{cases} \\ \vdots \\ f_n \quad \{\ \ldots \end{cases}.$$

在对控制平面不同目标节点的标识进行注册时,与相应的函数g_{ij}绑定在一起.不失一般性,假设控制平面有m个节点.用一个长度是m的向量位图来表示数据流的目标节点分布状况,记作$B=(b_1,\ldots,b_m)$.当某个数据平面节点收到一个上行数据流,根据特征域依次取出相应的字段,用特征域下的匹配函数计算是否对绑定的目标节点选中该流.

为了加快模式匹配的处理过程,对多个特征域可以采用 NP(网络处理器)的多个微引擎并行处理.同时,可以进行规则的合并.

定义 1(简单匹配). 我们只需要进行一次模式匹配,且匹配范围为单一值的函数称为简单匹配.简单匹配函

数可以表示为一个二元组$g=\langle f,p\rangle$.其中,f是匹配的目标域,p是匹配的模式值.

对同一个特征域的 k 个简单匹配,可以通过一次哈希在一个匹配周期内完成.输出结果是一个长度为 m 的向量位图:

$$(b_1,\ldots,b_m)=Hash(field,(p_1,\ldots,p_k)).$$

这个操作可以在一个匹配周期内完成.

n个特征域的匹配结果分别记作$(B_1,\ldots,B_n)$,最终的模式匹配结果:

$$B=\wedge_{i=1,\ldots,n}(B_i) \tag{7}$$

其中,$\wedge$是按位与操作.

假设一个匹配周期为τ,一次$\wedge$操作的周期为α,如果存在 n 个特征域,则总的时间开销为

$$Cost_{serial}=n\times\tau+(n-1)\alpha.$$

如果可以在 u 个微引擎上进行并行匹配,然后在微引擎间按位的与进行归并,则并行过程总的时间开销为

$$Cost_{parallel}=\begin{cases}\tau+\alpha\log_2(n), & u\geq n\\ \tau\left\lceil\dfrac{u}{n}\right\rceil+\alpha(\log_2(u)-\log_2(n))\log_2(u), & u<n\end{cases} \tag{8}$$

3 模型的性能分析

传输适配子层的引入,就是要提高数据平面到控制平面的有效通信率.因此,我们对系统实际取得的有效流量率进行分析.但由于有效流量率是一个绝对值,并不能反映传输适配子层的性能改进,我们在固定的任务以及流量模式下,对传输适配子层的流量与洪泛模式的流量进行比较,进而对传输适配子层的性能进行衡量.

第 1 节中讨论了洪泛模式下的有效流量率公式(6),得到的是理想情况下有效流量能够达到的上限.事实上,由于在控制平面中存在多个控制平面节点任务接收相同的信息流的情况,也就是控制平面的节点在数据平面信息源上存在一定的相关性,所以在改进的传输适配子层下,对有效流量率的提高并不能达到理想的 m 倍(m 为控制平面节点的个数).

定义 2(面间流量). 数据平面到控制平面流经的信息流总量称为面间流量.

定义 3(面间流增量指数). 在固定的任务以及流量模式下,将洪泛模式面间流量与一个对比模型的面间流量的比值称作这个对比模型的面间流增量指数,记作II_c.

从定义可以看出,面间流增量指数表达了对比模型的面间流量对带宽的“节约程度”.面间流增量指数的倒数表达了一个对比模型要达到没有优化的洪泛模式的面间流量,还能额外容纳负载的能力.我们为系统的性能分析引入表 3 中的参数.

Table 3 The analytical parameter of inter-layer flow incremental index

表 3 面间流增量指数的分析参数

Parameter	Expression	A brief description
Distribution ratio	$D=(d_1,\ldots,d_t)$	The ratio of the flow between multi-nodes in a task to the total flow
Spread number	$E=(e_1,\ldots,e_t)$	The number of nodes that a task performs on
Flow ratio	$C=(c_1,\ldots,c_t)$	The ratio of a task's flow to the total flow

不失一般性,假定系统有 t 个任务,分布在 m 个控制平面节点上,此外定义如下的系统描述量:

任务i的流量$\mathrm{II}(i)=c_i+d_i(e_i-1)c_i=(1+d_i(e_i-1))c_i$,则面间流量:

$$\mathrm{II}=\sum_{i=1}^{t}\mathrm{II}(i)=\sum_{i=1}^{t}(1+d_i(e_i-1))c_i \tag{9}$$

式(9)是一个通用的公式.取$d_i=1,e_i=m$,则得到洪泛模式的面间流量.因此,可以得到传输适配子层模型的面间流增量指数:

$$\mathrm{II}_c=\left(\frac{\sum_{i=1}^{t}(1+d_i(e_i-1))c_i}{\sum_{i=1}^{t}mc_i}\right)^{-1}=\left(\frac{\sum_{i=1}^{t}\frac{1+d_i(e_i-1)}{m}c_i}{\sum_{i=1}^{t}c_i}\right)^{-1}$$

根据定义有 $\sum_{i=1}^{t}c_i=1$,则有

$$\mathrm{II}_c=\left(\sum_{i=1}^{t}\frac{1+d_i(e_i-1)}{m}c_i\right)^{-1} \tag{10}$$

令 $\beta_i=\frac{1+d_i(e_i-1)}{m}$,它反映了一个任务的流量分布特征对系统的面间流增量指数的影响能力,称为任务的面间流增量因子.从式(10)可以看出,面间流增量指数的倒数就等于任务的面间流增量因子与流量率的内积.图4 显示了这种关系,图中以两个任务的情形为例,横、纵坐标分别代表两个任务的面间流增量因子.(O,C)是两个任务的流量率向量.根据面间流增量因子的定义可知:

$$0\le d_i\le 1,\ 1\le e_i\le m,$$

所以

$$\frac{1}{m}\le\beta_i\le 1,\forall i\in[1,\ldots,t].$$

因此,图中点A是面向流增量因子取值范围的下界,点B是面向流增量因子取值范围的上界.阴影部分就是面间流增量因子向量可以取值的范围.在只考虑相对大小的概念下,阴影部分任意点(β_1,β_2)所对应的面间流增量指数的倒数就是其在向量(O,C)上的投影长度.(β_1,β_2)的投影为D',A,B的投影分别为A',B'.面间流增量指数的倒数取值范围就是(A',B').不同的面间流增量因子组合对应的投影就表征了对比模型的带宽节约程度.由于洪泛模型就是点B所对应的面间流增量因子,因此,从图中还可以看出对比模型中对新加入负载的接纳能力,也就是B的投影与(β_1,β_2)投影的长度之比$\frac{|(O,B')|}{|(O,D')|}$.

另外,从公式(10)还可以得出:面间流增量指数的最小值、最大值与任务的流量率分布状况无关.这一点也可以从图4 中看出.但是,实际系统的面间流增量指数仍然与任务的流量率分布情况相关.这也意味着,评价一个实际系统中各个任务对面间流增量的影响时,需要兼顾任务的分散数、分布率和流量率3 个参数.

式(10)也可以写作 $\mathrm{II}_c=\left(\sum_{i=1}^{t}\beta_ic_i\right)^{-1}$,其中,$\beta_ic_i$描述了一个任务的各个参数对面间流增量指数倒数的综合影响.也就是说,面间流增量因子与任务的流量率的乘积就是一个任务对面间流增量指数倒数的贡献.从图 5 可以看到,深色阴影区域是面间流增量因子的优化区域.当点(β_1,β_2)向这个区域移动时,系统的面间流增量指数倒数会进一步减小,系统可以容纳更多的负载.但是,任务的面间流增量因子的变化规律并不相同,即图中l_1和l_2的移动和任务的面间流特征以及分布特征相关,这个关系由面间流增量因子和任务的流量率共同确定.

通过上面的分析可以计算在引入传输适配子层后,系统能够新增多大面间流量负载,也就对应了能够增加各个任务的相应正规化负载.下面的问题就是如何在传输适配子层模型下,根据各个任务的面间流量特性来对性能进行优化.在图 5 中,也就是要确定l_1,l_2在不同参数下的移动速率.通常,任务的流量率是系统固有的特性之一,也和负载环境相关,而与采用了什么样的任务分配模式无关,因此不以此作为改进的条件.这里,我们讨论分布率与分散数对面间流增量指数倒数的影响.

根据式(9),可以得到面间流增量指数倒数对任意任务 i 关于分布率的偏导数与关于分散数的差分:

$$\frac{\partial(\mathrm{II}_c)^{-1}}{\partial d_i}=\frac{(e_i-1)c_i}{m} \tag{11}$$

$$\frac{\Delta(\mathrm{II}_c)^{-1}}{\Delta e_i}=\frac{d_ic_i}{m} \tag{12}$$

式(11)、式(12)给出了面间流增量指数倒数关于每个任务参数的变化速率.由此可以综合考虑分散数和分

布率对面间流增量指数倒数的影响,选择使得面间流增量指数变化最快的参数作为改进任务分布模式的确定方向.

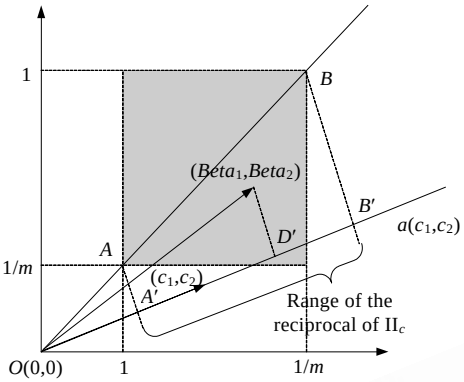


Fig.4 The map of inter-layer flow incremental index

图 4 面间流增量指数示意图

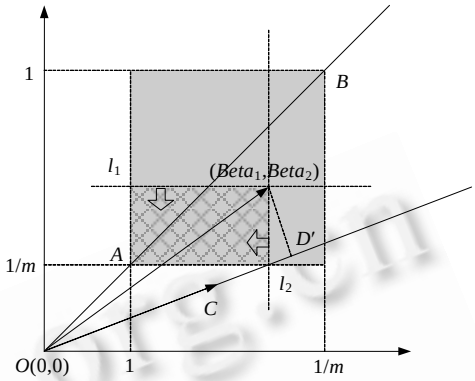


Fig.5 The changes of inter-layer flow incremental factor

图 5 面间流增量因子的变化

4 实验结果与分析

实验采用基于树型结构的BGP路由迭代模型[12],取k=2 的 4 个节点迭代树,任务除了BGP的模拟任务以外,还有一个管理任务.BGP任务分布在迭代树的 3 个次叶节点上,管理任务分布在所有节点上.这个分布式模型限定了任务的分配模式,因而任务的分散数是固定的.我们的实验改变BGP任务的流量分布率,分别测量洪泛模型与不同参数下的传输适配子层模型的面间流量,从而得到面间流增量指数.各个参数的设定值见表 4.

Table 4 Parameter configuration in experiment

表 4 实验参数设定

Parameter	BGP task	Management task
Distribution rate (d_i)	5%~100%	5%
Spread number (e_i)	3	4
Flow rate (c_i)	0.9	0.1

在实际的实验中,BGP来自数据平面的控制流数据只到达次叶节点,所以,任务分布节点数没有取m=4,而是对两个任务给定了不同的任务分配节点数m_{BGP}=3 和m_{MGMT}=4,这与实际的路由器系统更为接近.BGP任务的面间流量分布率为 5%~100%,递增步长取 5%.实验的结果如图 6 所示.

在不同参数下测定得到的面间流增量指数在图中用*标出.从图中可以看到,沿 BGP 任务的面间流分布率减小的方向,面间流增量指数逐渐增大,这意味着可以容纳更大的面间流量负载.当BGP任务的面间流分布率只有 5%时,可以容纳相当于洪泛模式极限约 4.5 倍的负载.这与 BGP 的路由更新报文仅发送给对应的非叶节点,其他为广播的控制消息的实际情形相当.另一方面,图中用虚线给出了在对应参数下面间流增量指数理论值.实验测定的结果与理论值在趋势上基本吻合,在数值上略小于理论值,这是由于在实验过程中存在额外的面间流量开销所致.

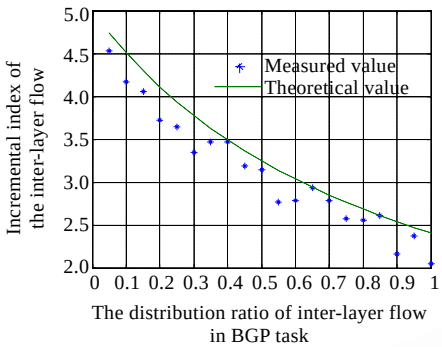


Fig.6 The experimental result

图 6 实验结果

5 结论和下一步工作

可扩展的路由结构带来了数据平面到控制平面的数据流瓶颈问题,已经成为制约路由器软件体系结构大规模扩展的关键因素.为解决此问题,我们在可扩展路由器软件体系结构中引入了一个介于数据平面和控制平面之间的传输适配子层.本文集中描述了传输适配子层的目标、结构、交互流程和模式匹配方法,并对性能进行了理论分析,定义了面间流量和相关的面间流增量指数,作为模型的性能评价依据.通过控制平面任务的分布率、分散数和流量率 3 个特征参数对模型的性能进行了分析,此外,基于模型性能的理论分析还为任务分配模式的优化提供了一定的依据.理论分析表明,传输适配子层可以消除平面间的冗余流量,消除面间通信的可扩展性瓶颈.实验结果表明了模型的有效性和理论分析的正确性.

此外,在传统通信模型中引入传输适配子层之后,也必然引入了额外的任务注册、数据流特征提取、模式匹配等通信和运算开销,文中已经就这些部分的优化实现进行了描述.分析和实验结果表明,这些开销与模型带来的通信效率的提高相比是值得的.

最后,由于模型抽象简单性的需要,在文章分析过程中不可避免地与实际系统存在一定的差别,不能完全反映所有细节,下一步工作需要对任务的流量进行更准确的描述,从而得到与实际系统更接近的评价模型.此外,需要进一步深入研究负载均衡、高可靠性以及可扩展结构之间的关系等问题.

References:

[1] Schaller RR. Moore's law: Past, present and future. IEEE Spectrum, 1997,34(6):52−59.

[2] Gupta P. Algorithms for routing lookups and packet classification [Ph.D. Thesis]. Stanford: Stanford University, 2000.

[3] 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 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/445.htm

[4] Iyer S, McKeown N. Analysis of the parallel packet switch architecture. IEEE/ACM Trans. on Networking, 2003,11(2):314−324.

[5] Keslassy I, Chuang ST, Yu K, Miller D, Horowitz M, Solgaard O, Mckeown N. Scaling Internet routers using optics. In: Proc. of the 2003 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York: ACM Press, 2003. 189−200.

[6] Chao HJ, Deng K, Jing Z. PetaStar: A petabit photonic packet switch. IEEE Journal on Selected Areas in Communications, 2003, 21(7):1096−1112.

[7] Cheyns J, Develder C, Colle D, De Truck F, Lagasse PM, Demeester P. Clos lives on in optical packet switching. IEEE Communications Magazine, 2004,42(2):114−121.

[8] Kohler E, Morris R, Chen B. The click modular router. ACM Trans. on Computer Systems, 2000,18(3):263−297.

[9] Mosberger D, Peterson L. Making paths explicit in the scout operating system. In: Proc. of the 2nd USENIX Symp. on Operating System Design and Implementation (OSDI). 1996. 153−168.

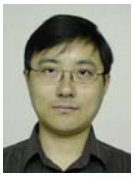
[10] Decasper D, Dittia Z, Parulkar G, Plattner B. Router plugins: A software architecture for next generation routers. IEEE/ACM Trans. on Networking, 2000,8(1):2−15.

[11] Gottlieb Y, Peterson L. A comparative study of extensible routers. In: Proc. of the IEEE Open Architectures and Network Programming. 2002. 51−62.

[12] Wu K, Wu JP, Xu K. A tree-based distributed model for BGP route processing. In: Porc. of the Int'l Con. on High Performance Computing and Communications 2006. LNCS 4208, 2006. 119−128.

附中文参考文献:

[3] 张晓哲,卢锡城,朱培栋,彭伟.一种集群路由器转发表同步框架及关键算法.软件学报,2006,17(3):445−453. http://www.jos.org.cn/1000-9825/17/445.htm



徐恪(1974－),男,江苏洪泽人,博士,副教授,主要研究领域为新一代互联网体系结构,高性能路由器体系结构.

王青青(1984－),女,硕士生,主要研究领域为 Overlay 网络,高性能路由器.

吴鲲(1976－),男,博士,主要研究领域为高性能路由器体系结构.