基于 Click 系统的优先级转发软件路由器①

龙 腾, 荀 鹏

(国防科学技术大学 网络工程系, 长沙 410073)

摘 要: 在云计算环境中, 利用软件技术调度内部网络流量成为云计算环境下的一个重要组成部分. Click 是一种模块化的软件路由器, 可以很好的接入到任何网络中, 并通过扩展其模块实现丰富的调度策略. 论文主要对 Click 软件路由器在调度模块上进行研究, 设计并实现一个能够基于优先级转发的路由器, 将软件路由器应用在云服务的请求调度入口, 当出现多种服务同时请求时, 实现按优先级调度以达到服务性能保障的目的, 解决峰值时重要服务响应缓慢的问题.

关键词: 云计算; 软件路由器; click 路由器; 服务; 服务优先级

Service Priority Software Router Based on Click System Computing Environment

LONG Teng, XUN Peng

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

Abstract: In the cloud computing environment, the use of software component become an important part for forwarding inside–network packets. Click is a modular software router and can be very goodly accessed to any network. By extending its modules, click router can achieve any complicated scheduling strategy. This thesis researches on how to design and implement a priority-based forwarding router and use it in the cloud computing environment. In the event of multiple service request at the same time, the router can forward these packets by priority. This method is used to solve the problem of important service with slow response time.

Key words: cloud computing; software router; click router; service; service priority

随着互联网的发展, 基于虚拟化的云计算规模越来越大, 云计算环境的建立也成为各大 IT 巨头研究建设的重点. 无论是国内外著名的 AWS、阿里云, 还是以开源为主的 OpenStack, CloudStack 等云计算平台[1], 内部的网络环境都是不可或缺的重要组成部分. 云计算环境中网络数据的转发调度策略决定了云环境中大量资源调度的性能, 也直接影响着云平台内部各种服务的响应时间. 在云计算环境下的众多服务中, 大量服务的接入路由器作为服务请求入口, 承受着严峻的考验. 例如: 当大量请求涌入时, 请求调度不及时导致服务请求等待时间过长, 影响服务性能; 当恶意流量过多时, 未经检测转发会造成大量计算资源用于处理非法请求, 导致服务性能降低.

一般来说, 服务性能保障可通过对服务所需网络资源, 计算资源两方面进行研究. 在网络资源方面, 文献[3–6]分别从服务所需网络资源以及外部存储资源进行研究, 从网络带宽利用角度及外部资源角度优化服务性能; 在计算资源方面, 文献[7-12]从云计算虚拟化技术进行研究, 有效地分配任务和虚拟资源调度来保障用户服务质量.

云计算数据中心实质上是由大量虚拟机构成的一个网络集群环境, 而一个网络环境的构建必然少不了网络核心设备路由器. 从性能上来考虑, 目前大部分云计算环境中的核心路由器仍由专用的硬件电路模块构成, 用以提供高速稳定的路由转发保障. 但是, 由于云数据中心相比传统物理机房具有更强的动态性

① 基金项目:国家自然科学基金(61170285)
收稿时间:2015-01-20;收到修改稿时间:2015-03-23

(如按需分配, 动态调控等), 因此在云计算环境中使用功能固定的硬件路由器会制约整个云计算网络的拓展. 如今软件定义网络也开始走向成熟, 软件路由器由于其高度扩展性, 价格低廉等特点, 也逐步在一些云环境建设中得到广泛使用.

本文立足于云计算环境, 以服务性能保障为目标, 从请求调度上进行相关研究, 设计和实现了一种基于Click[2]系统的软件路由器, 为云计算环境中的服务请求调度提供更好的保障.

1　相关研究

Click 是一种模块化的软件路由器, 它的设计目标是使路由器软件更加灵活并且易于配置和管理. Click 软件路由器由美国 MIT 大学 MIT 计算机技术系并行与分布式操作系统实验室开发完成. Click 软件路由器包括三个基本单元: 组件、连接、配置文档. 组件是路由器的一个功能处理单元, 也是Click 软件路由器的核心组成部分, Click 路由器也即是由一系列有序的组件模块进行连接而成. 组件的连接方式主要有下推(push) 、上拉(pull) 和不定方式(agnostic) 三种方式, push 连接表示组件执行结束后主动地将该数据报推向它之后的组件; pull 连接则表示该组件主动地向它之前的组件请求数据报, agnostic 连接由相邻组件连接方式决定. 组件之间的工作模型如图 1 所示.

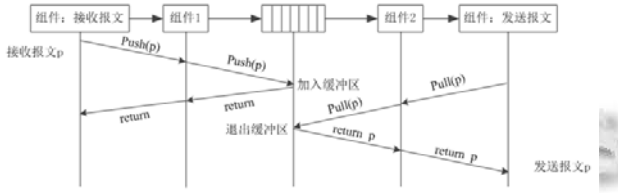

图 1　Click 软件路由器组件之间的工作模型

配置文挡描述一组组件和连接之间的关系, 通过在配置文件中组织各个组件逻辑关系形成一个软件路由器. 例如, 定义配置: FromDevice(eth0) −>Queue−> ToDevice(eth1), 表示一个从 eth0 接收报文, 经过组件 Queue 组件处理后从 eth1 端口转发的简易路由器, 其中 FromDevice, Queue, ToDevice 为这个简易路由器中的三个组件.

2　优先级软件路由器的设计

2.1 路由器模型设计

优先级路由器主要面向服务请求调度, 用来接收服务请求, 对请求报文进行分析, 确定目标服务, 再结合管理员设定的服务优先级, 按请求优先级进行转发. 服务优先级由管理员根据实际情况动态修改, 如系统监测到某一服务在一段时间内有大量用户访问, 管理员可调节优先级使该类服务请求优先转发, 甚至制定相应规则, 由程序动态调节服务优先级, 降低高并发服务请求的等待调度时间, 提高服务性能.

为更好的突出软件路由器的性能, 避免不同组件进程之间的影响, 将所有组件分类, 最后得到由三个软件路由器共同组成的优先级路由器. 优先级转发的调度主要由这三个软件路由器共同完成, 基本模型如图 2 所示.

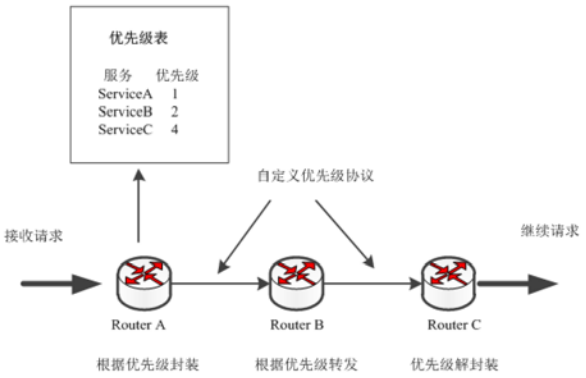

图 2　优先级软件路由器模型

Router A 负责接收分析报文, 识别目标服务, 并根据目标服务查询优先级表, 重新封装报文, 交由下一个软件路由器进行优先转发.

Router B 接收到带优先级的报文后, 根据优先级存入自身缓存队列, 正常情况下, 接收速率小于转发速率, 按照先进先出方式转发, 在出现高并发请求时, 接收速率会大于转发速率时, 此时每次转发选取缓存队列中优先级最高的请求进行转发.

Router C 主要对带优先级的报文解封装后恢复原有标准IP报文继续转发, 并记录服务调度日志等操作, 用于分析服务调度等待时间.

为了实现上述模型, 还需对服务识别、协议设计、优先级转发组件等进行具体分析.

2.2 软件路由器对服务的识别

目前针对服务的调用主要采用 SOAP、Post、Get、RESTful API[13-14]等方式, 其请求传输都是通过 Http 协

议完成, 不同的是具体响应数据格式各有区别, 常见的响应格式以 xml, json 为主. Http 协议的请求、响应在传输层上采用面相连接的传输协议 TCP 协议作为支撑, 因此每次在对服务进行一次请求时, 都会先建立连接再提交 Http 请求. 因此, 对于服务的请求识别, 可以通过分析一次请求报文来确定具体的服务.

所谓 HTTP 请求, 也就是 Web 客户端向 Web 服务器发送信息. 一个 HTTP 请求报文由请求行(request line)、请求头部(header)和请求数据(data)3 个部分组成, 图 3 给出了一次请求传输的实际报文.

图 3 利用 WireShark 抓取的 HTTP 请求报文

服务识别的目标主要是通过分析 http 请求报文, 获取 URL 地址. 因此首先需要对报文过滤, Mac 层协议号识别 0x0800(ip 协议); IP 层协议号识别 6(TCP 协议); 传输层上识别数据层偏移量及端口 80(http 常用端口), 由于有些 http 并不是部署在 80 端口, 因此 80 端口识别仅作参考. 在获取 http 请求头部数据偏移量后即可得到具体的 URL 地址. URL 地址即为一个目标服务请求地址, 服务识别进程利用该地址与优先级缓存中的优先级表进行对比, 获取此次请求调度的优先级.

优先级表主要为<key,value>键值对, key 为存储服务 URL, value 为服务的优先级. 管理员通过监控对服务的请求情况, 调节 value 数据, 用以管理服务优先级表. 为提高查表效率, 可由 Router A 定期查询数据库后写入软件路由器内存中. 在得到服务优先级数值后, 根据优先级协议重新封装报文, 转发至 Router B 进行排队转发.

2.3 优先级协议设计

为了让优先级路由器更好的对服务进行调度, 在标准 IP 数据报文上进行了简单协议设计, 并且为便于协议测试, 精简了优先级报文头格式, 主要突出优先级数值.

定义优先级范围为 0–255(1 个字节)之间的整数, 定义链路层类型为 0x0518,定义网络层协议号为 140, 用于对数据包的过滤. 具体数据报文格式如下表所示:

表 1 链路层服务优先级协议标识

目的 MAC 地址	源 MAC 地址	0x0518

表 2 网络层服务优先级协议标识

……		
TTL	Protocol(140)	Header Checksum
源 IP 地址		
目的 IP 地址		

表 3 优先级请求报文自定义优先级层

Priority(0–255)	Data

在图 2 模型中, Router A 即可通过上述协议格式构造符合优先级调度的数据报文, 转发给 RouterB 进行具体转发. RouterA 发往 Router B 的报文相比传统 IP 报文多了一个“自定义优先级层”, 如表 4 所示. 单独拆分为数据设计一个层, 可为后续协议改进提供更好的支持.

表 4 带优先级报文

Mac 头	IP 头	自定义优先级层	传输层	应用层

3 优先级软件路由器实现

3.1 路由器转发原理

优先级转发路由器主要原理为 Router B 在接收到数据报文后, 通过优先级协议中的优先级数值将报文按优先级存入缓冲区, 每次转发时选区缓冲区优先级最高的报文进行转发. 缓冲区结构如图 4 所示.

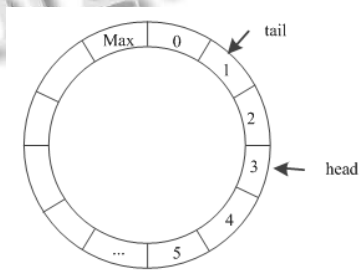

图 4 软件路由器缓冲区结构

缓冲区的数据结构为一个环形队列, 保存的数据区域由 head, tail 两个指针进行维护, head 指向队头, tail 指向队尾, 初始时两个指针均指向 0 处. 每次接收到一个报文时进行 Push 操作, 每次转发一个报文时进行 Pull 操作.

Push 操作: (1)判断缓冲区是否已满, 若满则丢弃报文并写入日志, 否则继续; (2)根据优先队列插入算

法, 将报文存入对应缓冲区域, 此时 head 指针加 1.

Pull 操作: (1)判断缓冲区是否为空, 若空则忽略, 非空则继续; (2)将 head 头指针指向的优先级最高的数据包弹出进行转发, 此时 tail 指针+1.

Head, Tail 两个指针依次循环计数通过“Max+1”模除取余后固定在[0, Max]之间. 当 head == tail 时表示缓冲区为空, head–tail == Max 时表示缓冲区已满. 由于是软件路由器, 缓冲区 Max 大小可根据软件路由器所在虚拟机配置具体设置.

3.2 路由器组件的设计

Click 软件路由器的一大特点即是由组件定义路由器的各个逻辑功能模块, 每个模块在实现上均为一个继承 Click 组件类的 C++程序. 软件路由器的设计, 事实上也是对路由器内部各个组件及逻辑关系的设计. 在文中的优先级调度路由器上, 将组件拆分, 分别置于三个不同路由器中, 也是为了突出组件灵活定制的特点.

在图 2 的模型中, 由于 Router A 将收到的普通请求报文重新封装成带优先级的报文, 因此可定义 Router A 为发送路由器, Router B 负责转发, 定义为转发路由器, Router C 负责解封装后继续转发, 定义为接收路由器. 为了进行更好的实验测试, 设计各个路由器中主要组件如图 5 所示.

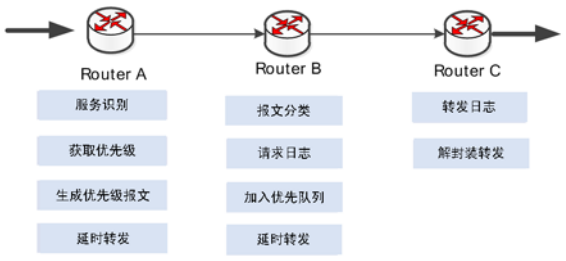

图 5　路由器中主要组件

“服务识别”组件主要用户对报文进行分析, 提取服务目标 URL 地址; “获取优先级”组件主要是与外部数据库进行通信, 定期将管理员定义的“服务–优先级”记录存入路由器缓存; “生产优先级报文”组件则是将普通请求报文生产带优先级的数据报文; “延时转发”组件用于控制路由器的转发速率; “报文分类”组件对收到的报文进行过滤, 负责转发只带有优先级协议的报文; “请求日志”组件按时间片记录请求到达日志; “加入优先级队列”组件则是将收到的报文按优先级存入路由器的待转发缓存队列中; “转发日志”组件用于记录经过优先级转发后的请求记录; “解封装转发”组件用于将报文恢复为正常请求, 发送给服务所在主机.

其中优先队列组件为核心, 包括优先转发相关的数据结构及算法实现, 并且为满足不同测试的需要, 优先队列的缓冲区还应设计为可配置, 在 click 配置文件中以参数形式设置缓冲区大小. 延时转发等组件用于控制路由器的转发速率, 实验时通过该组件调节 Router B 接收/转发速率, 产生便于分析的日志记录.

请求报文在不同路由器之间调度时, 按状态可划分为标准报文, 优先级报文两个状态. “标准报文”为接收时与解封装时的报文状态; “优先级报文”为路由器之间数据传递的报文状态. 两种状态在优先级路由器中体现为图 6 所示过程.

图 6　请求报文的状态变化

4　实验验证

4.1 实验拓扑

为了便于实验演示, 利用两台虚拟机作为软件路由器的实验主机, 再其中 1 台虚拟机上配置 2 台软件路由器, 分别模拟发送优先级报文及接收优先级报文的路由器(Router A 和 Router C), 在另 1 台上配置 1 台软件路由器, 模拟转发路由器(Router B). 实验拓扑如图 7 所示.

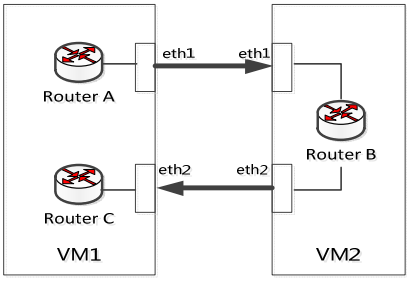

图 7　实验拓扑

4.2 实验测试

4.2.1 测试方法

VM1 中的 Router A 从 eht1 端口, 按一定频率生成带服务请求报文并按优先级封装, 数据转发到 VM2 中的 Router B 路由器时, Router B 打印接收到的数据并将数据按优先级存入队列缓存, 转发时将数据转发到 VM1 中的 eth2 端口, 由 Router C 进行处理, 打印接收到的相关数据.

Router A 按每秒发送 4 个报文的速度随机发出 200 个对 a/b/c 的请求, “a/b/c”模拟 3 种不同服务标识, 分别对应 ServiceA/ServiceB/ServiceC, 请求比例为 2:1:1. Router A 中延时组件为每秒转发 1 个报文, Router B 中延时组件为每 2 秒转发 1 个报文, 保证 RouterB 的接收速率大于转发速率.

4.3 实验结果

仿真实验每秒随机产生 4 个服务请求, 服务请求 A、B、C 其中之一, 共生成 200 个请求, 产生的三种请求比例为 2:1:1. 路由器转发时按每 2 秒执行一次转发调度. 得到对服务 A, B 的等待时间曲线如图 8 所示, 对于服务 C, 由于优先级最低, 不做分析.

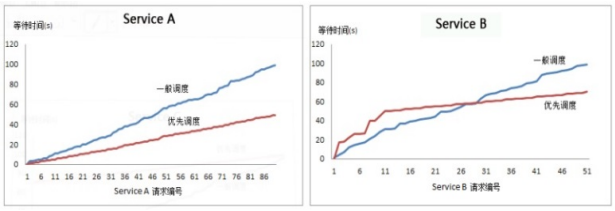

图 8 实验结果

结果说明: 对于优先级最高的 A 服务, 在路由器调度时能够根据优先级进行优先调度, 降低 A 服务请求的等待时间; 对于优先级次之的 B 服务, 在优先保障对 A 的请求时, 相比一般调度有较多等待时间, 但在 A 服务完成调度后, 仍能保证 B 服务的优先调度, 降低 B 服务的等待时间. 优先级调度路由器的优先转发功能验证成功.

5 结束语

实验演示测试表明, 选择局部最优的策略, 可以更好的保障优先级高的服务性能, 并且利用软件路由器同样可以达到硬件路由器的转发调度效果, 功能订制上以基于组件的方式进行编程开发, 可以按需自定义组件, 比硬件路由器更新固件的方式更方便. 在未来基于虚拟化的网络环境中, 软件路由器作为网络中的核心数据调度设备, 必将在云计算上有更加明显的优势.

参考文献

1 林利,石文昌.构建云计算平台的开源软件综述.计算机科学,2012,11:1–7

2 Kohler E. The Click modular router. http://www.read.cs.ucla.edu/click.

3 曹家鑫.数据中心中的一种可扩展和高效的可靠组数据传输方法[学位论文].合肥:中国科学技术大学,2013

4 冯振乾.云计算数据中心的网络带宽隔离技术研究[学位论文].长沙:国防科学技术大学,2012

5 Sun DW, Chang GR, Shang G. Modeling a dynamic data replication strategy to increase system availability in cloud computing environments. Journal of Computer Science & Technology, 2012, (2)

6 郝向涛.基于 Hadoop 的分布式文件系统技术分析及应用[学位论文].武汉:武汉理工大学,2013

7 林伟伟,齐德昱.云计算资源调度研究综述.计算机科学,2010,10:1–6

8 储雅,马廷淮,赵立成.云计算资源调度:策略与算法.计算机科学,2013,11:8–13

9 李冰.云计算环境下动态资源管理关键技术研究[学位论文].北京:北京邮电大学,2012

10 李建敦.私有云中虚拟资源的节能调度研究[学位论文].上海:上海大学,2011

11 Shao J, Wei H, Wang QX, Mei H. A runtime model based monitoring approach for cloud. 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD). 2010. 313–320.

12 刘永.云计算环境下虚拟机资源调度策略研究[学位论文].济南:山东师范大学,2010.

13 周荃.云计算资源管理中 QoS 保障机制研究[学位论文].北京:北京交通大学,2014

14 楚宇原.基于 SOAP 的 Web 服务安全模型研究[学位论文].哈尔滨:哈尔滨理工大学,2012

15 刘畅,孙连英,彭涛等.基于 RESTful 面向资源的 Web 服务研究.数学的实践与认识,2013,3:124–128