一种基于分段路由的多路径流传输机制

黄建洋,兰巨龙,胡宇翔,马 腾
(国家数字交换系统工程技术研究中心,河南郑州 450002)

摘 要: 针对传统网络多路径流量调度时存在的负载均衡效能差、路径部署困难的问题,利用软件定义网络的集中控制优势,设计了一种基于分段路由的多路径流传输(Segment Routing based Multipath Flow Transmission,SRMFT)机制. 首先,以实现数据流的协同最优调度为目标,建立了SRMFT最优化模型;其次,采用分段路由技术和最简段标识序列(Segment IDentify sequence,SIDs)生成算法将多路径流调度问题转化为最简SIDs的选择问题,并设计了流调度算法求解;最后,试验结果表明,同等网络流量模型下,与较典型的多路径流传输机制相比,SRMFT有效提高了网络的对分带宽,降低了短流的传输时延,同时具有较低的流表存储开销.

关键词: 分段路由; 软件定义网络; 多路径流; 负载均衡; 数据中心
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2018)06-1488-08
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.06.031

A Segment Routing Based Multipath Flow Transmission Mechanism

HUANG Jian-yang, LAN Ju-long, HU Yu-xiang, MA Teng
(*National Digital Switching System Engineering & Technology Research Center, Zhengzhou, Henan* 450002, *China*)

Abstract: To solve the problem of poor load balancing performance and difficult path deployment in the traditional networks, a SRMFT (Segment Routing based Multipath Flow Transmission) mechanism is designed, which utilizes the centralized control advantage of software-defined networking. Firstly, the SRMFT optimization model is set up to achieve the co-optimal scheduling of the data stream. Secondly, the multi-path flow scheduling problem is transformed into the simplest SIDs (Segment IDentify sequence) selection problem with segmented routing algorithm and the simplest SIDs generation algorithm. And the traffic flow scheduling algorithm is proposed. Finally, experimental results show that, under the same network traffic model, compared with the typical multi-path streaming mechanism, SRMFT effectively improves the bisection bandwidth of networks, reduces the transmission delay of short stream, and also have a lower flow table storage overhead.

Key words: segment routing; software defined networking; multipath flow; load balance; data center

1 引言

近年来,随着Web请求、电子商务、云计算和云存储等网络应用的飞速发展,作为信息服务基础设施的数据中心得到了广泛部署和应用. 如今的数据中心网络中通常包含百万服务器节点,并通过数据中心网络实现服务器间的通信[1]. 为了高效地互联数量如此庞大的服务器节点,研究人员提出了Fat-Tree, DCell, Bcube和VL2等高连通度网络拓扑[2]. 这些网络架构使得任何两台服务器间存在多条物理路径,便于网络管理员通过在多条可达路径间均衡网络负载以期提高网络整体性能.

针对网络中多路径间的流调度问题,国内外研究人员提出了不同的解决方案[3,4]. 等价多路径路由(E-qual-Cost MultiPath routing,ECMP)通过对数据包头进行哈希运算,将数据流映射到不同的转发路径上,并以随机化方法实现流的调度[5],然而由于非均匀流量在不同路径间的非均匀分布,该方案可能导致严重的网络性能降低. 多路径传输控制协议(MultiPath Transmission Control Protocol,MPTCP)方式通过使用多网卡主机为端

收稿日期:2016-11-18;修回日期:2017-04-12;责任编辑:梅志强
基金项目:国家"973"计划资助项目(No. 2012CB315901, No. 2013CB329104);国家自然科学基金资助项目(No. 61572519, No. 61502530);国家"863"计划资助项目(No. 2013AA013505, No. 2015AA016102)

到端的数据建立多条传输连接,将用户间的流量分为多条子流进行传输[6],可有效提高网络的吞吐量,但其实现过程复杂且加大了短流的完成时间. 上述方案均根据局部路由信息进行流量的调度,极易造成网络中链路或节点拥塞进而影响网络性能.

软件定义网络(Software Defined Networking, SDN)[7]是一种基于网络抽象思想的新型网络架构,其将网络控制平面和数据平面分离开来,提供对分布式网络的集中管理和软件编程能力. 参考 SDN 控制器的全局网络状态信息去调度信息流,为优化网络性能带来了便利. Hedera 是一个可扩展、动态的网络流管理系统,其利用集中控制器检测并评估网络中出现的大流,并通过首次适应算法将其调度到满足带宽需求的低负载 ECMP 路径上[8]. Mahout 也是通过优化大流的管理进而提高带宽利用率,与 Hedera 不同的是该方案优先选取拥塞度最小的 ECMP 路径进行流调度[9]. 除此之外,还有基于通配符的流量调度(DevoFlow, DIFANE)、面向 QoS 的流量调度(OpenQoS)和基于负载均衡决策的流量调度(LABERIO, FSEM)等相关研究方案被提出,充分显示了 SDN 在资源管控方面有着不可替代的优势.

2013 年,Internet 工程任务组(Internet Engineering Task Force, IETF)提出了分段路由(Segment Routing, SR)技术[10,11]. SR 是一种基于源路由的隧道技术,通过将数据转发路径信息保存在数据包头去控制包的路由. 由于数据包中包含了路径转发信息,可以根据需要控制数据包沿任何路径转发,因此 SR 是一种具有极强服务路径定制能力的路由技术. 利用 SDN 的全局网络视图能力,可参照当前网络资源利用状况为 SR 源节点选择基于特定路由调度策略的数据转发路径,因此将 SDN 与 SR 进行结合已成为信息基础网络发展的一个新研究点[12,13]. 此外,由于 IP 和 MPLS 依然是当前网络的主流路由方式且已得到广泛部署,为了与现有网络体系架构实现较好的兼容,这就使得现有的 SDN 控制研究必须支持区分路由形式. 目前 SDN 主要通过 OpenFlow 多级流表的方式支持区分路由形式,这也为通过 MPLS 数据平面来实现 SR 技术带来了便利.

本文通过借鉴已有的新研究思路,利用 SDN 全局网络视图和集中控制的优势,结合具有极强服务路径定制能力的 SR 技术,提出了一种基于分段路由的多路径流传输(Segment Routing based Multipath Flow Transmission, SRMFT)机制. 针对多路径流调度问题,SRMFT 首先以网络效应最大化为目标生成源、目的对间的等价最优路径集合;然后采用 SR 技术将路径集合中的各路径映射为 SR 段序列(SIDs),并提出最简 SIDs 生成算法将各 SIDs 压缩为最简 SIDs;最后将源、目的节点对之间的多路径流调度问题转化为最简 SIDs 的选择问题. 流调度时,SRMFT 优先为时延敏感的短流分配资源,但也通过更新优先级权重的方式对长期未被调度的长流进行优先调度. 为进一步减小流的完成时间,SRMFT 首先选择 SID 数最小的最简 SIDs 进行流的传输.

2　SRMFT 机制及其最优化模型

2.1　SRMFT 机制

SR 是一种部署灵活、高扩展和可兼容的路由机制,其通过一串有序段标识(Segment Identify, SID)序列表示任意端到端的路由路径,其中每个 SID 表示一个节点(或者链路和服务). 流传输时,数据包从源节点出发,以最短路径方式依次从前一个 SID 对应节点被转发到后一个 SID 对应节点,直至最终到达目的节点. 如图 1 所示,各链路上的数字表示 IGP 链路权重,若想要在 R1 与 R8 之间建立一条流传输路径,则会优先选择路径 R1→R3→R5→R7→R8. 假设链路 R1→R3 和 R3→R5 均发生拥塞,我们想让数据包沿路径 R1→R2→R4→R5→R7→R8 转发,SR 会以表 1 所示的过程控制数据包到达 R8 节点. 由此可见,SR 仅在源节点处保存单流状态,中间节点只完成 SID 处理和包转发工作. SR 可与现有网络协议体系实现较好的兼容,不需要对 MPLS 数据平面做任何修改即可实现 SR 的部署. 此外,对目前普遍使用的 IGP(如 ISIS 和 OSPF)进行扩展[10,11],可实现各节点 SID 在全网范围内的通告.

表 1　SR 控制数据包转发过程

节点	栈顶 SID	操作
R1	SID(R4)	转发给 R2
R2	SID(R4)	转发给 R4
R4	SID(R4) SID(R7)	弹出栈顶 SID 转发给 R5
R5	SID(R7)	转发给 R7
R7	SID(R7) SID(R8)	弹出栈顶 SID 转发给 R8
R8	SID(R8) ∅	弹出栈顶 SID 接收数据包

将 SR 与 SDN 进行结合,可以很好地解决当前网络面临的一些资源规模扩展受限、组网灵活性差和满足业务需求困难等问题[14]. 例如:利用 SDN 控制器完成 SR 数据转发路径的计算,然后将计算好的 SIDs 下发给源节点,可解决 SR 的源节点计算能力受限问题.

基于以上描述,提出面向 SDN 的 SRMFT 机制总体结构图,如图 2 所示,该结构主要包括:应用层、控制层和数据层. 应用层基于不同的流调度目标,结合控制器

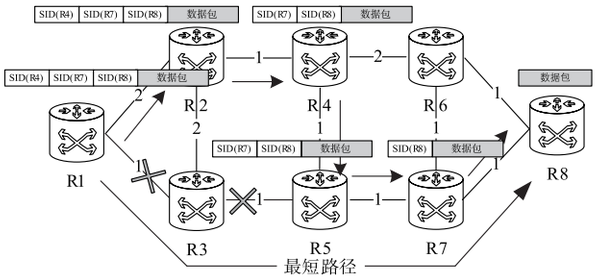


图1 SR网络拓扑范例

的资源信息状态,形成数据流的协同最优调度决策. 控制层是网络的管理中心,作用主要有:存储和更新网络资源视图、配置网元支持 SR 功能和接收应用层流调度决策并完成各流传输路径 SIDs 的构建与下发. 数据层是由标准化的硬件设备(如 OpenFlow 交换机)构成的网络数据报文处理和传输通道. 其主要作用包括:接收控制层下发的流传输 SIDs,对数据包执行具体的转发操作.

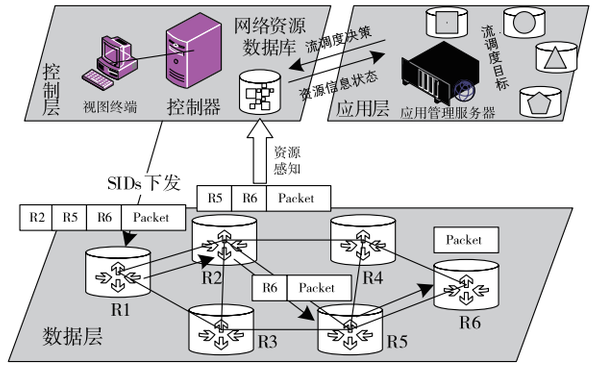


图2 SRMFT机制总体结构图

2.2 SRMFT 最优化模型

用有向图 $G=(N,E)$ 表示一个网络,其中 N 为节点集合,E 为链路集合. SRMFT 最优化模型是一种网络效用最大化模型,与已有的网络效用最大化模型的不同点在于:(1)已有的网络效用最大化是一种流速率控制模型,而 SRMFT 是一种流调度路径选择模型,其特点是网络中的注入流量是固定的,由网络管理者选择源、目的节点之间的数据传输路径进而获得最大的网络效应. (2)已有网络模型的效应函数都是链路负载的函数,而 SRMFT 最优化模型的网络效应函数是关于链路剩余容量的严格凹函数,因为当链路产生故障时,选取剩余容量非零的链路进行路由重计算更加便利[15]. SRMFT 模型的设计目标是:为实现网络资源的最优利用,由网络管理者去协调调度源、目的节点对之间的数据流,将全网范围内各源、目的节点之间的流调度问题看做多商品流问题,且要求最后解得的流分布 $\boldsymbol{f}=(f_{ij},(i,j)\in E)$ 是一个可行流分布.

定义 1 (可行流分布)如果存在流分布向量 $\boldsymbol{f}^M=(\boldsymbol{f}^m,m\in M)$ 满足式(1)的约束条件,则称该流分布为可行流分布:

$$\begin{cases} f_{ij}=\sum\limits_{m\in M}f_{ij}^m\leqslant c_{ij},\forall(i,j)\in E \\ \sum\limits_{i:(n,i)\in E}f_{ni}^m-\sum\limits_{j:(j,n)\in E}f_{jn}^m= \\ \quad \begin{cases} d_m,\text{if}(n=\text{Source}(m)) \\ -d_m,\text{if}(n=\text{End}(m)) \\ 0,\text{otherwise} \end{cases},\forall m\in M,\forall n\in N \\ f_{ij}^m\geqslant 0,\forall m\in M,\forall(i,j)\in E \end{cases} \tag{1}$$

上式中,第一行为容量约束;第二行为流守恒约束,其中 $\sum\limits_{i:(n,i)\in E}f_{ni}^m$ 表示从节点 n 流出的属于源、目的节点对 m 中的流量总和,$\sum\limits_{j:(j,n)\in E}f_{jn}^m$ 表示流入节点 n 中的属于源、目的节点对 m 中的流量总和.

根据以上描述,本文建立的 SRMFT 最优化模型如式(2)所示,模型中用到的符号及含义如表 2 所示.

$$\max U(\boldsymbol{r})$$
$$\text{s.t.}\begin{cases} \boldsymbol{r}+\sum\limits_{m\in M}\boldsymbol{f}^m=c \\ A\cdot\boldsymbol{f}^m=\boldsymbol{d}^m,\forall m\in M \\ \boldsymbol{r}\geqslant 0,\boldsymbol{f}^m\geqslant 0 \end{cases} \tag{2}$$

表 2 SRMFT 最优化模型中的符号及含义

符号	含义
M	源、目的节点对集合,元素是各节点对 $m\in M$
$\boldsymbol{c}$	链路容量向量,$\boldsymbol{c}=(c_{ij},(i,j)\in E)$
$\boldsymbol{r}$	剩余容量向量,$\boldsymbol{r}=(r_{ij},(i,j)\in E)$
$\boldsymbol{r}^*$	最优剩余容量分配向量
$\boldsymbol{f}^m$	节点对 m 的流量分布向量,$\boldsymbol{f}^m=(f_{ij}^m,(i,j)\in E)$
$\boldsymbol{f}^{m*}$	最优的节点对 m 的流量分布向量
$\boldsymbol{d}_m$	源、目的节点对 m 之间的请求带宽
$\boldsymbol{d}^m$	节点向 m 注入流量值向量,$\boldsymbol{d}^m=(d_n^m,n\in N)$
$U(\cdot)$	网络效用函数,是剩余容量向量 $\boldsymbol{r}$ 的函数

式(2)的目标是最大化网络的聚合效用,而网络的聚合效用依赖于每条链路获得的总的剩余带宽分配;第一行限定了网络链路的资源指标约束;第二行限定了流守恒条件,其中 A 是一个 $N\cdot E$ 的关联矩阵,实现了式(1)中第二行中的等式;第三行表明任何链路剩余容量和承载流量都不能为负.

3 核心算法设计

本节将对 SRMFT 机制的相关算法进行详细的设计,主要包括 OD 对等价最优路径集合的生成、最简 SIDs 的生成和流的路由调度.

3.1 OD 对等价最优路径集合的生成

3.1.1 SRMFT 最优化模型分析

为了便于求解 SRMFT 最优化模型,本文引入了 Lagrange函数:

$$L(\boldsymbol{r},\boldsymbol{f}^m,\boldsymbol{\lambda}) = U(\boldsymbol{r}) - \boldsymbol{r}\cdot\boldsymbol{\lambda} - \sum_{m\in M}\boldsymbol{\lambda}\cdot\boldsymbol{f}^m + \boldsymbol{c}\cdot\boldsymbol{\lambda} \quad (3)$$

其约束条件为 $A\cdot\boldsymbol{f}^m=\boldsymbol{d}^m, \forall m\in M$. 其中,$\boldsymbol{\lambda}=(\lambda_{ij}(i,j)\in E)$为拉格朗日乘子向量,$\lambda_{ij}$可以理解为网络管理者对每单位剩余容量支付的价格. 根据凸优化理论,若$(\boldsymbol{r}^*,\boldsymbol{f}^{m*})$是式(2)的最优解,则存在非负的拉格朗日乘子向量$\boldsymbol{\lambda}$,使得$(\boldsymbol{r}^*,\boldsymbol{f}^{m*})$也是下面式(4)的最优解[15].

$$\begin{aligned} D(\boldsymbol{r},\boldsymbol{f}^m,\boldsymbol{\lambda}) &= \max_{\boldsymbol{r}\geqslant 0,\boldsymbol{f}^m\geqslant 0} L(\boldsymbol{r},\boldsymbol{f}^m,\boldsymbol{\lambda}) \\ &= \sum_{(i,j)\in E} A_{ij}(r_{ij},\lambda_{ij}) + \sum_{m\in M}\sum_{(i,j)\in E} B_{ij}(\lambda_{ij}) \\ &\quad + \sum_{(i,j)\in E} C_{ij}(\lambda_{ij}) \end{aligned} \quad (4)$$

其中:

$$A_{ij}(r_{ij},\lambda_{ij}) = \max_{r_{ij}\geqslant 0} u_{ij}(r_{ij}) - \lambda_{ij}r_{ij} \quad (5)$$

$$B_{ij}(\lambda_{ij}) = \min_{f_{ij}^m\geqslant 0} \lambda_{ij}f_{ij}^m \quad (6)$$

$$C_{ij}(\lambda_{ij}) = \max\ \lambda_{ij}c_{ij} \quad (7)$$

拉格朗日乘子λ_{ij}可以理解为链路(i,j)上附加流量的影子价格. 在计算 OD 对之间的最优路径集合时,本文将λ_{ij}看作路由计算时的链路权重.

3.1.2 最优链路权重的获取

假设$(\boldsymbol{r}^*,\boldsymbol{f}^{m*})$是问题式(2)的最优解,由于$(\boldsymbol{r}^*,\boldsymbol{f}^{m*})$也是问题(4)的最优解,则根据 Karush-Kuhn-Tucker (KKT)[16]条件可知,存在乘子向量$\boldsymbol{\lambda}^*$使得下式成立:

$$\nabla_r D(\boldsymbol{r}^*,\boldsymbol{f}^{m*},\boldsymbol{\lambda}^*) = \nabla_r\Big(\sum_{(i,j)\in E}\max_{r_{ij}\geqslant 0} u_{ij}(r_{ij}^*) - \lambda_{ij}^* r_{ij}^*\Big) = 0 \quad (8)$$

由式(8)解得,若已知问题式(2)的最优解$(\boldsymbol{r}^*,\boldsymbol{f}^{m*})$,则存在$\boldsymbol{p}^*\in\partial U(\boldsymbol{r}^*)$,使得$\boldsymbol{\lambda}^*=\boldsymbol{p}^*$. 在设置最优链路权重$\lambda_{ij}$时,本文规定:

$$\lambda_{ij}\begin{cases} = p_{ij}, & \text{if}(r_{ij}^*>0) \\ \geqslant p_{ij}, & \text{if}(r_{ij}^*=0) \end{cases} \quad (9)$$

式(9)中,当$r_{ij}^*=0$时,说明链路已饱和,本文设置较大的链路权重值,使得新加入流量时链路代价较大;当$r_{ij}^*>0$时,链路还未饱和,链路权重值设定为效用函数对于最优剩余容量的偏导.

定义 2　若$u_1(\boldsymbol{r}),u_2(\boldsymbol{r}),\cdots,u_m(\boldsymbol{r})$是$R^E\to R$上 m 个连续可微的效用凹函数,则设定:$u(\boldsymbol{r})=\min\{u_1(\boldsymbol{r}),u_2(\boldsymbol{r}),\cdots,u_m(\boldsymbol{r})\}$,其有效集为$I(\boldsymbol{r})=\{i\mid u_i(\boldsymbol{r})=u(\boldsymbol{r}),i=1,2,\cdots,m\}$,偏导为:$\partial u(\boldsymbol{r})=\Big\{\sum_{i\in I(\boldsymbol{r})}\theta_i\nabla u_i(\boldsymbol{r})\mid\theta_i\geqslant 0, i\in I(\boldsymbol{r}), \sum_{i\in I(\boldsymbol{r})}\theta_i=1\Big\}$.

当获取到所有链路的最优权重后,对于任意源、目的节点对(s_m,t_m),可计算出对应的等价最优路径集合.

3.1 最简 SIDs 的生成

假设上一小节中为(s_m,t_m)之间计算的一条最优路径为$P=(n_1,\cdots,n_k)$,其中$n_1=s_m,n_k=t_m$. 为了以最少 的分段路由 SID 数表示该路径,本文提出了一种最简 SIDs 生成算法,如算法 1 所示.

算法 1　最简 SIDs 生成算法

输入:路径$P=(n_1,\cdots,n_k)$
输出:P 的最简 SIDs
1. $S_P=\{s_1,s_2,\cdots,s_k\}$ = convertSeg(P) /* 将 P 中各节点转化为对应 SID */
2. initStack(S)
3. push(S,s1)
4. top = 1, pointer 1 = 2, pointer2 = 3
5. for pointer2 = 3:k
6. if onlyShortestPath($\{s_{\text{top}},\cdots,s_{\text{pointer2}}\}$) = = ture /* 若$\{s_{\text{top}},\cdots,s_{\text{pointer2}}\}$是唯一的最短路径 */
7. pointer1 = pointer2, pointer2 + +
8. else
9. push(S,s_{pointer1})
10. top = pointer1, pointer1 = pointer2, pointer2 + +
11. end if
12. end for
13. push(S,s_k)
14. popAll(S)　//输出 P 的最简 SIDs

算法 1 的复杂度为 $O(k)$,其中 k 为网络中所有路径的平均节点数. 经算法 1,可将计算出的所有候选流调度路径转化为对应的最简 SIDs.

3.2 流的路由和调度

对于任何一条流,SRMFT 路由机制在其等价最优路径集合中找到一条 SR 段(SID)数目最少的路径作为该流的传输路径. 具体实现过程为:当一条流的报文到达控制器,控制器首先根据全局网络视图信息为该流计算等价最优路径集合;然后通过算法 1 将集合中的各路径转化为最简 SIDs;优先选择段数目最小的最简 SIDs 为该流传输,当有多个这样的最简 SIDs 时,控制器优先选择流数目最少的最简 SIDs. 此外,每次在进行流的调度时,SRMFT 优先为短流分配传输路径,同时也通过更新优先级权重的方式对长期未被调度的长流进行优先调度,描述 SRMFT 的路由调度过程如算法 2 所示.

算法 2　SRMFT 路由调度

输入:$F=(f_1,\cdots,f_m)$/* 所有请求传输的流 */
输出:$S_F=(s_{f_1},\cdots,s_{f_m})$/* 分配给各流的最简 SIDs */
1. $W=\varnothing$

```
2.  for f_i in F do
3.    w_i = f_i
4.    W = W + w_i
5.  end for
6.  w_j = min(W) /* 优先调度权重最小的流 */
7.  f = f_j
8.  while F ! = ∅ do
9.    P = calECOP(f)  /* 求 f 的等价最优路径集合 */
10.   S_p = ∅
11.   for p in P do
12.     s_p = Algorithm1(p)  /* 计算 p 的最简 SIDs */
13.     S_p = S_p + s_p
14.   end for
15.   S_p^min = min(S_p)
16.   s = minFlowNum(S_p^min) /* 获得流数目最少的最简 SIDs */
17.   s_f = s
18.   Adjust(F,W)  /* 更新 F 和 W */
19.   for f_i in F do
20.     if isNewFlow(f_i) = = true /* 设置新流的优先级权重 */
21.       w_i = f_i
22.       W = W + w_i
23.     else
24.       w_i = 1/2 * w_i/ * 旧流的优先级权重减半 */
25.     end if
26.   end for
27.   w_j = min(W) /* 获取各流优先级的最小权重值 */
28.   f = f_j
29. end while
```

若网络中需要传输的流数目为 m,平均每条流的等价最优路径数为 n,则算法 2 的时间复杂度为 $O(mn)$. 又因为在一般网络内部,n 的数目远小于 m,所以算法 2 的时间复杂度为 $O(m)$.

4 实验结果与分析

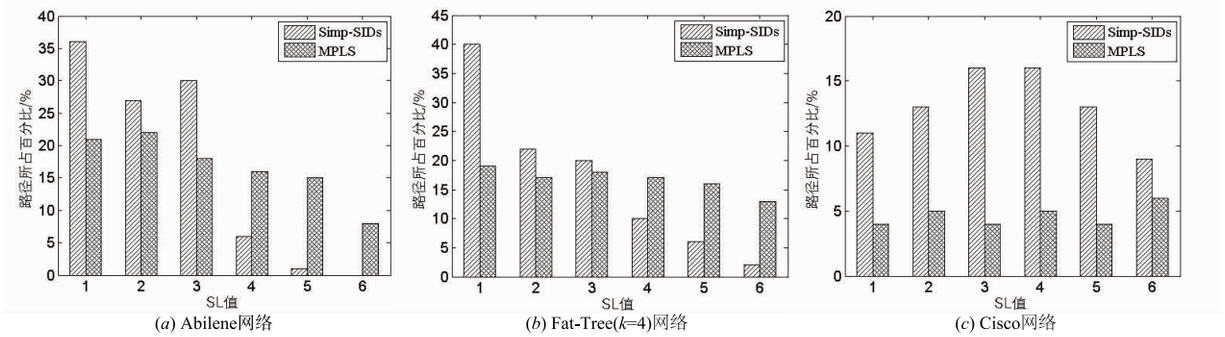
4.1 算法性能评估

(1)实验设置:通过 Matlab 软件评估本文所提的最简 SIDs 生成算法(记为 Simp-SIDs). 所对比的算法是:MPLS 标签映射法(记为 MPLS),该算法在源节点处直接将路径中的所有节点映射为对应 SID. 实验时,利用生成随机数的方式产生用于通信的源、目的节点对以及二者之间的传输路径,然后比较不同算法的平均段序列长度(记为 SL).

(2)实验结果及分析:表 3 给出了两种算法在不同拓扑中的平均 SL. 其中,三个拓扑中 Simp-SIDs 的平均 SL 分别为 MPLS 的 68.3%、67.87% 和 52.99%,表明 Simp-SIDs 算法有效地压缩了源节点处 SR 段的数目,且随着网络越复杂压缩率越高. 图 3 显示了两种算法在不同 SL 值下的路径分布,其分析结论与表 3 结果相一致.

表 3 两种算法在不同拓扑中的平均 SL

拓扑	节点数	链路数	最大 SL	Simp-SIDs	MPLS
Abilene	11	14	6	2.09	3.06
Fat-Tree(k=4)	20	32	6	2.26	3.33
Cisco	76	160	14	4.6	8.68



(a) Abilene网络 (b) Fat-Tree(k=4)网络 (c) Cisco网络

图3 各网络不同SL值下的路径分布

4.2 试验验证

试验选用 ONOS + Minnet 作为验证平台. 其中,ONOS 是一款具备高扩展性、高可用性和高性能的网络操作系统,在试验中作为 SDN 控制器,且为完成本文验证,我们在其原有基础上实现了 SR 功能扩展;Mininet 作为 SDN 网络仿真器,在试验中用于虚拟出所有交换节点. 试验选用 k = 4 的 Fat-Tree 拓扑(如图 4 所示),网络中包含 20 台交换机和 16 台服务器,设定所有链路的带宽为 10Mbit/s,链路时延为 3ms,传输数据量小于

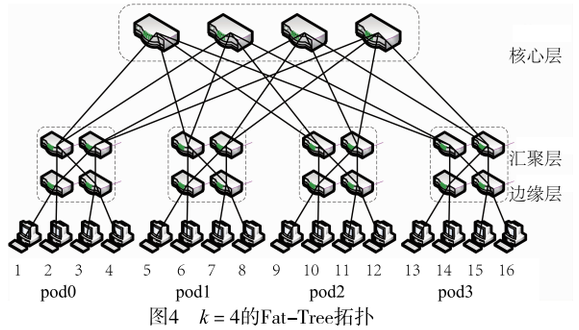


图4 k = 4的Fat-Tree拓扑

100KB 的流为短流,大于 100KB 的流为长流,同时,大于 1MB 的流占总流数的 30%,并且网络中 95% 的字节由这些流产生.

试验时,每台主机根据某种通信模式来选取目的主机,并且目的主机只接收网络中某一台主机的流量数据. 参考文献[1],本文选用的通信模式有以下三种:

(1)间隔方式(Stride(x)):每个主机与距离其位置 x 的其它主机通信.

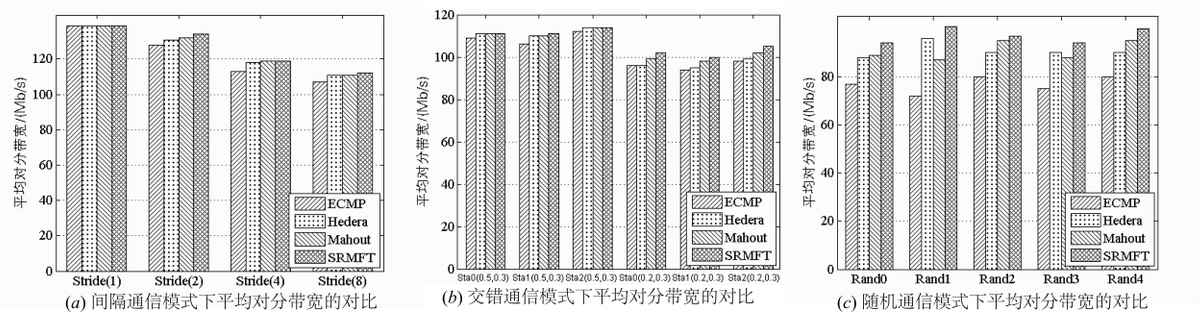
(2)交错方式(Staggered(pEdge,pPod)):每个主机以概率 pEdge 向在同一个边缘交换机的主机发送数据,以概率 pPod 向在同一 Pod 内的主机发送数据,以概率 1-pEdge-pPod 向核心交换机发送数据.

(3)随机方式(Random):每个主机以等概率随机向网络中其他主机发送数据.

4.2.1　数据传输性能对比

为验证 SRMFT 的数据传输性能,本文选择 ECMP、Hedera 和 Mahout 作对比试验. SRMFT 的效用函数选择最大化网络吞吐量,为了提高试验的准确性,每组试验重复 50 次,并取平均值作为试验结果.

图 5 为四种传输机制在不同通信模式下的网络对分带宽测试结果.



(a) 间隔通信模式下平均对分带宽的对比　(b) 交错通信模式下平均对分带宽的对比　(c) 随机通信模式下平均对分带宽的对比

图5　各网络不同SL值下的路径分布

由图 5(a)可知,随着间隔值 x 的增大,各传输机制的平均对分带宽都有所下降. 其原因为:ECMP 是分布式静态路由机制,难以对拥塞链路进行路由再调整,因此平均对分带宽下降的最快;Hedera、Mahout 和 SRMFT 均是集中式动态路由机制,但 SRMFT 设定以最大化网络吞吐量为效应函数,可根据当前网络资源利用情况对网络流量进行调整,获得最大的网络吞吐量,因此具有较大的平均对分带宽.

图 5(b)为 Staggered(0.5,0.3)和 Staggered(0.2,0.3)两种交错通信模式下各传输机制的平均对分带宽对比图,试验时每种通信模式下取 3 组统计值. 在 Staggered(0.2,0.3)交错模式下,Pod 间的流数量增加,发生冲突的流数量也随之增加,各机制的平均对分带宽都有所下降. 又因为此时跨 Pod 间的流数量依旧很少(占总流的 50%),ECMP 和 Hedera 都以随机化方法进行流调度,并未对流冲突作出规避,因此平均对分带宽下降较快;Mahout 以局部优化的方式能在一定程度对流冲突进行规避,但优化效果有限;SRMFT 是一种以最大化网络吞吐量为效应函数的全局最优传输机制,可实现全局网络流量的最优分布,因此具有比其它三种算法更高的平均对分带宽.

图 5(c)为随机通信模式下各传输机制的平均对分带宽对比图,试验时每种通信模式下取 5 组统计值. 该通信模式下,跨 Pod 间的流量占大多数,发生流冲突的概率也高于前两种通信模式. ECMP 以随机化方法进行流调度,容易发生流冲突现象,因此平均对分带宽最小;Hedera 通过模拟退火算法得到流优化路径,Mahout 优先选取拥塞度最小的 ECMP 路径进行流调度,但二者对流的调度优化都不太明显;SRMFT 通过全局的方式对流进行调度优化,可有效避免冲突流的数量,具有较高的平均对分带宽.

图 6 为四种传输机制在随机通信模式下短流的平均传输时延测试结果. 由图中数据可得,SRMFT 较其它三种传输机制具有较低的短流平均传输时延,即使网络负载大于 70% 时,短流的完成时间依然保持在 50ms 以内. 其原因在于 SRMFT 在进行流调度时,优先为时延敏感的短流分配传输路径;此外,在进行传输路径选取时,SRMFT 优先选择 SID 数最小的最简 SR 段序列,进一步减小了短流的传输时延.

4.2.2　资源开销对比

SDN 交换机中的流表大部分都是由 TCAM 构成的,同时控制器单位时间可以处理的流请求消息数目也是有限的,因此 SDN 网络中的流表资源是受限的,需要进行有效的管理. 由于 SRMFT 是一种 SDN 网络架构下的多路径流传输机制,所以为验证 SRMFT 的资源开销,本文选择传统 SDN 机制(记为 SDN)作对比试验.

图 7 为两种机制在随机通信模式下的网络资源开销对比结果,其中,横坐标代表测试时间,间隔为

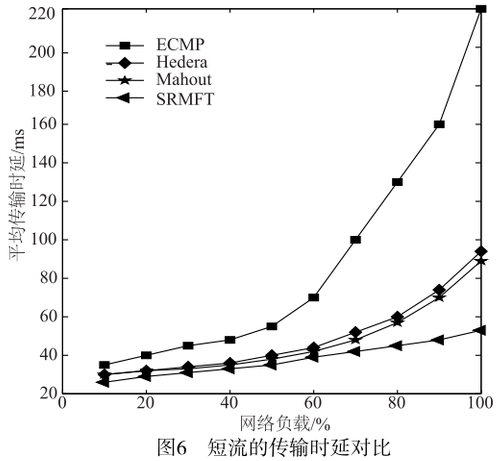


图6 短流的传输时延对比

100ms,纵坐标代表四台核心层交换机的总流表条目数.SDN 机制在整个测试时间内四台核心层交换机的总流表条目数介于 153 ~ 183(均值为 168)之间,而 SRMFT 机制的一直是 144.其原因在于:传统 SDN 机制在部署流传输路径时需要向所有相关节点下发流表,而 SRMFT 机制采用了分段路由模式,其通过 openflow1.3 协议以及多级流表的方式仅在源节点处将流传输路径的 MPLS 标签加在数据包头,中间节点只需根据包头的标签对数据包进行处理,因此在整个数据传输过程中没有新流表的加入.

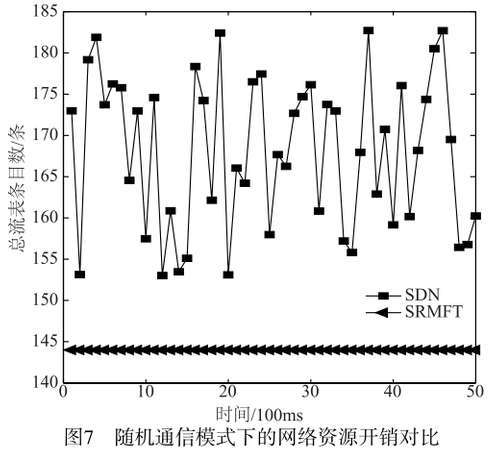


图7 随机通信模式下的网络资源开销对比

5 结束语

针对传统网络中多路径流调度时负载均衡效能差、路径部署困难等问题,本文采用软件定义网络架构,同时基于 IETF 提出的分段路由技术,提出了一种基于分段路由的多路径流传输(SRMFT)机制.该机制采用集中控制调度方式,在全局网络视图下通过 SRMFT 最优化模型实现网络流的最优调度.同时,为实现快速灵活的路径部署与切换,SRMFT 采用 SR 技术实现了流传输路径的部署.系统试验测试结果表明:数据传输性能方面,SRMFT 具有较高的平均对分带宽,同时缩短了短流的传输时延,使短流的传输时延维持在较低的时间范围内;资源开销方面,SRMFT 较传统的 SDN 机制节省了流表存储空间.目前,SDN 和 SR 有机结合的相关研究正在不断兴起,下一步将继续完善本文的研究,实现在真实网络环境中对 SRMFT 进行性能验证分析.

参考文献

[1] Al-Fares M, Loukissas A and Vahdat A. A scalable, commodity data center network architecture [J]. ACM SIGCOMM Computer Communication Review, 2008, 38(4): 63 – 74.

[2] 林智华,高文,吴春明,等.基于离散粒子群算法的数据中心网络流量调度研究[J].电子学报,2016,44(9):2197 – 2202.
LIN Zhi-hua, GAO Wen, WU Chun-ming, et al. Data center network flow scheduling based on DPSO algorithm [J]. Acta Electronica Sinica, 2016, 44(9): 2197 – 2202. (in Chinese)

[3] Wang N, Ho K, Pavlou G, et al. An overview of routing optimization for internet traffic engineering [J]. IEEE Communications Surveys & Tutorials, 2008, 10(1): 36 – 56.

[4] 牛志升,段翔,刘进. MPLS 网络中保证服务质量的多径路由选择策略[J].电子学报,2001,29(12):1638 – 1641.
NIU Zhi-sheng, DUAN Xiang, LIU Jin. A QoS-guaranteed mult-ipath routing policy for mpls networks [J]. Acta Electronica Sinica, 2001, 29(12): 1638 – 1641. (in Chinese)

[5] Chiesa M, Kindler G, Schapira M. Traffic engineering with Equal-cost-multipath: An algorithmic perspective [A]. Proceedings of IEEE INFOCOM [C]. Toronto: IEEE, 2014. 1590 – 1598.

[6] Raiciu C, Barre S, Pluntke C, et al. Improving datacenter performance and robustness with multipath TCP [J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 1 – 12.

[7] I F Akyildiz, A Lee, P Wang, et al. Research challenges for traffic engineering in software defined networks [J]. IEEE Network, 2016, 30(3): 52 – 58.

[8] Alfares M, Radhakrishnan S, Raghavan B, et al. Hedera: dynamic flow scheduling for data center networks [A]. Proceedings of N-etworked Systems Design and Implementation [C]. San Jose: NS-DI, 2010. 19 – 19.

[9] Curtis A R, Kim W, Yalagandula P, et al. Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection [A]. Proceedings of IEEE INFOCOM [C]. Shanghai: IEEE, 2011. 1629 – 1637.

[10] Filfils C, Nainar N K, Pignataro C, et al. The segment rou-

ting architecture[A]. Proceedings of IEEE Global Communications Conference[C]. Austin, Texas, USA: IEEE, 2014. 1 – 6.

[11] Filfils C, Nainar N K, Pignataro C, et al. Segment Routing with MPLS Data Plane[DB/OL]. https://tools. ietf. org/html/draft-ietf-spring-segment-routing-mpls-08, 2013.

[12] R Bhatia, F Hao, M Kodialam, et al. Optimized network traffic engineering using segment routing[A]. Proceedings of IEEE Global Communications Conference[C]. Kowloon: IEEE, 2015. 657 – 665.

[13] S Bidkar et al. Field trial of a software defined network (SDN) using carrier ethernet and segment routing in a tier-1 provider[A]. Proceedings of IEEE Global Communications Conference [C]. Austin: IEEE, 2014. 2166 – 2172.

[14] Hartert R, Schaus P, Vissicchio S, et al. Solving segment routing problems with hybrid constraint programming techniques[A]. Principles and Practice of Constraint Programming, CP215[C]. Cork, Ireland: Springer, 2015. 592 – 608.

[15] K Xu, M Shen, H Liu, et al. Achieving optimal traffic engineering using a generalized routing framework[J]. IEEE Transactions on Parallel and Distributed Systems, 2015, 27(1): 1 – 1.

[16] H Yaiche, R R Mazumdar, C Rosenberg. A game theoretic framework for bandwidth allocation and pricing in broadband networks[J]. IEEE/ACM Transactions on Networking, 2000, 8(5): 667 – 678.

作者简介

黄建洋　男, 1991 年出生, 陕西合阳人. 2014 年毕业于北京师范大学计算机科学与技术专业, 其后进入国家数据交换系统工程技术研究中心攻读硕士学位, 主要研究方向为新型网络体系结构、网络空间拟态防御.
E-mail: m15136143225@163. com

兰巨龙　男, 1962 年出生, 河北张北人. 国家数据交换系统工程技术研究中心总工程师、教授、博士生导师, 主要从事新一代信息网络关键理论与技术的研究工作, 目前作为首席科学家主持国家“973”项目“可重构信息通信基础网络体系研究”.