

基于合作博弈的数据中心骨干网带宽分配策略

孟 飞 兰巨龙 胡宇翔

(国家数字交换系统工程技术研究中心 郑州 450002)

(mf472933350@126.com)

A Cooperative Game Based Data Center Backbone Network Bandwidth Allocation Policy

Meng Fei, Lan Julong, and Hu Yuxiang

(National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002)

Abstract Currently, traffic engineering is typically deployed to improve the utilization of data centers (DC) backbone networks, which usually belongs to the same online service providers. Although the efficiency is remarkable, the bandwidth allocation fairness of different aggregate flow isn't considered. Hence, the QoS guarantee is restricted. Because the bandwidth resource is expensive and packet loss is typically thought unacceptable, the bandwidth utilization should be maximized, at the same time, the QoS guarantee of different flow should be improved. In this paper, the problem of contending the share bandwidth is modeled as a cooperative game, and different aggregate flow competes the share bandwidth and maximizes the overall bandwidth resource utilization simultaneously, and the optimal bandwidth allocation policy, called cooperation game based bandwidth allocation (CGBA), is obtained through searching the Nash bargaining solution (NBS) of the game and balancing the tradeoff between minimum bandwidth guarantee and bandwidth allocation fairness. Simulation on a Mininet testbed shows that the proposed policy can effectively guarantee minimum bandwidth of each aggregate flow while ensuring the allocation fairness, compared with three other classical bandwidth allocation policies.

Key words data centers (DC); cooperative game theory; Nash bargaining solution (NBS); bandwidth allocation; Mininet

摘 要 数据中心(data center, DC)之间通过部署流量工程来提高连接各个数据中心骨干网的利用率,虽然效率提升显著,但对不同类型汇聚流的带宽分配的公平性没有考虑.将多个汇聚流对带宽分配的竞争行为建模为一个合作博弈,通过寻求此博弈的纳什谈判解(Nash bargaining solution, NBS)来确定优化的带宽分配策略 CGBA(cooperation game based bandwidth allocation),权衡各汇聚流的最小带宽保证与带宽分配的公平性.在 Mininet 平台上进行实验仿真并和典型的带宽分配策略对比,结果表明 CGBA 不但可保证各汇聚流的最小带宽需求,还确保了各类流对带宽资源竞争的公平性.

收稿日期:2014-12-23;修回日期:2015-03-19

基金项目:国家“九七三”重点基础研究发展计划基金项目(2012CB315901,2013CB329104);国家自然科学基金项目(61372121);国家“八六三”高技术研究发展计划基金项目(2013AA013505)

This work was supported by the National Basic Research Program of China (973 Program) (2012CB315901,2013CB329104), the National Natural Science Foundation of China (61372121), and the National High Technology Research and Development Program of China (863 Program) (2013AA013505).

关键词 数据中心;合作博弈理论;纳什谈判解;带宽分配;Mininet

中图法分类号 TP391

随着云计算技术的迅速发展,数据中心(data center, DC)已成为一种重要的信息通信基础设施,它采用虚拟化技术将海量的计算、存储、网络等物理资源高度整合为一个共享虚拟资源池^[1],实现资源的高效共享.为了提高数据中心服务的性能和可靠性,数据中心通常分布在地理位置相距很远的世界各地,彼此之间通过高速骨干网络互连^[2-3].这些骨干网络通常属于同一个在线服务提供商(online service providers, OSPs),如 Google 的 G-scale^[4],这些网络的建设成本巨大且其中发生数据丢包是不可接受的^[5],因此,高效合理地分配利用数据中心骨干网带宽资源且保证数据流的传输服务质量(quality of service, QoS)十分必要.

目前,对于数据中心骨干网带宽分配的研究已经成为学术界的一个重要研究课题,并取得了大量研究成果. LBAPS^[6], NetStitcher^[7] 都是通过感知带宽使用状况,前者优先把待传输数据块上传到空闲带宽大的节点,即通过占用空闲带宽来减少传输时延;后者使用存储-转发算法调度数据块,并根据带宽使用状况而实时调整变化. GRESE^[8] 是 1 个可以减少峰值带宽消耗的调度算法,在流量高峰时段传输实时或时延敏感数据,而在非流量高峰时期传输非时延敏感流量,通过在不同时段对带宽进行分配,可使带宽的使用代价显著减少. 以上 3 种方案均是通过调度策略对带宽在时间轴上进行的分配,虽可减少数据流传输的带宽开销,但是没有考虑不同类型数据流的带宽需求差异,即没有对不同数据流的差异需求区分对待. Ghosh 等人^[9] 提出了一种数据中心骨干网可扩展多类流管理策略,通过网络分层实现管理可扩展,根据需求的不同为每类流进行带宽分配,定义每类流的效用函数并以整体效用最大化为目标,但是它没有考虑不同流之间带宽分配的公平性. OSPs 通常通过部署流量工程合理布局流量来提高数据中心骨干网的链路带宽利用率,例如 Google 的 B4^[5] 通过使用基于 OpenFlow 的 SDN 架构实施流量工程,把应用流分隔部署到多条路径上来均衡流量,链路的平均带宽利用率高达 70%,它使用最大最小公平(max-min fairness)^[10] 算法为各类流分配带宽,提供了较高的公平性. 但鉴于最大最小公平算法固有的缺陷, B4 对各类流的带宽需求差异性考虑不足,因而使得对带宽需求较大的流的

QoS 保障受限. 在数据中心骨干网链路带宽分配过程中,公平和效率是要综合考虑的 2 个方面,只有这样才能在保证带宽资源高效利用的同时为各类流提供可预测的传输性能,提供较高水平的 QoS 保障.

博弈论^[11] 是应用数学的一个分支,适合于研究具有竞争或对抗性质的各种行为. 在带宽分配过程中,不可避免地会存在多个任务对有限带宽资源的竞争,而博弈论又恰好能有效地解决多个自私个体之间的竞争问题,从而达到全局任务效用值最优. 合作博弈纳什谈判解(Nash bargaining solution, NBS)已广泛用于解决资源分配中效率和公平的权衡^[12]. 受文献[12-14]启发,本文将数据中心骨干网链路上多类汇聚流对共享带宽的竞争分配问题建模为一个合作博弈,各汇聚流之间竞争带宽并以最大化整体效用为目标,通过设计集中式的带宽分配算法来寻求该博弈问题的纳什谈判解,得到优化的带宽分配策略,并将该策略称为基于合作博弈的带宽分配(cooperation game based bandwidth allocation, CGBA). 最后基于 Mininet^[15] 对所设计的模型进行实验仿真验证,并与现有典型策略的实验结果对比以表明其优势.

1 基于合作博弈的带宽分配模型

1.1 网络模型

如图 1 所示,本文的网络架构采用 2 层控制架构,该结构目前已被大多数基于 SDN 的数据中心网络所普遍采用^[4-5,14]. 网络中有 2 种类型控制器:多个

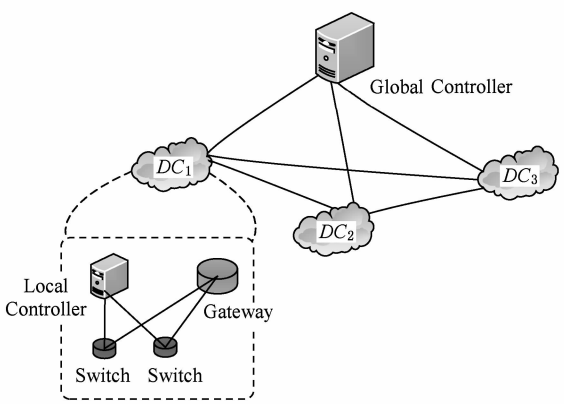


Fig. 1 Network architecture.

图 1 网络架构

本地控制器和 1 个全局控制器,前者对所属单个数据中心(DC)进行管理,不同 DC 之间相互独立,地位相同;1 个集中式全局控制器管理整个网络.数据中心间通过可编程边界网关(programmable border gateways)互连,数据中心内部虚拟机(virtual machine, VM)流量在边界网关汇聚分类,两 DC 间的通信可简化为分属的两网关间的通信.本文基于博弈理论的思想对数据中心骨干网链路上多个汇聚流的带宽分配问题进行建模分析.带宽资源被多条流共享,在进行带宽分配时除了要使带宽资源整体利用率最大化外还要对各个不同类型数据流提供最小带宽保障及保证分配的公平性,合作博弈的解——纳什谈判解(NBS)可有效权衡公平性和效率的关系,因此基于合作博弈模型分析解决此问题更为合适.

定义 1. 图 2 中,数据中心骨干网中共有 M 条链路 $L = \{link_1, link_2, \dots, link_M\}$,其带宽分别为 $C = (C_1, C_2, \dots, C_M)$. 此 M 条链路上共有 N 条汇聚流 $Flow = \{flow_1, flow_2, \dots, flow_N\}$,定义各流的带宽需求向量和分配向量分别为 $D(t) = (D_1(t), D_2(t), \dots, D_N(t))$, $b(t) = (b_1(t), b_2(t), \dots, b_N(t))$,其中, $D_i(t)$, $b_i(t)$ 分别表示时刻 t 汇聚流 $flow_i$ 的带宽需求和所分配的带宽.由于只考虑某一特定时刻的瞬时带宽需求的分配情况,为简便起见,时间 t 可省去.定义基带宽向量 $B^0 = (B_1^0, B_2^0, \dots, B_N^0)$ 作为带宽分配时的基准值,以使得每条流均能获得一个最小带宽,提供最低限度的 QoS 保障.分布于同一链路上的所有流的基带宽之和要小于此链路的带宽容量,以保证当所有流的带宽需求均高于其基带宽时,有多余的带宽可供使用,不至造成所有流的带宽

需求都得不到满足,都处于带宽“饥饿”状态.基带宽的选取可根据网络中某条流的历史需求信息决定.

定义 2. 矩阵 $F = (f_{ij})_{N \times M}$ 表示 DC 骨干网中各汇聚流的分布,二进制变量 f_{ij} 定义如下:

$$f_{ij} = \begin{cases} 1, & flow_i \text{ 在 } link_j \text{ 上,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

带宽分配的原则是若某条流的带宽需求小于其基带宽,则满足其需求;反之则分配其满足带宽分配公平性要求的带宽,定义带宽分配的上限和下限分别为 $b_i^u = \min(D_i, \sum_{j=1}^M f_{ij} C_j)$, $b_i^d = \min(B_i^0, D_i)$,则 $b_i \in \{b_i^d, b_i^u\}$.

1.2 博弈建模

设 $X \subset R^N$ 是 N 个汇聚流的带宽分配向量空间, R^N 是所有可能的分配结果的集合.给每条流分配一个初始带宽,初始带宽向量为 $b^0 = (b_1^0, b_2^0, \dots, b_N^0)$, b_i^0 即为 $flow_i$ 的最小效用值.带宽分配集合为 $B = \{b | b \in X, b_i \geq b_i^0, i \in \{1, 2, \dots, N\}\}$,则效用集合 B 是 R^N 上的一个非空、闭、有界的凸子集. (B, b^0) 是一个 N 人合作博弈.

定义 3. 若映射 $\phi: (B, b^0) \rightarrow R^N$ 满足: $\phi(B, b^0) \in B$ 、个体理性、可行性、帕累托最优性、不相关选择独立性、线性变换独立性和对称性,则此映射为纳什谈判博弈问题.定义向量 $b^* = (b_1^*, b_2^*, \dots, b_N^*)$ 是此合作博弈的 NBS,即 $b^* = \phi(B, b^0)$,则 b^* 满足:

$$b_j^* \in \arg \max_{b_j \in B} \phi_j(b_1^*, b_2^*, \dots, b_{j-1}^*, b_j, b_{j+1}^*, \dots, b_N^*).$$

(2)

该解的意义可解释为在 $(b_1^*, b_2^*, \dots, b_{j-1}^*, b_{j+1}^*, \dots, b_N^*)$ 给定的条件下,任意 $flow_i$ 不能通过改变其带宽分配策略 b_j ($b_j \neq b_j^*$),使自己获得更大的效用即分配的带宽,且同时其他参与人的效用又不被降低.

定理 1. 存在唯一满足定义 3 中条件的解,且是下述问题的最优解:

$$\max_{b_i \in B, \forall i \in \{1, 2, \dots, N\}} \prod (b_i - b_i^d).$$

(3)

NBS 的存在性与唯一性证明见文献[11].将目标函数取对数,式(3)等价于:

$$\max_{b_i \in B, \forall i \in \{1, 2, \dots, N\}} \sum \ln(b_i - b_i^d).$$

(4)

由定理 1 可知,CGBA 博弈的 NBS 是下述全局最优问题的解:

$$\begin{aligned} & \max_{b_i \in B, \forall i \in \{1, 2, \dots, N\}} \sum \ln(b_i - b_i^d), \\ & \text{s. t. } b_i \leq b_i^u, \forall i \in \{1, 2, \dots, N\}, \end{aligned}$$

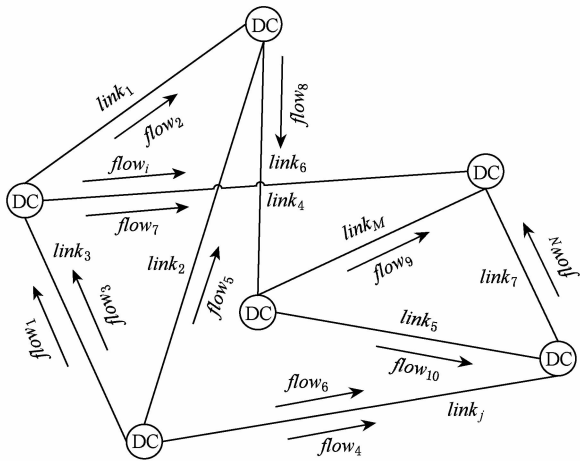


Fig. 2 N flow contend the share bandwidth.
图 2 N 条汇聚流竞争链路带宽资源

$$b_i \geq b_i^d, \forall i \in \{1, 2, \dots, N\},$$

$$\sum_{flow_i \in F_{L_m}} b_i \leq C_m, \forall link_m \in L. \quad (5)$$

定理 2. 存在 $\alpha_i \geq 0, \beta_i \geq 0 (i \in \{1, 2, \dots, N\})$ 和 $\gamma_m \geq 0 (m \in \{1, 2, \dots, M\})$, 使得 $\forall i \in \{1, 2, \dots, N\}, \forall m \in \{1, 2, \dots, M\}$ 有

$$\frac{1}{b_i^* - b_i^d} + \alpha_i - \beta_i - \sum_{m=1}^M \gamma_m f_{mi} = 0, \quad (6)$$

$$\alpha_i (b_i^d - b_i^*) = 0, \quad (7)$$

$$\beta_i (b_i^* - b_i^u) = 0, \quad (8)$$

$$\gamma_m ((b \times F)_m - C_m) = 0. \quad (9)$$

证明. 利用拉格朗日乘子方法分析最优化问题式(5), 其拉格朗日函数定义为

$$L(b', \alpha, \beta, \gamma) = \max_{b_i \in B, \forall i \in \{1, 2, \dots, N\}} \ln(b_i - b_i^d) -$$

$$\sum_{i=1}^N \alpha_i (b_i^d - b_i) - \sum_{i=1}^N \beta_i (b_i - b_i^u) -$$

$$\sum_{m=1}^M \gamma_m ((b \times F)_m - C_m), \quad (10)$$

其中, $\alpha_i, \beta_i, \gamma_m$ 为非负的拉格朗日乘子. 求关于 b_i 的一阶导数, 并根据 Karush-Kuhn-Tucker (KKT)^[16] 条件, 则有式(6)~(9)成立. 证毕.

“互补问题”作为一类新的数学模型, 是 1964 年美国 Cottle^[17] 在其博士学位论文中提出的. 非线性互补问题 (nonlinear complementarity problem, NCP) 是研究非线性互补条件的问题, 其定义为: 映射 $S: \mathbb{R}^n \rightarrow \mathbb{R}^n$, 求向量 $x \in \mathbb{R}^n$, 满足 $x \geq 0, S(x) \geq 0, x^T S(x) = 0$, 非线性互补问题记为 $NCP(S)$, 其中 $\mathbb{R}^n$ 表示 n 维欧氏空间, $\mathbb{R}^n$ 中的矢量均为列矢量, $x^T y$ 表示 x 与 y 的内积.

NCP 函数的定义为: 函数 $\varphi: \mathbb{R}^2 \rightarrow \mathbb{R}$, 对 $\forall x, y \in \mathbb{R}^2, \varphi(x, y) = 0$, 当且仅当 $x \geq 0, y \geq 0, x^T y = 0$. 因为 NCP 函数可自动保证非线性互补条件的成立, 利用某些 NCP 函数, 我们可将互补问题转化为等价的方程组. 对于任意一个 NCP 函数 $\varphi(\cdot, \cdot)$, 非线性互补问题 $NCP(S)$ 可等价地转化为方程组:

$$\Phi(x') = \begin{bmatrix} \varphi(x'_1, S_1(x')) \\ \varphi(x'_2, S_2(x')) \\ \vdots \\ \varphi(x'_n, S_n(x')) \end{bmatrix} = 0. \quad (11)$$

本文选择 Mangasarian^[18] 提出的 NCP 函数 $\varphi(a, c) = |a - c|^3 - a^3 - c^3$, 这个函数关于 a, c 是光滑的. 利用这个 NCP 函数将不等式约束问题的 KKT 条件, 即式(6)~(9), 转化为等价光滑方程组:

$$S(z) = \begin{bmatrix} \nabla_{x'} L(b', \lambda) \\ \varphi(b_1^d - b_1, \alpha_1) \\ \varphi(b_2^d - b_2, \alpha_2) \\ \vdots \\ \varphi(b_N^d - b_N, \alpha_N) \\ \varphi(b_1 - b_1^u, \beta_1) \\ \varphi(b_2 - b_2^u, \beta_2) \\ \vdots \\ \varphi(b_N - b_N^u, \beta_N) \\ \varphi(\sum_{i=1}^N b_i f_{i1} < C_1, \gamma_1) \\ \varphi(\sum_{i=1}^N b_i f_{i2} < C_2, \gamma_2) \\ \vdots \\ \varphi(\sum_{i=1}^N b_i f_{iM} < C_M, \gamma_M) \end{bmatrix} = 0. \quad (12)$$

后续的问题就是如何选择简单适当的算法去求解这个光滑方程组. 带一维搜索的 Newton 法克服了传统 Newton 法的局部收敛缺点, 使此方法具有全局收敛性的同时保持了传统 Newton 法的简单、收敛快的优点. 共轭梯度法 (conjugate gradient, CG)^[19] 是带有特殊性质的共轭方向法, 仅需利用一阶导数信息, 在产生共轭向量计算新向量时, 只要用一个相邻的前面的向量, 所需存储和计算量很少. 为了不失效率而又保持收敛性, 对 Newton 方程运用 CG 法迭代, 在达到系统的某些非精确近似解时终止迭代, 用非精确一维搜索 Newton-CG 法来求解大型非线性光滑方程组, 算法描述如下:

算法 1. 基于 Newton-CG 算法的带宽分配计算.

输入: 初始点 $z_0 \in R^{2N+M}$ 、流分布矩阵 $F = (f_{ij})_{N \times M}$ 、带宽需求向量 $D = (D_1, D_2, \dots, D_N)$ 、链路带宽容量 $C_m (m \in \{1, 2, \dots, M\})$ 、迭代次数 $k=0$ 、参数 η_k 和精度 ϵ ;

输出: 带宽分配计算结果 $b = (b_1, b_2, \dots, b_N)$.

① if $\|F(z_k)\| \leq \epsilon$, 结束, else 转②;

② 用 CG 算法解线性方程组

$\nabla F(z_k)^T \nabla F(z_k) p_k = -\nabla F(z_k)^T F(z_k)$, 令 $r_k = \nabla F(z_k) p_k + F(z_k)$, 当 $\|r_k\| \leq \eta_k \|F(z_k)\|$ 时, 得 p_k ;

③ 令 $x_{k+1} = x_k + \alpha_k p_k$, 其中 α_k 满足

Wolfe 或 Armijo 后退准则等;

④ $k = k + 1$, 转①.

本算法的计算复杂度主要集中在步骤②③, 在

使用 CG 算法求解线性方法组时,终止条件 $\|r_k\| \leq \eta_k \|F(z_k)\|$ 使得算法计算量减少,计算复杂度降低;在进行迭代求解时,仅需知道前面一个向量,存储空间可重复利用,算法空间复杂度低.非精确的一维 Newton 搜索算法简单,全局收敛速度快,计算复杂度低.因此,本算法的计算复杂度和空间复杂度均较低.本算法的实质是用共轭梯度法解 Newton 方程,具有单步收敛性,高稳定性,关于共轭梯度法的收敛性证明可参见文献[20].

2 集中式带宽分配算法

在进行带宽分配时,控制器需要监测底层网络资源使用状况,网络状态信息的获取是由各交换机上报本地控制器,然后再上报全局控制器而得到的. CGBA 部署在集中式全局控制器上计算全网的带宽分配结果,其中,算法 1 计算得到带宽分配结果,算法 2 控制带宽分配的迭代收敛.得到结果后全局控制器向特定的某条流的源、目的 DC 本地控制器下发结果,本地控制器再向相关交换机下发配置命令,通过对特定队列的流表的配置实现数据包转发速率控制.

算法 2. 集中式带宽分配控制.

输入:基带宽 $\mathbf{B}^0 = (B_0^0, B_1^0, \dots, B_N^0)$ 、总迭代次数 S 、精度 Δ ;

输出:带宽分配向量 $\mathbf{b} = (b_1, b_2, \dots, b_N)$.

- ① for 所有的流: $\forall i \in \{1, 2, \dots, N\}$, do
- ② 初始化 b_i^d 和 b_i^u ;
- ③ $b_i^{(0)} = b_i^d$;
- ④ while $s < S$ or $b_i^{(s)} - b_i^{(s-1)} > \Delta$ do
- ⑤ 在集中式控制器中运行算法 1;
- ⑥ end while
- ⑦ end for

由于要保持全网信息的一致性,SDN 网络中控制器会周期性地向交换机节点下发询问信息,因此通过在下发询问信息时捎带控制器对节点的带宽配置信息则避免了对节点配置时的一部分额外开销.

本算法的计算复杂度主要在步骤④⑤,算法最多迭代 S 次,因此其计算复杂度最大为 S 倍的算法 1 的计算复杂度, S 为定值且算法 1 复杂度低,则本算法计算复杂度较低.

3 CGBA 的仿真实验与性能分析

3.1 实验场景建立

类比 Fallocc^[12] 建立实验仿真验证环境,在基于 OpenFlow 的 Mininet 仿真平台上对 CGBA 进行仿真验证.建立同图 1 所示架构的网络拓扑,为不失一般性同时为简便分析起见,建立 3 个相互连通的 DC 域,即存在 3 条链路;此网络中共有 5 条流,则 $M=3, N=5$.如图 3 所示,3 个交换机相互连通,每个交换机连接 4 个主机对外等价于一个 DC 域,链路的带宽向量为 $\mathbf{C} = (1 \text{ Gbps}, 2 \text{ Gbps}, 3 \text{ Gbps})$.在进行带宽分配时,要从效率和公平 2 方面综合比较分配策略的性能,效率主要从对各流的最小带宽的保障考虑,根据对需求的满足程度衡量效率的高低,定义平均带宽分配满足度(satisfaction degree, SD)为

$$SD = \frac{1}{N} \sum_{i=1}^N \frac{b_i}{D_i}, \tag{13}$$

其中, SD 越接近于 1,说明对各流的带宽分配满足较好,效率越高;反之则效率越低.对公平性的考量,选用文献[21-22]中的公平指数(fairness index, FI)来进行评价,本文公平指数的定义为

$$FI = \frac{(\sum_{i=1}^N b_i)^2}{N \times (\sum_{i=1}^N b_i^2)}, \tag{14}$$

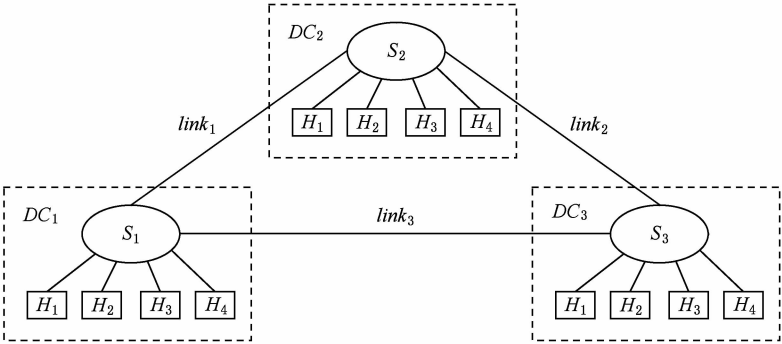


Fig. 3 The simulation topology.
图 3 仿真拓扑图

其中, FI 越接近于 1, 说明公平性越高; 反之, 则说明公平性越低。

在资源饱和场景(所有链路均能满足其上的全部流的带宽需求之和)中, 全部流需求都会获得满足, 没必要进行带宽博弈, 按需分配即可, 因此本文重点研究资源匮乏(至少存在一条链路的带宽容量不能满足其上的所有流的带宽需求之和)的情况, 并将 CGBA 和其他策略进行对比。

3.2 数值结果与讨论

基带宽的选取是根据每条流的历史需求信息选取的, 越接近博弈结果则求解得到最终结果的速度越快, 博弈时间越短。不同的流分布对 CGBA 本身不产生影响, 即不会影响 CGBA 的性能, 因此随机取定流分布矩阵

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

本文把 CGBA 和按比例带宽分配策略(proportional policy, PP)、基于优先级的带宽分配策略(priority based policy, PBP)及最大最小公平分配(max-min fairness, MMF)进行比较。这 3 种策略的描述如下:

1) 按比例带宽分配策略(PP). 若分布在带宽容量为 C_i 的同一链路 $link_i$ 上的流为 $flow_1, flow_2, \dots, flow_j$, 其带宽需求分别为 $D_1, D_2, \dots, D_j$, 则为每条流分配的带宽为 $b_j = C_i \times \frac{D_j}{\sum_{s=1}^j D_s}$ 。

2) 基于优先级的带宽分配策略(PBP). 将同一链路 $link_i$ 上的流按优先级排序为 $flow_1, flow_2, \dots, flow_j$. 若其总需求不大于带宽容量, 即 $\sum_{s=1}^j D_s \leq C_i$, 则每条流分配的带宽等于其需求; 否则, 存在 $k (0 < k < j)$ 使得 $\sum_{s=1}^k D_s \leq C_i$, $\sum_{s=1}^{k+1} D_s > C_i$, 则前 k 条流的分配带宽为其需求, 其他流的分配带宽为 0, 即优先保障高优先级数据流的带宽需求, 只有在高优先级的需求得到满足之后才去考虑低优先级部分。

3) 最大最小公平分配(MMF). 若分布在带宽容量为 C_i 的链路 $link_i$ 上的流为 $flow_1, flow_2, \dots, flow_n$, 其带宽需求从小到大排序为 $D_1, D_2, \dots, D_n$, 那么, 我们初始把 C_i/n 带宽分配给 $flow_1$, 这可能会

超过 $flow_1$ 的需求, 将超出 $flow_1$ 需求的那部分加上剩余没有分配的带宽平均分给其他流, 依次按此

方法处理, 则 $b_i = \min \left[\frac{C - \sum_{j=1}^{i-1} b_j}{n - i + 1}, D_i \right] (1 \leq i \leq n)$ 。

该过程结束时, 每条流得到的带宽不会比自己要求的多, 而且, 如果其需求得不到满足, 得到的带宽也不会比其他得到最多带宽的流得到的少。

图 4、图 5 表示基带宽为 1 Gbps, $D_3 = D_5 = 1$ Gbps, $D_2 = D_4 = D$ (为仿真分析简便, 其他情况仅是变量个数的增加、维度的扩展)时, 运用 4 种策略分别进行带宽分配的所得结果。在带宽资源充足(带宽需求 D 在 0 到 1 Gbps 之间)时, 即资源饱和场景中, 4 种策略结果相同。在带宽资源匮乏场景中, CGBA 通过合作博弈对带宽进行最优分配, 满足各流的最小带宽需求之后, 剩余带宽按照需求差异按比例分配, 可以提供一唯且公平的帕累托最优解, 兼顾了不同的需求和最小性能保证, 具有很高的公平性, 效率和公平 2 方面性能均优于其他策略。MMF 是一种对低需求用户按需保障、高需求用户“尽力而为”保障的分配策略, 因其过于追求流量公平, 没有考虑各个流之间的需求差异性, 所以当各流的需求差异较大时, 对需求大的流的保障则会造成限制, 反而造成 MMF 公平性不是很好; 同时因为没有对流需求差异加以区分, 导致需求较大的流的带宽分配满足度较低(随需求的增大呈反比例降低), 故整体 QoS 保障水平受限, 效率较低。PP 实质上是一种按需求大小加权的分配策略, 对不同需求用户机械地“一视同仁”对待, 仅根据需求的占比决定分配的大小, 因此平均带宽分配满足度和公平指数均

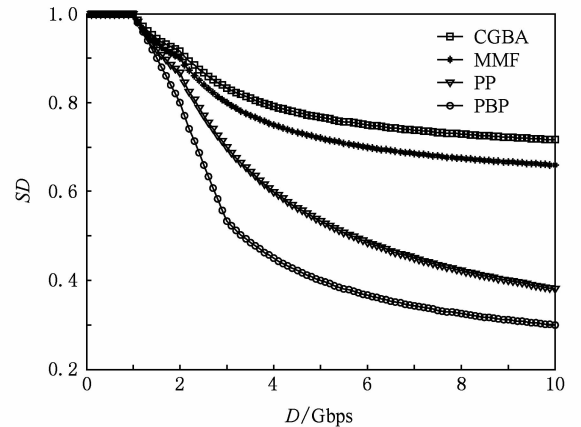


Fig. 4 The results of average satisfaction degree.

图 4 平均带宽分配满足度

较低. PBP 仅考虑各类流的优先级的不同, 对高优先级的流按需满足, 而对优先级低的流可能会造成带宽分配“饿死”现象, 平均带宽分配满足度和分配公平性在 4 种策略中最低.

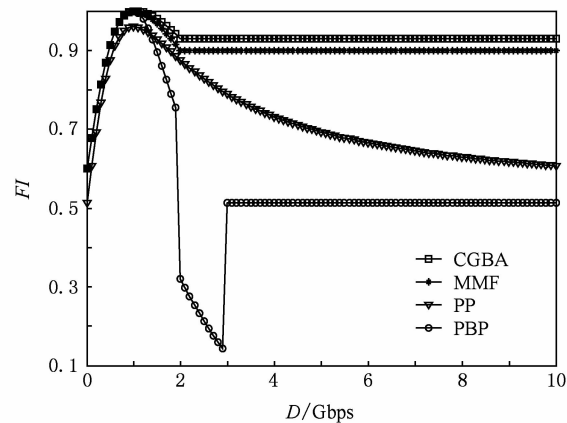


Fig. 5 The results of fairness index.
图 5 公平指数

按照“天下没有免费的午餐 (no free lunch, NFL)”^[23]理论, 一种算法不可能在任何方面或在任何问题上都能够提供比其他所有算法更好的性能. 在我们的实验结果中也观察到了这点, 如图 6 所示, 考察在给定基带宽为 1 Gbps, $flow_2$ 和 $flow_4$ 的带宽需求分别为 1.5 Gbps, 2.5 Gbps, 3.5 Gbps 时 3 种策略的迭代收敛情况. 图 6 表明: 选取的基带宽合适与否影响迭代的周期, 即带宽分配博弈的时间, 越接近最优值则所需寻优迭代次数越少, 收敛越快; CGBA 的收敛速度比 MMF 快, 较 PP 和 PBP 慢, 并

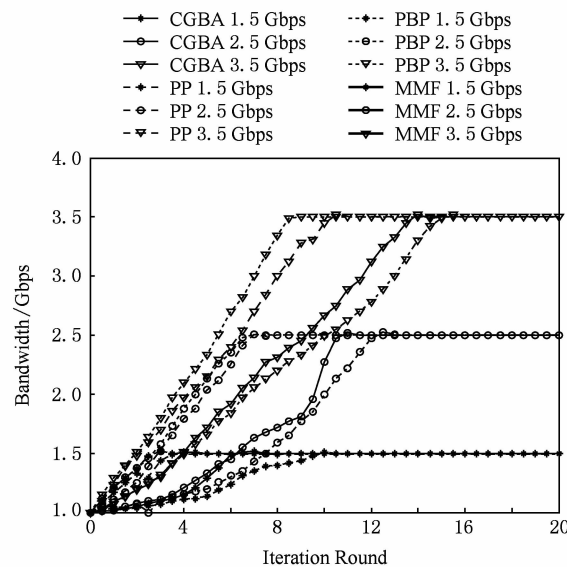


Fig. 6 Comparison of convergence speed.
图 6 收敛速度比较

将 4 种策略分别进行 50 次实验, 每种策略达到博弈稳定所需的平均计算时间如图 7 所示:

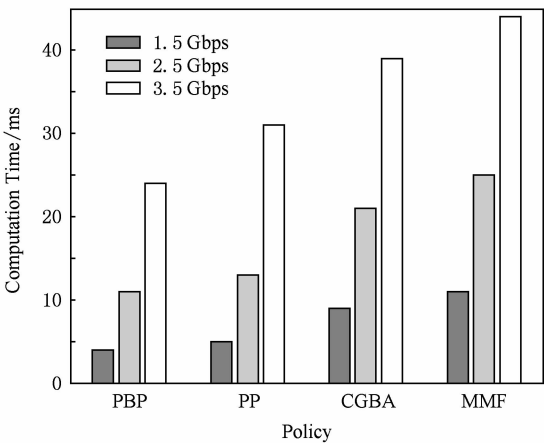


Fig. 7 Comparison of computational complexity.
图 7 不同策略计算复杂度比较

从图 7 可以看出, CGBA 的计算复杂度要明显高于 PP 和 PBP, 但是却低于 MMF. 这表明, CGBA 具有较高带宽分配满足度和较高公平性是以牺牲一定计算复杂度为代价的, 并进一步验证了文献[23]中 NFL 理论的正确性.

4 结束语

本文深入研究了数据中心骨干网络中的带宽分配问题, 将各条流对共享带宽的竞争分配问题建模为一个合作博弈模型, 并通过寻求该博弈模型的 Nash 谈判解 NBS 得到优化的带宽分配策略 CGBA. 将 CGBA 与 3 种典型的带宽分配策略进行了详尽的实验对比. 实验结论表明, CGBA 在各种网络场景中都能较好地提高带宽分配的满足度, 为每条流提供最小带宽保障, 效率较高; 同时可保证较高的带宽分配公平性. 本文提出的带宽分配策略为设计性能更佳的数据中心骨干网络带宽分配系统提供了借鉴和理论支持.

参 考 文 献

[1] Bari M F, Boutaba R, Esteves R, et al. Data center network virtualization: A survey [J]. IEEE Communications Surveys & Tutorials, 2013, 15(2): 909-928

[2] Mahimkar A, Chiu A, Doverspike R, et al. Bandwidth on demand for inter-data center communication [C] //Proc of the 10th ACM Workshop on Hot Topics in Networks. New York: ACM, 2011: 24:1-24: 6

- [3] Greenberg A, Hamilton J, Maltz D A, et al. The cost of a cloud; Research problems in data center networks [J]. ACM SIGCOMM Computer Communication Review, 2008, 39(1): 68-73
- [4] Google. Inter-datacenter WAN with centralized TE using SDN and OpenFlow [EB/OL]. 2012[2014-12-22]. <https://www.opennetworking.org/images/stories/downloads/misc/googlesdn.pdf>
- [5] Jain S, Kumar A, Mandal S, et al. B4: Experience with a globally-deployed software defined WAN [C] //Proc of the ACM SIGCOMM 2013. New York; ACM, 2013: 3-14
- [6] Huang Yongfeng, Dong Yongqiang, Zhang Sanfeng, et al. Leftover bandwidth-aware peer selection algorithm for inter-datacenter content distribution [J]. Journal on Communications, 2013, 34(7): 24-33 (in Chinese)
(黄永锋, 董永强, 张三峰, 等. 数据中心间空闲带宽感知的内容分发算法[J]. 通信学报, 2013, 34(7): 24-33)
- [7] Laoutaris N, Sirivianos M, Yang X, et al. Inter-datacenter bulk transfers with NetStitcher [J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 74-85
- [8] Thyaga N, Krishna P N P. Lowering inter-datacenter bandwidth costs via bulk data scheduling [C] //Proc of the 12th IEEE/ACM Int Symp on Cluster, Cloud and Grid Computing. New York; ACM, 2012: 244-251
- [9] Ghosh A, Ha S, Crabbe E, et al. Scalable multi-class traffic management in data center backbone networks [J]. IEEE Journal on Selected Areas in Communications, 2013, 31(12): 2673-2684
- [10] Danna E, Hassidim A, Kaplan H, et al. Upward max min fairness [C] //Proc of the IEEE Int Conf on Computer Communications (INFOCOM 2012). Piscataway, NJ: IEEE, 2012: 837-845
- [11] Osborne M J, Rubinste A. A Course in Game Theory [M]. Cambridge, MA: MIT Press, 1994: 12-30
- [12] Guo Jian, Liu Fangming, Tang Haowen, et al. Falloc: Fair network bandwidth allocation in IaaS datacenters via a bargaining game approach [C] //Proc of the 21st IEEE Int Conf on Network Protocols (ICNP). Piscataway, NJ: IEEE, 2013: 1-10
- [13] Vitaly A, Ruslan S. Global network modelling based on mininet approach [C] //Proc of the 2nd ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking. New York; ACM, 2013: 145-146
- [14] Liu Zhongjin, Li Yong, Su Li, et al. M2cloud: Software defined multi-site data center network control framework for multi-tenant [J]. ACM SIGCOMM Computer Communication Review, 2013, 43(4): 517-518
- [15] Lantz B, Heller B, McKeown N. A network in a laptop: Rapid prototyping for software-defined networks [C] //Proc of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks. New York; ACM, 2010: 19:1-19:6
- [16] Yaïche H, Mazumdar R, Rosenberg C. A game theoretic framework for bandwidth allocation and pricing in broadband networks [J]. IEEE/ACM Trans on Networking, 2000, 8(5): 667-678
- [17] Cottle R W. Nonlinear programs with positively bounded jacobians [J]. SIAM Journal on Applied Mathematics, 1964, 14(1): 147-158
- [18] Mangasarian O L. Equivalence of the complementarity problem to a system of nonlinear equations [J]. SIAM Journal on Applied Mathematics, 1976, 31(1): 89-92
- [19] Hager W W, Zhang Hongchao. A survey of nonlinear conjugate gradient methods [J]. Pacific Journal of Optimization, 2006, 2(1): 35-58
- [20] Chen Yu. The convergence research on conjugate gradient method [D]. Jingzhou, Hubei: Yangtze University, 2012 (in Chinese)
(陈禹. 共轭梯度法的收敛性研究[D]. 湖北荆州: 长江大学, 2012)
- [21] Jain R, Chiu D M, Hawe W. A quantitative measure of fairness and discrimination for resource allocation in shared computer system, TR-301 [R]. Saint Louis, Missouri: Washington University, 1984
- [22] Chiu D M, Jain R. Analysis of the increase and decrease algorithms for congestion avoidance in computer networks [J]. Computer Networks and ISDN Systems, 1989, 17(1): 1-14
- [23] Wolpert D H, Macready W G. No free lunch theorems for optimization [J]. IEEE Trans on Evolutionary Computation, 1997, 1(1): 67-82



Meng Fei, born in 1989. Master candidate. His main research interests include new network architecture, QoS.



Lan Julong, born in 1962. Professor and PhD supervisor. His main research interests include new network architecture, information security, and computer network.



Hu Yuxiang, born in 1982. Lecturer. His main research interests include new network architecture, information security.