同构流媒体集群系统优化内容部署

卫 星　杨 坚　奚宏生
(中国科学技术大学自动化系　合肥　230027)

摘　要：该文研究了在固定节目流行度的情况下，如何进行内容优化部署以最小化流媒体集群系统拒绝率和降低复制存储消耗的问题。首先运用排队理论知识分析得出优化目标和服务器访问概率之间的数值联系，并且通过某些数值方法确定出系统最小拒绝率情况下的最优服务器访问概率。由于内容部署属于 NP-Hard 问题且完全决定每台服务器的访问概率，该文设计了副本交换和对等副本访问概率调整两种启发式策略来进行内容部署，以满足在优化内容分布下每台服务器访问概率和最优值之间的差异最小，从而实现降低系统拒绝率和存储代价的目标。最后分别采用数值分析和离散事件仿真验证了模型的正确性和算法的有效性。

关键词：流媒体集群系统；内容部署；存储均衡；拒绝率
中图分类号：TP393　　**文献标识码**：A　　**文章编号**：1009-5896(2009)09-2232-05

Optimal Content Distribution on Clustered Streaming Media System Consisting of Homogeneous Configuration

Wei Xing　Yang Jiang　Xi Hong-sheng
(*Automation Department*, *University of Science and Technology of China*, *Hefei* 230027, *China*)

Abstract: The optimizing problem of content distribution which minimizes the blocking probability and storage consumption on clustered streaming media system is discussed, in the case of knowing every program's unchanged popularity. Firstly, the queuing theory is adopted to analysis the relationship between the server's access probability and the optimizing goal. The ideal access probability of every server can be obtained by some numerical methods, under the circumstance of minimal blocking probability. Content distribution determining each server's access probability, has been proved to be NP-Hard. The whole content distribution process consists of two strategies, i.e. duplicate swapping and peer duplicate's access probability adjusting. All the heuristic arithmetic is designed to perform the content distribution in order to minimize the distance between the result of optimization and the ideal one, minimize the storage consumption and reduce the blocking probability. Finally, the correctness of system modeling and the efficiency of proposed arithmetic are verified by numerical analysis and discrete event simulation.

Key words: Clustered streaming media system; Content distribution; Storage balancing; Blocking probability

1　引言

宽带化高速网络的迅速发展、动态影像压缩解码技术与大容量存储技术的成熟，促成了流媒体技术的诞生和发展。与传统基于超级计算机架构的集中方式相比，通过分立的服务器子系统构建能够满足大规模点播请求的“虚拟”服务器集群，具有可扩展性、高性价比和高可用性的优势[1−3]。集群式流媒体服务系统的一般结构如图 1 所示，集群功能的实现过程如下：集群由负载均衡器、流化服务器、存储子系统和高速网络等部分组成；客户端通过 Internet/Intranet 发送服务内容请求，整个集群共享一个虚拟 IP 地址，只有负载均衡器对于客户端可见，首先通过握手协议与客户端建立 TCP 连接，然后根据系统内容部署和实时负载情况将该 TCP 连接“透明”地迁移到某台流化服务器，流化服务器

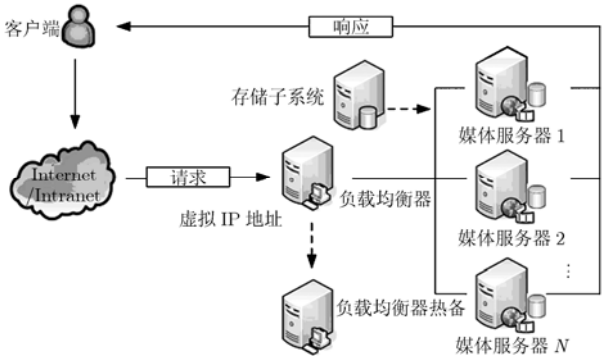

图 1 集群式流媒体服务系统一般结构

2008-10-06收到，2009-04-06改回

通过单播的方式对客户端请求进行服务响应；集群中所有服务类型的流媒体文件存储于第 3 方的存储子系统，每台流化服务器配有独立的硬盘空间用以存储部分流媒体文件副本，集群内部负载均衡器、流化服务器和存储子系统通过分立的高速网络连接。

内容部署问题用以解决如何将庞大数量的不同种流媒体文件合理分布到存储子系统中并且满足性能指标的要求，文件副本的放置状态直接影响到请求调度的方式，影响到系统的运行方式和整体性能，诸多文献[4−8]在特定目标下提出了相应的解决方案。文献[5]提出面向点播热度的最优同构服务节点副本生成/放置算法 MMPacking，由于算法本身对于文件复制条件的严格限制，必然会导致存储空间的浪费和较差的容错性。文献[9]假设点播率服从某种统计规律，提出了最优的副本生成算法和一种贪婪的放置策略，然而没有对存储和性能指标的联系进行深入阐述，并且其放置策略效果比较粗糙。

本文应用排队理论对流媒体服务系统进行建模和分析，给出系统拒绝率和服务器访问概率的联系，并且由数值方法确定每台服务器理论的最优访问概率。由于内容部署决定每台服务器的实际访问概率，并且属于 NP-Hard[3]问题，因此将最小化系统不均衡度的问题转化为最小化每台服务器访问概率和其最优值之间距离的问题。继而在现有贪婪初始放置的基础上提出副本位置交换和对等副本访问概率调整两种优化部署策略，使得在满足优化目标的同时提升了系统容错性能降低了存储消耗。

2　系统模型

设集群系统支持 M 种不同内容的服务，任意流媒体文件均为恒定码率编码且需要存储空间相等，以相同码率采用单播的方式进行流化服务，考虑到用户交互式操作的影响，设每个服务流的持续时间为一般分布，其均值为 d。集群系统由 N 台同构服务器组成，每台服务器所能缓存的文件副本总数和容纳的最大服务流数分别为 C 和 L。设流媒体文件 j 的流行度为 p_j，显然有 $\sum_{j=1}^{M} p_j = 1$。用户请求的选择性倾向导致了不同服务内容的相异流行度，为了减弱或者消除这种不均匀访问模式带来的影响，将热门文件产生多个副本且分布于多个服务器的复制策略是一种有效途径。设文件 j 在系统中的副本总数为 r_j，文献[9]研究了最优副本数问题，并且给出了 Adams Divisor 算法，本文借用这种算法的结果，将文件 j 最优的副本数表示为 r_j^*。

定义 $\boldsymbol{A} = [a_{ij}]_{N\times M}$ 为系统的内容分布矩阵，$a_{ij} = 1$ 表示文件 j 有某个副本放置于服务器 i，否则 $a_{ij} = 0$。如果文件 j 的某个副本被放置于服务器 i，其一部分服务请求会由服务器 i 承担，定义 $\boldsymbol{W} = [w_{ij}]_{N\times M}$ 为集群系统负载分布矩阵，w_{ij} 表示服务器 i 上文件 j 被访问的概率。当 $a_{ij} = 0$ 时，必然有 $w_{ij} = 0$。由此可见，系统的内容部署完全由内容分布矩阵 $\boldsymbol{A}$ 和负载分布矩阵 $\boldsymbol{W}$ 决定，且有如下约束：

(1)任意文件不可能在同一台服务器存有两个副本，即

$$a_{ij} \in \{0,1\},\ i = 1,\cdots,N,\ j = 1,\cdots,M \tag{1}$$

(2)文件 j 在所有服务器上的存储标记 a_{ij} 之和等于其副本数，即

$$\sum_{i=1}^{N} a_{ij} = r_j^*,\ j = 1,2,\cdots,M \tag{2}$$

(3)文件 j 所有服务器上的副本访问概率之和等于其流行度，即

$$\sum_{i=1}^{N} w_{ij}\cdot a_{ij} = p_j,\ j = 1,2,\cdots,M \tag{3}$$

当用户点播了某个流媒体进程，集群系统却没有空余的服务容量接纳，则会产生拒绝。拒绝率 B 作为系统 QoS 的重要衡量指标，直接影响到系统的吞吐量和系统资源的利用率，应当予以降低乃至最小化。设系统服务请求为参数 λ 的 Possion 到达过程，请求被分配到服务器 i 的概率为 G_i，则服务器 i 的访问请求为参数 λG_i 的 Possion 到达过程，当某个请求到达时，如果服务器 i 拥有空闲的服务进程，以平均服务率 $1/d$ 响应该请求；如果所有的 L 个服务进程被占用，则该请求被拒绝。因此由排队理论知识，服务器 i 的服务行为服从多服务窗损失制排队模型。当系统稳定时，服务器 i 的拒绝率即为所有服务进程被占用的概率[10]，从而 B 可以如下表示：

$$B = \sum_{i=1}^{N} G_i b_i = \sum_{i=1}^{N} \left(\frac{\rho^L}{L!} G_i^{L+1} \Big/ \sum_{k=0}^{L} \frac{\rho^k}{k!} G_i^k \right) \tag{4}$$

其中 $\rho = \lambda d$。

由于式(4)是 Erlang loss 函数，而 ρ 和 L 固定，故 B 是 N 维向量 $\boldsymbol{G} = (G_1, G_2, \cdots, G_N)$ 的凸函数，故可以通过数值方法确定出 B 满足约束 $0 \le G_i \le 1$ 和 $\sum_{i=1}^{N} G_i = 1$ 的全局最小值 B^* 和相应的最优解 $G^* = (1/N, 1/N, \cdots, 1/N)$。因此由内容分布矩阵 $\boldsymbol{A}$ 和负载分布矩阵 $\boldsymbol{W}$ 的定义可以得出，任意服务器 i 被访问的概率完全取决于内容部署，即 $G_i = \sum_{j=1}^{M} w_{ij}\cdot a_{ij}$。从而最小化系统拒绝率的优化目标等价于确定最优的内容部署参量 $\boldsymbol{A}$，$\boldsymbol{W}$，使得各服务器的访问概率达到或者接近最优值：

$$\text{Min}\quad D = \left\| G - G^* \right\|^2 = \sum_{i=1}^{N} \left(G_i - 1/N \right)^2 \tag{5}$$

其中 $G_i=\sum_{j=1}^{M} w_{ij}\cdot a_{ij}$, $i=1,\cdots,N$ ，且满足约束式式(1)-式(3)。

3 优化策略

上述优化部署问题的 NP-Hard 性，可采用文献[3]的类似过程加以证明，因此启发式策略能够达到接近其最优解。在未进行内容部署时，内容分布矩阵 $\boldsymbol{A}$ 和负载分布矩阵 $\boldsymbol{W}$ 均为空，服务器访问概率为0，首先需要进行初始放置，将生成的最优文件副本放置于各服务器，从而将不同热度的点播负载合理分配到系统中。未放置任何副本前式(5)的偏差 D 最大，当副本被放置于服务器时，其访问概率会随之累加至所在服务器，引起偏差 D 的变化，为了使得每一步放置操作最大程度的减小 D ，需要优化副本放置顺序和每步放置的目标服务器选择方式。文献[4]不加证明地给出一种贪婪放置策略——最小负载优先放置算法，虽然在一定程度上具有减小式(5)的效果，但是存在较大的偏差和不完备，本文设计了两种启发式内容部署策略进行优化，分别称之为副本位置交换、对等副本访问概率调整。

3.1 副本位置交换

由于贪婪初始放置只是在局部做出最优选择而不具有全局的优化效果，为了修正或减小贪婪策略的误差，在副本初始放置的基础上可以进行服务器间文件副本的交换操作，以进一步减小目标距离 D 目的，改变副本所在的位置。

如果将服务器 m 上的副本 x 和服务器 n 上的副本 y 进行互换， $m\neq n$ 且 x,y 不是同一文件的副本。则交换前后目标距离的差值为

$$\Delta D=D-D'=2\left(w_{mx}-w_{ny}\right)\times\left[\left(G_m-G_n\right)-\left(w_{mx}-w_{ny}\right)\right] \tag{6}$$

交换前后的目标函数差值越大表示交换的效果越好，不妨设 $G_m>G_n$ ，利用二次函数的性质，当 $(w_{mx}-w_{ny})=(G_m-G_n)/2$ 时， $D-D'$ 有最大 $(G_m-G_n)^2/2$ 。因此得出如下的交换规则：选择具有较大 G_m 的源服务器和较小 G_n 的目标服务器；副本 x,y 满足 $w_{mx}-w_{ny}<G_m-G_n$ 且接近于 $(G_m-G_n)/2$ 。副本交换算法略。

3.2 对等副本访问概率调整

由于最优副本生成算法的结果有 $p_{j_1}/r_{j_1}^{*}\approx p_{j_2}/r_{j_2}^{*}$, $\forall j_1\neq j_2$ ，即任意文件的不同副本平摊总的点播负载，并且任意副本的访问概率大致相同，初始副本放置和副本位置交换策略的结果很大程度上依赖于副本生成的“平滑”效果。如果点播概率之间差异很大，为了达到整体的平衡效果，就不得不增大存储消耗来减小不同文件副本之间的访问概率差异。而实际系统可以通过调度服务器随机分发请求方式方便地控制不同副本承担的负载总量，例如文件 j 具有两个副本 v_{j1} 和 v_{j2} ，分别放置于服务器 i_1 和 i_2 ，如果调度服务器将文件 j 的到达请求按照 1:2 的比例分发至 i_1 和 i_2 ，则可认为副本 v_{j1},v_{j2} 的访问概率分别是 $p_j/3$ ， $2p_j/3$ 。因此，内容部署的优化可以在不改变内容分布矩阵 $\boldsymbol{A}$ 的情况下，将对等副本的访问概率进行重新分配，即重新构建 $\boldsymbol{W}$ 来达到优化式(5)的目的。如图 2 和图 3 所示的 3 台服务器组成的示例系统，经过初始副本放置和副本交换之后，各服务器之间的访问概率依然不均衡。如果欠载服务器和过载服务器之间存在着“对等”副本，可以直接调整两者共有的单个或者多个对等副本访问概

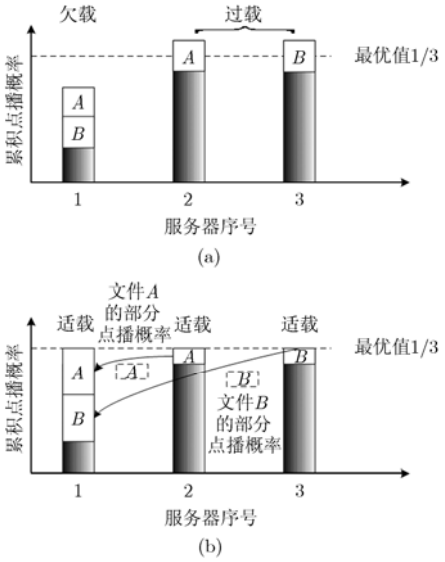

图 2 对等副本访问概率直接调整

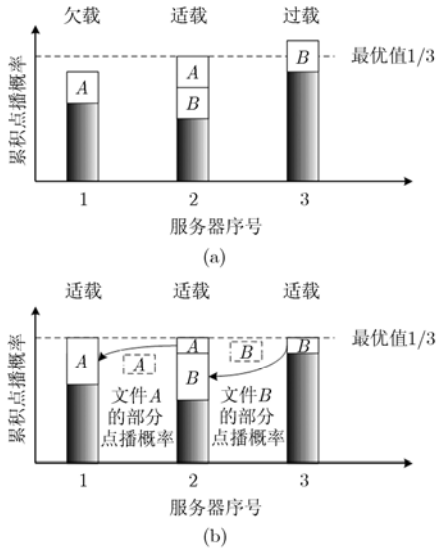

图 3 对等副本访问概率过渡调整

率达到全局均衡，见图 2；如果欠载服务器和过载服务器之间没有对等副本，则可以通过适载服务器上的“对等”副本进行“过渡”交换，得到同样的均衡结果，见图 3。对等副本访问概率的调整算法略。

4　数据和仿真分析

考虑模型所描述的同构流媒体集群系统。令集群系统文件种类 $M=400$，文件 j 的点播概率符合 Zipf-like 分布：

$$p_j=\left(1/j^{-\theta}\right)\Big/\sum_{k=1}^{M}\left(1/k^{-\theta}\right),\ j=1,2,\cdots,M \qquad (7)$$

θ 为常数，称为 Zipf 倾斜因子。倾斜因子 θ 越大，文件点播率越不平均，即热门文件的点播概率占有的份额越大，经验公式通常取 $\theta=0.729$，为了研究文件热度差异对于内容部署的影响，本文考虑 $\theta\in[0,1.5]$ 的情况。集群系统服务器数 $N=20$，每台服务器能够容纳的实时流数 $L=60$。则系统容纳请求的能力为 NL，这里只考虑满载情况，此时 $\rho=1200$。为了衡量文件复制程度或者系统存储消耗，定义复制因子 $\beta=NC/M$，显然 $\beta\in[1,N]$。

首先考察完成内容部署之后，根据式(4)计算的拒绝率指标。图 4 所示为单副本放置情况下的多种算法结果。由于此时文件不进行复制，$\beta=1$，本文的第 2 种策略是无效的，退化为初始副本放置算法。由图可见，随着倾斜因子的增加，各种算法对于拒绝率的优化结果均逐渐偏离最优值。θ 越大，文件点播概率之间的偏差越大，例如当 $\theta=1.5$ 时，前 10 个热门文件的点播概率之和就已经达到 0.7942。当 θ 很大时单个热门文件的点播概率，可能接近或超过访问概率理想值 $1/N$。MMpacking 放置的优化效果最差，因为在不进行文件复制时，其文件放置顺序是和本文的初始放置顺序相反，而服务器选择依据也是轮询方式。最小负载优先的放置算法结果类似于初始放置，但由于缺少了副本交换的优化过程，其整体优化效果比本文的算法略差。图 5 所示为进行复制部署情况下的多种算法结果。各种算法对于拒绝率的优化结果均随着倾斜因子的增长也逐渐偏离最优值，但是比单副本部署的总体效果好很多，例如当 $\beta=1.5$ 时，θ 在 0.8 左右依然能够达到或接近最优值，整体最差情况也只有 11%左右。随着复制因子的增加，各种算法优化效果愈加显著，例如当 $\beta=2.0$ 时，本文优化部署算法在倾斜因子全变化范围内均能够达到或者非常接近最优值，和复制情况下的 MMpacking 算法最优效果大致相似。其次，考察达到最优部署的存储消耗。显然当 $\beta=N$ 时，即每台服务器存储所有文件副本时，无论如何部署都能达到均衡目标。绘制出不同倾斜因子下达到最优部署时的复制因子曲线，如图 6 所示。由于本文提出的对等副本访问概率调整策略，能够利用副本之间的协作平衡，使得在相同复制因子下的总体部署效果具有优势，从而使得存储消耗大大低于最小负载优先放置算法和最大连通度放置算法[6,8]。最后，考察在故障时系统内容部署的容错性。为此我们采用排队理论中的离散事件仿真作为仿真平台，令每个副本单位时间内以相等概率随机发生故障，绘制出在 $\beta=2.0,\theta=0.729$ 时到达率和拒绝率的曲线，如图 7 所示。由于 MMpacking 算法只将流行度最高的一个或者多个文件进行复制，所以如果热门副本发生故障其承担的流负载全部会被拒绝，因此容错性相对较差。本文的优化部署算法兼顾了复制对于不均衡热度的“平滑”效果和对等副本之间的协调效果，故而具有最好的容错性。

5　结束语

本文研究了给定节目流行度的同构流媒体集群优化内容部署问题，以最小化流媒体集群系统拒绝率为目标，兼顾了复制存储的消耗，设计了副本位置交换和对等副本访问概率调整两种启发式策略来

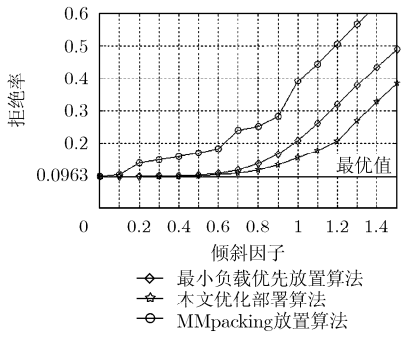

图 4 单副本部署情况下拒绝率指标

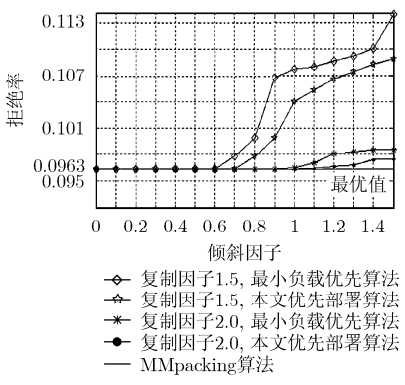

图 5 复制部署情况下拒绝率指

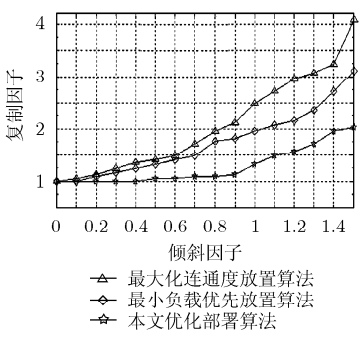

图 6 最优部署目标下的倾斜因子和复制因子曲线

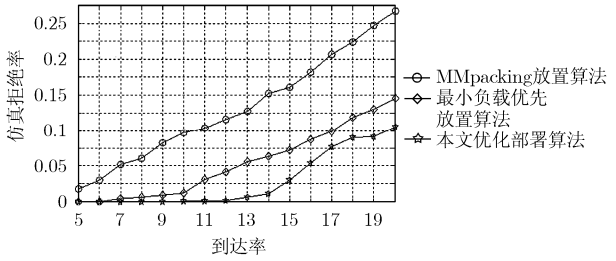

图 7 仿真环境下的到达率和拒绝率曲线

进行全局的内容部署。数值和仿真分析均表明，所提出的策略能够很好地适用于单副本放置和复制部署两种场合，并且能够具有较高的容错性。进一步的工作，一是寻找比贪婪算法更优的优化策略，二是研究基于文件点播率动态变化的副本更新问题，三是将动态负载迁移机制引入系统中作为调度策略的补充，将单个独立的服务器子系统通过负载的相互转移形成共同的协作体系。

参 考 文 献

[1] Dakshayini M, Guruprasad H S, and Masheshappa H D, *et al.*. Load balancing in distributed VoD using Local Proxy Server Group(LPSG). IEEE International Conference on Computational Intelligence and Multimedia Application, Sivakasi, Tamil Nadu, 13–15 Dec. 2007, 4: 162–168.

[2] Zhang Ming-long and Feng Bo-qin. A novel migration algorithm based-on “states-balancing” in a distributed multimedia services system. International Conference on Multimedia and Ubiquitous Engineering, Busan, 24–26 April 2008: 336–341.

[3] Leung Yiu-wing and Hou Yuen-tan. Assignment of movies to heterogeneous video servers. *IEEE Transactions on Systems, Man, and Cybernetics–Part A*, 2005, 35(5): 665–681.

[4] Jun Guo, Wong E W M, and Chan S, *et al.*. Combination load balancing for video-on-demand system. *IEEE Transactions on Circuits and Systems for Video Technology*, 2008, 18(7): 937–948.

[5] Serpanos D N, Georgiadis L, and Bouloutas T. MMPacking: A load and storage balancing algorithm for distributed multimedia servers. *IEEE Transactions on Circuits and Systems for Video Technology*, 1998, 8(1): 13–17.

[6] Zhao Yin-qing and Kuo C C J. Scheduling design for distributed video-on-demand servers. IEEE International Symposium on Circuits and System, Kobe, Japan, 2005, 2: 1545–1548.

[7] Zhao Yin-qing and Kuo C C J. Design issues on request migration for video-on-demand services. Proceedings of the 2004 International Symposium on Circuits and Systems, 23–26 May 2004, 2: 49–52.

[8] Tang Kit-sang, Ko King-tim, and Chan Sammy, *et al.*. Optimal file placement in VOD system using genetic algorithm. *IEEE Transactions on Industrial Electronics*, 2001, 48(5): 891–897.

[9] Zhou Xiao-bo and Xu Cheng-zhong. Optimal video replication and placement on a cluster of video-on-demand servers. IEEE International Conference on Parallel Proceeding, Vancouver Canada, 18–21 Aug. 2002: 547–555.

[10] Gross D and Harris C M. Fundamentals of Queueing Theory. New York, Wiley, 1985: 294–304.

卫 星： 男，1980年生，博士生，研究方向为计算机网络、离散事件动态性能优化.

杨 坚： 男，1977年生，副教授，研究方向为多媒体信号处理与多媒体通信.

奚宏生： 男，1950年生，教授，博士生导师，主要研究方向为离散事件动态、通信网络的性能分析与优化等.