doi:10.3969/j.issn.1000-436x.2014.05.022

基于带宽请求预测和云资源预留的视频移植策略

丛鑫[1,2]，双锴[1]，苏森[1]，杨放春[1]，訾玲玲[2]
（1. 北京邮电大学 网络与交换技术国家重点实验室，北京 100876；2. 辽宁工程技术大学 电子与信息工程学院，辽宁 葫芦岛 125105）

摘 要：提出云辅助 P2P-VoD 架构下的视频移植策略。首先根据 P2P 特性预测每视频频道的用户带宽请求数，为视频移植提供依据。而后设计了最小预留带宽的云带宽资源申请算法，利用较小的开销代价满足 VoD 服务实时性要求。最后设计了移植策略，决定如何进行视频移植。实验结果表明，提出的策略能够在费用开销和用户满意度之间取得较好的平衡。

关键词：带宽请求预测；资源预留；移植策略；云辅助 P2P-VoD 架构
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2014)05-0167-08

Video migration strategy based on bandwidth demand prediction and cloud resource reservation

CONG Xin[1,2], SHUANG Kai[1], SU Sen[1], YANG Fang-chun[1], ZI Ling-ling[2]
（1. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China）

Abstract: A video migration strategy which is used in P2P-VoD architecture is presented. First it predicts the bandwidth demands of users in each video channel according to the P2P characteristic. And it gives a basis for migrating videos. Then an algorithm of applying cloud resources which is based on the minimum bandwidth reservation is proposed. It satisfies the real-time of VoD service with the costs as low as possible. Finally the video migration strategies are designed to determine the methods of migration videos. The simulation results indicate that the proposed strategy can get good trade-off between costs and user satisfaction.

Key words: bandwidth demand prediction; resource reservation; migration strategy; cloud-assisted P2P-VoD

1 引言

视频点播服务(VoD)系统每天吸引大量的用户同时在线观看视频[1,2]。为了保证用户的观看体验，视频服务提供商必须申请足够的带宽来保证用户观看的流畅度[3]，从而造成 2 个严重的问题：带宽开销过大和带宽利用率过低。这在传统的 P2P-VoD 架构(当前应用的)下难以改进，而云平台的出现为解决此问题提供了契机。利用云平台的“基础设施即服务”的特点，向其申请存储和带宽资源用以满足日益增长的用户需求，避免本地基础设施的巨大投入；利用云平台的“即付即用”特性，在用户请求较小时释放部分非必需云平台资源，从而提高资源(包括带宽)利用率。另外，考虑到视频点播服务供应商已有的基础设施的投入及人员设置，完全利用云平台进行视频点播服务是不合适的，需要一种

收稿日期：2013-04-08；**修回日期**：2014-03-04
基金项目：中央高校基本科研创新计划专项基金资助项目(2013RC1102)；国家自然科学基金资助项目(61121061)；“新一代宽带无线移动通信网”科技重大专项基金资助项目(2012ZX03002008-002-03)；国家重点基础研究发展计划(“973”计划)基金资助项目(2011CB302506)
Foundation Items: The Fundamental Research Funds for the Central Universities (2013RC1102); The National Natural Science Foundation of China (61121061); Important National Science & Technology Specific Projects: Next-generation Broadband Wireless Mobile Communications Network (2012ZX03002008-002-03); The National Basic Research Program of China (973 Program) (2011CB302506)

过渡的架构与之适应。这种混合云平台和视频提供商服务器的架构称为云辅助的 P2P-VoD 架构。

视频点播服务要保证服务的实时性，即视频内容要及时推送到用户设备上来保证播放平滑性。但当前的云资源是基于在使用时申请的模式。在申请的过程中会造成一定的延迟(或许非常短，如 1~2 s)，但仍会造成服务的中断，尤其是遇到请求高峰的时候。因此，可以在云平台上申请预留一部分带宽来满足实时性要求。通过追踪 YouKu[4]中的数据，发现每个视频频道请求达到峰值的时刻是不同的(用户请求不均衡性)，因此为每个频道都预留足够的带宽是不必要的。另外，由于请求不均衡性，视频提供商的服务器负载在不同时刻是变化的，从而导致了不同时刻移植视频内容会产生不同的开销及造成不同的用户观看体验(QoE)，因此移植策略需要从开销和请求满意度 2 个角度考虑。

另外，视频类型不同也会导致开销不同。流行度高的视频由于其请求数量巨大而放到云平台上较之于流行度低的视频是节省开销的。因此提前获得每个视频的带宽请求数从而决定哪些视频可以移植到云平台上是先决条件。文献[5]的作者利用评估了基于时间的请求模型并将其应用于 IPTV 网络以达到视频内容在服务器上的优化放置。文献[6]利用协同过滤算法和早期发布的相似视频来预测视频的流行度。文献[7]利用时序分析技术和时不变系统来预测服务器请求，其数据集能体现一天之内的请求变化，能够预测单位为小时的数据，但仅此不能满足视频移植要求。

经如上分析，云辅助 P2P-VoD 架构需解决的问题是：对于在 VoD 提供商服务器上的视频内容，如何判断哪些视频需要移植到云平台上并采用何种策略移植。本文贡献如下。

1) 依据 P2P 技术特性和用户请求数据变化规律，设计了每频道视频请求的预测算法。利用基于时序的算法和历史数据预测在线节点数量；利用高斯过程回归方法和搜索数据曲线来预测在线节点数量的初始值；利用多元线性回归算法预测种子节点的上传带宽。从而计算出未来时间内每个频道上的带宽请求数。

2) 详细分析了用户请求数据的变化规律，发现每个频道的视频请求在一天之中不能同时达到峰值。从而设计了云带宽资源预留的优化算法，尽可能减少由于预留而产生的开销。

3) 以开销和用户满意度为不同的着重点，设计了不同的视频移植策略，使得视频内容能够以较小的代价从本地服务器移植到云平台中。

4) 设计了详细的实验来评估预测算法和移植策略的有效性。实验结果表明，设计的策略能够取得开销和用户满意度之间较好的平衡。

2 视频频道带宽请求预测

本节详细介绍了每视频频道带宽请求的预测方法，为视频移植到云平台提供了参考依据。

2.1 P2P 贡献请求平衡模型

在 P2P 系统中，存在 3 种类型的节点：服务器节点、种子节点和下载节点。服务器节点位于视频点播提供商处，是视频源所在的位置，其产生的开销是提供商每月费用的首要组成部分；种子节点是特殊的观看视频用户，其特点是不消耗 P2P-VoD 系统的资源，却能为其他用户提供视频数据；下载节点是正在观看视频的用户，其特点是消耗 P2P-VoD 系统的资源，同时也为其他用户提供观看视频中已下载的数据。P2P-VoD 系统能够正常的运行，其要满足上传的数据能够和下载数据保持平衡，即不能获取超过系统提供能力的数据。在 t 时刻，设定 P2P-VoD 贡献请求平衡模型参数如表 1 所示。

表 1 P2P 贡献请求平衡模型参数

参数符号	参数含义
N_t	某频道上的在线节点数量
DP_t	下载节点的平均下行带宽
UD_t	下载节点的平均上行带宽
SP_t	对服务器节点请求带宽(上行带宽)
US_t	种子节点的平均上传带宽
R	视频频道的播放速率

由以上分析可以得到，P2P-VoD 系统理想情况下，数据的下行带宽应该和上行带宽相匹配，如果下行带宽过大，会导致系统崩溃；如果上行带宽过大，会造成带宽浪费。t 时刻，节点 i 在视频频道 c 的带宽平衡为

$$DP_{ti} = UD_{ti} + US_{ti} + SP_{ti} \tag{1}$$

考虑到所有的正在观看视频频道 c 的节点，可以得到

$$\sum_{i=1}^{N_t} DP_{ti} = \sum_{i=1}^{N_t} UD_{ti} + \sum_{i=1}^{N_t} US_{ti} + \sum_{i=1}^{N_t} SP_{ti} \tag{2}$$

变换式(2)，可以得到服务器带宽请求为

$$SP_t = N_t DP_{ti} - UD_t - US_t \tag{3}$$

在 P2P-VoD 系统中，为了防止提前完成下载的用户离开正在观看的频道而不上传数据，限定用户下载速率为视频播放速率，同时为了保证视频的播放平滑性，下载速率要略高于视频播放速率，设定参数 λ（$\lambda \geq 1$）。变换式(3)，得到

$$SP_t = N_t \lambda R - UD_t - US_t \tag{4}$$

依据式(4)，需要分别预测在线节点数量和节点平均上传带宽。

2.2　在线节点数量预测

2.2.1　基于历史数据的节点数量预测

云辅助 P2P-VoD 架构中，在线节点序列$\{N_t\}$的历史数据能够从本地追踪服务器和云监视服务器[8](例如，Amazon CloudWatch[9])中获得到，而后利用 Box-Jenkins[10,11]算法得到未来一段时间的数据集。经过测算，$\{N_t\}$是非稳态序列，因此设定后退转移算子 B 和差分算子∇，即 $BN_t = N_{t-1}$，$\nabla N_t = N_t - N_{t-1}$。通过对数和差分消除趋势从而得到稳态序列$\{N_t'\}$。接着采用自回归移动平均模型($ARMA(pq)$)预测$\{N_t'\}$。

$$\begin{aligned} N_t' - \phi_1 N_{t-1}' - \phi_2 N_{t-2}' - \cdots - \phi_p N_{t-p}' \\ = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \end{aligned} \tag{5}$$

其中，e_t 是不相关的白噪声，其遵从 $WN\left(0,\sigma^2\right)$ 分布。简化式(5)，得到

$$\phi(B) N_t' = \theta(B) e_t \tag{6}$$

其中，$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$，$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$。而后用 Gauss-Newton 法或者 Nelder-Mead 法评估 $\phi(B)$ 和 $\theta(B)$。实验结果如图 1 所示。

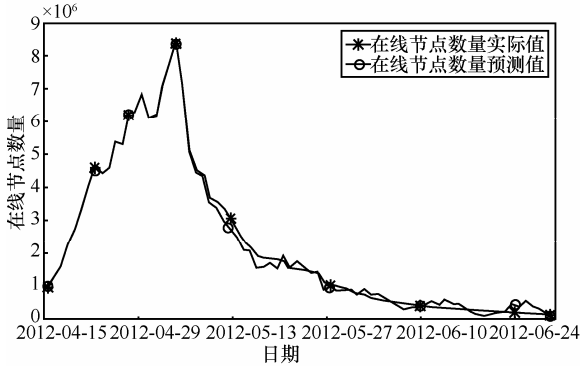

图 1　在线节点数量预测(依据历史数据)

2.2.2　节点数量初值预测

通过追踪 YouKu 中相同视频请求序列$\{N_t\}$和搜索序列$\{N_s\}$，发现其曲线有相同的趋势，因此搜索序列可以用来预测请求序列的初始值。

给出判断 2 个序列是否相似的方法如下：2 个序列 $Z_1=\{z_{11},z_{12},\cdots,z_{1n}\}$ $Z_2=\{z_{21},z_{22},\cdots,z_{2n}\}$ 随机误差因子 $er=\{z_{11}-z_{21},\ z_{12}-z_{22},\ \cdots,z_{1n}-z_{2n}\}$ 误差因子均值计算如下

$$er' = mean(er) = (1/n)\sum_{i=1}^{n}\left(z_{1i} - z_{2i}\right) \tag{7}$$

相关因子 cf 为

$$cf = (1/n)\sum_{i=1}^{n}\left|er - er'\right| \tag{8}$$

cf 越接近于 0，则 2 个序列越相似。

搜索序列 N_s 经过数据统计，其也符合时序序列特征，能够被 2.2.1 节的方法预测(实验结果如图 2 所示)。如果 N_s 和 N_t 之间的距离 d 能够被计算出来，那么 N_t 的初始值即可通过 N_s 和 d 计算得到。而 d 的数值需要利用 N_t 具有相似特征视频的数据来估计。相似特征视频是有相似的释放时间(同日或不同日的同一时间范围)及搜索数的一组视频集合。通过上述分析可知，高斯过程回归可用于对 d 进行估计。经过对真实数据的追踪及参考文献[8]，N_s 和 N_t 是符合高斯分布的序列。依据高斯过程回归[12]，得到

$$f\left(N_s\right) \sim GP(\mu, \boldsymbol{C}) \tag{9}$$

$$\mu = \mathrm{E}(f(N_s)) \tag{10}$$

$$\mathrm{cov}\left(N_{si}, N_{sj}\right) = \mathrm{E}\left[\left(f(N_{si}) - \mu_i\right)\left(f(N_{sj}) - \mu_j\right)\right] \tag{11}$$

其中，N_s 是输入向量，f 是值向量，$\boldsymbol{C}$ 是 $n \times n$ 矩阵，且 $C_{ij} = cov(N_{si}\ N_{sj})$。$N_t$ 的高斯回归过程模型为

$$N_t = f\left(N_s\right) + \varepsilon, \varepsilon \sim N\left(0, \sigma^2\right) \tag{12}$$

训练集 N_t 和测试集 N_t' 的联合分布为

$$\begin{bmatrix} N_t \\ N_t' \end{bmatrix} \sim \left(0, \begin{bmatrix} \boldsymbol{C}(N_s, N_s) + \sigma^2 \boldsymbol{I} & \boldsymbol{C}(N_s, N_s') \\ \boldsymbol{C}(N_s', N_s) & \boldsymbol{C}(N_s', N_s') \end{bmatrix}\right) \tag{13}$$

$\boldsymbol{I}$ 是单位矩阵，解式(13)得到

$$N_t' \sim N\left(\mu', \sigma'^2\right) \tag{14}$$

$$\mu' = \boldsymbol{C}\left(N_s', N_s\right)\left[\boldsymbol{C}\left(N_s, N_s\right) + \sigma_n^2 \boldsymbol{I}\right]^{-1} N_t \tag{15}$$

$$\sigma'^2 = \boldsymbol{C}\left(N_t', N_t'\right) - \boldsymbol{C}\left(N_t', N_t\right)\left[\boldsymbol{C}\left(N_t, N_t\right) + \sigma_n^2 \boldsymbol{I}\right]^{-1} \boldsymbol{C}\left(N_t, N_t'\right) \tag{16}$$

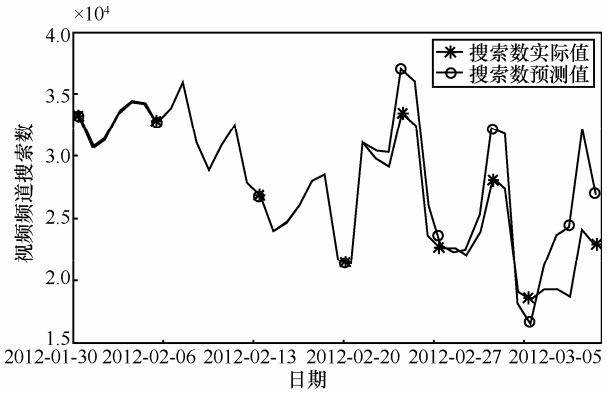

图 2 搜索数预测(依据历史数据)

实验结果如图 3 所示。

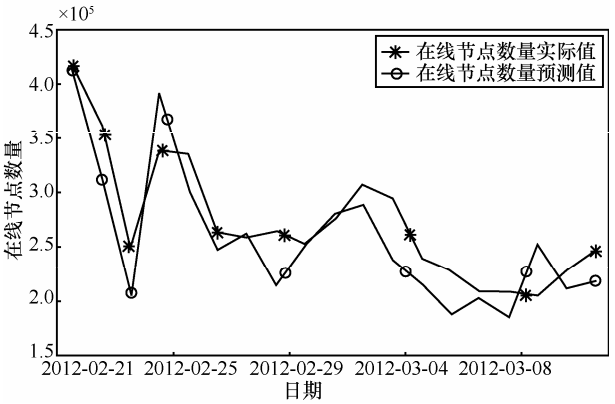

图 3 在线节点数量预测(初始值预测)

2.3 节点上传带宽预测

本节中,有 2 种类型节点的上传带宽需要预测,文献[7]的作者已经证明 UD_t 能够通过 $AR(p)$过程得到。而 US_t 估计可以通过多元线性回归得到，证明如下。

定理 1 US_t 可以通过由在线节点数量和不相关的白噪声为主导向量的多元线性回归方法估计。

证明 在 t 时刻，在线节点数量为 N_t，平均上传带宽为 u，相关系数为 $\gamma_t \in [0,1]$，随机误差为 e'_t。

$$US_t = \gamma_t N_t u + e'_t \tag{17}$$

其中，$N_t u$ 是系统的上传带宽，γ_t 是 US_t 占据的比例。

将 N_t 展开得到

$$US_t = \gamma_t u(\phi_1 N_{t-1} + \phi_2 N_{t-2} + \cdots + \phi_p N_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}) + e'_t \tag{18}$$

其中，N_t 是稳态序列，u 是固定值，替换 $\gamma_t u\alpha_i$ 为 φ_t，替换 $\gamma_t u\theta_i$ 为 δ_t，得到

$$US_t = \theta_1 N_{t-1} + \theta_2 N_{t-2} + \cdots + \theta_p N_{t-p} + \varphi_0 e_t - \varphi_1 e_{t-1} - \varphi_2 e_{t-2} \cdots - \varphi_q e_{t-q} + e'_t \tag{19}$$

简写为：$US_t = \theta^T N + \varphi^T e + e'_t$

初始条件为 N_0=0，即 t=0 时刻，系统中没有节点。由 N_t 预测时得知，q=0。因此，变换式(19)得到

$$\begin{cases} US_1 = e'_1 \\ US_2 = \theta_1 N_1 + e'_2 \\ US_3 = \theta_1 N_2 + \theta_2 N_1 + e'_3 \\ US_4 = \theta_1 N_3 + \theta_2 N_2 + \theta_3 N_1 + e'_4 \\ \cdots \\ US_p = \theta_1 N_{p-1} + \theta_2 N_{p-2} + \theta_3 N_{p-3} + \cdots + \theta_{p-1} N_1 + e'_p \end{cases} \tag{20}$$

式(20)可以由最小平方法求解。

在文献[7]中，线性时不变系统被应用于预测 US_t。其与$\{US_{t-h}\}$和$\{N_{t-h}\}$序列相关，但 US_t 的初始数据预测方法没有提及，而多元线性回归方法只与$\{N_t\}$相关。应用多元线性回归方法预测 US_t 结果如图 4 所示。

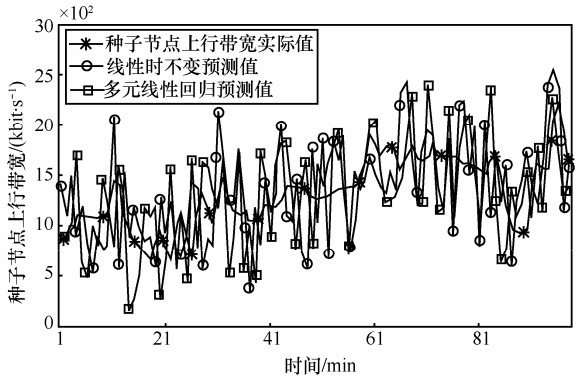

图 4 种子节点上行带宽预测

2.4 每频道节点请求预测

经过 2.1 节和 2.2 节的分析及收集的真实数据集，利用式(4)可以得到服务器请求的预测数据，选取 R=50 bit/s 和 λ=1.3，实验结果如图 5 所示。

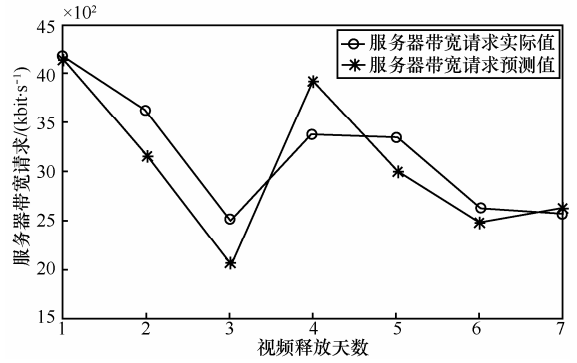

图 5 服务器带宽请求预测

从实验结果得知，尽管预测数据和真实数据间存在部分差异，但是作为此视频是否值得移植到云平台上的依据已经足够；此外预测算法的计算速度很快，但收集训练数据集的速度很慢，因此更新和统计的开销很大。

3　云带宽预留优化算法

本节中，在已知 t 时刻每个视频频道上的服务器请求带宽$\{SP_t\}$时，解答如何计算出在云平台上需要预留带宽最小值的问题。

3.1　带宽预留模型

云辅助 P2P-VoD 架构下，用户的请求由视频点播服务提供商服务器和云平台共同承担。如图 6 所示。

从图 6 中可以发现，如果云平台欲参与请求分发过程，则视频内容需要预先移植到云平台中。此过程会产生开销，即视频内容从视频点播服务提供商服务器(本地服务器)发出时产生的费用；视频内容由云平台接收时产生的费用；云平台存储的费用；视频请求由本地服务器应答的费用；视频请求由云平台应答的费用。参数设置如表 2 所示。

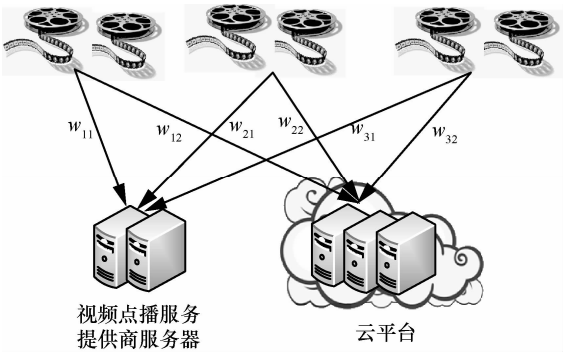

图 6　云辅助 P2P-VoD 架构请求

表 2　云辅助 P2P-VoD 带宽预留模型参数

参数符号	参数含义
p_1	本地服务器上行带宽单位费用
p_2	云平台下行带宽单位费用
p_3	云平台上行带宽单位费用
p_4	云平台存储单位费用
SP_t(同表 1)	对服务器节点请求带宽(上行带宽)
$w_{i1}\in[0,1]$	请求分配给本地服务器的权重
$w_{i2}\in[0,1]$	请求分配给云平台的权重
b_i	给定 SP_{ti} 需要在云平台预留的带宽
$B_s(s=1,2)$	本地服务器和云平台的最大带宽
v_i	视频 i 的大小
N	需要移植到云平台的视频个数

3.2　最小化带宽预留算法

参考文献[8]指出，SP_t 的分布符合期望为 μ_i，方差为 σ_i^2 的高斯分布，从 SP_t 预测过程及文献[13]可知，其是彼此相关的，这是由于相似的释放时间和搜索数会导致相似的请求数，相关系数为 ρ_{ij}。设定 p 为在云平台预留带宽和请求数之间的关系，描述如下

$$b_i = f_p\left(SP_{ti}\right) \tag{21}$$

其含义为给定请求 SP_{ti}，在云平台预留带宽为 b_i 时，能够被完全满足的概率为 p。

由图 6 可知，本地服务器负载和云平台负载分别计算为

$$l_s = \sum_{i=1}^{n}\left(w_{is}SP_{ti}\right) \tag{22}$$

其中，s=1 代表本地服务器负载，s=2 代表云平台负载。计算期望和方差，得到

$$\mathrm{E}[l_s] = \mu_1 w_{1s} + \mu_2 w_{2s} + \cdots + \mu_n w_{ns} = \mu^{\mathrm{T}} w_s \tag{23}$$

$$\mathrm{Var}[l_s] = \sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{ij}\sigma_i\sigma_j w_{is}w_{js} = w_s \psi w_s^{\mathrm{T}} \tag{24}$$

其中，$\mu=\{\mu_1,\mu_2,\cdots,\mu_n\}$，$w_s=\{w_{1s},w_{2s},\cdots,w_{ns}\}$，且 $\psi=\begin{bmatrix}\sigma_{11} & \cdots & \sigma_{1n}\\ \cdots & \cdots & \cdots\\ \sigma_{n1} & & \sigma_{nn}\end{bmatrix}_{n\times n}$　$\sigma_{ii}=\sigma_i^2$，$\sigma_{ij}=\rho_{ij}\sigma_i\sigma_j$。

效仿式(21)，为了能以概率 p 满足本地服务器和云平台上的请求为 l_s 时，需要预留的带宽计算如下

$$b_s = f_p\left(l_s\right) \tag{25}$$

依据 SP_t 的高斯分布特性，可以得到

$$f_p\left(x\right) = \mathrm{E}[x] + \theta\sqrt{\mathrm{Var}[x]}\ \ \theta = \Phi^{-1}\left(1-p\right) \tag{26}$$

其中，$\Phi\left(x\right)=\dfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\dfrac{x^2}{2}\right)\mathrm{d}x$。

由式(25)和式(26)可以得知

$$b_s = f_p\left(l_s\right) = \mu^{\mathrm{T}} w_s + \theta\sqrt{w_s\psi w_s^{\mathrm{T}}} \tag{27}$$

经过以上分析，得到最小化带宽预留问题描述如下

$$\min\left(\sum_{s=1}^{2}\left(\mu^{\mathrm{T}} w_s + \theta\sqrt{w_s\psi w_s^{\mathrm{T}}}\right)\right) \tag{28}$$

满足

$$\mu^{\mathrm{T}} w_s + \theta\sqrt{w_s \psi w_s^{\mathrm{T}}} \leqslant B_s \tag{29}$$

$$\sum_{s=1}^{2} w_{is} = 1, i \in [1, n] \tag{30}$$

定理 2 最小化预留带宽可以通过分配权重 w_{is} 得到：$0 \leqslant w_{is} \leqslant \min\left(1, \dfrac{bu_s}{\mu^T \mathbf{1} + \theta\sqrt{\mathbf{1}\psi\mathbf{1}^{\mathrm{T}}}}\right)$。

证明 假设 $g(w_i) = \sqrt{w_i \psi w_i^{\mathrm{T}}}$，这是凸函数，由凸函数的性质可以得到

$$\sqrt{\left(\frac{w_1 + w_2}{2}\right)\psi\left(\frac{w_1 + w_2}{2}\right)^{\mathrm{T}}} \leqslant \left(\sqrt{w_1 \psi w_1^{\mathrm{T}}} + \sqrt{w_2 \psi w_2^{\mathrm{T}}}\right)\Big/2$$

$$\Rightarrow \sqrt{(w_1 + w_2)\psi(w_1 + w_2)^{\mathrm{T}}} \leqslant \sqrt{w_1 \psi w_1^{\mathrm{T}}} + \sqrt{w_2 \psi w_2^{\mathrm{T}}} \tag{31}$$

利用迭代方式，整合所有视频请求，如下

$$\sqrt{\sum_{i=1}^{N}(w_i)\psi\left(\sum_{i=1}^{N}(w_i)\right)^{\mathrm{T}}} \leqslant \sum_{i=1}^{N}\left(\sqrt{w_i \psi w_i^{\mathrm{T}}}\right) \tag{32}$$

当带宽请求 SP_t 完全被满足时，其条件为 $w_i = w_{i1} + w_{i2} = 1$，式(29)转换为

$$\mu^{\mathrm{T}} \mathbf{1} + \theta\sqrt{\mathbf{1}^{\mathrm{T}} \psi \mathbf{1}} \leqslant B_s \tag{33}$$

因此由上述不等式，可以得到

$$0 \leqslant w_{is} \leqslant \min\left(1, \frac{B_s}{\mu^{\mathrm{T}} \mathbf{1} + \theta\sqrt{\mathbf{1}^{\mathrm{T}} \psi \mathbf{1}}}\right)$$

经过验证，其能够满足式(30)，且能够取得式(29)的等式需求，因此是式(28)的最优解。

4 视频移植策略

经过对真实服务器带宽请求的分析，发现一天之中不同时刻服务器带宽负载是不同的，即早晨负载大量空闲，傍晚和晚间负载巨大，因此何时进行视频移植才能得到较好的收益是需要考虑的问题。考虑服务器负载是否超过预期(超过需多支付费用)和用户的满意度，设定移植策略如下。

预先移植策略。当服务器空闲时(早晨)，将视频内容移植到云平台中。此策略的优势在于本地服务器负载超过预期带宽的可能性非常低。缺点是不能保证用户的满意度，这是因为视频请求的变化是非常频繁的。经过上述分析，预先移植策略设置如下：大多数的高流行度(SP_t高)视频内容在服务器空闲时移植到云平台中，替换云平台上流行度最低的一组视频内容。

智能移植策略[14]。当前最流行的视频移植到云平台上，替换当前最不流行的视频，一天之中此策略运行一次，并且已经证明此策略要好于预先移植策略。

按需移植策略。当服务器繁忙时(傍晚和晚间)，移植视频内容到云平台中，替换最不流行的视频内容。此策略的优势在于能够收到很好的用户满意度，这是由于未预测出视频的及时移植能够使云平台参与到服务用户请求当中。缺点是服务器的带宽请求很容易超过预期，造成视频点播服务提供商的费用开销的增大。经过上述分析，按需移植策略设置如下：当突发请求出现且视频还未移植到云平台时，立刻执行移植，替换云平台上最不流行的视频内容或者追加云平台的存储空间(云平台存储费用极低)。

混合移植策略。由于视频的请求随时间变化是非常大的，例如，通过跟踪 YouKu[4]的视频数据，发现视频访问量随天数有激增的情况且很频繁。此部分视频应该被及时移植到云平台中，反之，应被及时替换出云平台。为了处理突发视频请求，要提升服务器请求高时的用户满意度。经过上述分析，设置混合移植策略如下：当服务器空闲时，移植当前最流行的视频(包括预测出来的视频)到云平台中，替换最不流行的视频；在接近服务器繁忙时，进行再次预测，并执行替换过程；在有突发视频请求时，进行小规模的视频移植。

5 模拟实验和性能分析

5.1 视频频道带宽请求预测结果

利用 2012/4/15 到 2012/6/29 从 YouKu 收集到的某视频频道在线节点数量数值对基于历史数据的在线节点数量预测方法进行评估。取训练集数据 22 个，设置 p=5 q=0 且$\nabla = 1$，图 1 中可以看出，预测数据曲线和实际曲线有较好的重合度。图 2 是从 YouKu 上收集的从 2012/1/30 到 2012/3/10 某视频频道的搜索数，利用时序方法对未来一段时间搜索数的预测结果。图 3 是在线节点数量初始值预测结果，从图中可以看出，基于相似形和高斯回归过程的预测方法能够较准确地预测距离释放日期较近的几天的数值，但对较远日期的预测则有较大差距，因此，当初始值数足够的情况下，利用基于历史数据的在线节点数量预测方法是更好的选择。

图 4 是在不同时间点上的上行带宽预测结果，实验结果显示，基于多元线性回归的方法预测的数值较之基于线性时不变系统和实际值曲线有相似

的重合度，但前者利用的主导序列较之后者少。

图 5 是服务器带宽请求预测结果，从图中看出，预测的曲线较之真实曲线有一定的差距。但能够反映出未来一段时间的服务器请求变化趋势和在数值上提供参考。

5.2　视频移植策略评估

在评价视频移植策略前，需要定义 2 个评价参数。1)归一化开销：云辅助 P2P-VoD 架构下带宽总开销除以 P2P-VoD 下带宽总开销。2)归一化请求拒绝率：云辅助 P2P-VoD 架构下总请求拒绝数除以 P2P-VoD 架构下总请求拒绝数。

在本节中，参考 Amazon Cloud [15]的价格来制定实验模拟的参数。

如图 7 所示，随着从云平台申请存储空间的增大，成本在增大，这是由于租用和预留费用的增大。但从图中可以看出，相比于按需移植策略，智能和混合策略在成本上增量很少，这得益于后 2 种策略对 VoD 提供商预期带宽的影响较小。相比于智能策略，混合策略的归一化开销要高，这是由于有预留开销的存在。图 8 显示的是在不同存储大小和移植策略下，归一化请求拒绝率的变化。相比于按需策略，混合策略表现欠佳，但从图 7 中发现，相比于取得的优势，按需策略在开销上付出了很大的代价。

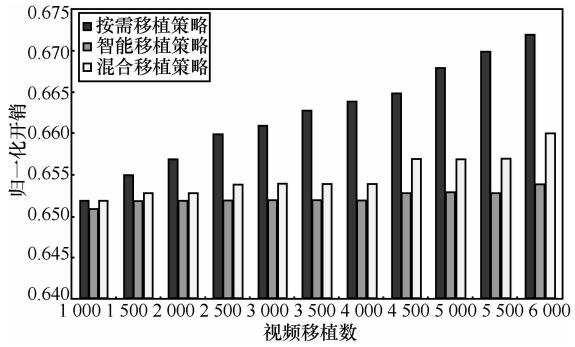

图 7　不同视频移植数下的归一化开销

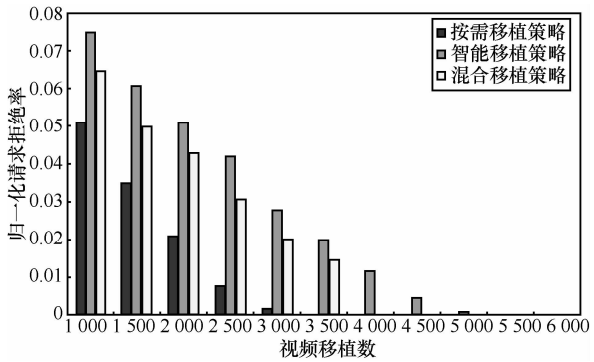

图 8　不同视频移植数下的归一化请求拒绝率

图 9 显示了不同服务器能力、不同移植策略、不同云价格下的归一化开销。实验结果显示，在不同的云价格体系下，申请较多的云存储空间并分配较多的请求到云中对费用开销减少不会产生影响。经过观察，分配因子设置在 0.45~0.6 之间较为恰当。

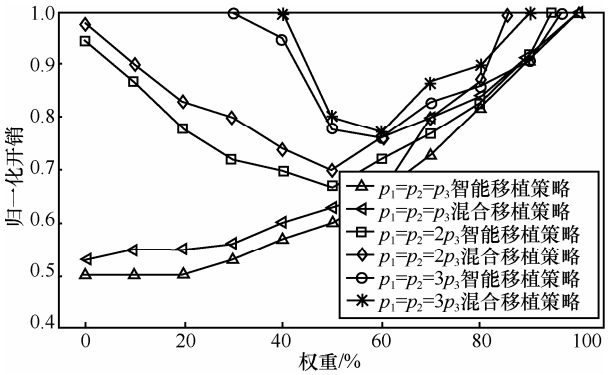

图 9　不同本地服务器能力及单位带宽价格下的归一化开销

图 10 显示，混合策略和智能策略有相似的趋势，即放置到云平台上的请求越多，请求拒绝率就越低。实验显示混合策略下，当请求拒绝率等于 0 时，系统所需带宽只是智能策略下的 80%。

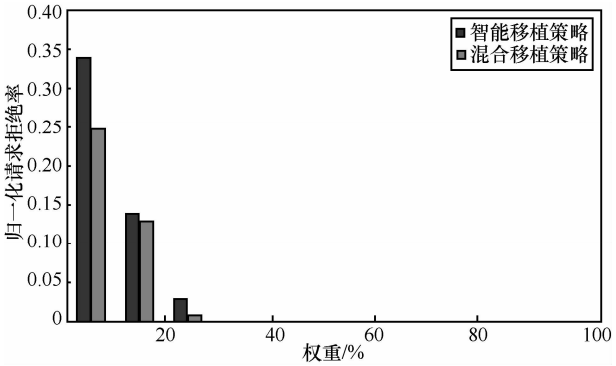

图 10　不同本地服务器能力下的归一化请求拒绝率

图 11 显示在不同策略下对突发请求的归一化开销。云辅助 P2P-VoD 架构的开销增长在合理的范围之内。原因如下：由于视频内容的替换机制导致存储开销增加幅度较小；突发请求需要带宽的增加是从云平台处申请到的，对 VoD 提供商期望带宽影响较小。

图 12 显示不同策略下突发请求的归一化请求拒绝率。结果显示，随着越来越多的视频被移植到云平台上，请求拒绝率越来越低。通过预留机制，混合策略的请求拒绝率要低于智能策略。但当不考虑开销代价时，按需策略是最好的。

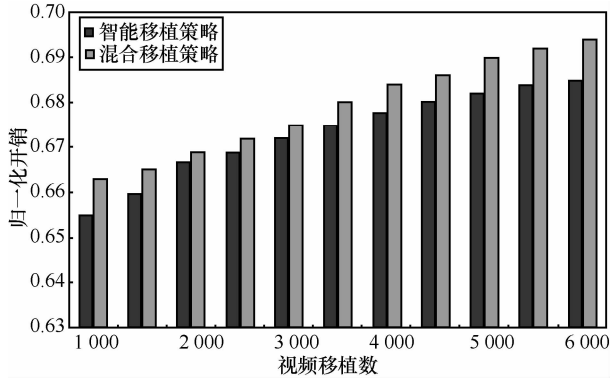

图 11　不同视频移植数下的处理突发请求时归一化开销

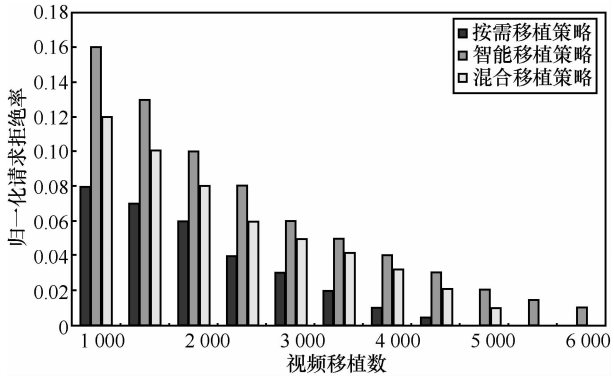

图 12　不同视频移植数下的处理突发请求时归一化请求拒绝率

6　结束语

本文主要解决的问题是如何降低 VoD 提供商的开销。针对传统 VoD 架构在带宽上存在的很难解决的问题，使用一种基于云平台的 VoD 架构。在新架构下，需解决视频内容移植问题。

针对视频内容移植问题，首先提出了每视频频道服务器带宽请求预测方法，为移植哪些视频到云平台上提供了参考依据；而后为了保证 VoD 服务的实时性，提出了基于服务器请求数的最小化预留带宽的算法，能够减少带宽申请时间，降低请求拒绝率，提升用户体验；最后提出了本文的移植视频策略。实验结果证明，混合视频移植策略能够在增加 0.5%的开销的基础上降低 1%的请求拒绝率，该方法能够在开销和请求拒绝率上取得较好的平衡。

由于本文的架构是基于目前 VoD 状况的过渡性架构，下一步需要研究在基于云平台的 P2P-VoD 架构下如何进一步降低 VoD 供应商服务器开销。

参考文献：

[1] NIU D, LI B C, ZHAO S Q. Understanding demand volatility in large vod systems[A]. Proceedings of the 21st International Workshop on Network and Operating Systems Support for Digital Audio and Video[C]. 2011.39-44.

[2] Youtube homepage[EB/OL]. http://www.youtube.com, 2012.

[3] YING Q, GUO Y, CHEN Y, *et al*. Understanding users' access failure and patience in large-scale P2P VoD systems[A]. Wireless Mobile & Multimedia Networks (ICWMMN 2011) 4th IET[C]. 2011. 283-287.

[4] YouKu[EB/OL]. http://index.youku.com/vr/, 2012.

[5] APPLEGATE D, ARCHER A, GOPALAKRISHNAN V. Optimal content placement for a large-scale VoD system[A]. ACM CoNEXT[C]. 2010. 1-12.

[6] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[A]. Proc of the 14th ACM SIG-KDD Conference[C]. 2008.426-434.

[7] NIU D, LI B C, ZHAO S Q. Demand forecast and performance prediction in peer-assisted on-demand streaming systems[A]. IEEE Infocom Mini-Conference[C]. 2011. 421-425.

[8] NIU D, CHEN F, LI B C. A theory of cloud bandwidth pricing for video-on-demand providers[A]. IEEE Infocom[C]. 2012. 711-719.

[9] Amazon Web services[EB/OL]. http://aws.amazon.com/.

[10] BOX G E P, JENKINS G M, REINSEL G C. Time Series Analysis: Forecasting and Control[M]. Wiley, 2008.

[11] CHAN K S, CRYER J D. Time Series Analysis with Applications in R[M]. Springer, 2008.

[12] HE H, SIU W C. Single image super-resolution using gaussian process regression[A]. IEEE Computer Vision and Pattern Recognition Conference[C]. 2011.449-456.

[13] GURSUN G, CROVELLA M, MATTA I. Describing and forecasting video access patterns[A]. INFOCOM Proceedings IEEE 2011[C]. 2011.16-20.

[14] LI H, ZHONG L, LIU J, *et al*. Cost-effective partial migration of VoD services to content clouds[A]. Cloud Computing (CLOUD) 2011 IEEE International Conference on[C]. 2011. 203-210.

[15] Amazon S3 Pricing[EB/OL]. http://aws.amazon.com/s3/pricing/,2012.

作者简介：

丛鑫（1982-），男，辽宁阜新人，博士，辽宁工程技术大学讲师，主要研究方向为 P2P 技术、云计算。

双锴[通信作者]（1977-），男，辽宁铁岭人，博士，北京邮电大学副教授、硕士生导师，主要研究方向为下一代网络技术。E-mail: shuangk@bupt.edu.cn。

苏森（1971-），男，山东菏泽人，博士，北京邮电大学教授、博士生导师、交换与智能控制研究中心主任，主要研究方向为下一代网络和新一代互联网服务。

杨放春（1957-），男，北京人，博士，北京邮电大学教授、博士生导师，主要研究方向为新一代通信网络、服务计算、云计算、车联网。

訾玲玲（1981-），女，辽宁阜新人，辽宁工程技术大学讲师，主要研究方向为智能软件、图形图像与多媒体。