数字图书馆服务器及其镜像的容量规划探讨

曾艳兰

（湖南大学图书馆，湖南 长沙 410082）

摘 要：对如何做好数字图书馆的服务器及其镜像系统的出口容量规划问题进行了探讨，重点研究了服务器群网络出口总容量的规划设计问题。提出以呼叫损失概率来衡量读者满意度的方法，将读者群按照上网速率进行分类，并采取为各类不同网速的读者预留适量带宽的措施以得到公平的服务质量。给出了主要服务质量指标和达到预期质量所需的容量的计算方法，并通过计算实例表明了方法的可行性。

关键词：服务器；服务质量；容量；负荷

中图分类号：TP393 **文献标志码**：A **文章编号**：1673-9833(2009)03-0054-04

On the Plan of the Server′s Capacity and Its Images in a Digital Library

Zeng Yanlan

（Library of Hunan University，Changsha 410082，China）

Abstract：The problems on how to plan on the server′s network capacity and its network images in a digital library are considered. The focus is on the design of the total outer network capacity of the servers. A method of call blocking probability is proposed to evaluate the readers' satisfication about the provided services. The readers are classified in accordance with their network information transferring rates and an appropriate bandwidth is reserved for each kind of readers in achieving fair service quality. The methods of calculating the quality indexes of service and the capacity for the expected service are both presented and the feasibility of the method is shown through numerical examples.

Keywords：server；quality of service；capacity；load

0 引言

数字图书馆在给读者带来便利的同时，也提高了文献利用率[1]。读者在数字图书馆上进行借阅和电子资源下载等活动时希望得到良好的服务，衡量这种服务质量的指标类似于人们在公众通信网上从事其它通信活动时的质量指标。电信网中一个最常用的指标为呼叫损失概率，简称呼损率，指用户在进行拨号时因线路忙而无法接通的概率。这一指标可以推广到数字图书馆的读者服务质量指标中，这里可以将呼损率定义为读者尝试从数字图书馆获得某项图书资料时因数字图书馆的服务器容量有限、瞬时并发用户数过多而暂时不能为读者提供服务的概率。显然，提高数字图书馆服务器的处理器速度、增大存储容量及其网络出口速率、增加数字图书馆镜像服务器个数可以降低呼损率而提高对用户的服务质量[2-3]，但需要增加软硬件成本和租用通信专线的成本。另一方面，用户一般能承受一定范围内的呼损，即只要呼损率低于某个门限值，读者就会对数字图书馆的服务表示满意。因此如何在经济成本和用户满意度间进行权衡，是在进行数字图书馆服务器及其镜像服务器规划时要解决的一个

收稿日期：2009-03-05

基金项目：湖南省社会科学基金资助项目（05YB37）

作者简介：曾艳兰（1968-），女，湖南武岗人，湖南大学图书馆助理馆员，主要从事图书信息管理系统方面的研究，E-mail：zengyanlan815@163.com

关键问题。

数字图书馆读者群呈现多样化的特性，主要体现为专业背景、阅读兴趣和阅读时间、上网速率等方面的多样性。考虑到读者群的上述多样性，需要将由若干个镜像服务器构成的数字图书馆网络服务系统建模为一个多速率系统[4]。可以通过问卷调查或对数字图书馆历史记录进行分析等手段得出各类不同网络速率用户的读者数量、他们访问数字资源的频率、平均每次访问需要下载的信息量或持续时间等有关通信业务量的数据。多速率系统中不同速率类型连接请求的呼损率不仅与各类读者访问服务器所构成的负荷有关，还与所采用的资源预留策略有关[4]。通信网络中关于这类问题的代表性研究有文献[5-7]。

提高数字图书馆服务器及其镜像的网络出口总容量可以提高对读者的服务质量，但同时要增加经济成本。而如何对数字图书馆服务器及其镜像的总容量进行规划以保证满足读者需求的同时尽量降低成本是本文探讨的主要问题。

1　服务器镜像规划需要考虑的因素

对数字图书馆服务器及其镜像进行规划时，需要考虑的主要因素包括：电子资源的种类及数量；读者数量及通过网络访问服务器的流量（负荷）；服务器及其镜像的软硬件成本；读者满意度的衡量。

1.1　电子资源的种类及数量

数字图书馆为读者提供的各种电子图书、电子期刊、电子报纸等可以按与纸质图书统一的方法分类，以便于图书馆管理人员管理，也便于熟练的读者查找所需要的资料。

1.2　读者数量及通过网络访问服务器的负荷

为便于数字图书馆经营者做好服务器及其镜像的规划，需要对读者进行分类。要对读者按专业需求、喜欢阅读的时间段、喜欢阅读哪些图书或报刊等进行分类。这种分类越细致越好。传统纸质图书馆所积累的大量研究成果有助于我们作出较准确的分类。由于我们对读者分类的主要目的是统计出各种不同上网速率的读者访问数字图书馆所形成的业务量负荷，因此还要对每个读者的上网速率进行登记。对读者进行分类需要的主要信息如表1。

表1　读者信息统计表的主要字段

Tab. 1　The principal fields of the table of readers' information

读者编号	专业	喜欢的阅读时段	喜欢的图书	上网速率

读者的实际上网速率与网络的背景负荷有关，本文仅考虑读者所采用的上网方式、预订的上网速率和数字图书馆服务器的服务能力的影响，而不考虑复杂的广域网的影响。典型的上网速率有100 kb/s，2 Mb/s和10 Mb/s等。在进行服务器镜像规划时，最终读者分类是按上网速率进行的。需要统计或估计各类上网速率的读者访问数字图书馆所形成的负荷，这里某类速率的业务量负荷是指为一个该类读者服务的连接持续时间内，其它同类读者平均发起连接请求的次数。

1.3　软硬件成本

服务器的成本包括硬件成本和软件成本，此外还有通信成本。硬件成本主要决定于CPU的速度、内存容量、硬盘容量和网络出口速率等；软件成本包括操作系统、数据库系统、ILLAS系统及其它辅助软件的成本；通信成本主要指维持每个服务器或其镜像与Internet的连通需要缴纳的通信费用。本文考虑影响软硬件成本及通信成本的一个关键因素是服务器的网络出口速率，因而进行服务器镜像规划时以服务器群的网络出口总容量为成本的直接体现者。

1.4　读者满意度的衡量

读者通过网络获取数字图书馆提供的图书资料，每次访问活动体现为一个下载或在线阅读电子资源的过程。实际上，这是一个包含了连接建立、信息传输、释放连接的过程。当所有正在接受服务的并发连接和一个新发起的连接所需的网络出口速率未超出服务器的服务能力或其网络出口容量容许时，该连接请求会被接纳，否则被拒绝。被拒绝的概率称为呼损率，用B表示。读者的满意度S可以用呼损率来度量。呼损率越小，读者满意度越高，我们可以用连接接纳概率$S=1-B$来表示用户的满意度。

2　排队模型

信息在通信线路、网络设备和服务器等公共设施上都要经历排队过程，这种排队过程一般都是随机过程。运用排队论来对计算机系统或网络的性能进行分析评价是学术研究中的通用方法。要解决数字图书馆服务器及其镜像规划问题，首先要解决的一个基本问题是建立衡量服务质量的主要指标与镜像服务器的主要参数、读者群访问数字图书馆所形成的网络负荷等因素间的函数关系。或者说，就是要确定各类网速的读者的连接请求被接纳的概率与数字图书馆服务器群的网络出口总容量、各类不同上网速率的读者群访问服务器或其镜像所形成的负荷之间的关系。下面从排队论角度对问题作进一步描述。

2.1　有关参数

服务器群的网络出口总容量C，常用单位为Mb/s，它是规划问题中要求解的量，是数字图书馆经营者决定需要添置的服务器台数、其CPU及内存配置、软件系统的关键依据。

各类上网速率的读者访问服务器所形成的负荷。设读者上网速率有 N 种，我们用 b_i（i=1, 2, …, N）表示各类读者的网速，分类的序号按网速大小从小到大排列，即 $b_1<b_2<\cdots<b_N$；用 a_i（$i=1, 2, \cdots, N$）表示第 i 类网速的读者群随机访问网络所构成的平均负荷。其定义是在一个该类网速的连接持续时间内，同类读者发起的连接请求数的平均值。为了便于沿用排队论中已有的方法，按惯例对读者发起连接请求的随机过程及读者每次访问服务器的连接持续时间作如下假定[4-8]：

第 i 类读者发起新连接请求的过程是参数为 λ_i 的泊松过程，连接请求的持续时间服从参数为 μ_i 的负指数分布，因此 $a_i=\lambda_i/\mu_i$。

各类网速的负荷需要根据对所有读者按表 1 所需的各项信息作长时间统计分析才能得出。

2.2 服务质量指标

本文考虑的服务质量指标是呼损率。某类网络读者的呼损率 B_i 是指当该类网速的读者发起连接请求时，恰逢服务器的剩余可用带宽已不足以为该读者提供其所期望的网速（带宽）的概率，即连接被拒绝的概率。一般而言，不同网速的读者可能有不同的服务质量即呼损率，这与服务器所采用的带宽共享策略有关。共享策略可分为完全共享和部分共享两大类[4]。采用完全共享策略会导致网速越高的读者用户感受到的呼损率越高，因此一种比较常用的带宽共享策略是服务器为每类不同网速的用户预留 1 个合适的带宽，使所有不同网速要求的读者具有相同的呼损率[5]。

2.3 呼损率的计算

在带宽完全共享策略下，各种不同网速的读者的呼损率的计算公式为：

$$B_k=\sum_{j=1}^{b_k-1}\left(G(C-j)/G_\Sigma\right)\ , \qquad (1)$$

其中：

$$G_\Sigma=\sum_{i=0}^{C}G(i), \qquad (2)$$

$$G(i)=\begin{cases}1 & \text{for} \quad i=0,\\ \dfrac{1}{i}\sum_{k=1}^{N}a_k b_k G(i-b_k) & \text{for} \quad i=1,2,\cdots,C,\\ 0 & \text{other},\end{cases} \qquad (3)$$

在采用恰当的带宽预留策略以使各类网速的读者具有公平的呼损率的情况下，一种为多数学者采用的计算呼损率的方法是[5]：

$$B_1=B_2=\cdots=B_N=\sum_{j=1}^{b_N-1}\left(G(C-j)/G_\Sigma\right), \qquad (4)$$

而下式中 $G(i)$ 的计算与式（3）略有不同，

$$G(i)=\begin{cases}1 & \text{for} \quad i=0,\\ \dfrac{1}{i}\sum_{k=1}^{N}a_k D_k(i)G(i-b_k) & \text{for} \quad i=1,2,\cdots,C,\\ 0 & \text{other},\end{cases} \qquad (5)$$

其中：

$$D_k(i)=\begin{cases}b_k, & i\le C-b_N+b_k;\\ 0, & i>C-b_N+b_k。\end{cases} \qquad (6)$$

此时 G_Σ 的计算法与式（2）相同，仍是所有 $G(i)$ 之和。

需要说明的是，式（3）、（5）的含义是它们正比于服务器的容量中被所有在线的连接占用的带宽恰为 i 个带宽单位的概率，而文献[8]表明，按式（5）计算的这一概率是有误差的，因此它只能看作是一种近似方法。要较准确地计算出系统的带宽被占用的概率分布，需要从多维 Markov 链的全局平衡条件出发进行求解[8]。但从工程应用角度而言，按式（4）~（6）计算呼损率与精确结果相比，其典型相对误差在 5 % 左右，属于可接受的范围，同时考虑到精确计算将使计算量显著增加，因此本文采用式（4）~（6）来计算呼损率。

3 规划问题的求解

本文所讨论的数字图书馆服务器及其镜像的规划问题主要是服务器的网络出口总容量的规划问题，即如何根据读者群访问各种数字资源所形成的网络流量负荷及预期的服务质量来确定服务器的出口总容量，它是决定服务器群的硬件配置及软件系统的关键依据。由于读者访问数字图书馆的负荷具有较大的随机性，而在总容量不变及其它类型负荷不变的前提下，数字图书馆服务器群提供的服务质量显然是随着任何一类读者负荷的增加而下降的，因此这种规划应当以数字图书馆较为繁忙时段的统计数据为依据，这样所确定的容量能保证（即使在繁忙时段也能保证）为读者提供的服务质量不低于所承诺的最低水平。

设 B_0 表示数字图书馆经营者承诺的呼损率上界，即读者所能容忍的呼损率的最高值。在采用具有公平呼损率的带宽预留策略时，式（4）给出了根据指定负荷和容量情况下求呼损率的方法，认为该式给出了呼损率 B 与服务器群网络出口总容量 C 及负荷 a_i（$i=1, 2, \cdots, N$）的函数关系。因此求给定负荷 a_i（$i=1,2,\cdots,N$）及预期呼损率 B_0 所对应的容量 C_0 的问题就是由式（4）表达的函数的求逆问题，可以记为 $C_0=B^{-1}(B_0)$。进一步的分析表明，B 是 C 的单调递减函数，而服务器的网络出口容量一般为某一最小单位的整数倍，因此得出服务器群的总容量 C 应将 C_0 向上取整，即

$$C=\left\lfloor B^{-1}(B_0)\right\rfloor。 \qquad (7)$$

根据式（4）计算呼损率的过程是 1 个迭代过程，它并不是 1 个解析表达式，因此其求逆问题只能采用数值解法进行。本文后续的算例中，利用 B 与 C 之间关系的单调性，先用试探法确定 1 个初始的正确搜索范围，再采用二分法找到最终的容量 C。

4　算例

按照前面给出的根据已知负荷和服务器群网络出口总容量计算呼损率的方法，以及根据已知负荷和承诺的服务质量下限确定容量的方法，作者编制了有关程序，进行了大量的数值计算。图 1 给出了在具有 3 种不同网速、承诺的呼损率为不超过 3 % 的情况下，按本文容量规划方法求出的网络出口容量与负荷之间的关系的 1 个例子。3 种网速分别是 100 kb/s，1 Mb/s，10 Mb/s，它们分别是其中最小网速 100 kb/s 的 1 倍、10 倍、100 倍。3 种网速的读者的负荷按 200 ： 30 ： 3 的比例同时增大或减小，因此图中横坐标标示值对应的 3 种网速的负荷是标示值分别乘以200 Erlang、30 Erlang、3 Erlang。图 1 中所示的最大负荷标示值为 10，意味着 3 种网速的网络负荷分别为 2 000 Erlang 、300 Erlang 、30 Erlang，此时使呼损率不超过 3 % 所需的网络出口容量最少要 8 453 个单位带宽即 845.3 Mb/s，接近 1 Gb/s，需要用很高档次的单台服务器或多台中档以上服务器才能够胜任。

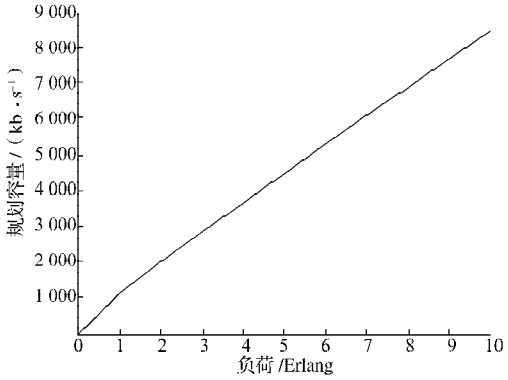

图 1　规划容量与负荷之间的关系图

Fig. 1　The relation between the planning capacity and loads

5　结语

本文对数字图书馆资源服务器及其镜像的规划问题进行了探讨，重点研究了服务器群网络出口总容量的规划问题，提出了以呼损率为服务质量评价指标的观点，归纳总结了进行数字图书馆资源服务器群规划需考虑的主要问题及解决办法，给出了服务质量指标的计算方法和服务器群出口总容量的求解方法，并给出了 1 个实例。

参考文献：

[1] 程　娟. 网络环境下文献资源共享效率评价指标研究[J]. 图书馆论坛，2008，28(2): 69-70，107.
Cheng Juan. Research on the Efficiency Evaluation of Reference Resources Sharing under Network Environments[J]. Library Forum，2008，28(2): 69-70，107.

[2] 张志彬. 高校数字图书馆光盘资源管理系统的设计与实现[J]. 图书馆论坛，2006，26(3): 110-112，139.
Zhang Zhibin. Design and Implementation of the Management System of Compact Disks of the Digital Library of Universities [J]. Library Forum，2006，26(3): 110-112，139.

[3] 徐伟强. 基于光盘镜像服务器的大型光盘信息系统[J]. 情报检索，2001(10)：70-72.
Xu Weiqiang. Large Compact Disks Information System Based on Compact Disk Imaging Servers[J]. Information Indexing，2001(10)：70-72.

[4] Kaufman J S. Blocking in a Shared Resource Environment [J]. IEEE Trans. on Commun.，1981，29(10)：1474-1481.

[5] Roberts J W. Teletraffic Models for the Telecom 1 Integrated Services Network[C]//Proc. 10th International Teletraffic Congress. Montreal：IEEE Press，1983：11-17.

[6] Siebenhaar R. Multi-Service Call Blocking Approximations for Virtual Path Based ATM Networks with CBR and VBR Traffic[C]//Proc. INFOCOM’95. NewYork：IEEE Press，1995：321-329.

[7] Awater G A，Van De, Vlag H A B. Exact Computation of Time and Call Blocking Probabilities in Large，Multi-Traffic，Multi-Resource Loss Systems[J]. Performance Evaluation , 1996(25)：41-58.

[8] 蒋红艳，林亚平. 多速率系统呼叫级和分组级的损失率分析[J]. 通信学报，2007(9)：15-21.
Jiang Hongyan，Lin Yaping. Analysis of the Loss Probabilities on the Call Level and Packet Level for Multi-Rate Systems [J]. Journal of Communications，2007(9)：15-21.

（责任编辑：罗立宇）