运营系统设备需求预测模型研究

The Forecast Model for Device Requirement of Operational System

胡茂伟，苏运霖

Hu Maowei，Su Yunlin

（广西大学梧州分校，广西梧州 543002）

（Guangxi University Wuzhou Branch，Wuzhou，Guangxi，543002，China）

摘要：在分析IT企业运营系统的设备预算问题的基础上，提出设备需求预测的鉴定标准，给出一套运营系统设备需求预测模型. 实施效果表明，预测模型有较高的设备需求预测精确度和系统载荷预测精确度，达到了鉴定标准.

关键词：运营系统 设备需求 预测模型

中图法分类号：TP338.6 **文献标识码**：A **文章编号**：1002-7378(2005)04-0228-04

Abstract：The assessment criteria of facility requirement is presented based upon the analysis of the facility budget of operational system in IT enterpvices. A forecast model for facility requirement of operational system is developed. The implementation of the model shows that it has higher accuracy in prediction of facility demand and system load.

Key words：operational system, device requirement, forecast model

随着互联网应用的日益广泛，单机或者局域网软件逐渐被互联网运营性软件替代，但是IT企业的运营带来另外一些问题. 首先，运营就意味着IT企业需要自己的服务器设备和带宽使业务正常运行，如腾讯的QQ游戏业务在100万人在线时该业务的服务器数量就达到了400多台. 由于增加一台服务器，就需要增加一定的带宽来满足业务需求，因此大量服务器设备的支出已经成为IT企业的最大成本之一. 其次，设备到位需要一定时间，如果等业务上涨了再增加设备，这个时候已经来不及了. 所以业界领先的企业都提出了设备预算的制度，但大型运营实施过程中发现设备需求和设备提供总有很大的差距. 如果需求满足不了，则严重影响业务发展. 如果设备供给过多，则引起资金过度消耗，严重影响公司整体的发展. 为了既不浪费设备又能不因为设备不足而影响系统性能，希望能将设备需求预测提高到一定的精确程度. 为此本文研究一套需求预测模型，并将模型进行实施，实施过程中收到了良好的效果. 因为设备预算与运营告警、业务优化变更息息相关，本文在分析设备预算过程前先介绍相关理论.

收稿日期：2005-08-11

作者简介：胡茂伟(1979-)，男，湖南岳阳人，硕士，主要从事软件开发和网络维护工作。

1 相关理论

本文解决问题的思路基于以下假设：

假设1 在高负载，多服务器的情况下，各个服务器的负载是均衡的. 在负载不均衡状态下，要引入负载变化模型，这里暂不考虑.

假设2 我们所有的推理是建立在一个确定的业务的前提上，即业务的功能形式都已经确定，而且自身的(收费策略等)和外部的游戏规则也已经确定. 因为虽然这些游戏规则几乎是时刻变化着的，但是它的变化只会引起各类指数之间的关系系数的变化，对下面的推理没有任何影响.

本文要建立模型必须在一定程度上先对逻辑进行简化，就像力学研究中简化摩擦力一样. 下面所有的引理和定理里都以对一个确定的业务为前提，所指也就是这个意思.

定义1 性能指数：单位时间内服务器性能表现出的最大值。例如CPU、内存使用率、Web服务器数目、cache数目、DB数目、故障次数、pageview.

定义2 业务指数：单位时间内业务发展相关的事件发生的次数。例如注册次数(人数)、贺卡发送次数、邮件发送次数. 它分为共有部分和特有部分. 共

有部分是几乎所有项目都需要涉及的一些指数,例如登陆次数、在线人数、注册人数等. 特有部分是一个特定的项目才有的一些指数,例如贺卡有贺卡的发送量统计,qqshow 有保存形象数目统计等.

引理1 对一个确定的业务,存在某种条件关系的业务指数之间肯定是存在某种近似关系.

证明 设对一个确定的业务,有事件 $C1$ 和事件 $C2$,而且事件 $C1$ 发生是以事件 $C2$ 发生为前提的. $C1$ 事件对应的业务指数为 $M1$,$C2$ 事件对应的业务指数为 $M2$.

根据概率统计学定义[1],事件 B 的发生是以事件 A 的发生为前提的,那么条件概率为:

$$P(B|A) = P(AB)/P(A). \tag{1}$$

对网络运营业务,完全符合概率事件中大样本、随机的两个条件. 根据定义 2,上面的定理也同样适用于存在某种条件关系的业务指数之间.

所以可得对于业务指数 $M1$ 和业务指数 $M2$ 之间存在的近似关系为:

$$M1 \approx f(M2) \approx M2 * P(M1|M2). \tag{2}$$

证毕.

对于一个业务,如果找到这种近似关系,那么通过一个指数 A 就能求得其他具有条件关系的指数 B,C 的近似值. 如果这个计算出来的理论值 B 和实际值 B' 差别超出一定范围,那么肯定直接影响 B 值的程序出了问题. 例如:一个系统(qlove)中普通用户注册量每天都是 4 万多,高级用户每天都 4 千多(是除10 的关系),如果某天普通用户注册是3 万多,高级用户注册突然变成了 500 多,虽然曲线没有陡升陡降,但是系统肯定也有问题,因为 $B - B' > B/2$(B 为理论值,B' 为实际值). 这个时候就应该触发系统拨测或报警. 所以业务指数这里只需要以一项关键指数作为主导即可,也就是说一项关键指数能够代表整个系统的业务情况.

定理1 一个业务的优化实际就是提高收费指数和其他指数之间的条件概率.

证明 根据上面的推理,可得对于业务指数 $M1$ 和业务指数 $M2$ 之间存在的近似关系为:

$$M1 \approx f(M2) \approx M2 * P(M1|M2). \tag{3}$$

当 $M2$ 一定时,$P(M1|M2)$ 越大,C 越大.

实际中这种关系也是显然的. 例如,QQ 交友中心系统的收费用户注册以免费用户注册为条件,也就是要成为收费用户必须先成为免费用户. 我们业务优化实际也是提高免费用户愿意注册成为收费用户的概率,也就是我们说的条件概率.

引理2 对一个确定的业务,业务指数大小能够代表整个系统的载荷大小.

定理2 对于一个确定的业务,设备需求预测以业务的预测为基础,也就是预测业务会在未来一段时间达到某个值的前提下,预测需要增加多少设备才能使在这个规模业务量的情况下没有设备浪费,也没有设备告急情况发生.

引理 2 的证明 系统载荷大小的直接表现是在业务终端打开的次数. 对于所有网络业务都是以用户打开业务终端的次数为条件的,根据引理 1,有业务指数 M 和系统载荷 D 之间存在的近似关系为:

$$M \approx f(D) \approx D * P(M|D). \tag{4}$$

对于一个确定的业务,$P(M|D)$ 是一定的,M 越大 D 越大. 所以有:对一个确定的业务,业务指数大小能够代表整个系统的载荷大小. 同理可以证明定理 2.

在现实中,业务指数上涨,整个系统的载荷上涨,如果服务器数目不变,那么每台服务器的载荷就上涨了. 如果服务器的载荷上涨超过了他能承受的阀值,那么系统就会有故障.

定理3 对于一个确定的业务,一项关键指数能够代表整个系统的载荷情况.

证明 由引理 1 和引理 2 可以得证.

定义3 业务量:服务器载荷大小的一个衡量尺度.

定义 4 设备性能指数:从设备本身采集的,能够准确表示系统当前的忙闲状态的参数. 根据以往的经验,对各种服务器,包括 apache server,mysql server,cache server,我们发现单位时间内最大进程数是最能表现系统闲忙的一个指标. 所以在我们的模型里我们也准备用单位时间内该设备里服务的最大进程数作为性能指数.

定义5 设备闲状态阀值:当设备性能指数低于这个值时,设备处于很闲的状态.

定义6 设备忙状态阀值:当设备性能指数高于这个值时,设备处于很忙的状态,也就是进入很容易引起故障的状态.

定义7 设备性能标准指数:当设备性能指数处于这个值附近时,设备处于最佳工作状态.

定义8 故障率:系统单位时间内出现的故障次数. 故障频率是一个可统计,又可直接表明系统服务质量的一个指数.

2 设备需求预测的鉴定标准

设备需求预测首先要解决和定义一个问题是鉴

定标准,以及问题解决好坏的比较标准.在一定业务量的载荷下,通常说的设备满足需求,就是在这些设备上运行的系统能保持的良好性能,包括服务器的反映速度和利用率,通常用设备需求预测精确度和系统载荷预测精确度来描述.

定义9 设备需求预测精确度:按照预测业务量和实际业务量的比率调整当前的服务器数目后,处于闲和忙之间状态的服务器与服务器总数之比.各类服务器预测精确度的积为设备需求预测精确度.

定义10 系统载荷预测精确度:实际服务器的载荷和预计的服务器载荷之间比值.

3 设备需求预测模型描述

设备需求预测的总体描述是:

(1)通过历史的业务量和性能指数数据得到一个曲线拟合方程,通过这个方程曲线可以得到相应业务量下系统性能指数的值.

(2)根据历史的系统性能指数和服务器台数得到导数方程.通过这个方程曲线可以得到相应业务下的服务器的设备性能标准指数.

(3)通过(1)和(2)得到的系统性能指数预测值和设备性能标准指数,就可以得到指定时间点需要的服务器数量.

(4)当前点的数据也会当作以后预测的输入数据来达到自我学习的功能,并且根据实际和预测的偏差来调整预测方程系数.

(5)记录历史时间点的指定业务指数之间的条件概率值,通过比较这些值来得到优化率.

(6)记录每个时间点的设备需求预测精确度和系统载荷预测精确度,对整个模型进行评估,同时也能看出该模型自学习的能力.

设t表示时间点,d表示该时间点系统实际总性能指数,d_i表示时间点i的系统载荷,m表示该事件点的业务量,m_i表示时间点i的业务量.设k_1,k_2,k_3,k_4,…,k_i,为拟合方程系数

自学习的预测模型为[3]:

$$d = f(m) = k_1m^1 + k_2m^2 + k_3m^3 + k_4m^4 + \ldots + k_im^i.$$

将各个时间点的数据带入方程中有:

$$d_1 = k_1{m_1}^1 + k_2{m_1}^2 + k_3{m_1}^3 + k_4{m_1}^4 + \ldots + k_i{m_1}^i;$$

$$d_2 = k_1{m_2}^1 + k_2{m_2}^2 + k_3{m_2}^3 + k_4{m_2}^4 + \ldots + k_i{m_2}^i;$$

…

$$d_i = k_1{m_i}^1 + k_2{m_i}^2 + k_3{m_i}^3 + k_4{m_i}^4 + \ldots + k_i{m_i}^i.$$

以上方程可通过解多元一次方程组得到系数,也可用矩阵求解.

在本文的模型里,根据6sigma策略[2],引入了一个很重要的参数——故障率.故障频率是一个既可统计,又能直接表明系统服务质量的一个指数.

下面几点是关于系统性能要求预测的说明:

(1)一个时刻、业务指数、性能指数和故障率组成一个多维点.许许多多的点组成多维曲线,用曲线捏合来预测未来的系统性能.

(2)根据实际情况决定1个月为一个点,或者是1个星期为一个点.得到多个点的拟合曲线后,输入时间轴的值、期望故障率的值和在改时间点期望达到的业务指数,就可以得到该故障率相应的服务器的预期台数[3].

(3)服务器的承载能力隐含在故障率和服务器台数的指标内.如果服务器台数太少,载荷过重,自然就会引起故障率增加,所以一个合理的故障率很重要.

(4)故障率和服务器成对出现,例如DB故障率和DB服务器数目是一对,cache故障率和cache服务器数目也是一对.所以各种服务器要求的预测也是成对的.

(5)服务器数目实际上可以看成服务器各项性能指数综合值和服务器载荷阀值的整除结果.所以多维点组成里的性能指数实际上是有多项值组成.但是刚开始,为了简化模型可以以一项关键指标作为主导,例如apache server,mysql server,cache server都可以分别以httpd,mysqld,cached的进程数目均方差为主导,后期再加入其他因素来优化模型,均方差越大说明系统瓶颈越多.

4 预测模型的实施

我们在一个大型网络应用系统中进行了实际实施工作.实施背景如下:业务量计算以该业务的最高在线人数计算;设备以设备金额计算.

经过3次迭代得到的设备预算模型为:

$$Y = 3658 - 47780 * 1/X + 237300 * 1/X^2 - 396800 * 1/X^3, \quad (5)$$

其中,Y表示金额;X表示从2004年10月开始计算的月数,具体运营和计算数据如表1所示,曲线拟合过程中得到运营曲线图如图1所示.结果表明,预测模型有较高的设备需求预测精确度和系统载荷预测

精确度,达到了鉴定标准.

表1 运营和计算数据

日期	理论值(千元)	实际值(千元)	业务量理论值(万人)	业务量计算	设备预测精度	系统载荷预测精度
2004-10	1771.38	1800	29.52	30		
2004-11	1836.33	1860	30.60	31		
2004-12	1929.60	1920	32.16	32		
2005-01	2015.22	1980	33.58	33		
2005-02	2094.38	2040	34.90	34		
2005-03	2168.09	2100	36.13	35		
2005-04	2237.15	2160	37.28	36	0.636	0.745
2005-05	2302.25	2220	38.37	37	0.646	0.893
2005-06	2363.94	2280	39.39	38	0.712	1.108

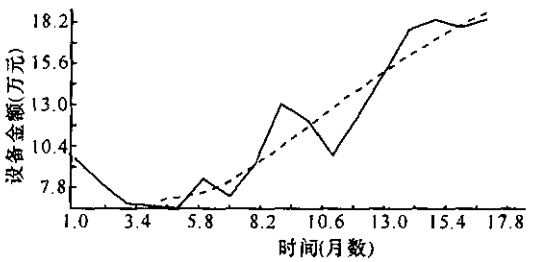

图1 运营曲线拟合

——:实际运营曲线;……:模型曲线

5 结束语

运营设备需求预测是IT企业控制成本的关键,许多大型IT企业为此提出了预算制度.本文在分析相关理论的基础上,提出了运营系统设备需求的鉴定标准,给出了相应的预测模型.从目前的3次迭代来看,准确率一直在提升,整个运营设备预算方案将会更为流畅,更工具化,这也是我们后续要继续推进的工作.

参考文献:

[1] 李懋和.概率论和数理统计[M].长春:吉林大学出版社,1999.

[2] 张 弛.6sigma实战[M].广州:广东经济出版社,2001.

[3] Hans Van Vliet. Software Engineering Principles and Practice[M]. Second Edition. New York: John Wiled & Sons, 2002.

(责任编辑:黎贞崇)

(上接第227)

[2] 潘谦红,王 炬,史忠植.基于属性论的文本相似度计算[J].计算机学报,1999,22(6):651-655.

[3] Feng J L. The research on decision supports system of nuclear accident emergency and its computer realization [D]. Beijing: Chinese Atomic Energy Institute, 2001, 97-118.

[4] 冯嘉礼.基于属性抽取与整合的感觉神经检测模型[J].计算机研究与发展,1997,34(7):481-486.

[5] Gerard Salton, Chris Buckley. Improving retrieval performance by relevance feedback [J]. J of the American Society for Information Science, 1990, 41(4):288-297.

[6] Salton G, Buckley C. Term-weighting approaches in automatic retrieval. Infor-mation [J]. Processing and Management, 1988, 24(5):513-523.

(责任编辑:黎贞崇)