基于类均值向量及指数分布的流量行为特征分析

董仕丁伟

(东南大学计算机科学与工程学院 江苏 南京 210092) (华东(北)地区网络中心 江苏 南京 210092)

摘要 引入类均值向量度量及 $\alpha\beta$ 指数分布方法,旨在提高分类正确率的基础上,克服由于抽样而带来的对分类结果的影响。 利用流记录 NOC_SET 为 DATASET,并以 NETFLOW 固有的测度和少量扩展测度为属性,利用所提出的 FBRI (Flow behavior identification) 属性选择算法对经典的机器学习算法进行流量识别。实验结果表明:任意比例的抽样对于采用 FBRI 属性选择的评估结果基本一致,并且利用 FBRI 属性选择算法可以提高应用识别正确率。

关键词 类均值向量 αβ指数分布方法 NETFLOW 机器学习

中图分类号 TP3 文献标识码 A

CLASS MEAN VECTOR AND EXPONENTIAL DISTRIBUTION TRAFFIC BEHAVIOR FEATURE ANALYSIS

Dong Shi Ding Wei

(School of Computer Science and Engineering, Southeast University, Nanjing 210092, Jiangsu, China) (Eastern China(North) Regional Network Center, Nanjing 210092, Jiangsu, China)

Abstract The introduction of class mean vector measurements and α β exponential distribution methods is to base on improving the correct classification ratio to overcome the classification result influence brought by sampling. Taking flow records NOC_SET as dataset, in addition taking NETFLOW's inherent measure and a few extended measures as features, the method uses the proposed FBRI (Flow behavior identification) attribute selection algorithm to identify the traffic for classic machine learning algorithms. Experimental results show that the evaluation results to FBRI feature selections for arbitrary proportions of sampling are of the same; additionally, by using FBRI feature selection algorithm proposed in the thesis, the correct identification ratio can be improved.

Keywords Class mean vectors α β exponential distribution method NETFLOW Machine learning

0 引言

随着网络带宽不断增长,网络行为模式日益复杂,各种新网 络应用产生,网络流量识别作为网络管理中一个研究热点方向 逐渐受到国内外研究人员的关注。在流量识别中为了获取更加 精准的识别精度、更高的识别效率,就需要对分类的数据进行流 量的特征选择,把对分类精度影响很大的特征属性通过量化以 及有效评估的方式选择出来。目前常见的属性选择方法主要有 三类:信息增益[6]、卡方分布[6]和互信息 MI[7],信息增益容易 受样本比例的影响,从而导致在样本的选择上遭遇局限性。而 互信息 MI 方法也仅考虑属性和类之间的相关性,互信息量越 大表明属性和类之间的相关性越强。但是忽略了属性之间的相 关性。目前很多研究者都是基于全报文的数据进行分析,这种 数据由于采集起来要付出很大的代价,而且对于在线的流量识 别是很大的一个障碍,因此本文以基于 NETFLOW 抽样的流记 录和扩展流记录作为研究对象,提出并深入分析了其流记录的 几种测度,并依据分析的流记录测度作为特征属性。通过引入 类均值向量及 α β 指数分布等概念而提出了 FBRI 属性选择算 法,并采用机器学习[1-5] C4.5 分类算法进行分类,结果表明,该 属性选择算法可以提高分类结果的正确率,并且采用基于 NET-FLOW 固有流记录和扩展流记录组成的特征属性进行分类的总体效果和采用全报文采集数据的分类结果基本相同。而且通过对报文进行抽样进一步验证了 FBRI 选择算法能有效地消除传统方法所带来的易受抽样比影响的因素。因此采用该方法可以运用到在线的网络流量识别中,并且获取比较好的识别效果。

1 类均值向量及 α β 指数分布

1.1 类均值向量式分析

引入类均值向量主要是对类内和类间的距离有一个度量, 并且通过各测度对类内和类间的距离的影响来评估测度的重要 性及测度与类之间的相关度。

类均值向量相关的概念如下:

定义1 使用 h; 表示第 i 类样本集的均值向量式为:

$$h_i = \frac{1}{n} \sum_{k=1}^{n_i} x_k^{(i)} \tag{1}$$

收稿日期;2011-09-01。2011中国计算机大会论文。董仕,博士生,主研领域:网络测量与网络行为学。

定义 2 用 h 表示所有各类的样本集的总平均向量式为:

$$h = \sum_{i=1}^{c} P_i h_i \tag{2}$$

$$S_b = \sum_{i=1}^{c} P_i (h_i - h) (h_i - h)^{\mathsf{T}}$$
 (3)

$$S_{c} = \sum_{i=1}^{c} P_{i}(x_{k}^{(i)} - h_{i})(x_{k}^{(i)} - h_{i})^{T}$$
 (4)

式中, S_a 为类间之间的模糊距离。 S_a 为类内的模糊距离。因此根据类内和类间距离,有如下公式:

$$A(x) = \frac{|S_b + S_c|}{|S_c|}$$
 (5)

A(x) 值越大代表类间越大,类内距离越小。

计算所有的 A(x) 值,去掉 A(x) 为 0 的特征,并抽取特征 i 的 $A(x_i)$ 满足式子组成的特征子集。

$$A(x_i) > \frac{1}{n} \sum_{k=1}^{n} A(x_k)$$
 (6)

下面给出一个例子来说明计算方法:

例:设有两类问题,其先验概率相等,即 $p(x_1) = p(x_2) = \frac{1}{2}$,样本均值向量分别为:

$$h_1 = [4,2]^T$$
 $h_2 = [-4,-2]^T$

协方差矩阵分别是:

$$\sum_{1} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \quad \sum_{2} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

为进行降维处理,把维数从 2 维压缩到 1 维,需求出 S_c 的 值及 S_c 的特征向量,因此:

$$S_c = \frac{1}{2} \sum_{1} + \frac{1}{2} \sum_{2} = \begin{bmatrix} 3.5 & 1.5 \\ 1.5 & 3.5 \end{bmatrix}$$

并求得它的特征值矩阵 Λ 和特征向量 U:

$$\Lambda = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix} \quad U = \begin{bmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{bmatrix}$$

已知
$$S_b = \begin{bmatrix} 16 & 8 \\ 8 & 4 \end{bmatrix}$$

可计算得 $A(x_1) = 3.6$, $A(x_2) = 1$ 。

通过每组特征对类内和类间距离的影响度,可以分析测度和类别之间的相关性。而测度之间的相关度就不能被评估。为此我们在下一小节中引入 α β 指数分布的概念。

1.2 αβ指数分布

数据集合本身是一张二维表,表中的每一行相当于一个样本,每一列相当于一个属性或测度,总体表空间为数据的样本空间。

定义α指数:

$$I_{\alpha}(A_i, A_j) = \sum_{k=1}^{n(A_i)} \sum_{l=1}^{n(A_i)} |a(A_i, k, A_j, l)|$$

其中 $a(A_i,k,A_j,l)$ 的具体定义见文献[8]。

定义β指数:

$$I_{\beta}(A_{i},A_{j}) = \sum_{k=1}^{n(A_{i})} \sum_{l=1}^{n(A_{i})} | P\{A_{i} = c(A_{i},k),A_{j} = c(A_{j},l)\} | -P'\{A_{i} = c(A_{i},k),A_{j} = c(A_{j},l)\}$$

其中 P' 为 P' { $X = x_i, Y = y_i$ } $= p_i \cdot p_j; i, j = 1, 2, \dots; c(A_i, k)$ 代表 A_i 中第 k 个的取值。

利用概率的相关性理论来研究属性之间的关系的时候,根

据不同的数据类型可能有不同的分布特点,而一般研究对象的数据空间都是有限的,因此所得到的概率值和理想值会有抖动。但是这种抖动在可控制的范围内。当进行两两属性进行分布概率统计时,会出现如下的问题点。(1)区分的分布粒度对属性选择精度的影响。(2)抽样对分布概率的影响。

针对上述问题进行分析:因为 A_i 、 A_j 性之间会存在交叉和重叠以及不同的分布,那么如何区分。当 β 指数接近0的时候表示两属性是相互独立的。但是当分布区间的范围过大则会造成精度丢失,会导致概率相关性方法的失效。那就要提出区分粒度的评估概念。由于 α 指数为正值代表 A_i 、 A_j 会彼此减少对方的不确定性。相反为负值则会绝对值越大越增加对方的不确定性。为0时反而会彼此互相干扰。因此选择剔出 α β 指数为0的数值,然后考虑更多的 α 指数为正的值。而且要采用投票机制来完成对概率区分粒度影响的评估。具体的算法如下:

Algorithm vote (Ai, Aj)

For
$$(i = 0; i + + ; i < n)$$

For
$$(j=0;j++;j< n)$$
 If I_{α} or $I_{\beta}=0$ then

else

Set the number of the corresponding interval S

Compute MAX(Ia) in every interval

Compute MIN(IB) in every interval

Compute the number of selected properties i ,j in every interval t++;

compute number(i,j)/total number

应用此算法对评估结果的预测效果如图1所示。

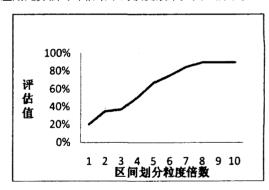


图 1 概率区分粒度对评估率的影响

从图1可以看出随着概率区分粒度的增加,评估效率会有不断的提高,在一定的区分极限会变得缓慢直至停滞。

通过上述两种算法的介绍,因类均值向量方法只侧重属性 对类之间的影响而 α β 指数分布则多考虑属性之间的相关性。 故提出 FBRI 算法,具体的算法思想如下:

先通过类均值向量选择对类影响比较大的属性。并通过对 两两属性之间的相关性进行评估,选出影响最大的属性集合。 此属性集合是根据投票算法中的投票数>均值投票数的集合。

伪代码如下所示:

Agorithm FBRI(i)

Properties by taking the mean vector greatest impact on the properties of class

While (flag)

The properties of the extract to further compare the α β exponential distribution

Compute vote(i) $> \frac{1}{n}$ total(i,j,k,...,n)

Choose i as new properties in all properties

2 测度属性

目前的研究大部分采用的数据是针对全报文采集的数据,这样可以获得更多的信息,更准确地对流量进行识别和分类。但是这些分类识别目前只能通过离线采集数据,然后再通过在线的方式进行识别,这样识别效率比较低。由于 NETFLOW 流的出现,我们尽量考虑采用 NETFLOW 固有流的属性来实现流量识别,这样可以减少很大的流量负载所带来的压力,又可以提高识别效率,真正实现在线的流量识别。鉴于此,本文考虑采用 NETFLOW 以及扩展的 NETFLOW 流记录作为研究对象,并提出如图 1 所表示的模型。

在具体介绍模型之前,先引人本文中所要使用的 NET-FLOW 流记录和扩展流记录以及识别的应用类型结果目标的描述:

定义3 NETFLOW 流记录和扩展的流记录样本描述: $x = (x1, x2, \dots, xt)$ 。

定义 4 应用类型识别目标集合的描述: $Y = F(x) = (y1, y2, \dots)$ 。

通过样本数据可以确定函数的参数,而分类器就是函数 F (x) 本身。

图 2 描述了基于 NETFLOW 流记录的在线流量识别模型。该模型分为数据收集、属性选择、流量分类模型建立、流量分类四个阶段。数据收集主要是全报文采集以及使用 DPI 工具进行打标签。属性选择主要是组流产生 NETFLOW 标准的流测度以及根据标准流测度进行的一些变形所产生的测度(详细见实验部分)。流量分类模型建立主要是根据机器学习算法建立相应的分类器。流量分类主要是运用建立的各分类器分别对 NETFLOW V5 格式及扩展的流记录进行识别分类,并得出相应的分类结果。

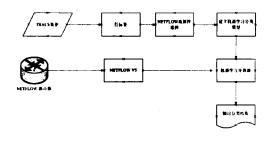


图 2 基于 NETFLOW 数据机器学习识别过程示意图 测度属性的描述如表 1 所示。

表 1 测度属性描述表

测度	测度描述	
双向报文数	前向和后向的报文数之和	
双向字节数	前向和后向的字节数之和	
平均报文长度	双向字节数/双向报文数	
持续时间	流结束时间-流开始时间	
TOS	NETFLOW 中双向 TOS 之 OR	
I_{α} TCPFLAGS1	某一方向流的 TCPFLAGS *	
TCPFLAGS2	另一方向流的 TCPFLAGS *	
传输层协议	NETFLOW 直接得到 *	

低位端口	NETFLOW 直接得到 *	
高位端口	NETFLOW 直接得到 *	
PPS	报文数/持续时间	
BPS	字节数/持续时间	
平均报文到达间隔	持续时间/报文数	
双向报文数比	流中双向报文数的比	
双向字节数比	流中双向字节数的比	
双向报文长度比	流中双向报文长度的比	

注:表1中共列出了16种测度,其中有5个测度(带*)可以在 NETFLOW中直接得到,而其余的部分需要进行相应的计算。

3 机器学习分类算法评估

目前流量识别算法有效性的评估标准有下面三个概念:

- (1) 真正 $TP(true\ positive)$: 实际类型为 i 的样本中被分类模型正确预测的样本数。
- (2) 假正 $FP(false\ positive)$:实际类型为非i 的样本中被分类模型误判为类型i 的样本数量。
- (3) 假负 FN(false negative):实际类型为 *i* 的样本中被分类模型误判为其他类型的样本数。

$$Precision(査准率) = \frac{TP}{TP + FP}$$
 (7)

$$Recall(査全率) = \frac{TP}{TP + FN}$$
 (8)

Overall accuracy(整体准确率) =
$$\frac{\sum_{i=1}^{n} TP_{i}}{\sum_{i=1}^{n} (TP_{i} + FP_{i})}$$
 (9)

4 实验结果与分析

4.1 本地抓包数据

本文主要研究了几种应用: WWW, Bulk, Mail, P2P, Service, Interactive, Multimedia, Voice。对 8 种应用分别抓包。利用 L7-filter 软件对这些数据进行打标签,并形成 NOC_SET 数据集。如表 2 所示。

表 2 NOC_SET

应用协 议标号	应用类别	所含协议举例	流数	比重 (%)
1	www	НТТР	4943	64.6
2	Bulk	FTP	39	0.5
3	Mail	IMAP, POP3, SMTP	91	1.19
4	P2P	BitTorrent , eDonkey , Gnutella , XunLei	1414	18.5
5	Service	DNS, NTP	433	5.7
6	Interactive	SSH, CVS, pcAnywhere	6	0.08
7	Multimedia	RTSP, Real	20	0.3
8	Voice	SIP, Skype	276	3.6
9	Others	games, attacks	431	5.6

注: "Flows"(流数量)采用 NETFLOW 组流规范对 Trace 组流得到,超时采用 16s, "Bytes"为每个报文按 68 字节计算的值。

4.2 IPTrace 数据

Trace 数据共 3 组:第 1 组采集于 2010 年 5 月 18 号 00:00 ~ 1:00,第 2 组采集于当天的 1:00 ~ 2:00,第 3 组采集于当天的 19:00 ~ 20:00,江苏省教育网边界到 CERNET 国家主干路由之间,每个报文均为 68 字节,前 8 字节为时间戳,后 60 字节为截取的报文长度;时间戳中 usec 的最后 1bit 为报文方向标志(0、1 分别表示出、进江苏省网)。由于信道吞吐量大,也为了保证 IP 流的完整性,报文抽样采用流抽样,抽样比为 1/4,抽样比的计算条件是所有被使用的 IP 地址的最后 3 位(bit)呈均匀分布。总体 Trace 的 3 组数据具体情况见表 3 所示。

表 3 IPTrace 数据

Trace	Pkts count	Bytes	Flows count
ALL_Tracel (0:00 - 1:00)	3.70E+8	2.52E + 10(24.0G)	3.74E+6
ALL_Trace2(5:00-6:00)	1.23E+8	8.39E+9(8.0G)	1.68E+6
ALL_Trace3(19:00 - 20:00)	9.78E+8	5.73E +11(533.57G)	1.12E+7

表 3 中三段不同的 Trace 采集于一天当中的不同时段,体现出了人们行为作息对流量的影响:比如晚上 7-8 点为上网高峰期,流量大,而凌晨 0:00-1:00 相对较少,5-6 点则更少,此时大部分用户正在睡眠中,因此流量极少。

4.3 实验平台和工具

本文的识别分类算法采用了 WEKA 套件中的 C4.5 算法, 采集打标签工具是采用基于 L7-filter 开源工具,并对由于该软件本身(存在大约5%的误报率)问题进行了改进。所采用的实验平台为两台 PC 机, Intel Core2 Duo CPU 2.80 GHz, 打标签采用系统为 Linux 系统,分类识别采用操作系统为 Windows XP系统。

4.4 实验结果与分析

本文分别对基于 NOC_SET 数据分别用 2 种选择算法进行实验: (1) 采用 FCBF 算法。(2) 采用本文提出的 FBRI 属性选择方法。另外所采用的分类算法为经典的 C4.5 分类算法,评估验证采用十折交叉验证对数据进行交叉验证。十折交叉验证法是常用的精度测试方法,它的基本思想是将数据集分成 10 份,轮流将其中的 9 份作为训练数据,1 份作为测试数据,进行实验。每次实验都会得出相应的正确率,采用 10 次结果正确率的平均值作为对算法精度的估计。通过对本文中数据进行交叉验证并通过分类算法进行评估后,所得结果如图 3、图 4、表 4 所示。

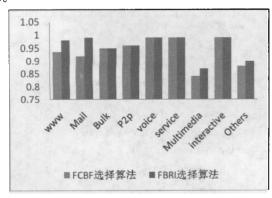


图 3 使用属性选择算法进行 C4.5 分类 的查准率 Precision 的结果

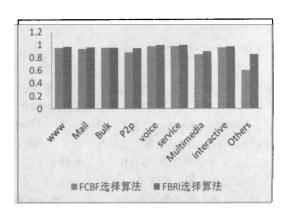


图 4 使用属性选择算法进行 C4.5 分类 的查全率 Recall 的结果

表 4 属性选择算法对 Overall 的影响情况

属性选择算法	Overall (%)
采用 FCBF 算法	92.36
采用 FBRI 属性选择算法	96.65

为了分析报文抽样对属性选择算法的影响,分别通过五种不同的抽样率来观察对最终总体正确率的影响,结果如图 5 所示。

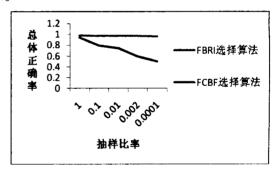


图 5 抽样对属性选择算法的影响

从图 3、图 4 可以看出,FBRI 属性算法比采用其他两种方式得到的属性有更高的识别率,从表 4 的总体准确率上也可以发现总体识别效果很好。图 5 中不同抽样比率采用 FBRI 选择算法总体正确率比较平稳,而传统的 FCBF 选择算法就出现随抽样比率的增加而出现总体正确率的下降现象。

综上可得,本文基于流记录并引入了基于 FBRI 属性选择算法进行属性选择,以识别效果最佳的 C4.5 为识别算法对流记录进行分类识别,识别结果表明:采用 FBRI 属性选择算法无论是在查准率还是在查全率上都比未使用选择算法以及使用FCBF 算法更高,另外从整体正确率方面也得到了验证。并且从结果上看,采用基于流记录进行流量识别,虽然相关的属性很少,但是识别结果却可以达到和采用全报文的数据集几乎差不多的分类效果,因此这样就为对于流量的在线分类提供了一种很好的方法。可以在基于 NETFLOW 现有的字段中加入上述的少量的属性特征也可以达到很好的分类效果,而且可以提高在线进行分类的效率。

5 结 语

本文在江苏省网边界以及本地分别获取数据,并利用 L7-filter 对数据进行打标签,得到基准数据集 NOC_SET。提出了

FBRI 属性选择算法对多维属性进行降维处理,并与常用的 FCBF 算法进行了对比,结果表明了本文提出的属性选择算法 在流量识别中可以使分类的结果更好,识别率更高,且不受因抽 样而带来的结果的抖动和不确定性。

本文的创新点在于:(1)基于江苏省网边界数据构建了 NOC_SET 标准数据集;(2)提出基于流的几种测度属性。(3) 提出 FBRI 属性选择算法。

基于本文的研究,下一步工作主要是:基于本文的流数据以及所提出的流测度属性,为后续研究提供数据支持,也能够通过改善属性选择算法来选择更好的测度以及提出更好的测度属性。

参考文献

- [1] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: Multilevel traffic classification in the dark [C]//Proc. Of the ACM SIGCOMM. Philadelphia, 2005:229-240.
- [2] Roughan M, Sen S, Spatscheck O, et al. Class-of-Service mapping for

- QOS: A statistical signature-based approach to IP traffic classification [C]//Proc. of the ACM SIGCOMM Internet Measurement Conf. Taormina, 2004:135-148.
- [3] Moore AW, Zuev D. Internet traffic classification using Bayesian analysis techniques [C]//Proc. of the 2005 ACM SIGMETRICS Int'l Conf. on Measurement and Modeling of Computer Systems. Banff, 2005:50-60.
- [4] 李君, 张顺颐, 王浩云, 等. 基于贝叶斯网络的 Peer to peer 识别方法[J]. 应用科学学报, 2009, 27(2):124-130.
- [5] 徐鹏, 刘琼, 林森. 基于支持向量机的 Internet 流量分类研究[J]. 计算机研究与发展, 2009,46(3):407-414.
- [6] 陈小莉. 基于信息增益的中文特征提取算法研究[J]. 重庆大学.
- [7] Kenneth, Ward Church, Patrick Hanks. Word association norms, mutual information and lexicography [C]//Proceedings of ACL27, Vancouver, Canada, 1989;76-83.
- [8] 曹源. 基于属性间相关性分类理论的属性选择方法研究[D]. 北京交通大学, 2007.

(上接第 244 页)

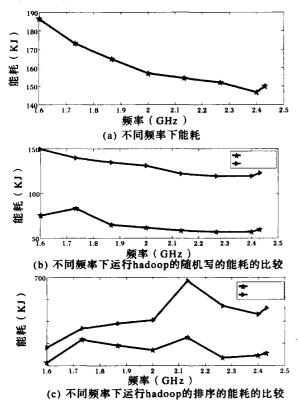


图 6 不同任务的实验结果

图 6(a) 描述了运行单一任务时的频率和能量消耗的关系。可以看出,随着节点频率的增加能耗有显著的降低,因为执行这个任务的时间减少了。也可以看出,频率增加到一定的时候,能耗开始增加。因此,对于不同的任务,应该将频率设置成这个转折点以达到最大程度的节能。

图 6(b)和(c)在分别运行随机写和排序程序下对能量消耗进行了比较。从图中可以看到,通过应用上节所描述的资源分配算法和节能方法,获得了高达50%的能源节约。

5 结 语

在文中,我们提出了一个对云计算平台进行能耗监控和能

源管理的系统框架。通过 ACPI、IPMI 实时收集系统信息和能源信息,并可以整合到云能源管理策略中。

本文又提出了一个在云计算环境下基于节能模式的资源自动分配算法。在云节点上运行一个任务时,我们发现并不是CPU的频率越高任务执行得就越快。其中存在一个转折点,在这个点上可能使我们达到频率、任务执行时间和能源消耗之间的平衡。不幸的是,多次实验后,我们发现这个转折点是浮动的,它依赖于具体的任务、硬件架构和网络。因此,我们部署了一个程序,它可以自动找到适合当时环境节能的 CPU 频率、内存模式和磁盘工作模式。

将来,我们计划通过收集和引入更多的信息,如温度或风扇转速,来完善算法。与此同时,一个强大的监控系统对云集群的能源管理很重要,为了进一步完善监控和管理系统,有很多工作需要做。

参考文献

- [1] Mahadevan P, Sharma P, Banerjee S, et al. A Power Benchmarking Framework for NetworkDevices [C]//Fratta F, et al. NETWORKING 2009, LNCS 5550,2009;795 - 808.
- [2] Feller E, Leprince D, Morin C. State of the art of power saving in clusters + results from the EDF case study [M]. INRIA Rennes-Bretagne Atlantique, France 31,2010.
- [3] Iannaccone N P, Wetherall G R. Reducing Network Energy Consumption via Rate-Adaptation and Sleeping [C]//Proceedings of NSDI (April 2008).
- [4] Elnozahy E N M, Kistler M, Rajamony R. Energy-Efficient Server Clusters [C]//Falsafi B, Vijaykumar T N, PACS 2002, LNCS 2325, 2003: 179-197.
- [5] Nurmi D, Wolski R, Grzegorczyk C, et al. The eucalyptus open-source cloud-computing system [C]//Cluster Computing and the Grid, 2009. CCGRID '09. 9th IEEE/ACM International Symposium on, 2009:24 -131.
- [6] Cooper B F, Silberstein A, Tam E, et al. Benchmarking Cloud Serving Systems with YCSB[C]//SoCC'10, June 10-11, 2010, Indianapolis, Indiana, USA.
- [7] Matthias E, Rukun M, Wang Xiaorui. Power Management for Main Memory with Access Latency Control M. 2009.