海量存储系统能耗评测模型的研究*

贺秦禄，李战怀，杨　雷，王惠峰，孙　鉴
（西北工业大学 计算机学院，西安 710129）

摘　要：针对海量存储系统能耗建模的问题，在理论研究的基础上用数学语言描述系统能耗的各个组成部分，并进行整合以得出系统整体的能耗模型。通过模型估算值与实际测试值的对比，验证了该能耗模型的有效性和可用性，为海量存储系统能耗预估以及制定相关能耗模型提供了思路。

关键词：海量存储系统；能耗模型；能耗预估
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2013)05-1419-04
doi:10.3969/j.issn.1001-3695.2013.05.034

Research on energy consumption evaluation model of mass storage system

HE Qin-lu, LI Zhan-huai, YANG Lei, WANG Hui-feng, SUN Jian
(*School of Computer Science, Northwestern Polytechnical University, Xi'an* 710129, *China*)

Abstract: For mass-storage system energy consumption modeling problems, in theory on the basis of the study and use mathematical language to describe each component of the system energy consumption, consolidation and energy consumption model to produce the system as a whole. Model comparison between estimated and actual test values to verify the effectiveness and availability of energy consumption model, it developed for the prediction of energy consumption, as well as mass storage system model provides a way of energy consumption.

Key words: mass storage system; energy consumption model; energy consumption estimate

随着信息技术广泛应用，数据在世界范围内共享传递，使得世界村的愿望真正实现。但是数据共享在给人们带来各种便捷和利益的同时，也无形中增加了资源的消耗，尤其是一些不合理的系统结构和数据中心的重复建设，使得能耗进一步增长。数据显示，能源成本的日益高涨使企业面临如何保持在较低能耗前提下，以更高效率运行其数据中心的问题。能源成本现在约占平均 IT 预算的 10%，除非企业采取根本性措施，否则可能在几年内上升到 50%。市场研究公司 Forrester Research 最新发布的 PC 市场数字预测称：从 IBM 的第一台电脑开始，全球达到 10 亿台电脑用了 27 年的时间。全球达到下一个 10 亿台电脑所用的时间会更快。到 2015 年，全球正在使用的电脑数量将达到 20 亿台，年增长率为 12.3%。Forrester 预测称：到 2015 年，仅中国就将新增加 5 亿台新电脑。目前，中国正在使用的电脑为 5 400 万台。大量电脑功耗造成的电费支出，已成为各级政府和企事业单位重要的财务支出[1,2]。以一般电脑 250 W 的功耗、一天使用 8 h、一年 300 d 来算，一个拥有 1 000 台电脑的单位，一年就要直接消耗 60 万度电，间接产生的二氧化碳排放达到 300 t 之多。随着能源价格的进一步升高，低能耗的产品将具有更强的竞争力。本文将能耗结构和具体测试数据进行整合，建立被测系统整体能耗的简单数学模型，通过理论值和实验数据的对比，证明模型的有效性，并阐述该模型的意义。

1　研究现状

目前，IT 业界已经普遍认识到存储产生的巨大能耗，企业用户也开始关注数据中心的电力成本，于是各存储厂商开始发布各种绿色认证和标准，目的在于标榜其产品的优势，但是作为买家却难以从这些所谓的绿色认证获取他们想要的信息，因为这些认证和标准大多数面向具体设备并且设定了各种条件，其测试条件脱离了实际应用环境，无法全面和真实地反映实际运行中存储设备的能耗状况。并且这些认证和标准的数据由于工作量、应用场合以及测试方法的不同而导致其不具备可比性[3]。然而，用户真正关心的是存储设备的能源利用率，尤其是正常运行状态下的能源利用率。由于存储能耗评测技术发展缓慢，现在还没有明确而具体的标准来给出用户想要的答案，但是已经有相关机构组织在进行存储能耗的研究项目，并取得了一定的成果。

SPEC(standard performance evaluation corporation，标准化性能评估组织)在 2012 年 6 月提出了首款针对服务器系统的能耗性能测试工具 SPECpower_ssj2008 v1.12 版[4]。该测试基准可以评估服务器系统在 10% 负载的待机模式到 100% 满负载状况下的功耗数据。这是业界第一项用于评测系统级别服务器的与运算性能相关的功耗的基准测试工具，该基准测试套件填补了服务器基准测试阵营中有关能效测试的空白。

IBM 在 2007 年提出了“ Project Big Green”项目，首次将能

收稿日期：2012-09-16；**修回日期**：2012-11-06　**基金项目**：国家“863”计划基金资助项目(2009AA01A404)；低能耗存储设备研制与产业化项目(2011BAH04B05)；国家自然科学基金资助项目(60970070，61033007)

作者简介：贺秦禄(1982-)，男，陕西西安人，博士研究生，主要研究方向为云存储、重复数据删除(luluhe8848@hotmail.com)；李战怀(1961-)，男，陕西旬邑人，教授，博士，主要研究方向为数据库、数据管理、海量存储；杨雷(1985-)，男，陕西旬阳人，硕士，主要研究方向为海量存储；王惠峰(1986-)，男，河北石家庄人，博士研究生，主要研究方向为海量存储；孙鉴(1982-)，男，山东烟台人，博士研究生，主要研究方向为海量存储.

耗问题列为数据中心设计的关键问题[5,6]。该项目于 2012 年 7 月发布了 4.0 版。在 Project Big Green 第二阶段,IBM 公司将通过模块化的数据中心产品和服务,包括 EMDC(enterprise modular data center,企业级模块化数据中心)、PMDC(portable modular data center,可移动的模块化服务器)、存储虚拟化产品等,力图将大企业或者中小企业的数据中心能耗降低 50%。

SNIA(storage networking industry association,全球网络存储工业协会) GSI(green storage initiative) 在 2011 年 8 月对外发布了供公众评议的绿色存储功耗测量规范 Green Power(green storage power measurement specification) 草案 v1.0 版,包括标准化功耗测量的指导方针和标准[7,8]。该规范包括了按照能源消耗与闲置率进行区分的绿色存储分级产品的特点和应用,对不同类型的设备在不同的 IO 负荷下进行功率的测定并取平均值得出设备的功耗基准。

SNIA Emerald 能耗测试规范对存储产品制定了活动和空闲状态下相关性能和能耗的评估方法和指标,并在提炼收集现有数据的基础上对该能耗测试规范作进一步的完善[9]。

以上评测规范都是对存储网络的能耗研究,都从设备角度出发,无法全面准确地反映系统的整体能耗状况。

2 能耗模型

2.1 组织架构分析

为了对存储能耗进行研究,需要进一步研究海量存储系统的组织构架,弄清楚其能耗的结构,为后面的测试工作打下基础。根据海量存储系统的功能体系结构,用其三大基础设备构建了相应的硬件组织结构,将其中的抽象层次具体化为本次研究的实际设备,将具体设备的功能放入相应的功能层次[9],如图 1 所示。

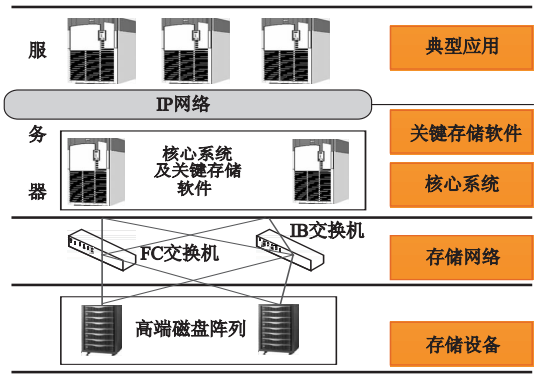

图1 海量存储系统设备层次图

SNIA 的“能耗测量规范”对系统进行了分类,直接从系统级别进行研究,但是这样的研究模式也导致了其进度十分缓慢,因为能源在系统内的流动方式与设备有着直接的关系,系统整体的能耗测试几乎不可能。本文的研究将系统能耗直接与设备挂钩,参照图 1 所示的海量存储系统结构模型,将设备和功能体系相结合,将存储设备具体化为高端磁盘阵列;将交换设备具体化为 FC 交换机、IB 交换机和以太网交换机,这样就明确了具体的目标设备。

电源将电能通过电路传输到系统中,系统中各个设备根据自身需求消耗部分电能以提供功能或者维持服务,各个设备消耗的具体电能数值不尽相同,与设备硬件构成和工作配置有着直接关系。本文用图 2 来反映海量存储系统的耗能状态。

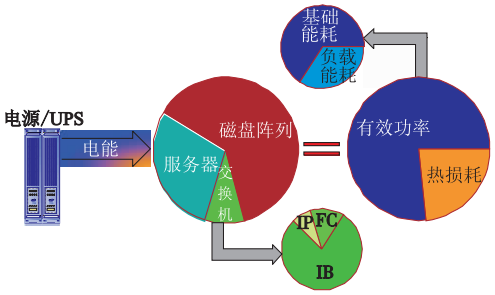

图2 海量存储系统电能利用情况

一般来说磁盘阵列的功耗较大,常常能占据整个系统总功耗的 60% 以上;其次是服务器,占整个系统总能耗的 20% ~ 30%;交换机的功耗需要根据实际的网络结构定,以太网交换机和 FC 交换机的功耗都较低,而 IB 交换机则较大,一台 IB 交换机的功耗常常大于等于一台普通服务器的功耗[10,11]。

2.2 相关定义

通过 2.1 节的分析,海量存储系统的整体能耗是由具体硬件设备的能耗组合而成,使用数学符号来描述各个设备的能耗,并定义相关函数描述影响能耗的各个因素,具体如下:

假设:当前有这样的海量存储系统,其中所有的可扩展设备都将按照保证系统消耗能耗最小的准则安装(最大限度地使用所有插槽和端口),同时,系统中只有以太网交换机和 IB 交换机两种交换设备,设备互连只影响以太网交换机,阵列均采用 RAID0,其他设备和部件不限。

E:能耗,表示能源消耗,数值等于设备的平均有效功率值,通过不同的下角标表示不同的能耗。

E_{SYS}:整体能耗,整个被测系统的能耗总值。

E_{base}:基础能耗,设备运行所要求的基本能耗,由设备的硬件组成决定,通过不同的下角标表示不同设备的基础能耗,数值上等于设备空闲状态的功率值。

$E_{workload}$:负载关联的能耗,设备上运行的负载所带来的能耗增加值,根据设备不同形式有所不同,通过不同的下角标表示不同设备的基础能耗,数值通常用与负载数值有关的函数表示。

IOPS:磁盘阵列每秒的 IO 数量,以此表示磁盘阵列负载。

CWS:CPU 的负载百分比,以此表示服务器的负载。

N:表示某种设备或者可扩展部件的数量。

C:表示设备的可连接数或者部件的容量。

$DPF(x,A)=A\times x$:(direct proportion function)正比例函数,用来表示负载与能耗的关系,A 为常量系数,x 为自变量,并且根据设备的不同,A 和 x 所代表的内容有所不同。

$$CNF(y,B)=\begin{cases} n,\dfrac{y}{B}\leqslant n<\left(\dfrac{y}{B}+1\right),n\in N \\ 0,y=0 \end{cases}$$

:(container number function)容器数量函数,用来表示可扩展部件与其相应容器的数量关系,y 为自变量,表示可扩展部件数量,B 为常量系数,表示容器的容量,数值等于容纳 y 个部件需要的容量为 B 的容器的最小数量。

对于某具体的符号“X_S”,下角标“S”表示该信息归属于磁盘阵列,下角标“DA”表示该信息归属于服务器,下角标“SWB”“IBSW”表示该信息归属于交换机设备,下角标“disk”表示该信息归属于硬盘,下角标“DAE”表示该信息归属于硬盘框,下角标“CON”表示该信息归属于磁盘阵列控制器,下角

标“Head”表示该信息来源于设备空闲工作状态,下角标“Idle”表示该信息来源于设备空闲工作状态,下角标“Port”表示该信息归属于实际启用端口,下角标“D-Port”表示该信息归属于数据传输的端口。

注:IB 交换机由于条件限制不进行估算,通过直接测试给出能耗 $\sum E_{\mathrm{IBSW}}$。

2.3　构建能耗模型

利用 2.2 节中的符号定义和函数关系来形式化地描述海量存储系统的能耗按照先整体再部分的顺序,逐步建立海量存储系统的能耗模型。

由图 2 可知,整个系统的能耗由组成系统的磁盘阵列、服务器和交换机的能耗组成,可以表示成:

$$E_{\mathrm{SYS}} = \sum E_{\mathrm{DA}} + \sum E_{s} + \sum E_{\mathrm{SWB}} + \sum E_{\mathrm{IBSW}} \tag{1}$$

继续分解式(1),将其各个组成部分进一步展开:

$$E_{\mathrm{DA}} = E_{\mathrm{DA-base}} + E_{\mathrm{DA-workload}} \tag{2}$$

$$E_{S} = E_{\mathrm{S-base}} + E_{\mathrm{S-workload}} \tag{3}$$

$$E_{\mathrm{SWB}} = E_{\mathrm{SWB-base}} + E_{\mathrm{SWB-workload}} \tag{4}$$

磁盘阵列的基础能耗由其可扩展模块的能耗组成:

$$E_{\mathrm{DA-base}} = \sum E_{\mathrm{disk}} + \sum E_{\mathrm{DAE}} + \sum E_{\mathrm{CON}} + \sum E_{\mathrm{Head}} \tag{5}$$

服务器和交换机的基础能耗用其空闲状态的能耗表示:

$$E_{\mathrm{S-base}} = E_{\mathrm{S-Idle}} \tag{6}$$

$$E_{\mathrm{SWB-base}} = E_{\mathrm{SWB-Idle}} \tag{7}$$

设备能耗随着负载变化而变化,因而可以通过函数关系来表示:

$$E_{\mathrm{DA-Workload}} = \mathrm{DPF}(\mathrm{IOPS}, \mathrm{PPIO}) \tag{8}$$

$$E_{\mathrm{S-Workload}} = \mathrm{DPF}(\mathrm{CWS}, \mathrm{PETC}) \tag{9}$$

$$E_{\mathrm{SWB-Workload}} = \mathrm{DPF}(N_{\mathrm{Port}}, \mathrm{PEPT}) + \mathrm{DPF}(N_{\mathrm{D-Port}}, \mathrm{PEDT}) \tag{10}$$

同时,设备之间的连接关系将决定部分可扩展部件和端口的数量,可以通过函数关系表述如下:

$$N_{\mathrm{DAE}} = \mathrm{CNF}(N_{\mathrm{disk}}, C_{\mathrm{DAE}}) \tag{11}$$

$$N_{\mathrm{SWB}} = \mathrm{CNF}(N_{\mathrm{Port}}, C_{\mathrm{SWB}} - 1) \tag{12}$$

$$N_{\mathrm{Port}} = N_{S} + N_{\mathrm{Head}} \tag{13}$$

通过式(1)~(13)可以比较完整地描述海量存储系统的能耗状况,现将这些公式进行整理合一,得到整个系统的能耗模型:

$$\begin{aligned} E_{\mathrm{SYS}} &= \sum E_{\mathrm{DA}} + \sum E_{S} + \sum E_{\mathrm{SWB}} + \sum E_{\mathrm{IBSW}} = \\ &\sum (E_{\mathrm{DA-base}} + E_{\mathrm{DA-workload}}) + \sum (E_{\mathrm{S-base}} + E_{\mathrm{S-workload}}) + \\ &\sum (E_{\mathrm{SWB-base}} + E_{\mathrm{SWB-workload}}) + \sum E_{\mathrm{IBSW}} = \\ &N_{\mathrm{disk}} \times E_{\mathrm{disk}} + \mathrm{CNF}(N_{\mathrm{disk}}, C_{\mathrm{DAE}}) \times E_{\mathrm{DAB}} + N_{\mathrm{CON}} \times E_{\mathrm{CON}} + \\ &N_{\mathrm{Head}} \times E_{\mathrm{Head}} + \sum_{1 \to N_{\mathrm{Head}}}^{i} \mathrm{DPF}(\mathrm{IOPS}_i, \mathrm{PPIO}_i) + \\ &N_{S} \times E_{\mathrm{S-Idle}} + \sum_{1 \to N_{s}}^{j} \mathrm{DPF}\left(\frac{\mathrm{CWS}_j}{10\%}, \mathrm{PETC}_j\right) + \mathrm{CNF}((N_{S} + N_{\mathrm{Head}}), \\ &C_{\mathrm{SWB}} - 1) \times E_{\mathrm{SWB-Idle}} + \mathrm{DPF}((N_{S} + N_{\mathrm{Head}}), \mathrm{PEP}') + \\ &\mathrm{DPF}(N_{\mathrm{D-Port}}, \mathrm{PEDT}) + \sum E_{\mathrm{IBSW}} \end{aligned} \tag{14}$$

式(14)组合了影响系统能耗的各种元素,将繁杂多变的能耗状况归纳成定量的数学公式,形式化地描述了整个海量存储系统的能耗状况,使得系统的能耗结构更加清晰,更具可控性。同时,该模型将降低大型系统能耗测试的工作量,通过模型计算配合少量测试便可预测大型系统的能耗状况,这在实际应用中具有十分重要的意义。

3　验证模型

数学建模的一个重要环节就是模型的验证工作,为了能保证所建模型的可用性。只有返回实际,将模型应用于实际的系统进行计算,方可验证其可用性[12,13],将通过模型估算值与实际测试数据的比较得出结论。

根据第 2 章开始提出的假设条件,使用文献[14,15]的被测硬件设备,搭建三个与假设同构的存储系统,测试拓扑采用软/硬件结合的方式,软件控制测试所需的负载组合,从硬件取得数据并计算系统的各个能耗指标。图 3 是能耗测试系统拓扑的示意图。

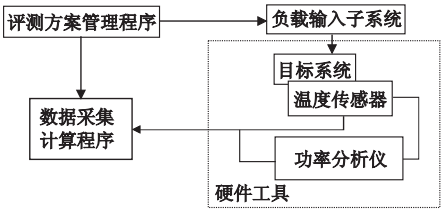

图3　能耗测试系统拓扑图

根据文献[14~16]测试海量存储系统的测试环境以及测试用例,分别对磁盘阵列、交换机和服务器的能耗进行测试,由于篇幅限制,列举几个关键指标测试的结果,如图 4~6 所示。

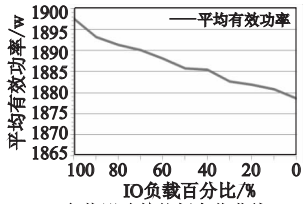

图4　磁盘阵列单位IOPS的能耗

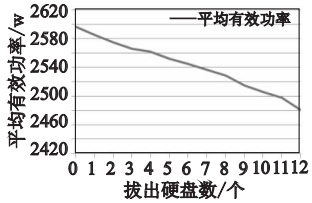

图5　磁盘块对能耗的影响

3.1　磁盘阵列能耗测试

$$\frac{\sum_{i=0}^{10} \mathrm{Watts}_i}{\sum_{i=0}^{10} \mathrm{IOPS}_i} = 0.08386\,(\mathrm{Watts/IOPS})$$

通过计算得出平均每块硬盘消耗 9.7 W。

通过计算得出平均每个控制器消耗 165.44 W。

3.2　交换机能耗测试

交换机空闲状态下能耗测试数据如表 1 所示。

表 1　测试数据

测试用例	平均电压/V	平均电流/A	有用功/W
1	222.19	6.957 8	1499.3
2	223	8.078 1	1757.6
3	222.62	7.075 3	1528.6

结果分析:

a)交换机内部全交换对交换机能耗有较显著影响。

b)交换机光模块有较大能耗,平均每块 IB 光模块消耗 1.5 W。

交换机满负荷下能耗测试数据如表 2 所示。

表 2　测试数据

测试用例	平均电压/V	平均电流/A	有用功/W
1 组服务器	220.63	7.138 1	1 529.8
2 组服务器	219.8	7.163 9	1 530
3 组服务器	222.37	7.090 8	1 530.4
4 组服务器	221.53	7.112 5	1 529.8
5 组服务器	222.72	7.078 2	1 529.9

结果分析:服务器间传输数据对交换机能耗需求变化不大,交换机能耗主要集中在交换机维持正常工作的消耗和光模块消耗。

3.3 服务器能耗测试(图 7)

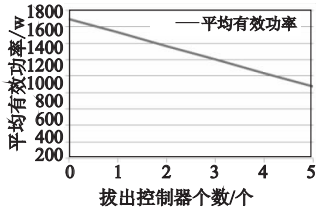

图6 控制器对能耗的影响

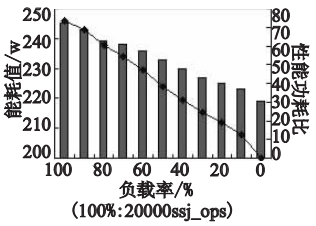

图7 AS300N服务器能耗

平均功率 219.1 W, ∑ssj_ops/ ∑power = 39.7

三个系统主要在设备的数量和负载上发生变化,具体如表 3 所示。

表 3 被估算系统具体配置

编号	项目						
	机头数量	控制器数量	服务器数量	硬盘块数量	磁盘阵列 IOPS	服务器 CPU 占用率/%	数据传输端口数
1	1	2	4	96	40000	均 80	4
2	1	1	1	50	20000	均 40	1
3	1	4	10	150	60000	均 60	10

将表 3 所示参数带入能耗模型,结合上面已经给出的部分测试数据以及文献[14]的测试数据,利用模型对整个系统的能耗进行估算,得出估算值:

E_{SYS-1} = 4118.747 W,其中 $\sum E_{IBSW-1}$ = 629.13 W。

E_{SYS-2} = 2173.411 W,其中 $\sum E_{IBSW-2}$ = 627.36 W。

E_{SYS-3} = 7712.651 W,其中 $\sum E_{IBWS-3}$ = 631.08 W。

通过具体测试的系统功耗为:

1)系统 1

磁盘阵列总功耗为 2561.794 W,服务器总功耗为 922.63 W,交换机总功耗为 21.972 W,IB 交换机总功耗为 629.13 W;系统 1 总功耗为 4135.526 W,比估算值高 16.779 W。

2)系统 2

磁盘阵列总功耗为 1307.154 W,服务器总功耗为 222.31 W,交换机总功耗为 21.856 W,IB 交换机总功耗为 627.36 W;系统 2 总功耗为 2178.68 W,比估算值高 5.269 W。

3)系统 3

磁盘阵列总功耗为 4761.372 W,服务器总功耗为 2321.7 W,交换机总功耗为 22.374 W,IB 交换机总功耗为 631.08 W;系统 3 总功耗为 7736.526 W,比估算值高 23.875 W。

通过估算和实际测试两种方式得出了系统能耗,图 8 将三组数据进行比较,以验证能耗模型的可用性。

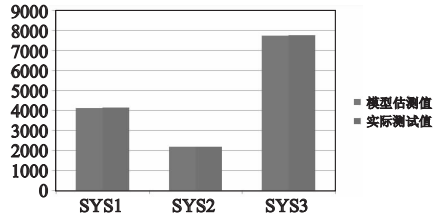

图8 估算值与实测值的比较

比较估算值与实测值显示,估算值与实测值十分接近,在 SYS1 中估算值低了 16.779 W,失真率为 0.41%;在 SYS2 中估算值低了 5.269 W,失真率为 0.24%;在 SYS3 中估算值低了 23.875 W,失真率为 0.31%。这样的失真率完全可以接受。同时,比较还显示系统规模越大,失真率越高,因为该能耗模型当前还是初级、大尺度地反映系统能耗,需要进行更多、更细致的研究,如考虑环境温度、功率因素、RAID 结构以及工作时间分布等。

应当看到,小规模测试辅助以建模估算将是一种适合于海量存储系统能耗测试的方法,这种测试方法既避免了大量重复的测试工作,又在一定范围内保证了测试的精度,而且还能通过能耗模型分析系统能耗结构,有助于系统硬件结构的优化,因而值得继续研究。

4 结束语

本文在研究海量存储系统能耗的理论和实际测试基础之上,建立了海量存储系统的能耗模型,并应用模型对实际系统能耗进行估算,将估算值与实际测试结果比较,证明了模型的可用性。该模型在系统级别上研究海量存储系统的能耗,清晰地反映了系统能耗结构,对于控制能耗、提高能源利用率以及优化系统硬件配置有着重要意义。同时,该模型的建立对于规范海量存储系统硬件组织结构、制定相应规范和标准有着重要意义。

本文是在假设的基础上,建立能耗模型以形式化地反映系统的整体能耗,但是,由于假设的条件限制,模型的失真率会随着系统规模变大而逐步增大,需要进一步完善。后面需要将 RAID 级别和设备限制放宽,并考虑温度因素和功率因素对整体能耗的影响,将设备每个状态的时间作为权值加入模型中,以便全面而精确地反映系统的整体能耗。

参考文献:

[1] 郭兵,沈艳,邵子立. 绿色计算的重定义与若干探讨[J]. 计算机学报,2009,32(12): 2311-2319.

[2] 林闯,田源,姚敏. 绿色网络和绿色评价:节能机制、模型和评价[J]. 计算机学报, 2011,34(4):593-613.

[3] LIU Hai-kun, JIN Hai, XU Cheng-zhong, *et al*. Performance and energy modeling for live migration of virtual machines[C]//Proc of the 20th International Symposium on High Performance Distributerd Computing. 2011:171-182.

[4] SPEC, SPECpower_ssj2008 benchmark[EB/OL]. (2012-07). http://www.spec.org/power_ssj2008/.

[5] EBBERS M, GALEA A, SCHAEFER M. The green data center: steps for the journey[EB/OL]. http://www.redbooks.ibm.com/redpapers/pdfs/redp4413.pdf.

[6] AINSWORTH P, ECHENIQUE M, PADZIESKI B. Going green with IBM systems director active energy manager[EB/OL]. http://www.redbooks.ibm.com/redpapers/pdfs/redp4361.pdf.

[7] SNIA[EB/OL]. http://www.snia.org/home/.

[8] The green storage and computing knowledge center, SNIA[EB/OL]. http://www.snia.org/forums/green/knowledge/.

[9] SNIA green storage power measurement technical specification [EB/OL]. [2010-09]. http://www.snia.org/tech_activities/publicreview/GreenPower_v018.pdf.

[10] JANG J W, JEON M, KIM H S, *et al*. Energy reduction in consolidated servers through memory-aware virtual machine scheduling[J]. IEEE Trans on Computers, 2011,60(4): 552-564.

[11] GRAUBNER P, SCHMIDT M, FREISLEBEN B. Energy-efficient management of virtual machine in EucalyPtus[C]//Proc of the 4th IEEE International Conference on Cloud Computing. Washington DC: IEEE Computer Society,2011:243-250

[12] KONSTANTINOU I, ANGELOU E, TSOUMAKOS D, *et al*. Distributed indexing of Web scale datasets for the cloud[C]//Proc of Workshop on Massive Data Analytics on the Cloud. New York: ACM Press, 2010.

[13] 杨硕,史仪凯,杨宁,等. 服务器关键能耗部件实时功率测量系统的设计与实现[J]. 计算机应用,2010,30(10):2846-2849.

[14] 海量存储系统能耗评测方案[R]. 2010.

[15] 海量存储系统评测体系[R]. 2010.

[16] 海量存储系统能耗测评报告[R]. 2011.