Achieving Energy Efficiency through Optimized Resource Usage

Rachit Shah
Computer Engineering Department,
Institute of Technology Nirma University, Ahmedabad 380009, India.
Email: rachitshah@hotmail.co.in

Prof. Vivek Kumar Prasad
Computer Engineering Department,
Institute of Technology Nirma University, Ahmedabad 380009, India.
Email: vivek.prasad@nirmauni.ac.in

***Abstract*— With an increasing amount of businesses shifting their point of operations from in-house applications to cloud solutions, the burden on cloud infrastructure is ever increasing. In this paper, I will discuss about ways to accentuate the efficiency of IaaS infrastructure resources like storage, network, computing (CPU cycles and GPU), time, applications, services, et cetera in terms of the energy consumption they utilize. By leveraging power with performance, an analysis of cloud applications will be done per the energy consumption while maximizing QoS constraints and minimize energy. We will discuss about a VM Scheduling algorithm which leverages resource utilization, availability and energy consumption. With the major goal of achieving energy efficiency which will not only lead to minimizing greenhouse gases but also achieving cost-efficient resource utilization, I will detail current industry practices and highlight possible solutions and their repercussions if they were brought in to cloud computing environment.**

***Keywords*-**Green Cloud, Energy Efficiency, VM Scheduling

I. INTRODUCTION

First of all, cloud computing can be defined as, according to Buyya et al, *"A Cloud is a market-arranged distributed computing framework comprising of an accumulation of interconnected and virtualized PCs that are progressively provisioned and introduced as at least one computing resource(s) in view of SLAs built up through transaction between the CSPs and customers." [3]*

In other words, cloud computing provides remote access to infrastructure like network, storage, computing, etc. or platforms or applications on a pay-per-use basis whenever the customer wants it. It provides an illusion of unlimited computing resources by optimal use of resources, dynamic scheduling and other algorithms.

A company can choose to deploy their application through various deployment models. These include:

1) **Private Cloud -** There is exclusivity in the use of infrastructure by a single organization for their own use, and can be managed and owned by a third party or the organization or both.
2) **Public Cloud –** The infrastructure is lent out to multiple people or organizations. It may be owned and managed by multiple organizations or third parties.
3) **Community Cloud –** A group of organizations with similar interests can deploy such cloud for easy sharing of data, research or test data.
4) **Hybrid Cloud –**Two or more different cloud models (public, private or community) are combined to form a hybrid cloud. A private cloud can provide interface to public clouds at times when its resources are insufficient. [2]

Green Cloud Computing is a new field which targets energy efficient use of cloud resources to reduce greenhouse gases as well as provide a cost-effective solution to avoid wasting resources.

To give an example, let's provide some statistics about energy consumption of data centers in US in 2014. With ever increasing reliance on IT technologies and the ever present and growing amount of data held by companies like Google, Amazon, Facebook, etc. at their data centers, and also the growing number of tech firms on cloud, the data centers which host these clouds are consuming electricity at an alarming rate. According to a study by US Government, about 70 million kilowatt-hours of electricity was consumed by US data centers in 2014 (2% of country's total energy consumption). To put this in to perspective, this is equivalent to amount consumed by 6.4 million average American homes per year. [8]

However, along with consumption, about 620 billion kWh will be saved by these energy efficiency improvements between 2010 and 2020.

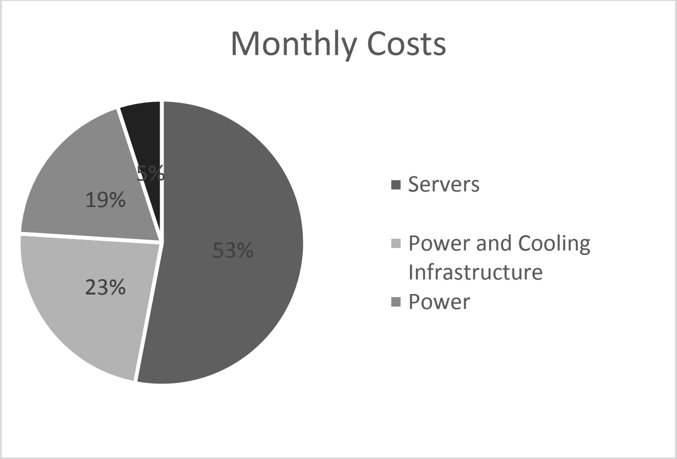

Figure 1 Monthly Costs of Data Centers

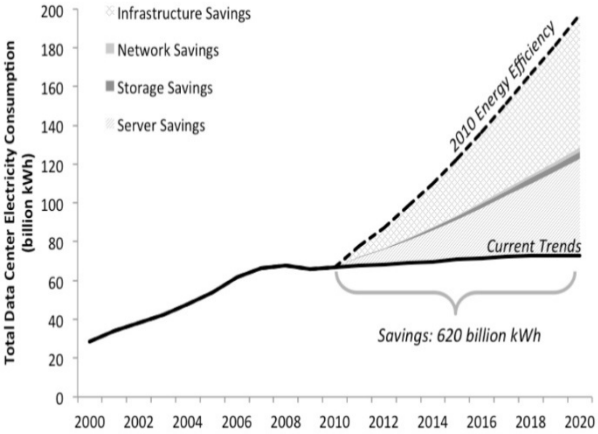

Figure 2 Energy Savings (Source: US Department of Energy, Lawrence Berkeley National Laboratory)

Despite this, cloud computing has actually made energy efficiency easier compared to traditional data centers. This is because of techniques like workload consolidation and resource virtualization. In traditional data centers, the demand for resources are sporadic and thus majority of resources are wasted. However, cloud computing with the help of virtualization makes use of the same physical host for multiple virtual machines. This makes it possible to consolidate tasks, making efficient use of the infrastructure and shutting down devices no longer in use to save energy. [1]

These days, application scheduling is generally utilized as a compelling energy conservation strategy. The energy utilization can be essentially decreased by merging applications such that the number of servers used are as less as possible and by turning off servers which are idle. By advancing the power utilization of the infrastructure like networks, few of existing application scheduling has been talked about. The network infrastructure is made to give high separation data transfer capacity to applications in regular large-scale frameworks. By sharing differing applications and each using a piece of the framework, each time a large portion of these systems stay unused.

The greater part of the present works does not describe about VMs during runtime. Because distributed computing condition workload differs with time, real time VM booking could diminish energy costs for calculation when workload decays and more VMs ought to be allotted when the workload increments. Arranging the various physical servers hierarchically in a server farm can expand the data transfer capacity utilization now and again when two groups impart through a hierarchical upper level connector. Some different works did not consider data transmission use amid the physical server of VMs. As huge overhead of cost occurs due to transfer speed, considering correspondence energy and data transmission utilization amid VMs amid VM planning could lessen the utilization of energy and data transfer capacity necessity as it were. None of them endeavors neighborhood asset accessibility in booking and in this way gives poor framework execution and expands transfer speed cost as it were. [1]

We will talk about in this paper about VM application scheduling its effect on the utilization of energy in the cloud framework in a server data center. The model of each bunch association and division of work among the parts inside the groups is additionally outlined. To plan VMs inside a bunch and among the groups inside the entire server farm, we have created two algorithms of scheduling VMs. The intra and inter cluster scheduling algorithms lessen utilization of energy by killing servers in excess and keeping less used clusters in rest mode. Inclination in relocating imparting VMs to same bunch guarantees data transfer capacity use and decreases for communication.

II. LITERATURE SURVEY

TABLE 1 – Literature Survey

Author Name	Paper Name and Year	Information	Open Research Issues
Jayant Baliga, Robert W. A. Ayre, Kerry Hinton, and Rodney S. Tucker [9]	Green cloud computing: Balancing energy in processing, storage, and transport, 2011	Energy consumption in switching and transmission, as well as data processing and data storage.	
Anton Beloglazov and Rajkumar Buyya [10]	Energy efficient resource management in virtualized cloud data centers,2010	live migration of VMs and propose heuristics for dynamic reallocation of VMs to minimize the number of physical nodes serving current workload	consider multiple system resource in reallocation decisions, such as network interface and disk storage.
Andreas Berl, Erol Gelenbe, Marco di Girolamo, Giovanni Giuliani, Hermann de Meer, Minh Quan Dang and Kostas Pentikousis [11]	Energy-efficient cloud computing,2010	reviews the usage of methods and technologies currently used for energy-efficient operation of computer hardware and network infrastructure	o Energy-aware data centers o Energy savings in networks and protocols o the effect of Internet applications
Saurabh Kumar Garg and Rajkumar Buyya [12]	Environment-conscious scheduling of HPC applications on distributed cloud-oriented data centers	•Deployment models •Cloud Software Stack for SaaS, PaaS, IaaS Level •Features of Clouds enabling Green computing – dynamic provisioning, multi tenancy, server utilization.	•designing software at various levels (OS, compiler, algorithm and application) that facilitates system wide energy efficiency.

III. SCHEDULING OF NETWORKING INFRASTRUCTURE

A. Assumptions

The CSPs regularly fail to adapt to the expanding costs for rendering administrations and to relieve the expanding costs; now and again, they need to violate the service level agreements. Even though many models for overseeing servers and conveying VMs is predominant now-a-days, energy productive server administration model is rare and a significant number of them are not appropriate for useful arrangement. Again, the clear majority of the models don't consider the data transfer capacity cost. Arrange gadgets contribute generally to the vitality cost of server farms. Communication mindful VM planning can lessen the operational energy uses and data transmission necessities. We foresee utilization of resources for the input requests utilizing LPF. Exchanging information between two hubs associated straightforwardly to a similar change tends to utilize less energy than that of the nodes associated indirectly. Our proposed algorithm organizes information exchanges among those hubs which are on a similar switch, in support of energy conservation and postpone minimization.

B. Network Cluster Architecture

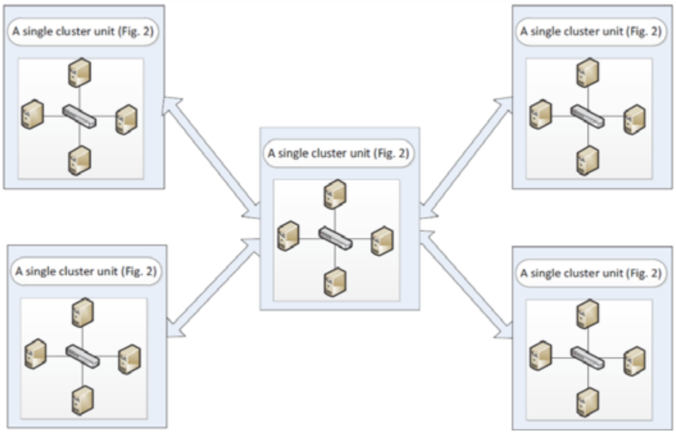
A single cluster unit (Fig. 2)
A single cluster unit (Fig. 2)
A single cluster unit (Fig. 2)
A single cluster unit (Fig. 2)
A single cluster unit (Fig. 2)

Figure 3 CSPs arranged as clusters

All cloud service provides having their own computing, storage and network resources can be considered as a cluster forming a network. The adjacent clusters can share information with each other. It becomes easy in such an environment for VM migration or distributing jobs. Downtime is reduced as migration can be done on-the-fly. The clusters can be in 2 states:

- **Active Mode** – Clusters are functional fully and all the resources and servers are working while serving requests.
- **Sleep Mode** – The cluster is shut down or inactive and its computational and other resources are not working except for the communication server.

An Information Record Data Base (IRDB) exists in each cluster that stores how much resource is required in past assignments and record it. At whatever point new demand shows up at the cluster, IRDB checks for the resource requirement of the demand. It then sends the resource data found for the task to the Resource Predictor (RP). The assignment of RP is to evaluate the sort and amount of resources required to serve the given demand. At that point, the resource administrator checks for the asset utilization at the bunch and decides if it ought to give resources to the demand from the group or take computational assistance from one of its neighbors. If resource necessity of the given demand can be provided by the resource limit of the group, then it asks for the resource allocator to relegate resources to the given task. Resource limit determines the measure of resources in the group that can be provided to the clients for their demand while the resource use of the cluster is inside as far as possible. Resource utilization uncovers the measure of resources utilized as a part of a cluster. At long last, the resource allocator doles out Virtual Machines (VM) to the requests. The resource allocator keeps up a dynamic pool of VMs for provisioning to new demands. The resource necessity for the requests is checked constantly by the segment, resource screen and at whatever point the prerequisite changes, the entire procedure of VM allotment is reused. At whatever point two VMs in two diverse group exchange information consistently and the correspondence interface between the clusters get congested, two VMs are situated in a solitary group, which can serve the two VMs and whose resource use is less. [1]

C. Scheduling Algorithms

The following is a VM provisioning algorithm.

```
I/P: IRDB info, past LPF value
O/P: VM provisioning using resource optimization
1. while request coming are new do
2. Find past resource utilization in IRDB
3. Find LPF and IRDB using equations
4. if available resources are present then
5. if there exists a VM possessing required resources
then
6. assign that VM to the client who requested it
7. else
8. Make a new VM and assign it to the customer
9. end if
```

```
    else
        Send the request to a neighboring cluster
    end if
end while
```

At whatever point the total resource prerequisite of a cluster goes beneath an edge level, it tries to relocate the greater part of its serving requests to a neighboring cluster. The cluster attempting to share its business to its neighbors is known as 'power saver (PS)' cluster and the cluster that gets the share of the tasks of its neighbors is known as 'neighbor server (NS)' cluster. Here a neighboring cluster completes the requests of the PS cluster on the grounds that along these lines the data transfer capacity cost can be diminished and the tasks can be served quicker. Accordingly, the execution proportion increments as the postponement for information exchange diminishes. [1]

$$Ratio\ of\ Performance\ (PR) = \frac{\delta_{ctpr}}{\sigma_{cttr}}$$

Where,

δ_{ctpr}= predicted resources computation time

σ_{cttr} = total resources computation time

The PS cluster chooses a neighboring NS cluster based on its occupancy ratio.

$$Ratio\ of\ Occupancy\ (\mu) = \frac{\emptyset_{ru}}{\varphi_{ar}},$$

Where,

$\emptyset_{ru}$= resource usage

φ_{ar}= allocated resources

PS cluster will pick the NS cluster which has the minimum μ value from its neighboring NS clusters. At whatever point the PS cluster moves all its working VMs to the NS cluster, it can change to rest mode. Thus, enormous measure of energy can be spared and subsequently the cost of calculation likewise diminishes. At whatever point a cluster goes to rest method of operation, if new demand comes to it at that period, it basically advances the demand to the NS cluster. At whatever point the estimation of the NS cluster transcends a predefined upper limit level, it basically stirs the PS cluster to dynamic mode and exchanges the occupations it has gotten for the PS cluster to that PS cluster which is currently working in dynamic mode. A NS cluster can't go to rest mode at whatever point it is serving the requests from some of its neighbors however its utilization level goes beneath the lower limit level.

We will discuss about 2 algorithms for VM migration while considering energy-efficiency and distributed clusters:

- Intra-cluster
- Inter-cluster

1) Intra-cluster VM Scheduling

Algorithm:

INPUT: *Each servers' info about resources*

OUTPUT: efficient *VM scheduling of intra-cluster*

```
for i=number of servers in cluster to 1 do
    Find value of $\omega i$
end for
for i=number of servers in cluster to 1 do
    if $\omega i$ < Lower boundary then
        for j=number of servers in cluster to 1 do
            Find $\omega i + \omega j$ such that it is less than upper boundary and return $\omega i$
        end for
        VMs of server I should be migrated to server j
        sleep(server i)
    end if
end for
```

For accomplishing computational advantages, we have considered every single physical server are of same limit, i.e., every single physical server of same sort has proportionate amount of servable resources. Subsequently, if there should be an occurrence of VM relocation or moving in administration requests starting with one server then onto the next inside a cluster, the prerequisite of resources does not change and the new workload proportion for the servers can be registered by utilizing basic summation and subtraction. [1]

$$Ratio\ of\ workload(\omega) = \frac{\varsigma_{ru}}{\sigma_{ar}} \times 100\%$$

Where,

ς_{ru} = server resource usage

σ_{ar}= server allocable resources

Resource Allocator (RA) controls the Intra-cluster scheduling of a cluster. At whatever point new requests come, RA gives free VMs to the requests. Amid peak hours, resource usage increments and subsequently the workload proportion (ω) additionally increases. In any case, if the servers keep on providing a similar measure of resources amid off peak hours, colossal amounts of energy wastage will occur. Consequently, RA will move VMs to a couple working servers and keep whatever is left of the servers in rest mode amid off peak period to save energy. At starting, RA computes the ω value for every one of the servers inside the cluster. From that point forward, RA will attempt to exchange every one of the requests under process from the servers with ω value lower than a predefined edge, Lower Work Threshold (LWT), to servers having higher ω value and which can suit the entire work. This server, which serves the requests of both servers, ought to have ω value lower than Upper Work Threshold (UWT). At long last, RA will keep the discharged server in rest mode to spare more energy.

2) Inter-cluster VM Scheduling

```
I/P:
$\mu$ : Ratio of Occupancy
OccMatrix :occupancy matrix of neighborhood servers,
Nn :amount of neighbors
O/P: Energy conserving inter cluster job scheduling
1. Find value of $\mu$
2. if $\mu$ < lower boundary then
3. for $i$ = 1 to Nn do
4. From OccMatrix , find minimum value of $\mu$
5. end for
6. if (OccMatrix[j] + $\mu$) < Upper boundary then
7. OccMatrix[j] = OccMatrix[j] + $\mu$
8. sleep(cluster)
9. end if
10. end if
```

Each cluster first figures the OR (occupancy ratio) for itself. It additionally gets OR from its neighbors to top off the Neighborhood Occupancy Matrix (OMat). At any time of operation, at whatever point the OR of a cluster gets down of lower limit (LTH), it checks from its neighbors which have the base OR. It then checks whether the OR of the chose cluster is more prominent than the upper edge (UTH). In the event that it stays beneath UTH, then the PS cluster will relocate all its working VMs and coming requests to the chose NS cluster. At last, it goes to rest mode.

D. Energy Consumption

Based on the ongoing computation of each cluster, the energy model can be made.

$$E_c(t) = \left(1 - \frac{\sum_{i=1}^{m} \omega_i}{m \times 100}\right) \times E_{idle} + (n - m) \times E_{sleep}$$

Where, $E_c(t)$ is a measure of the energy conserved from each cluster at an instance of time t , n is the total quantity of servers in a cluster and m is the quantity of On, i.e. active + idle servers in the cluster.

$\frac{\sum_{i=1}^{m} \omega_i}{m \times 100}$ provides the ratio of average workload in all servers within a cluster.

E_{idle} is energy saved when all servers are idle. Multiplying above two gives energy saved by the section of servers operating in sleep mode.

E_{sleep} is energy saved when servers operate in sleep mode and that multiplied by $(n-m)$ provides energy saved by servers in sleep mode.

$$Total\ Energy\ (E_c) = \int E_c(t)\ dt$$

Based on resource utilization, total energy consumed by whole cloud data center,

$$E_i = \alpha E_{i-1} + (1-\ \alpha)\ \times\ E_{max}\ \times CCR$$

Where, α is the weight variable and E_(i-1) is the energy utilization in the past stage. E_max is the most extreme measure of energy that could be consumed by the entire server data center. CCR is the calculation cost ratio, which is the proportion of the calculation cost of as of now dynamic servers to the calculation cost that is acquired when every one of the servers in the cluster is in dynamic state. The estimation of CCR is taken at the period of energy calculation. Since CCR is a division and estimation of CCR is under 1, Emax × CCR gives the present energy utilization in the server farm. That is Emax × CCR gives a small amount of most extreme measure of energy utilization in a server farm which is utilized for using calculation resources in the server farm at the season of calculation. [1]

$$CCR = \sum_{i=1}^{N} \mu i$$

The estimation of weight variable α can be balanced with the dynamic way of resource prerequisite and energy utilization in the server data centers.

E. Exactly Allocating VMs Algorithm

This is an algorithm similar to the Bin-Packing approach which includes valid conditions having inequalities and constraints. We can pack VMs(items) in a set of servers(bins) according to their power consumptions. Consider q to be the number of VMs that are requested and p to be the number of servers in a data center. $P_{i,max}$ is the power consumption limit in each server. $P_{i,current}$ is the current power consumption in the server. e_j is a variable which is 1 if a server is selected and 0 if it's not. $X_{i,j}$ denotes VM_i is in server j. [7]

Thus, to minimize number of servers to be used,

$$\min Y = \sum_{i=1}^{p} e_i$$

Constraints:

- *$P_{i,max}$ is the power consumption limit in each server*

$$\sum_{j=1}^{q} p_j x_{j,i} \le P_{i,max} e_i -\ P_{i,current}, \forall i = 1, \dots., p$$

- SLAs should be followed by the cloud provider in fulfilling all requests.

$$\sum_{i=1}^{p} x_{j,i} = 1, \forall j = 1, \dots \dots.., q$$

- The lower bound for all servers whose power consumption currently is less than max power consumption and the current is not zero, is

$$\left\lceil \frac{\sum_{i=1}^{p} P_{i,current}}{P_{i,max}} \right\rceil$$

- The resources like CPU, storage and memory can also be represented as an inequality like

$$\sum_{j=1}^{q} res_j x_{j,i} \le RES_i e_i$$

  Where res is resources like CPU, memory and storage.

IV. Conclusion

We talked about reducing energy consumption by scheduling VMs efficiently by thinking of a cloud service provider as a cluster which form a network. Tasks can be transferred to a neighboring cluster if the occupancy matrix is lower than the lower boundary while maximizing the utility of one machine (inter-cluster scheduling). Tasks can also be distributed inside a cluster in its nodes while maximizing the utility of one node and shutting down or sleeping the rest of the nodes. We stated formulas to calculate energy saved by the idle and sleep state servers. We also talked about Exactly allocating VM algorithm which uses an approach similar to the Bin-packing approach where VMs can be allocated to servers while minimizing the number of servers required and maximizing the number of VMs allocated. These algorithms can help reduce the energy consumption costs in data centers. By allocating more VMs, we can serve more customers and thus customer satisfaction is also increased. This also helps in maximizing the utility of one server instead of letting it idle away and thus help in energy efficiency.

REFERENCES

[1] Adhikary, Tamal, et al. "Energy-efficient scheduling algorithms for data center resources in cloud computing." High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), 2013 IEEE 10th International Conference on. IEEE, 2013.

[2] Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." (2011).

[3] Buyya, R., Yeo, C.S. and Venugopal, S. 2008. Market-oriented Cloud computing: Vision, hype, and reality for delivering it services as computing utilities. Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications, Los Alamitos, CA, USA.

[4] Guazzone, Marco, Cosimo Anglano, and Massimo Canonico. "Energy-efficient resource management for cloud computing infrastructures." Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on. IEEE, 2011.

[5] Kumar, Aman, Emmanuel S. Pilli, and R. C. Joshi. "An efficient framework for resource allocation in cloud computing." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.

[6] Yuan, Yuan, and Wen-Cai Liu. "Efficient resource management for cloud computing." System Science, Engineering Design and Manufacturing Informatization (ICSEM), 2011 International Conference on. Vol. 2. IEEE, 2011.

[7] Ghribi, Chaima, Makhlouf Hadji, and Djamal Zeghlache. "Energy efficient vm scheduling for cloud data centers: Exact allocation and migration algorithms." Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on. IEEE, 2013.

[8] Sverdlik, Yevgeniy, and Drake Wauters. "Here'S How Much Energy All US Data Centers Consume | Data Center Knowledge". Data Center Knowledge, 2017, http://www.datacenterknowledge.com/archives/2016/06/27/heres-how-much-energy-all-us-data-centers-consume/.

[9] Baliga, Jayant, et al. "Green cloud computing: Balancing energy in processing, storage, and transport." Proceedings of the IEEE 99.1 (2011): 149-167.

[10] Beloglazov, Anton, and Rajkumar Buyya. "Energy efficient resource management in virtualized cloud data centers." Proceedings of the 2010 10th IEEE/ACM international conference on cluster, cloud and grid computing. IEEE Computer Society, 2010.

[11] Berl, Andreas, et al. "Energy-efficient cloud computing." The computer journal 53.7 (2010): 1045-1051. Garg, Saurabh Kumar, et al. "Environment-conscious scheduling of HPC applications on distributed cloud-oriented data centers." Journal of Parallel and Distributed Computing 71.6 (2011): 732-749.