An energy efficiency of cloud based services using EaaS transcoding of the multimedia data.

Harshal P. Ganvir
Computer Science and Engineering
Vidarbha Institute of Technology
Nagpur, India
harshal.ganvir7@gmail.com
8600581510

Abstract— Network-based cloud computing is rapidly expanding all over as an alternative to conventional office-based computing. Cloud computing has become widespread and the energy consumption of the network and computing resources will grow cloud. This happens at a time when there is increasing attention being paid to the need to manage energy consumption across the entire information and communications technology (ICT) sector. Also data center energy use have much attention, as there has been less attention paid to the energy consumption of the transmission and switching networks. This paper, presents an analysis of energy consumption in cloud computing. The analysis will consider both public and private clouds. We show that energy consumption in transport and switching can be a significant percentage of total energy consumption in cloud computing. Cloud computing provides more energy efficiency and use of computing power. Computing tasks are of low intensity or infrequent. Thus, under some circumstances cloud computing may consume more energy than conventional computing where each user performs all computing on their own personal computer (PC).

Keywords — Cloud Computing, Vdata, Energy Efficiency,Transcoding as a Service, QoS, Lyapunov, EaaS, Data centre.

INTRODUCTION

New network-based services has increasing availability of high-speed Internet and corporate IP connections which are enabling the delivery. While Internet-based mail services have many years operating. Service offerings have expanded recently to include network-based storage and network-based computing. Corporate and individual end users both are offered by these new services. This type of services have been generically called B cloud computing services. Service provider are involved by the cloud computing service model of large pools of high performance computing resources and high-capacity storage devices that are shared among end users as required. Many cloud service models and end users subscribing to the service have their data hosted and have computing resources allocated on demand from the pool. The service provider's offering may also extend to the software applications required by the end user. To be successful, the cloud service model also requires a high-speed network to provide connection between the end user and the service provider's infrastructure. Financial benefits are potentially offered through cloud computing, in that end users share and manage pool of storage and computing resources, instead of owning and managing their own systems. In spite of using existing data centers as a basis, cloud service providers invest in the infrastructure and management systems. Thus in return receive an usage-based fee from end users. The service provider reaps the benefits of the economies of scale and from statistical multiplexing, and receives a regular incoming stream from the investment by means of service subscriptions. The end user in turn sees convenience benefits from having data and services available from any location, from having data backups centrally managed, from the availability of increased capacity when needed, and from usage-based charging. The last point is important for many users in that it averts the need for a large one off investment in hardware, sized to suit maximum demand, and requiring upgrading every few years. There are many definitions of cloud computing, and discussion within the IT industry continues over the possible services that will be offered in the future. The broad scope of cloud computing is succinctly summarized. In this paper, we present an overview of energy consumption in cloud computing and compare this to energy consumption in conventional computing. For this comparison, the energy consumption of conventional computing is the energy consumed when the same task is carried out on a standard consumer personal computer (PC) that is connected to the Internet but does not utilize cloud computing. We consider both public and private clouds and include energy consumption in switching and transmission, also data processing and data storage. Specifically, we present a network-based model of the switching and transmission network, a model of user computing equipment, and a model of the processing and storage functions in data centers. We examine a variety of cloud computing service scenarios in terms of energy efficiency. In essence, our approach is to view cloud computing as an analog of a classical supply chain logistics problem, which considers the energy consumption or cost of processing, storing, and transporting physical items. The difference in our case is that, the items are bits of data. As with classical logistics modeling, our analysis allows a variety of scenarios to be analyzed and optimized according to specified objectives. We explore a number of practical examples in which users/customers outsource their computing and storage needs to a public cloud or private cloud. As the name implies, storage as a service allows users to store data in the cloud. Storage as a service allows users to store data in the cloud. Processing as a service gives users the ability to outsource selected computationally intensive tasks to the cloud. Software as a service combines these two services and allows users to outsource all their computing to the cloud and use only a very-low-processing-power terminal at home. We show that energy consumption in transport and switching can be a significant percentage of total energy consumption in cloud computing. Cloud computing provides more energy-efficient use of computing power, only when the users' predominant computing tasks are of low intensity or arise infrequently. However, we show that under some circumstances cloud computing can consume more energy than conventional computing on a local PC. Our broad conclusion is that cloud computing can offer significant energy savings through techniques such as visualization and consolidation of servers and advanced cooling systems. However, cloud computing is not always the greenest computing technology.

- Overview Of Cloud Computing

Cloud computing is defined as a model enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g .networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction[3].

Over the Internet the applications are delivered as service. The hardware and systems software in the data centers that provide services that is software as a service. Its characteristics includes a broad network access which has the ability to access the network via heterogeneous platforms, when it provision the computing power automatically.

PROBLEM STATEMENT

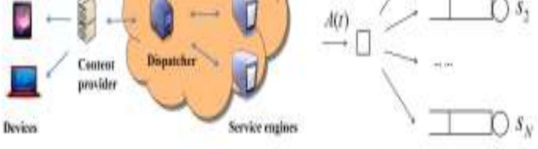
• Adopting the framework of Lyapunov optimization, we propose the control algorithm REQUEST to dispatch transcoding jobs. We characterize the energy-delay tradeoff of the REQUEST algorithm numerically and derive the performance bounds theoretically.

• We study the robustness of the REQUEST algorithm. Numerical results show that, given the inaccuracy of estimating the transcoding time, the error of the time average energy consumption and queue backlog is small. Therefore, the REQUEST algorithm is robust to inaccuracy of the transcoding time estimation.

• We compare the performance of the REQUEST algorithm with Round Robin and Random Rate algorithms using simulation and real trace data. The results show that by appropriately choosing the control variable, the REQUEST algorithm outperforms the other two algorithms, with smaller time average energy consumption while achieving queue stability.

METHODOLOGY

System Model Architecture:-



(a) System architecture. (b) System model.

A. Arrival Model

We consider a discrete time slot model[1]. The length of a time slot is τ . Here τ is small such that there is at most one transcoding job arriving to the dispatcher for each time slot. We denote p as the probability of one arrival to the dispatcher for each time slot and $1-p$ if there are no arrivals. We assume the transcoding time needed for an arriving job at each time slot is associated with the CPU speed of the service engine. Suppose that we have N service engines for transcoding. Each service engine can operate in different CPU speed si , where $i = 1, 2, ..., N$. Without loss of generality, we assume $s1 \leq s2 \leq ... \leq sN$. The service engine in faster CPU speed can have less completion time for transcoding. We denote A(t) as the transcoding time needed for the arrival at time slot t by a baseline server, which has a CPU speed S.W.

B. Queuing Model

We model the service engines as a set of queues, as shown in Figure 1(b). To characterize the dynamics of these queues, we define queue length Q(t) as the unfinished transcoding time of jobs in each service engine at time slot t, i.e., $Q(t) = \{Q1(t), Q2(t), ..., QN (t)\}$. The queue of the ith service engine evolves according to $Qi(t + 1) = \max[Qi(t) - \tau, 0] + Ai(t)1\{u(t)=i\}$, (2) where Ai(t) is the transcoding time of an arrival at time slot t for the ith service engine. And here1 is an indicator function that is 1 if $u(t) = i$ and 0 otherwise. If $u(t) = i$, the arrival is dispatched to the ith service engine and the queue length is increased by Ai(t); or else, no arrival occurs.

C. Energy Consumption Model

We consider each service engine as a physical machine2. Particularly, we only consider the computation energy consumption in the service engine, which is a dominant term for the energy consumption in the distributed servers [18]. As such, we ignore other sources of energy consumption in the service engine, e.g., memory and network. We assume that each service engine operates in a constant CPU speed when processing transcoding jobs. Its resulted energy consumption is assumed to be a function of CPU speed. If the dispatcher dispatches the transcoding job to the ith service engine at time slot t, the energy consumption on the ith service engine is $Ai(t)\kappa s\alpha$ i , where Ai(t) is the transcoding time for the ith service engine and $\kappa s\alpha$ i is the power that is a convex function of CPU speed[1].

- MODELS OF ENERGY CONSUMPTION

In this section, we describe the functionality and energy consumption of the transport and computing equipment on which current cloud computing services typically operate. We consider energy consumption models of the transport network, the data center, plus a range of customer-owned terminals and computers. The models described are based on power consumption measurements and published specifications of representative equipment. Those models include descriptions of the common energy-saving techniques employed by cloud computing service providers. The models are used to calculate the energy consumption per bit for transport and processing, and the power consumption per bit for storage. The energy per bit and power per bit are fundamental measures of energy consumption, and the energy efficiency of cloud computing is the energy consumed per bit of data processed through cloud computing. Performing calculations in terms of energy per bit also allows the results to be easily scaled to any usage level.

a) User Equipment.

A user may use a range of devices to access a cloud computing service, including a mobile phone (cell phone), desktop computer, or a laptop computer. In this paper, we focus on desktop computers and laptops. These computers typically comprise a central processing unit (CPU), random access memory (RAM), hard disk drive (HDD), graphical processing unit (GPU), motherboard, and a power supply unit. Peripheral devices including speakers, printers, and visual display devices are often connected to PCs. These peripheral devices do not influence the comparison between conventional computing and cloud computing and so are not included in the model. In our analysis, we assume that when user equipment is not being used it is either switched off or in a deep sleep state (negligible power consumption).

b) Data Centers.

A modern state-of-the-art data center has three main components Vdata storage, servers, and a local area network (LAN). The data center connects to the rest of the network through a gateway router, as shown on the right-hand side of lists equipment typical of that used in data centers, as well as the capacity and power consumption of this equipment. Power consumption figures for the LAN switches, routers, and storage equipment are the figures quoted in their respective product data sheets. The power consumption data for each server was obtained by first calculating the maximum power using HP's power calculator, then following the convention that average power use for midrange/high-end servers is 66% of maximum power. In the following, we outline the functionality of this equipment as well as some of the efficiency improvements in cloud computing data centers over traditional data centers.

c) Network.

In this section, we describe the corporate and Internet IP networks in greater detail and outline the functionality of the equipment in those networks. Lists equipment used in our calculations of energy consumption in the corporate network and the Internet IP network as well as the capacity and power consumption of this equipment.

- ENERGY EFFICIENCY IN CLOUD INFRASTRUCTURES

Building an energy efficient cloud model does not indicate only energy efficient host machines. Other existing components of a complete cloud infrastructure should also be considered for energy aware applications. Several research works are carried out to build energy efficient cloud components individually. In this section we will investigate the areas of a typical cloud setup that are responsible for considerable amount of power dissipation and we will consolidate the possible approaches to fix the issues considering energy consumption as a part of the cost functions to be applied.

1) Energy Efficient Hardware

One of the best approaches to reduce the power consumption at data centre and virtual machine level is usage of energy efficient hardware's at host side. International standard bodies such as: European TCO Certification, US Energy Star are there to rate energy efficient consumer products. The rating is essential to measure the environmental impact and carbon footprint of computer products and peripherals. New electronics materials like solid-state drives are more power efficient than common hard disk drives but that are costly. The Intel's wireless technology to adjust CPU power dynamically based upon the performance demand. It works in five usage modes: voice communication, standby mode, multimedia, data communication.

2) Energy Efficient Clusters of Servers

Power dissipation is primarily reduced by optimal CPU utilization and tasks scheduling. However other cluster components such as memory, storage discs, network peripherals etc. also consume power and hence a VM having idle CPU may still use considerable amount of energy. Figure shows a typical cloud cluster structure. New approaches aim to reduce the energy consumption as a whole at clusters of servers while considering system's latency and throughput.

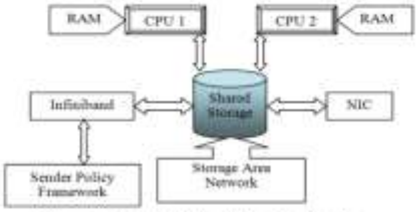


Figure 2. A Typical Cloud Cluster Structure

a) Resource management architecture:

At server side, the operating system of the machine hosts several separate instances of virtual machines. The Optimal resource management architecture should be built based upon the energy estimation in different nodes of a server cluster. Depending upon external situation and workload, the cluster can be easily affected by overloading and overheating despite sufficient cooling system. While a complete shutdown of cluster causes unwanted business downtime, it is the operating system that should take care of auto-scaling of power demand from different cluster components. Dynamic Thermal Management is a technique that controls power dissipation in high performance, server processing unit and provides low “worst case power consumption” with no or little impact on performance. Cluster’s network infrastructure is a major area of power dissipation that holds a substantial share of operating cost. Balancing of QoS and resource utilization during outage can also be a descent way of energy management in clusters. Policies are developed for resource management in economic way where cluster always checks for system’s work load and allocates resources by calculating their effects on system’s overall performance. A greedy allotment method can be used to estimate supply – demand tradeoffs for an efficient allocation of resources.

b) Dynamic Server Provisioning and load dispatching**:**

In order to save energy, Dynamic Server Provisioning method is useful in switching off unnecessary and idle hosts in a cluster. As the number of internet services is increasing rapidly, the servers are also increasing in number to host those services, resulting huge amount of power dissipation in form of heat. Dynamic Server Provisioning algorithms are designed to turn off extra servers and to allow the cluster to run on minimal number of host machines to satisfy the service load. Thereafter, load dispatching technique effectively distributes the current load among available servers. These techniques can be implemented in servers operating request-response type services (example: web services) as well as host machines connected to huge number of long lived TCP connections. For multi-tier internet services, queuing method can be implemented in dynamic provisioning technique that can be defined as a proactive and reactive approach to estimate short-term and long-term workload fluctuations in cluster. It can predict minimum capacity required for maintaining the required QoS and can also balance sudden surges in server load.

3) Energy efficient Network Infrastructure in cloud:

Minimizing energy consumption in various elements of cloud computing such as storage and computation has already been given importance by the researchers but the issue of energy minimization in network infrastructure is not given as much importance. Network in a cloud environment can be of two types - wireless network and wired network. According to ICT energy estimates in the radio access network consumes a major part of the total energy in an infrastructure and the cost incurred on energy consumption is sometimes comparable with the total cost spent on personnel employed for network operations and maintenance, provided a thorough study on routing protocols for saving energy consumption in sensor networks and wireless adhoc networks.

ACKNOWLEDGMENT

There are many persons in Vidarbha Institute Of Technology Nagpur College have supported me from the beginning of my Mtech project. Without them, the project work would obviously not have looked the way it does now. The first person I would like to thank is guide Prof. Pravin G. Kulurkar, Department of Computer Science and Engineering. His enthusiastic engagement in my project work never ending stream of ideas has been absolutely essential for the results, presented here. I am very grateful that he has spent so much time with me discussing different problems ranging from philosophical issues down to minute technical details.

CONCLUSION

In this paper we have investigated the need of power consumption and energy efficiency in cloud computing model.This work advances Cloud computing field in two ways. First, it plays a significant role in the reduction of data center energy consumption costs and thus helps to develop strong competitive. Cloud computing is facing an increasing attention nowadays, but it raises severe issues with energy consumption. The energy savings from cloud storage are minimal. In cloud software services, power consumption in transport is negligibly small at very low screen refresh rates.

REFERENCES:

[1] Weiwen Zhang, Yonggang Wen,Member 2013 IEEE, "Towards Transcoding as a service in Multimedia Cloud: Energy-Efficient Job Dispatching Algorithm."

[2] A Pliant-based Virtual Machine Scheduling Solution to Improve the Energy E_ciency of IaaS Clouds A. Kertesz _ J. D. Dombi _ A. Benyi.

[3] A Survey on Resource Allocation and Monitoring in Cloud Computing, vol 4, no 1, feb2014, Mohd Hairy Mohamaddiah, Azizol Abdullah, Shamala Subramaniam, and Masnida Hussin.

[4] Cisco Visual Networking Index: Forecast and Methodology, 2012–2017,2013.

[5] A. Vetro and C. W. Chen, "Rate-reduction transcoding design for wireless video streaming," inwork *Proc. Int. Conf. Image Process.*, 2002, vol. 1, pp. I-29–I-32.

[6] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010.

[7] A. Garcia, H. Kalva, and B. Furht, "A study of transcoding on cloud environments for video content delivery," in *Proc. ACM Multim. Worksh. Mobile Cloud Media Comput.*, 2010, pp. 13–18.

[8] Z. Li, Y. Huang, G. Liu, F. Wang, Z.-L. Zhang, and Y. Dai, "Cloud transcoder: Bridging the format and resolution gap between Internet videos and mobile devices," in *Proc. 22nd Int. Workshop Netw. Oper. Syst. Support Digit. Audio Video*, 2012, pp. 33–38.

[9] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures Commun. Netw.*, vol. 3, no. 1, pp. 1–211, 2010.

[10] J. Guo and L. N. Bhuyan, "Load balancing in a cluster-based web server for multimedia applications," *IEEE Trans. Parallel Distrib. Syst.*, vol. 17,no. 11, pp. 1321–1334, Nov. 2006.

[11] D. Seo, J. Kim, and I. Jung, "Load distribution algorithm based on transcoding time estimation for distributed transcoding servers," in *Proc.ICISA*, 2010, pp. 1–8.

[12] A. Garcia and H. Kalva, "Cloud transcoding for mobile video content. delivery," in *Proc. IEEE ICCE*, 2011, pp. 379–380.

[13] S. Ko, S. Park, and H. Han, "Design analysis for real-time video transcoding on cloud systems," in *Proc. 28th Annu. ACM Symp. Appl. Comput.*2013, pp. 1610–1615.

[14] Amazon Elastic Compute Cloud (EC2). [Online]. Available:http://www.amazon.com/ec2/

[15] Google App Engine. [Online]. Available: http://www.appengine.google.com