Hardware Loads and Power Consumption in Cloud Computing Environments

Razvan I. Dinita, *Member, IEEE,* George Wilson, Adrian Winckles, Marcian Cirstea, *Senior Member, IEEE*, Aled Jones

Anglia Ruskin University

Cambridge, UK

razvan.dinita@anglia.ac.uk, george.wilson@anglia.ac.uk, adrian.winckles@anglia.ac.uk, marcian.cirstea@anglia.ac.uk, aled.jones@anglia.ac.uk

Abstract-**This paper describes an optimised and novel approach to an Autonomous Virtual Server Management System in a 'Cloud Computing' environment and it presents a set of preliminary test results. One key advantage of this system is its ability to improve hardware power consumption through autonomously moving virtual servers around a network to balance out hardware loads. This has a potentially important impact on issues of sustainability with respect to both energy efficiency and economic viability. Another key advantage is the improvement of the overall end-user experience for services within the Cloud. This has been investigated through the configuration of a cloud-computing test-bed rig. The key features of this rig and some predictions of what may be achieved with it are described and evaluated.**

I. INTRODUCTION

Cloud computing is an emerging technology that devolves computing resources to the Internet [1]. This paper focuses on presenting a conceptually innovative approach to the resilience and optimization of Internet/network usage via mobility of virtualized servers within a cloud. The optimised running of a cloud has been investigated through the use of a virtual networking laboratory, developing new metrics that can be tested to quantify the energy efficiency (and therefore carbon footprint) of an optimised cloud network compared to a non-optimised cloud network and to a non-cloud infrastructure [2-3].

Although the protocols that enable current network topologies to interface in such a way as to support Cloud functionality are well established the effect of an unanticipated amount of people trying to access the same file/service is not yet well understood [4]. The authors of one study break down the energy consumption in a data centre and suggest different ways of reducing it overall [5]. Whilst management tools are available to enable Cloud Administrators to re-configure hardware loadings according to service demands, the approach is largely by-trial-and-error [6].

One consequence of the adoption of Cloud Computing in the commercial sector is that companies do not need to invest in their own hardware infrastructure. This has environmental consequences and their quantification is important for developing strategies for a sustainable environment [7]. Security is also a relevant issue. A botnet is a group of compromised computers connected to the Internet. Each compromised computer is called a bot, and could include individual virtual bots within a virtualised system. Whilst there are tools to protect a Cloud from such malware attacks, a less widely researched area is that of methods of identification of a successful hijacking of a Cloud's virtual operating system [8-9-10]. These and other relevant works [11-12] also discuss the hypervisor architecture and the possible vulnerabilities within them.

II. RESEARCH FOCUS

This research focuses on the optimization and security of Internet/network usage via mobility of virtualized servers within a cloud. A cloud-resident application-specific prototype utility such as a virtual networking laboratory has been developed. Potential applications of such a virtual laboratory will allow suitable metrics to be proposed and tested that can quantify the reduction in the carbon footprint of the IT sector compared to a non-cloud infrastructure.

Objective 1: Critically evaluate the pattern of disruption across a Cloud infrastructure as a result of an overloaded service request. The effect on the network infrastructure of a Cloud (not just the service itself) that is overloaded by a service request is of particular interest here.

Objective 2: Conceptually design a theoretical strategy by which a Cloud could autonomously manage the workloads placed on that infrastructure, and implement and test a software application to achieve this aim for a specific Cloud scenario. A theoretical strategy has been developed by which a virtualised operating system can autonomously optimize the location of virtualized servers within a network. A simple software application to establish this proof-of-principle is currently being built.

Objective 3: Innovatively develop metrics that quantify Cloud vs. centralized service provision in terms of environmental sustainability. Measures of sustainability (e.g. carbon-footprint) are currently being developed and effectively applied to Cloud vs. non-Cloud scenarios.

Objective 4: Develop an application that will identify virtualized system hijacking and undertake a range of appropriate activities from simple notification to service

suspension. Parameters indicative of the successful hijacking of a Cloud will be identified. A software tool is currently being developed to monitor the activities of a Cloud in such a way as to be able to identify the successful hijacking of the virtualized servers, and to mitigate the presence of a compromised Cloud.

III.   METHODOLOGY

There is very little current literature in the public domain as to how load balancing is achieved in data centres. So far, none of the most recent papers have explored the connection between processor loads, power consumption and VM migration. The authors of this work propose an original methodology to deal with this issue.

A small test bed has been constructed comprising seven physical servers (mix of HP Proliant, Dell R710 and Viglen brands), three Storage Area Networks (HP) and multiple routers and switches (CISCO and HP), and HP ILOs located within the Department of Computing and Technology[1] (Fig. 1 and Fig. 2). Two of the servers are 'public'-facing (that is, they are utilised for teaching purposes) whilst the others are 'private'-facing (i.e. they are utilised for research). The private-facing resource is isolated from the Internet and only accessible from departmental computers via an appropriate security protocol. All of the networking cards have Gigabit type interfaces (>=1Gbps, Gigabit per second, transfer rate).
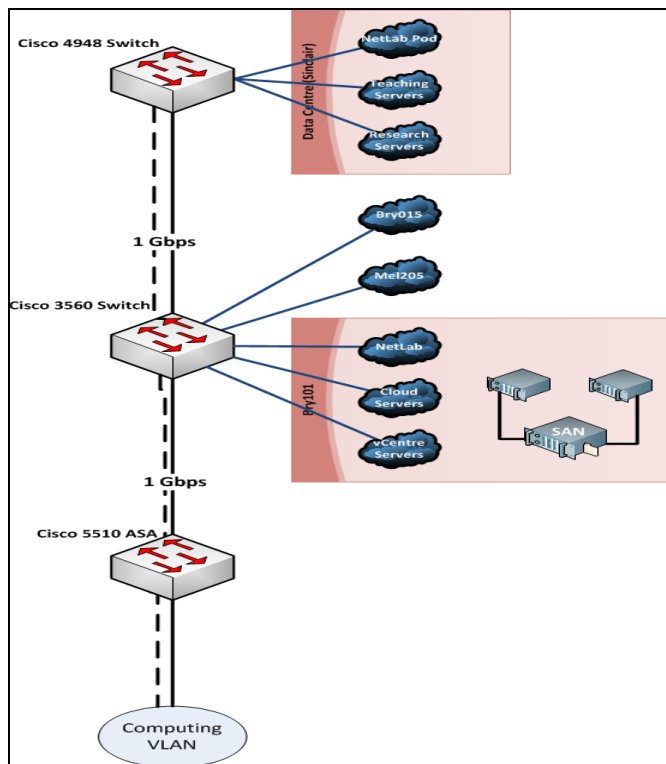

Figure 1 - Cloud Computing Test Bed – custom designed and built infrastructure.

In the context of the present research the development of this resource has focussed on the following:

[1] Anglia Ruskin University, East Rd, Cambridge CB1 1PT, UK

- Expansion of the private-facing hardware (addition of more PCs – some are already available).
- Development of new software modules to manage disparate hardware and operating systems across a network.
- Building of cloud optimization strategies via a top-down approach using a proprietary software virtualization product (VMWARE) [13].
- Building of cloud optimization techniques via a bottom-up approach using an appropriate mainstream programming language (Scala due to its library support tools for building bespoke network applications).
- Development of a software tool to detect and manage malware attacks on the virtualized servers.

An appraisal of the extent to which the given objectives have been met will now be discussed.

Objective 1: Has been achieved through experimental work using the test bed. Multiple virtual Machines (VMs) have been created on all four servers. Simultaneous requests have been made to all VMs to simulate a memory overload. Both hardware and network loads have been monitored (using a wide range of hardware appliances such as ILOs, PDUs and UPSes), logged (using Microsoft Excel spreadsheets) and analysed. Relevant methodology presented by the authors of paper [5] has also been taken into account which specifically describes issues related to increasing hardware loads on a few servers while at the same time shutting down unneeded ones to save power.

The authors have access to several test hardware that allow simple power consumption readings to be taken through proprietary User Interfaces: HP Integrated Lights-Out (ILO), APC Power Distribution Unit (PDU) and APC Uninterrupted Power Supply (UPS) devices. These devices are directly connected to the physical servers and as such produce relevant power metrics. The power consumption readings will be correlated with processor utilisation readings taken from within the VMWare vSphere Client. Tables will be produced detailing the connection between processor utilisation and power consumption.

Another way of gathering power consumption readings is through the SNMP protocol. All hardware presented in this paper is fully network enabled and as such have been provided with unique IP addresses to which connections can be made. The specialised hardware relevant to this point all support the SNMP protocol (communication protocol that allows two way data connections). A software module is currently being built using the Scala programming language capable of taking advantage of this feature. It is capable of retrieving real time power readings through the SNMP protocol and store them in a database. All collected data is analysed and the results are ultimately be presented in tables. The module also allows sorting and comparing the different results gathered up to and including the most recent ones. This allows a history of readings to be generated relevant to future research.

Objective 2: Achieved through socket programming using the Scala programming language. Sockets locally bind to specific port numbers, which allows direct communication to

other (virtual) machines over a (virtual) network. The resulting software application is able to communicate with each server included in the test bed. It is also be capable of directly communicating with other instances of itself and relaying information across the (virtual) network.

The software application has a modular design. As such, each task it is capable of performing represents one module. This design allows easy expansion of its capabilities to allow it to perform more complex operations within a cloud network. It also allows easy maintenance of each module and of the software application as a whole.

The software application will also be capable of autonomously making decisions based on current hardware and network loads to move VMs around from one physical location to another within the same virtual network. The main goal of the application is to reduce the overall power consumption of the cluster of computers it is directly overseeing.

Another approach to objective 2 is by looking at currently available commercial software such as the VMWare vMotion technology to achieve the same end result - either by using a pure VMWare solution or by creating a hybrid solution comprised of VMWare technologies and a set of custom built technologies.
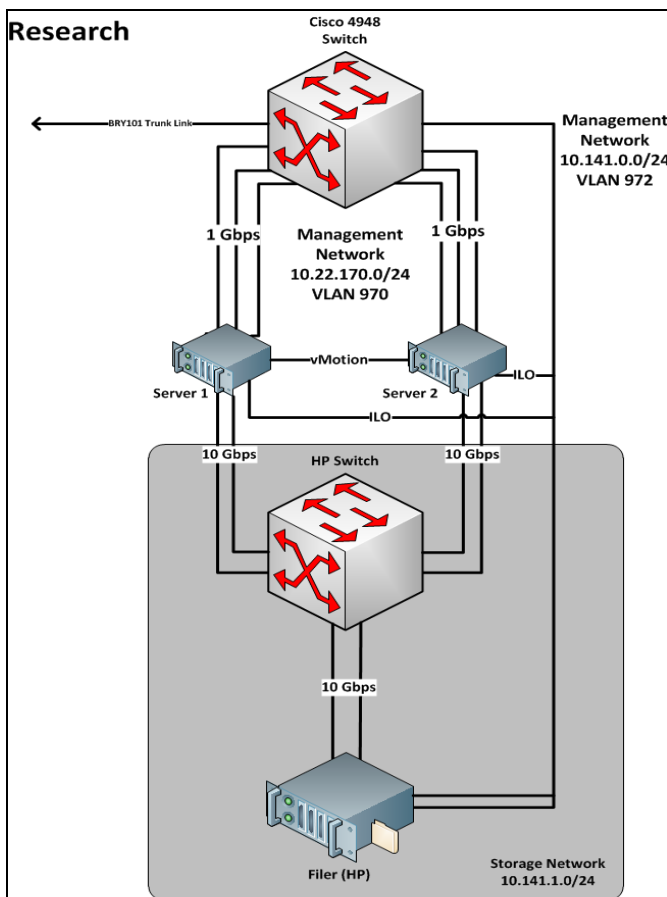
Moving a VM across a network to a different physical location raises the following difficulties concerns that are currently being addressed:

1) Maintaining and Updating the IP address tables across the network so that loss of service is avoided.

2) Ensuring the MAC addresses of the VMs are unique at the new physical location, refreshing them if clashes occur.

3) Refreshing the VM's Domain Name so that it matches the one at the new physical location.

Objective 3: Achieved by measuring server and network hardware power consumption and logging the results. The analyses of these results form the basis for developing the required metrics. An attempt is also being made towards identifying what levels of efficiency savings can be achieved through optimization of the cloud network.

Objective 4: To be achieved through testing of different approaches to service hijacking as presented here [14]. Some previous work on overcoming relevant security issues has been done [15]. That work will be investigated and, if relevant, a solution based on Public Key Infrastructure will be built. Otherwise, alternative approaches to the aforementioned security issues will be researched and developed.
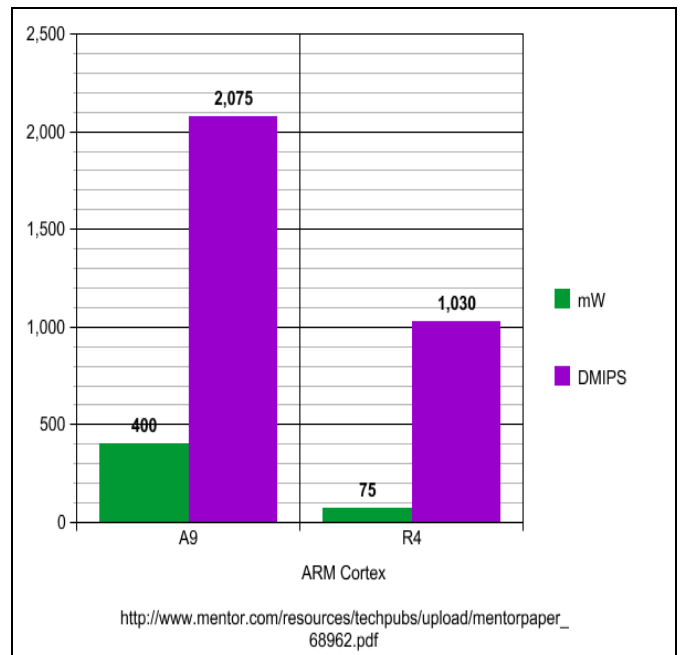

Figure 2 - Cloud Computing Test Bed – Datacentre view.

Figure 3 - Relation between miliWatts (mW) and the Millions of Instructions/Second (DMIPS). Data taken from paper [16]

IV.   TEST RESULTS

A.   Practical View

Power consumption is directly influenced by a great deal of factors [16], and [17]. The factors that this paper will focus on are the Processor, RAM and Hard Disk. These are most relevant in the context of cloud computing because the Processor does calculations, stores the results in the RAM and if applicable commits them to the Hard Disk. Network power consumption is also relevant but since it is necessary and

cannot be efficiently controlled its' effects have not been included in this work.

TABLE I. WATT SERVER POWER READINGS VS PROCESSOR LOADS

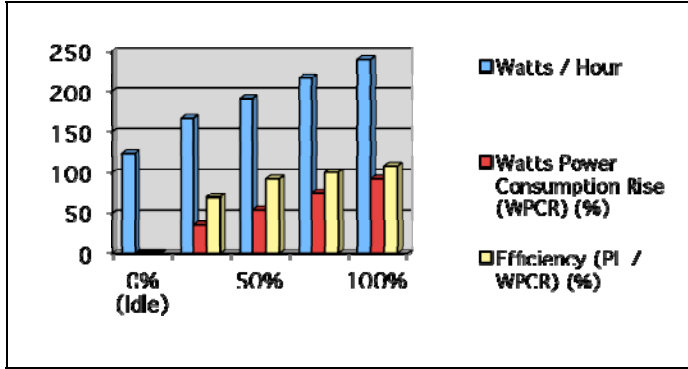| Processor Load (PL)	0% (Idle)	25%	50%	75%	100%
Watts / Hour	124	168	191	217	239
Watts Power Consumption Rise (WPCR) (%)	0	35.5	54	75	92
Efficiency (PL / WPCR) (%)	0	70	92	100	108

Figure 4 - Watt Server Power Readings vs. Processor Loads

TABLE II. AMPS SERVER POWER READINGS VS PROCESSOR LOADS

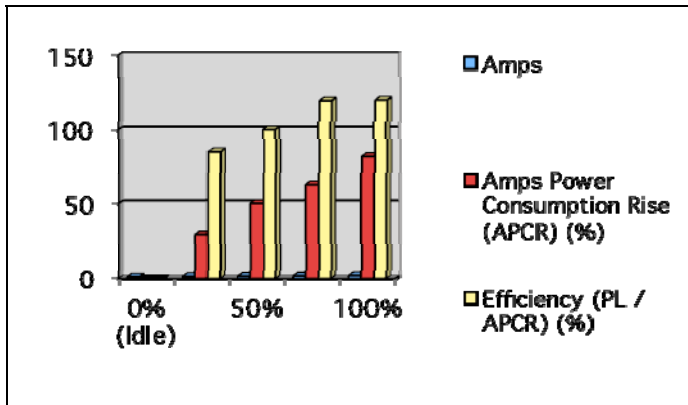| Processor Load (PL)	0% (Idle)	25%	50%	75%	100%
Amps	1.2	1.55	1.8	1.95	2.2
Amps Power Consumption Rise (APCR) (%)	0	29.1	50	62.5	83
Efficiency (PL / APCR) (%)	0	86	100	120	120.4

Figure 5 - Amps Server Power Readings vs. Processor Loads (PL)

According to [17] the Processor ("*CPU Quadcore*") and the RAM ("*Memory*") have the highest power consumption of all the hardware types in an actively used server system. There is also a strong relation between hardware loads and power consumption according to papers [16] and [17]. The less hardware load on a system, the more efficiently the server operates and the less power it consumes. Furthermore, paper [18] discusses the importance of CPU loads and takes the approach of making the distinction between the server's own CPU load and application generated CPU load. Thus, the approach this work takes helps reduce the hardware loads by attempting to move 'hot' data (data actively used at any given time) from a highly active server (many user service requests) to a less active server (few user service requests). This approach aims to balance out the hardware loads across the same virtual network and at the same reduce the overall power required to achieve the same results as before the move.

A second approach currently being considered is moving data from less active servers to highly active servers, maximising processor loads. This approach aims to reduce the number of active servers and as such cut down the overall power consumption of idle to moderately used servers.

A similar technique of energy efficient data transfer across a set of network nodes has been investigated by the authors of paper [19]. It describes a set of energy efficient rules that comprise the EARQ protocol which are being followed whenever the control application needs to decide which node to transmit the data next. These rules have been abstracted and considered for Objective 2's application logic.

B. Hardware Stress Tests

In order achieve Objective 1 the authors have put together a special Linux Virtual Machine template capable of performing server load stress tests. This is achieved by having the VM push the Virtual Processor to loads of up to 100%.

The VM comprises of the latest Ubuntu Linux operating system as well as the latest versions of the Apache Web Server and the Tsung open source multi-protocol distributed load testing tool.

Apache is a web service that upon start it listens on a predefined port (usually 80) on the server. It is capable of delivering multiple web pages to millions of clients simultaneously. A simple web page has been created and put in place for the purpose of these tests.

'Tsung' is a complex application that is capable of creating millions of simultaneous connections (also known as virtual clients) to any given web service. The configuration file allows fine control over the length of each connection time wise as well as how fast the number of simultaneous connections grows over a predefined time frame. This allows measurement of server power consumption while the processor load ranges from 0% to 100%. Other features of Tsung are beyond the scope of this research.

Upon launch the VM starts up both Apache and Tsung. Apache is set up to listen on port 80 for connections from anywhere on the network (in this case, from within the same location – localhost or 127.0.0.1). Tsung is set up to create virtual clients for Apache every 0.5 milliseconds over a time frame of 15 minutes. By the end of the given time frame the

TABLE III. POWER CONSUMPTION ESTIMATES WITH 100% UPTIME

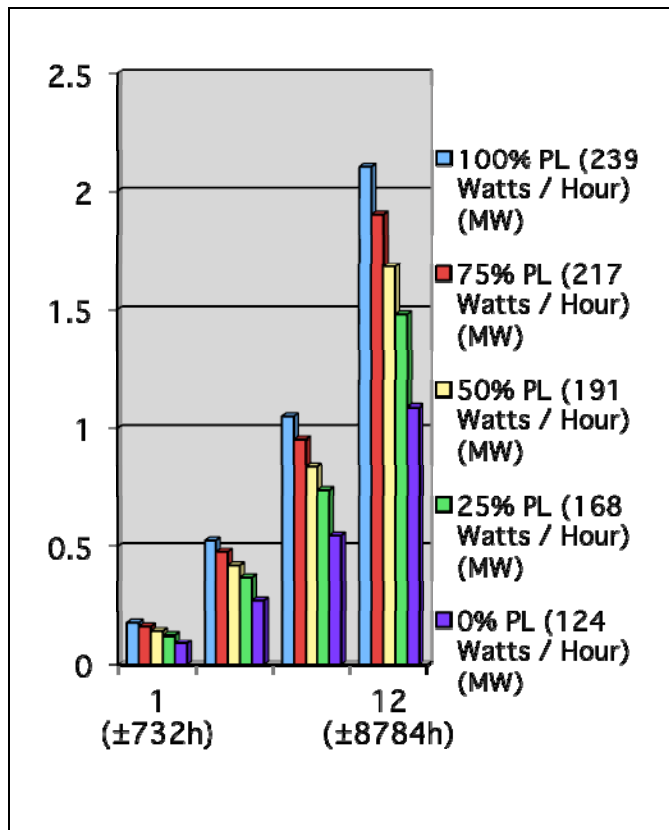| Period of Time (Months)	1 (±732h)	3 (±2196h)	6 (±4392h)	12 (±8784h)
100% PL (239 Watts / Hour) (MW)	0.175	0.524	1.05	2.1
75% PL (217 Watts / Hour) (MW)	0.159	0.477	0.953	1.9
50% PL (191 Watts / Hour) (MW)	0.14	0.42	0.839	1.68
25% PL (168 Watts / Hour) (MW)	0.123	0.369	0.738	1.48
0% PL (124 Watts / Hour) (MW)	0.91	0.272	0.545	1.09

Figure 6 - Power Consumption estimates (100% Uptime) vs. Processor Loads (PL)

number of simultaneous connections will mount up to 1.8 million. This forces the Virtual Processor to slowly go through the needed loads.

For the purpose of these tests the authors have deployed 40 VMs based on the original template VM and have launched them all at the same time to achieve the desired effect.

C. Results

The readings gathered from both the VMWare vSphere client and the relevant hardware devices have been put side by side in the Tables I and II and graphs have been generated based on them in Figures 4 and 5. Estimates of power consumption over longer periods of time have been generated in Table III with associated data graphed in Figure 6. The results show that when

When taking Processor Load percentages are compared with Watt Consumption Rise percentages the higher the Processor Load gets, the more efficient the Watt Consumption becomes. At 100% Processor Load there is only a 92% increase in Watt Consumption. The same applies for Processor Load percentages and Amps Consumption percentages. For 100% Processor Load there is only an 83% increase in Amps Consumption.

For optimal efficiency it appears to be desirable that the Processor Load be kept over 75% for the best Power Consumption efficiency.

From the perspective of Data Centres it appears to be more cost effective to maximise the Processor Load on each server and running less servers rather than having a higher number of servers but with lower Processor Load percentages. Furthermore, using the results presented in Table III the cost of operation from the energy efficiency point of view can be calculated. This was for the following Data Centre use case scenarios; Case 1 (UC1): one server running at 100% PL (based on new findings) versus Case 2 (UC2): four servers running at 25% PL each (original theory). Both use cases assume an operation time span of one year and the price per Kwh used in the calculations is 111.1 Euros / MW [2] (current at the time of writing). The calculated costs are presented in Table IV.

TABLE IV. COST ESTIMATES FOR TWO USE CASE SCENARIOS

Use Case 1 (UC1) One Server @ 100% PL	2.1 MW * 111.1 Euros / MW = 233.3 Euros
Use Case 2 (UC2) Four Servers @ 25% PL	One Server: 1.9 MW * 111.1 Euros / MW = 211.1 Euros Four Servers: 211.1 Euros * 4 = 844.4 Euros
| UC1 – UC2 | (Euros)	611.1 Euros UC2 shows a 261.9% cost increase compared to UC1

Table IV shows that UC2 exhibits a 261.9% operational cost increase compared to UC1. This is consistent with the results presented in Table III, which show a substantial economic benefit in Data Centre operation costs.

---

[2] Price taken from http://www.businesselectricityprices.com/kwh.php on 27th November 2012. Original price was shown in British Pence / KWh. http://www.xe.com/ucc/ was used to convert the price to Euro Cents / KWh. Result was then multiplied by 10 to get to Euros / MWh.

V. Conclusions

This paper has described novel development based on original concepts and ideas which optimise and improve the resilience of cloud computing environments. The authors have discussed how cloud computing networks can be improved to help reduce the carbon footprint of the associated hardware through implementing an autonomous solution to help manage the hosted virtual machines. The authors have also provided clear test results related to Power Consumption reduction at different Processor Load percentages and provided solutions, as well as calculated potential operation cost differences between two use cases.

According to Fig. 3, processors consume four times as much power running at full speed (GHz) as running at half speed. Thus, another approach might be to attempt to reduce the Processor clock speed in addition to moving virtual Machines across the network. This however would depend on whether the VMWare Hypervisor (operating system running on the servers) gives direct access to this type of setting. Another issue is that reducing the processor clock speed in order to reduce the power consumption by a factor of at least 3x is only practical in an environment where speed is not mission-critical.

Some future work can be explored on the causes of and whether solutions can be implemented to ensure the resilient operation of a cloud network.

Following to the results presented in this paper the authors have already started development of the aforementioned software application.

References

[1] Mirashe, S. P., and Kalyankar, N. V. (2010), "Cloud Computing," [*Communications of the ACM*], *51*(7), 9.

[2] Anderson, J. Q. (2010), "The future of cloud computing," Schubert, L., Jeffery, K. and Neidecker-Lutz, B. (Eds.) *Analysis*, *1*, 1-26, European Commission.

[3] Armbrust, M., Fox, A., Griffith, R., and Joseph, A. D. (2009), "Above the Clouds : A Berkeley View of Cloud Computing," *Science*, *53* (UCB/EECS-2009-28), 07-013, Citeseer.

[4] Williams, D., Weatherspoon, H., Jamjoom, H. and Liu, Y. (2011), "Overdriver: Handling memory Overload in an Oversubscribed Cloud," [*VEE '11 Proceedings of the 7th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments*].

[5] Berl, A., Gelenbe, E., Di Girolamo, M, Giuliani, G., De Meer, H., Quan Dang, M., Pentikousis, K. (2009), "Energy-Efficient Cloud Computing", The Computer Journal (2010) 53(7): 1045-1051 first published online August 19 2009, doi: 10.1093/comjnl/bxp080, Available at: http://comjnl.oxfordjournals.org/content/53/7/1045.full.pdf+html

[6] Moretti, C., Bulosan, J., Thain, D., Flynn, P.J. (2008), "All-pairs: An abstraction for data-intensive cloud computing," *Parallel and Distributed Processing, 2008*, IPDPS 2008. [*IEEE International Symposium*, vol., no., pp.1-11, 14-18 April 2008].

[7] Chou, D. C. and Chou, A. Y. (2011), "Seeking Sustainable Computing: The role of Cloud Computing," *Southwest Decision Sciences Institute Conference*.

[8] Chandrashekar, J. (2009), "The Dark Cloud: Understanding and Defending Against Botnets and Stealthy Malware," *Intel RTechnology Journal*, 13(2).

[9] Ke, A., Yu, Y., Chen, Y., Zhao, E., Xie, Y., Yu, F. and Gillum, Q. (2009), "BotGraph: large scale spamming botnet detection," [*Proceedings of the 6th USENIX symposium on Networked systems design and implementation* (NSDI'09)], USENIX Association, Berkeley, CA, USA, p321-334.

[10] Zeidanloo, H., Shooshtari, M., Amoli, P., Safari, M., and Zamani, M. (2010), "A taxonomy of Botnet detection techniques," Computer Science and Information Technology (ICCSIT), [*2010 3rd IEEE International Conference* v2, p158–162. IEEE].

[11] Murakami, J. (2008), "A hypervisor IPS based on hardware assisted virtualization technology," [*Black Hat USA 2008*].

[12] Rutkowska, J. and Tereshkin, A. (2008), "Bluepilling the Xen Hypervisor," [*Black Hat USA 2008*].

[13] Tsutomu, N., Yoshihiro, O., Hideki, E., Takahiro, S. and Kazuhiko, K. (2010), "Using a Hypervisor to Migrate Running Operating Systems to Secure Virtual Machines," [*2010 IEEE 34th Annual Computer Software and Applications Conference (Proceedings)* p37-46].

[14] Cloud Security Alliance (2010), "Top Threats to Cloud Computing V1.0," Available at: https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf

[15] Arora, P., Wadhawan, R. C., Ahuja, Er. S. P. (2012), "Cloud Computing Security Issues in Infrastructure as a Service," [*International Journal of Advanced Research in Computer Science and Software Engineering*] Vol. 2, Iss. 1, Jan. 2012, ISSN: 2277 128X.

[16] Minas, L., Ellison, B. (2009), "The Problem of Power Consumption in Servers," [*Dr. Dobb's Journal*, May 2009].

[17] Matalon, S., Klein, R., Walls, C. (2011), "Embedded System Power Consumption: A Software or a Hardware Issue?," Mentor Graphics, Available at: http://www.mentor.com/resources/techpubs/upload/mentorpaper_68962.pdf

[18] Warkozek, G., Drayer, E., Debusschere, V., Bacha, S. (2012), "A new approach to model energy consumption of servers in data centers", 2012 IEEE International Conference on Industrial Technology (ICIT), pp.211-216, 19-21 March 2012, doi: 10.1109/ICIT.2012.6209940, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6209940&isnumber=6209899

[19] Junyoung, H., Jiman, H., Yookun, C. (2009), "EARQ: Energy Aware Routing for Real-Time and Reliable Communication in Wireless Industrial Sensor Networks", 2009 IEEE Transactions on Industrial Informatics, vol. 5, no. 1, pp. 3-11, Feb. 2009, doi: 10.1109/TII.2008.2011052, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4799339&isnumber=4799331