Proceedings Paper:

Mohammad Ali, HM, Lawey, AQ, Elgorashi, TEH et al. (1 more author) (2015) Energy Efficient Disaggregated Servers for Future Data Centers. In: Networks and Optical Communications (NOC) 2015. The 20th European Conference on Network and Optical Communications, NOC 2015, 30 Jun - 02 Jul 2015, University College London. IEEE . ISBN 978-1-4799-8798-6

https://doi.org/10.1109/NOC.2015.7238630

Energy Efficient Disaggregated Servers for Future Data Centers

Howraa M. Mohammad Ali, Ahmed Q. Lawey, Taisir E. H. El-Gorashi, and Jaafar M. H. Elmirghani
School of Electronic and Electrical Engineering, University of Leeds, LS2 9JT, United Kingdom

***Abstract*— With the dawn of cloud computing, data centers' power consumption has received increased attention. In this paper we evaluate the energy efficiency potential of exploiting the concept of Disaggregated Server (DS) design in data centers for efficient resource provisioning. A DS, is a new approach for future racks where servers are disaggregated and resources, such as processors, memory and IO ports are arranged in resource pools constructing processing pools, memory pools and IO pools. We developed a mixed integer linear programming (MILP) model for energy minimization of the virtual machine (VM) placement problem in data centres implementing DS approach. The results show that the average power savings are up to 49% for the different VM types considered.**

***Keywords*—Resource provisioning, rack scale server, data center, VM placement, energy efficiency**

I. Introduction

With the rapid growth of data and processing intensive applications and the shift towards the cloud computing model, serious concerns are raised about the power consumption of data centers. To improve the energy efficiency of data centres, architecture design and hardware design must move in concert. Significant research efforts were dedicated to improve the design of switches and servers for data centres [1]-[10]. A number of energy efficient data centre and inter data centre network architectures were proposed and studied in [11]-[16].

Recently concerns have been raised about the "traditional single box" server design that has ruled the data centre industry for the last decades. These concerns have focused on the inflexibility and poor performance of "traditional single box" server designs in terms of resource utilization, scalability, energy efficiency and cost. Disaggregated servers are a departure from the norm that can address the limitations of conventional servers by disaggregating servers to memory, computational, storage and I/O resources and creating pools where each pool contains one type of these resources only. In conventional servers, a server may have a computationally intensive task resulting in the processor being fully engaged but with spare I/O resources that cannot be accessed by the data centre. Furthermore, a server with intensive I/O operation may have spare processor capacity that cannot be accessed as the I/O card of the server is fully engaged. In the DS design approach the pools of disaggregated resources can be used to provision the right amount of resources (processing, memory, I/O) to VM requests resulting in improved resource utilization, reduced resource wastage and reduced power consumption. These resources are typically connected using an optical networking infrastructure. In this paper we evaluate the energy efficiency of VM placement in the data center considering the DS concept as a new feature for the design of future data centers.

Various studies have considered improving the energy efficiency of data centers. Facebook and Open Compute [1]-[2], are setting forth to disaggregate the rack using Intel's new photonic architecture, and HP Moonshot software defined servers [3]-[4]. This approach can introduce a paradigm shift with significant improvement for the data center design where the previously "server dedicated components" are now shared among different servers.

The idea of disaggregated servers has been discussed in some previous work, either by investigating the network ability to support disaggregation at data center scale [5], or by presenting the memory and IO ports as separate blades from the server block [6]-[8], which allow resources to be disaggregated across a system to enable data center vertical elasticity.

Also and as mentioned earlier, Facebook and Open Compute are setting forth to disaggregate the monolithic server designs that have ruled the data center industry for the last decades, and to make them energy efficient, cheap, simple and independently upgradeable entities [1].

I. DS-Based Resource Provisioning Model

Figure 1 explains the concept of server disaggregation. Here each rack contains for example one type of resources only (i.e. CPU, memory or I/O cards). The resources have to be interconnected through a fast switching fabric. This fabric can be an optical backplane in each rack and fast switches must be used to interconnect the racks. Servers can therefore be constructed on the fly through a control plane with the appropriate number of processors, memory and I/O capability. We developed a MILP model to optimize the process of resource provisioning and VM consolidation in DS based data center. We used our MILP model to minimize the total power consumption and compare the results to resource provisioning in data centers implementing the "traditional single box" [17-20]. In this paper we only give a brief introduction to the model due to paper length limitations. Here, a data center comprises of a number of heterogeneous resources chosen from a set of known and well-characterized components (processors, memory and I/O cards). Note that here, and for comparison purposes, we calculate the power consumption of only these resources under the DS concept and under the conventional data centre, and leave an in-depth full data center power consumption evaluation to future. Given the requested VMs, the MILP model places VMs in the optimal location for minimum power consumption and tries to fully utilize the available resources by packing the resources (processors, memory and I/O) with as many VMs as they can hold, in order to minimize the data center's power consumption by reducing the number of working resources.

978-1-4799-8798-6/15/$31.00 ©2015 IEEE

(a) (b)

Figure 1: (a) Traditional single box server versus (b) disaggregated server

The power consumption of a resource is modelled in eq. (2), as the sum of two components: a variable power, (Power Factor (PF)), associated with each resource, linearly related to the utilization of the resource u, plus the power consumed by the resource in the idle state, thus

$$PF_j^x = \left(\frac{PMax_j^x - PMin_j^x}{Cap_j^x}\right) \quad (1)$$

where $PMax_j^x$, $PMin_j^x$ and Cap_j^x are the maximum active power, idle power and total capacity of the j^{th} resource of type x respectively; x can be processor, memory or IO port.

Note that the power consumed by cooling is not considered here, however we limit the resources utilization to 0.9 of its maximum capacity in order to avoid over heating these resources [21].

Below are the parameters and variables used in the model.

Sets:

PR	Set of processors
MR	Set of memories
IOR	Set of Network Interface cards (NIC) ports
VM	Set of virtual machines

Parameters:

NVM	Total number of VMs
P_j	The processing capabilities of processor j in Giga CPU cycles per second (GC/s)
M_j	The access rate of memory j, in GByte/s
IO_j	Total bit rate of NIC port j in Gb/s
$PMax_j^P$	Power consumption of fully utilized processor j
$PMin_j^P$	The idle power consumption of processor j
$PMax_j^M$	Maximum power draw of memory j
$PMin_j^M$	The idle power of memory j
$PMax_j^{IO}$	Maximum power of NIC port j
$PMin_j^{IO}$	The idle power of NIC port j
PF_j^P	Power Factor of processor j in W/(GC/s)
PF_j^M	Power Factor of memory j in W/(GB/s)
PF_j^{IO}	Power Factor of IO j in W/(Gb/s)
SLA	Agreed percentage of served VMs according to Service Level Agreement (SLA)

Variables :

θP_{ij}	The amount of processing capacity of processor j allocated to request i in Giga CPU cycles per second.
θM_{ij}	Portion of the BW of the memory j allocated to request i, in GByte/ s
θIO_{ij}	Portion of the transmission rate of the j^{th} port allocated to request i, in Gb/s
δP_j	The total processing capacity utilization of processor j
δM_j	The total utilization of memory j
δIO_j	The total utilization of IO port j
YP_{ij}	YP =1if processor j hosts request i, otherwise YP=0
YM_{ij}	YM =1if memory j hosts request i, otherwise YM=0
YIO_{ij}	YIO =1if port j hosts request i, otherwise YIO=0
KP_i	KP =1if request i processing requirement is being served, KP =0 if it is blocked
KM_i	KM =1if request i memory requirement is being served, KM =0 if it is blocked
KIO_i	KIO =1if request i IO requirement is being served, KIO =0 if it is blocked
XP_j	Indicates if processor j is active, XP =1, otherwise, XP =0

XM_j	Indicates if memory j is active, XM_j =1, otherwise, XM_j =0
XIO_j	Indicates if port j is being used, XIO_j =1, otherwise, XIO_j =0
Utl	The maximum allowed utilization of the resources of the data center

The power consumption of resources in a data center based on the DS architecture is composed of:

1) The power consumption of active processors

$$\sum_{j \in PR} ((XP_j \cdot PMin_j^P) + (PF_j^P \cdot \delta P_j))$$

2) The power consumption of active memories

$$\sum_{j \in MR} ((XM_j \cdot PMin_j^M) + (PF_j^M \cdot \delta M_j))$$

3) The power consumption of active IO ports

$$\sum_{j \in IOR} ((XIO_j \cdot PMin_j^{IO}) + \left(PF_j^{IO} \cdot \delta IO_j\right)$$

Each component power consumption is composed of two parts, a fixed factor which represents the idle power of the resource x, $xMin_j$ and a variable power term linearly related to the resource utilization δx_j [22].

Objective: minimize:

$$\sum_{j \in PR} ((XP_j \cdot PMin_j^P) + (PF_j^P \cdot \delta P_j)) + \sum_{j \in MR} ((XM_j \cdot PMin_j^M) + (PF_j^M \cdot \delta M_j)) + \sum_{j \in IOR} ((XIO_j \cdot PMin_j^{IO}) + (PF_j^{IO} \cdot \delta IO_j)) \quad (2)$$

Note that objective (2) minimizes the power consumption by consolidating resources into the minimum number of resources possible (due to the presence of an idle power component), so the number of working resources is minimized.

The model is subject to a number of constraints including:

Capacity Constraints:

$$\delta P_j = \sum_{i \in VM} \theta P_{ij} \leq Utl \cdot P_j \qquad \forall j \in NPR \quad (3)$$

$$\delta M_j = \sum_{i \in VM} \theta M_{ij} \leq Utl \cdot M_j \qquad \forall j \in NMR \quad (4)$$

$$\delta IO_j = \sum_{i \in VM} \theta IO_{ij} \leq Utl \cdot IO_j \qquad \forall j \in NIOR \quad (5)$$

Constraints (3)-(5) allocate the data center processing, memory and IO resources to the incoming requests, respectively, find the total utilization of each resource and ensure that each resource maximum utilization does not exceed a given threshold, Utl.

Slicing Constraints:

$$\sum_{j \in PR} YP_{ij} \leq 1 \qquad \forall i \in VM \quad (6)$$

$$\sum_{j \in MR} YM_{ij} \leq 1 \qquad \forall i \in VM \quad (7)$$

$$\sum_{j \in IOR} YIO_{ij} \leq 1 \qquad \forall i \in VM \quad (8)$$

Constraints (6)-(8) ensure that the model serves each VM using only one processor, one memory and one IO port respectively. If multiple VM copies or VM slicing is required, equations (6)-(8) should be upper bound by an appropriate number greater than 1.

Service Level Constraints:

$$\sum_{i \in VM} KP_i \geq NVM \times SLA \quad (9)$$

$$\sum_{i \in VM} KM_i \geq NVM \times SLA \quad (10)$$

$$\sum_{i \in VM} KIO_i \geq NVM \times SLA \quad (11)$$

Constraints (9)-(11) guarantee that the number of served VMs is greater than a pre-specified value according to SLA.

II. EVALUATION ENVIRONMENT, RESULTS AND DISCUSSION

We built a DS architecture for evaluation by disaggregating the IBM system X3650 M3 server [23]. The IBM X3650 M3 server supports 11 processor types with different number of cores, and power characteristics. Table I shows the 11 processor types with their maximum power draws. The IBM X3650 M3 server comes also with three standard memory bandwidth rates. The memory is a DDR3 SDRAM with three bandwidth values, where the evaluation in [24] for DDR3, gives the memory power consumption, see Table II.

In practice, and although the data centre may be fully disaggregated into resources (processors, memory, I/O), these resources are in practice not a single fully reconfigurable block. Instead to ease the resource handling, to ease the mounting of resources in racks and to ease the communication requirements, pools of resources can be defined. Note that a pool that has a single processor, a single I/O card and memory is a conventional server. We consider in this paper a medium size pool of processors made of 6 processors of each of the 11 types in Table I. We plan to optimize the size of the pool in future work. The active power consumption of the processors is set depending on the system described in [23]. The processor idle power consumption is set to be 0.7 of its power consumption when fully utilized [22].

A pool of memory is made of 3 memory types, see Table II. It contains a total of 66 memory units where 36% of the used memories operate at 4 GB/s, and 28% operate at 8 GB/s, while the remaining 36% operate at 24 GB/s. The system is completed with a pool of 66 NIC ports where half of them support a rate of 1 Gb/s and the remaining are 10 Gbps, see Table III. The numbers of the three used resources are equal in order to have a fair comparison with the CS data centers.

We considered two types of IO ports in the evaluation, 1 Gb/s and 10 Gb/s data rates, and their power consumption is based on the work in [25]. As a conservative case, we consider the situation where an idle port consumes 0.7 of the power consumed when it is fully utilized. The maximum power of each port type is given in Table III. Some of the input parameters of the model are given in Table IV.

A VM is characterised by three main requirements, CPU requirement CP_i, memory requirement CM_i, and IO requirement CIO_i. In view of the available resources capacity, the request types under consideration and the SLA violation avoidance needs, we have estimated the amount of resources that each VM type needs and proposed three types of VMs, memory intensive (MI) VMs, processing intensive (PI) VMs and IO intensive (IOI) VMs. The details of the resource requirements for each VM type are shown in Table V.

TABLE I POWER CONSUMPTION AND CAPACITY OF IBM X3650 M3 SERVER

Processors capacities (GHz)	Processors Max. power consumption (W)
3.46	130
3.6	130
2.93	130
2.66	95
3.2	95
2.4	80
2.53	80
2.13	80
2.26	60
2.13	40
1.86	40

TABLE II MEMORY RATE AND POWER CONSUMPTION

Memory Rate (GB/s)	Memory Max. Power Draw (W)
4	5.12
8	10.24
24	30.72

The results in Fig. 2 are based on our MILP optimization and compare the power consumption of memory intensive virtual machines if these VMs are implemented using a conventional data centre design (conventional servers (CS)) and if they are implemented using a disaggregated data centre design (DS). Fig. 2 shows that VMs implemented in a data center using the DS approach achieve a best average power saving of 49% (given our set of parameters) when all the VM requests are MI.

TABLE III NETWORK INTERFACE CARDS DATA RATES AND POWER CONSUMPTION

NIC Port Rate (Gb/s)	NIC Port Max. Power Draw (W)
1	1.9
10	21.4

TABLE IV INPUT PARAMETERS USED IN THE OPTIMIZATION MODEL

Number of processors	66
Number of memory Chips	66
Number of NIC ports	66
Number of processors types	11
Number of memory types	3
Number of NIC port types	2
Utl value	0.9
SLA value	100%

TABLE V RESOURCES REQUIRED BY EACH VM TYPE

Request Type	Processing (G CPU Cyclesps)	Memory (GBps)	IO (Gbps)
Processing Intensive	1-3.3	0.05-0.2	0.05-1
Memory Intensive	0.2-1	1-4	0.05-1
IO Intensive	0.2-1	0.1-0.5	1-4

Considering the MI VMs, the memory requirements are the highest and therefore in these VMs memory is used to a much larger extent compared to other processing and IO resources utilization, see Table V. Thus, in the “traditional single box” server, we call here CS, the memory requirement of the VM may cause a whole server to be dedicated to a single VM, as the memory of the server cannot accommodate more VMs. This comes at the expense of losing free space in the processor and the IO port, thus most of the servers will host only one VM. However, in DS, the memory, the processor and the IO port are limited by the server box boundaries, thus the spare processor and the IO ports capacities can be accessed allowing additional VMs to be accommodated and leading to improved resource utilization.

It is clear that processors consume the most power and memory resources consume the least power, while IO ports power consumption lies between the two. Thus, with MI VMs, the number of working processors and IO ports in the DS are much less than CS because they can be used efficiently in the DS architecture, which in turn will result in high power saving.

We optimized the DS infrastructure under processing intensive VM requests, and here we achieve about 11% power saving as shown in Fig. 3. With the PI VM requests the processing requirements are higher than memory and IO requirements, thus a large number of processors, which consume the most power, will be used in both CS and DS. Thus, the power saving in this case will come from the memories and the IO ports, which explains why we observe a smaller power saving.

For IOI VM requests about 24% of the power consumed by CS will be saved when implementing DS approach, see Fig. 4. With IOI VM type, the bottle neck is usually the IO

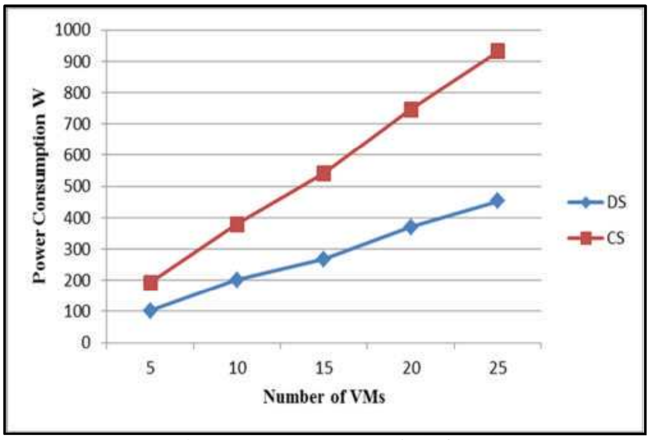

Figure 2: Power consumption of MI VMs

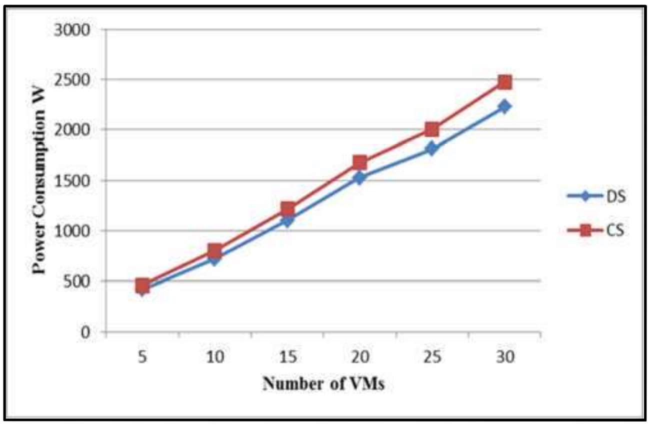

Figure 3: Power consumption of PI VMs

requirements, thus IO ports will not be used efficiently, which affects the use of the whole server in the CS. With IOI VM type, the bottle neck is usually the IO requirements, thus IO ports will not be used efficiently, which affects the use of the whole server in the CS. However, with the DS the number of working processors and memories is less than the number of working processors and memories in CS, which results in a good power saving. Nonetheless, this will be less than the power saving achieved when serving MI VMs, because the power saving will come from the latter's efficient use of processors and memories. Thus, serving MI requests will be the less efficient scenario with the CS, and this will lead to the maximum amount of power to be saved with the DS architecture. The IOI scenario will be an intermediate case and serving PI requests will result in the minimum power saving.

III. CONCLUSIONS

In this paper, we have investigated the energy efficiency of VM placement in data centres based on a disaggregated data centre approach and evaluated the power saving of the new approach. The approach considered enables the separation of the computing, memory, storage and network resources of the server leading to better resource utilization by "composing on

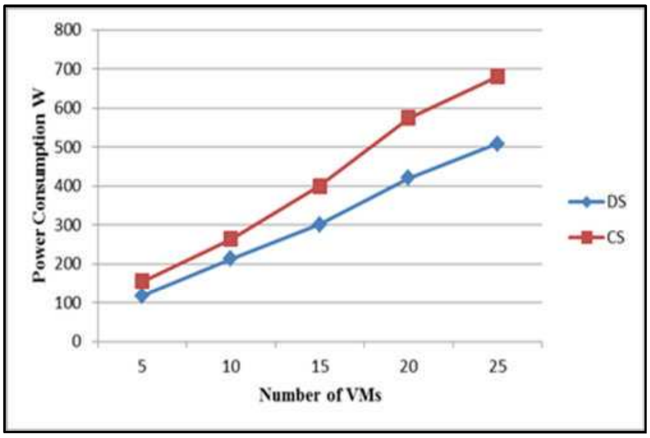

Figure 4: Power consumption of IOI VMs

the fly" servers with the exact required processing, memory and I/O capabilities to accommodate the virtual machines or tasks of interest.

We have developed a MILP optimization model which optimally places VMs in the disaggregated data center with the objective of minimizing the power consumption. We have compared a data center with DS architecture to a data center using the normal rack of servers units considering the VM placement and resource provisioning operations.

To gain a good view for the operation of the proposed approach, we have considered three types of VMs: PI, MI and IOI in the model. The results show that with MI applications, the DS approach achieves the maximum power saving. When serving MI requests the model achieves an average power saving of 49% and for IOI requests the average power saving is 24%, while serving PI request results in 11% average power saving under the set of typical parameters and conditions we considered.

IV. ACKNOWLEDGMENT

The authors would like to acknowledge funding from the Engineering and Physical Sciences Research Council (EPSRC), INTERNET (EP/H040536/1) and STAR (EP/K016873/1). Mrs. Howraa Mohammad Ali would like to acknowledge the Higher Committee for Education Development-IRAQ (HCED IRAQ), for funding her scholarship.

REFERENCES

[1] http://arstechnica.com/information-technology/2013/04/intel-wants-to-kill-the-traditional-server-rack-with-100gbps-links/.

[2] D. Version, "Design Guide for Photonic Architecture", Intel draft document, January, 2013.

[3] H. P. Moonshot, F. P. Led, H. P. Proliant, M. Server, S. Module, S. F. P. U. Module, C. Management, C. Manager, R. Fabric, and T. M. Fabric, "HP Moonshot System The world's first software defined servers", HP technical white paper, April, 2013.

[4] http://blog.pluralsight.com/hp-moonshot-software-defined-server.

[5] S. Han, N. Egi, A. Panda, S. Ratnasamy, G. Shi, and S. Shenker, "Network support for resource disaggregation in next-generation datacenters," Proc. Twelfth ACM Work. Hot Top. Networks - HotNets-XII, pp. 1–7, 2013.

[6] P. Sv, B. Hudzia, J. Tordsson, and E. Elmroth, “Hecatonchire : Enabling Multi-Host Virtual Machines by Resource Aggregation and Pooling”, a report published by the Department of Computing Science, Umeå University.

[7] K. Lim, J. Chang, T. Mudge, P. Ranganathan,S. K. Reinhardt, T. F. Wenisch, “Disaggregated Memory for Expansion and Sharing in Blade Servers”, ACM SIGARCH Comput. Archit. News, vol. 37, p. 267, 2009.

[8] C. Tu and C. Lee, “Marlin: A Memory-Based Rack Area Network”, Proceedings of the tenth ACM/IEEE symposium on Architectures for networking and communications systems, pp. 125–135, 2014.

[9] Kachris, C and I. Tomkos, “Power consumption evaluation of hybrid WDM PON networks for data centers. in Networks and Optical Communications (NOC),” in 16th European Conference IEEE, 2011.

[10] P. N. Ji, D. Qian, K. Kanonakis, C. Kachris, and I. Tomkos, “Design and Evaluation of a Flexible-Bandwidth OFDM-Based Intra-Data Center Interconnect”, IEEE JOURNAL OF SELECTED TOPICS IN QUANTUM ELECTRONICS, vol. 19, no. 2, 2013.

[11] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, “Green IP Over WDM Networks With Data Centers”, Journal of Lightwave Technology, VOL. 29, NO. 12, JUNE 15, 2011

[12] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, “On the Energy Efficiency of Physical Topology Design for IP Over WDM Networks”, Journal of Lightwave Technology, VOL. 30, NO. 12, JUNE 15, 2012

[13] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, “IP over WDM networks employing renewable energy sources”, Journal of Lightwave Technology, VOL. 29, NO. 1, JANUARY 1, 2013

[14] A. Q. Lawey, T. E. H. El-Gorashi, and Jaafar M. H. Elmirghani, “Distributed Energy Efficient Clouds Over Core Networks”, Journal of Lightwave Technology, VOL. 32, NO. 7, APRIL 1, 2014

[15] A. Q. Lawey, T. E. H. El-Gorashi, and Jaafar M. H. Elmirghani, “BitTorrent Content Distribution in Optical Networks”, Journal of Lightwave Technology, VOL. 32, NO. 21, NOVEMBER 1, 2014

[16] L. Nonde, T. E. H. El-Gorashi, and Jaafar M. H. Elmirghani, “Energy Efficient Virtual Network Embedding for Cloud Networks”, Journal of Lightwave Technology, VOL. 33, NO. 9, MAY 1, 2015

[17] B. Kantarci and H. T., Mouftah, "Optimal Reconfiguration of the Cloud Network for Maximum Energy Savings”, 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing 2012.

[18] H. Goudarzi and M. Pedram, “Energy-Efficient Virtual Machine Replication and Placement in a Cloud Computing System”, IEEE Fifth International Conference on Cloud Computing 2012.

[19] B. Kantarci and H. T. Mouftah, “Designing an Energy-Efficient Cloud Network [Invited]”, J. Opt. Commun. Netw., vol. 4, no. 11, p. B101, Oct. 2012.

[20] B. Kantarci and H. T. Mouftah, “Energy-Efficient Demand Provisioning in the Cloud,” Opt. Fiber Commun. Conf., no. 2, 2012.

[21] Juniper Networks Tech Library, available at "https://www.juniper.net/documentation/en_US/junos12.2/topics/concept/logical-system-security-cpu-allocation-control-understanding.html", accessed on 24 April, 2015

[22] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya, “A Taxonomy and Survey of Energy-Efficient Data Centers and Cloud Computing Systems”, Technical Report, CLOUDS-TR-2010-3, Cloud Computing and Distributed Systems Laboratory, The University of Melbourne, Australia, 2010.

[23] P. Guide, “IBM System x3650 M3”, June, pp. 1–21, 2011

[24] K. T. Malladi, F. A. Nothaft, K. Periyathambi, B. C. Lee, C. Kozyrakis, and M. Horowitz, “Towards Energy-Proportional Datacenter Memory with Mobile DRAM”, Proceedings of the 39th Annual International Symposium on Computer Architecture, pp.37-48, 2012.

[25] R. Sohan, A. Rice, A. W. Moore, and K. Mansley, “Characterizing 10 Gbps Network Interface Energy Consumption”, 35th Annual IEEE Conference on Local Computer Networks, pp. 268 - 271, 2010.