Architecture of an end-to-end energy consumption model for a Cloud Data Center

Kashinath Basu
School of Engineering, Computing and Mathematics
Oxford Brookes University
Oxford, United Kingdom
kbasu@brookes.ac.uk

Ali Maqousi
Faculty of Information Technology
Petra University
Amman, Jordan
amaqousi|@uop.edu.jo

Frank Ball
Frank Ball Consulting
Oxford UK.
frank_ball@ntlworld.com

***Abstract*— Estimates show that a significant proportion of future ICT related energy consumption will be from Cloud Computing. Based on detail analysis and survey of energy consumption and optimization trends in cloud computing, this research presents a comprehensive end-to-end energy consumption model of a cloud facility extending from the end-user equipment to the data center facility. The model is subdivided into three planes and four associated layers and depicts the cross-plane and cross-layer relationships between the components in terms of energy consumption and potential optimization areas and provides a reference framework for planning power optimization strategies at a cloud facility.**

Keywords—cloud, data center, energy efficiency, fog, frequency voltage scaling, SDN, power consumption

I. INTRODUCTION

Information and communication technologies (ICT) equipment accounted for around 2% of the total energy consumption (80 million megawatt-hours annually in recent years [1]. It is estimated that by 2025 Information and communication technologies will consume 2.7% of the total energy usage generating 1.43 gigatons of CO_2 [2]. Based on recent trend, it is predicted that a significant proportion of future ICT activities will be based on the cloud computing model and as a result the cloud computing infrastructure will account for up to 60% of ICT related energy usage [3]. Therefore, to reduce the overall ICT related power usage, power consumption in the Cloud Data Centre should be one of the main focus.

There are several benefits for optimizing the energy consumption in the cloud. Firstly, it will help to reduce the energy bills which are directly proportional to the energy usage. With the spiralling trend in energy prices, energy efficiency is significant for the long term competitiveness of the cloud model. Secondly, reduction in direct energy consumption will also reduce indirect energy usage and its associated energy cost. Thirdly, in the future government regulations on CO2 emissions will bring further stringent guidelines on level of energy usage on ICT facilities. Therefore, for the cloud model to be acceptable energy consumption has to be optimised. Finally, higher energy consumption leads to higher heat dissipation and reduces the life cycle and mean time between failures (MTBF) of cloud hardware thereby increasing the equipment cost and therefore the cost of cloud service. For all these reasons it is vital that a cloud solution is energy efficient.

The rest of the paper is organised as flows: Section II presents an analysis of energy consumption in a cloud facility; section III presents the overview of the proposed end-to-end energy consumption model; section IV presents the cross-sectional analysis of the three planes and four layers of the model; and finally, section V presents the a summary of the main features of the model and future work.

II. ENERGY CONSUMPTION IN A CLOUD FACILITY

The energy consumption of a cloud facility can be categorised based on direct and indirect energy usage [Fig. 1]. The direct energy consumption is related to ICT equipment that constitutes mainly of servers, but also includes storage and networking equipment. The nature of power consumption of these equipment can be classified as static and dynamic power dissipation. Static power dissipation is primarily caused by leakage current in the equipment circuitry and is independent of the load on the system. Server components such as memory, storage, network interfaces primarily contribute to this energy consumption. In a typical cloud server, this can constitute for up to 60% of the peak power [4]. In contrast, dynamic power dissipation is directly proportional to the workload on the system and depends primarily on the CPU utilisation of the servers, along with traffic load across the networking devices and read/write activities on the storage systems.

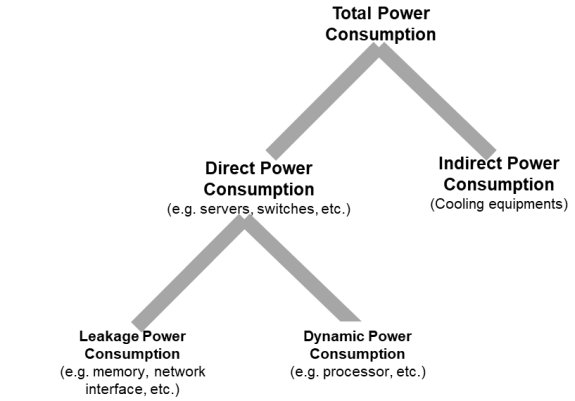

Fig. 1. Direct and indirect energy consumption in a Cloud

On the other hand, indirect energy consumption mainly involves powering the cooling equipment that is responsible for maintaining the temperature of the ICT equipment at optimal level. It is directly proportional to the direct power consumption. Studies show that indirect power consumes a larger share (40 – 70%) of the total energy consumption of a cloud facility [7, 8]. Therefore, reducing the amount of direct energy consumption will also reduce the amount of indirect energy usage in cooling equipment. Hence research in energy efficiency of a cloud has

to consider both direct and indirect energy usage and their correlation.

III. The Proposed End-to-End Energy Consumption Model

There have been various categories of research on energy optimisation covering different aspects of cloud and ICT in general. The proposed model provides a framework in which some of these individual solutions can fit together and work in synchronisation to provide an end-to-end seamless energy efficient cloud solution. The architecture of the model consists of three vertical planes covering the end-to-end path from the customer up to the cloud. These planes are the client or user equipment plane, the network plane and the data centre plane. Each of these planes is further subdivided into four horizontal layers. Each of the layers has different levels of abstraction of resources, energy usage and energy management. [Fig. 2]. This includes managing energy at the granularity of physical processor and memory at the lowest physical resource layer, followed by managing at the hardware level such as servers, routers and switches at the physical device layer, followed by virtual machines at the virtualization layer and finally the cooling and facilities system at spatial layer. A key factor that influences the energy usage and efficiency on an end-to-end basis is the service model of the cloud. There are several types of services available on the cloud. These could range from simple storage and infrastructure, to processing, to software, to full scale application hosting platform [7]. The type of service provisioned determines the energy requirement and therefore the suitability of the service over the cloud .

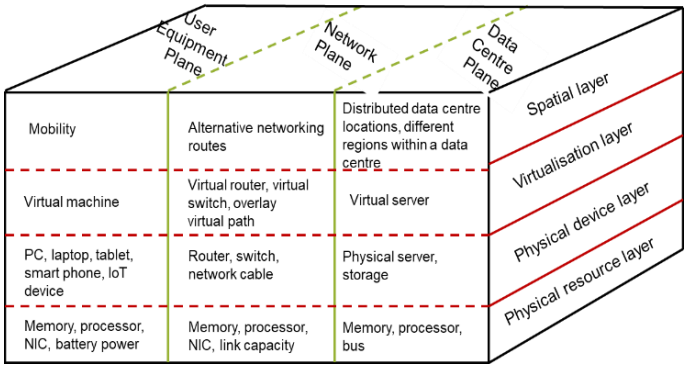

Fig. 2. End-to-end Energy Consumption Model

IV. Cross-Section of the Model: The Vertical Planes and the Horizontal Layers

A cloud provides performance, scalability and reliability by using redundant resources distributed across the cloud infrastructure. The goal is to ensure that the right resources are selected to meet the service level agreements (SLAs) of the tasks with the most optimal utilisation of energy. In this context, there are two key factors that are significant for energy optimisation across the vertical planes and the horizontal layers of the energy consumption model. Firstly, in each of the layers efficient scheduling mechanism is used to optimise the resource usage at the layer [8]. Secondly, this is synchronised with a cross-layer top down signalling framework to communicate and map energy related information across the layers [10]. For example when a new task is assigned to the cloud, the virtualization layer may decide based on the SLA whether to schedule the task to an existing virtual machine, to initialise a new virtual machine on an existing physical machine or to boot-up a new physical server. In order to make this scheduling decision, the virtualization layer relies on accurate information on current resource availability and their energy consumption level from the underlying physical device layer which in turn depends on information from the resource layer. Based on accurate capture, aggregation and signalling of these information across the underlying layers efficient scheduling of resources can be achieved on a particular layer. An analysis of the energy consumption across the three vertical planes and their associated horizontal layers is presented below:

A. Data Centre Plane

The main focal point for energy usage is the data centre plane. Here the main consumers for direct energy consumption are the servers, followed by networking and storage devices [11]-[13] [Fig. 3]. This vertical plane is divided into four horizontal layers based on the abstraction of energy usage [Fig. 2]. At the lowest layer are the physical resources. The main consumer here is the processor, followed by primary and secondary storage and peripheral devices. Research shows that the dynamic power consumed by a processor is proportional to the cubic power of the operating clock frequency. Therefore, power can be saved by scaling the clock frequency based on the load on the processor and the deadlines of the tasks. Over the years, processor manufacturers have introduced central processing units (CPU) based on frequency/voltage scaling (DVS) [11]. The full extent of DVS is exploited by using energy efficient processor level scheduling scheme in the multiprocessor environment of a cloud server. Some of these schemes also consider indirect energy costs in memory contention and DVS overheads [14, 15]. Similar approach to CPU clock gating have been extended to other hardware resources by powering down redundant hardware parts of chips (power gating), memories and peripheral devices to low power states [13, 16].

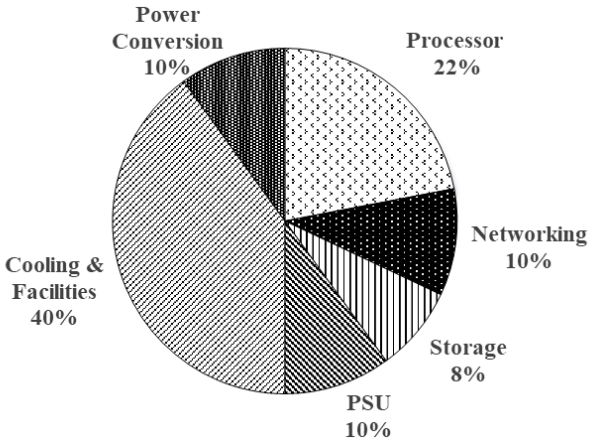

Fig. 3. Energy consumption at the Data Centre Plane

Studies shows that up to 40% of the power consumed by a typical server could be used up by the memory [17]. This memory usage can be optimised by partitioning memory into fine grained smaller units and then individually managing the power state of these units by a controller. However, there is a drawback in this approach which is the additional latency overhead in resynchronising back the memory units from idle to full operating state. The number of resynchronisation can be optimised to certain extent by concentrating the memory requests to a limited number of memory units and running them at a high utilisation level and switching on new units only when the existing capacity is full. In addition, the memory units could have different power state such as fully active, standby, nap and fully passive mode with each state having a reduced number of

active memory subcomponents and hence lower power state but at the cost of higher resynchronisation delay. A full active memory unit with active transistor-capacitor pairs, row and column decoders, sense amplifiers and bus drivers consume up to 40 times more energy than a passive memory unit with only the capacitor refresh circuitry on which is the minimum requirement for countering data loss against the leakage current. This huge difference in power consumption level makes optimisation of the resynchronisation latency important. This can however be handled by an energy efficient memory level controller which could gradually update the state of the memory units in advance based on the trend of the load on the servers and hence avoid the effects of resynchronisation latency from the system.

Secondary storage in a cloud computing data centre is generally based on hard disk arrays. Centralisation and consolidation of the storage space and hard disk usage can reduce energy consumption [18]. In addition, efficient disk array management system can reduce energy consumption at the disk access level by up to 65% [17]. These systems can either control the rate at which a disk spins or migrate the task to the disk with the right speed to meet the SLA. Furthermore, files that are not regularly used can be stored separately in low powered optimised disks to save energy. Recent technologies such as solid-state disks consume far less energy and are being used to replace the traditional hard disks. Cloud data centres also have significant amount of networking hardware for interconnection. Energy efficiency of the networking gears is discussed later in the networking plane section.

In the data centre, the physical device layer is dominated mainly by the energy consumption of the physical servers. As mentioned earlier an idle server consumes around 60% of its peak power. Similarly a server with a low utilisation consumes up to 70% of its maximum power consumption [4]. Surveys conducted in data centre facilities shows that on average 10% of the servers remain unused and the mean load on the operational servers is only 20%. This level of over provisioning by the cloud service providers is mainly to guarantee a high level of service and availability to its customers during peak load. It has been found that significant energy efficiency of up to 30% can be achieved by consolidating tasks into a limited number of physical servers and running them at higher utilisation and switching off or hibernating the remaining servers. The physical server layer has to work very closely with the virtualization layer above and the physical resource layer below to achieve its objectives.

Virtualization technique plays a key role in consolidation of the physical servers by aggregating the workload into a limited number of physical machines. Isolation between the workload environments is provided by logically partitioning the physical resources using virtual machines (VM). Virtualization can be at various levels and comes with different trade-offs [19]. Full virtualization provides virtualization of hardware and operating system (OS) but is comparatively slower whereas para-virtualization and hardware assisted virtualization require support from modified OS and system hardware respectively but is faster [20, 21]. The efficiency gain from virtualization is highly dependent on the scheduling scheme used to distribute workload among the VMs . Here the choice is between whether to assign the task to an existing VM, start a new VM on a running server or to start a new physical server. The decision is based primarily on the SLA of the task and the existing load and capacity of the servers and VMs [22]. Some virtualization technique allows setting VM level 'soft' virtual power budget in addition to the traditional physical server level 'hard' power budget [23]. In this approach, a more comprehensive fine grained energy optimisation scheme is feasible considering both 'hard' and 'soft' power budget to ensure both server and VM level energy targets are met. It has been demonstrated that by using virtual power management techniques up to 34% improvement in power consumption can be achieved in a heterogeneous server system. The utilisation level of a cloud infrastructure is dynamic and changes as new tasks joins and old tasks ends. Therefore, the cloud infrastructure has to be continuously monitored and tasks or VMs may have to be migrated midway to choose the most energy efficient option [24]. This can be achieved with a dynamic work conserving scheduling scheme that monitors the state of the cloud in real-time and migrate accordingly. However, the migration decision has to consider the cost of migration in terms of delay and energy usage in copying and transporting the workload against the SLA and energy budget.

Here, the performance of the scheduler is also dependent on its past knowledge of the characteristics of the tasks and their arrival pattern. It has been shown that prior knowledge of the different types of workload and their resource requirement can optimise energy usage by 20%. Here tasks can be efficiently distributed across a heterogeneous range of servers [25] that best matches the resource requirement of the tasks. In addition, any prior knowledge of the arrival pattern helps to prepare resources in advance of the task's arrival and thereby reduce the energy wastage due to dead-time during server/VM boot-strapping.

The topmost layer of the data centre plane is the data centre spatial layer. This layer mainly contributes to the indirect energy consumption mentioned earlier. Here, the energy usage is mainly due to the power consumption of the cooling equipment. Recent surveys show that there is a large variation (40-70%) across the industry on the amount of the data centre power that is spent on cooling equipment [Fig. 3] [26]. The energy consumption at this layer is directly proportional to the energy used at the lower layers. Recent studies [6] demonstrate that by improving the operational efficiency of a cloud facility a 20% savings in direct energy consumption of servers and networks facilities can be achieved. This consecutively will result in 30% savings in cooling cost. The increased performance and scalability demands have resulted in higher utilisation of the devices and reduction of server form factors. This has partially reduced the direct energy cost; however it has also created further engineering challenges in power delivery and cooling because of high voltage cable and hotspots on data centre floors. This effect can be partially offset by using a space and location aware scheduler that will distribute workload across different sites of a distributed cloud as well as at different regions of the floor space within a single cloud site using the temperature level as one of the decision parameter. In this way, hotspots can be avoided and the heat can be evenly distributed across floor space and sites. Studies show that by using techniques such as smart airflow management, variable speed cooling fans and by increasing the operating temperature to a slightly wider, yet safe range the cooling energy requirement can be reduced by up to 25%.

In a data centre, the power supply infrastructure is general based on AC (Alternating Current) power; however equipment such as standby UPS (Uninterruptible power supply) battery

units, microchips and various other IT components are based on DC (Direct current) power. During the conversion phase from AC to DC, up to 20% of the power is lost. This can be prevented by using a DC power supply system throughout the data centre. In addition, modern DC power supply technologies are up to 97% more efficient and can significantly reduce the energy consumption.

B. Network Plane

The role of the network plane has previously been mostly ignored in past analysis on the energy efficiency of the cloud infrastructure. However, the network is a vital factor while considering the energy efficiency of a cloud service. The scope of the networking plane extends from the communication interface of the end user devices all the way to the cloud servers. Some estimates rates the current annual energy consumption of the networking infrastructure across the globe at 167TWh [35] and an increasingly significant part of this is used to transport data for the cloud services. In the future this is likely to rise at a faster rate. It is estimated that soon the CO2 emission from networking infrastructures could reach up to 350 million tons [2]. A typical end-to-end connection from an end user to a data centre facility spread across three tiers of networks. These include the access, metropolitan and core networks. An overwhelming 94% of all networking equipment are located at the access network level. However, at the access level equipment have low capacity and also have a low utilisation (5%) and consume around 70% of the total energy of all the networking equipment. The metropolitan and core level makes up the remaining 6% of the equipment. However, they have higher capacity and consume around 30% of all networking equipment related energy. These equipment typically run at comparatively higher utilisation level of 30-40% [27].

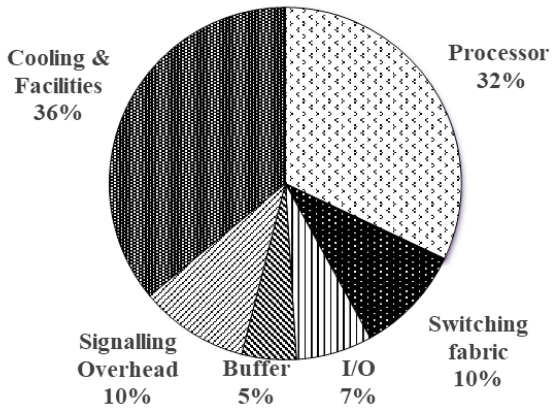

Fig. 4. Energy consumption at the Network Plane

Like the data centre plane, in the proposed model, the network plane is also divided into four layers with hierarchical level of abstraction of energy consumption. At the physical resource layer, the main consumer of energy are the processors, switching fabric, I/O and buffers [Fig. 4] [28]. Similar to servers, processors speed within a networking device could be scaled based on demand using DVS techniques discussed earlier [13, 14, 15] to provide an adaptive link rate. At the switching fabric, energy could be saved by using energy efficient silicon technologies [29] or pure optical switching architecture [30]. In addition, the use of optical fibre could be extended from the core to the access . Passive Optical Network (PON) implementations have shown to significantly reduce the energy and related cooling cost [31]. Buffers in a networking device have to be accessible on a real-time basis and are generally associated for each input and output port. Some energy could be saved by increasing the utilisation of buffers by using schemes such as shared memory management among the ports. At the I/O level, there has been substantial research in energy optimisation at the level of network interface cards (NIC). It has been demonstrated that by putting idle interfaces to sleep mode up to 20% energy can be saved [32]. However, on packet arrival this approach may result in loss of a small fraction of the data depending on the wake-up latency of the system. A solution to this problem is to use a dummy wake-up packet before transmitting the actual data [33] or keeping part of the circuitry awake to sense packet on an arrival line [34]. In network interconnections with several redundant links another proposal is to shut down some of the links during period of low utilisation. It has been demonstrated that this could reduce energy consumption of a networking device by as much as 80% [35] in certain cases. This however is only feasible at the core and metropolitan level, where there are several redundant connections. At the access level, where networks are generally organised in tree structure this approach is not feasible. In this case, links could be run at adaptive rates based on the demand and energy can be saved using DVS. Another alternative is to buffer data and send out at intervals in burst. Here energy is saved by powering down the line in-between the bursts. Any such scheme that requires changes in the state of the interconnection is a cross-layer issue and has to be coordinated with the upper layers.

Energy efficiency at the virtualization and physical device layer of a network are closely interdependent. In long off peak period partially or completely shutting down redundant networking devices will significantly reduce direct energy consumption. However, it may trigger huge signalling overhead in path reorganisation at the routing level. A solution to this approach is to ensure that layer-3 overlay paths are not affected during reorganisation. This can be achieved by remapping the changes via layer-2 tunnels and virtual paths.

There have been recent trend in softwarization, virtualization and soft partitioning of networks using Software Defined Networking (SDN), Network Function Virtualization (NFV) and network slicing technologies. In SDN, the physical network can be centrally managed by a programmable control plane thereby reducing the processing load on the networking devices whereas in NFV virtual networking devices can be spinned off on-demand basis optimizing their usage and thereby power consumption. Network slicing technologies logically partition the physical network into a number of transparent virtual networks each running its own customised pool of scalable virtual resources [36]. When additional resources are required at the physical level, they could be transparently added without disturbing the virtual network model. In this way by virtually portioning and customising the network architecture energy could be saved by both increasing the utilisation of the network and reducing unnecessary signalling overhead. In all optimisation at the network plane it has to be ensured that the network still provisions sufficient quality of service (QoS) to guarantee the SLA of the cloud service.

At the spatial layer, the challenges in optimising the energy consumption in heat management and on cooling equipment at a switching centre are similar to that of the data centre. The capacity per rack of the networking equipment have increased more steeply in comparison to development in heat and energy

management technology and this could be a bottleneck in the future [27].

1) Energy optimisation in emerging networking architectures

As mentioned earlier, it is believed that in the future cloud centric data will be one of the main traffic on the networks. Therefore, to provide end-to-end energy efficient cloud service, the network plane has to significantly reduce energy consumption. There have been several types of suggestions of architectural level changes. For example, in the context of distributed cloud, end user's request could be served from a localised edge cloud as in fog computing thus reducing the energy consumption in transport. Similar approaches are already being used in Content Delivery Networks (CDN) [37]. In the context of routing, the routing algorithms could be reengineered to concentrate traffic over a limited number of highly utilised links and powering down the remaining devices. This approach could optimise the wastage due to large redundant capacity on the networks. Another approach is to incorporate the energy usage level as a cost factor for route selection. This will ensure that the most energy efficient route is selected by the routing protocol .

As mentioned, the SDN networking model centralizes the control plane of a network in a separate server away from the forwarding devices. This optimises the signalling and orchestration overhead of a network. This has both direct and indirect positive implication on the energy usage of a network. By optimising the complexity in the control plane of a network, the network could be made more flexible, fast and robust. This indirectly reduces the amount of processing energy used during route setup. This approach also directly reduces the energy consumption in the forwarding devices. A control plane typically accounts for around 11% of the energy consumption of a forwarding device. Moving the control plane away will reduce the power consumption in the forwarding device to a more energy efficient centralised server. Moreover the overall power consumption across the control plane of the network will also reduce since it will not be linearly related to the number of devices. It is estimated that in the future a significant part of the networking plane of the cloud will be based on this model.

A significant part of the transport energy is used up by control overhead and signalling. This is mainly contributed by protocol complexities, some of which are unnecessarily repeated across the layers and some are inherited from the past but pointless for modern communication links and today's applications. It is estimated that by redesigning the protocol stack from scratch by focussing on cloud based services the transport overhead can be reduced by as much as 30-40%.

In the future, a significant proportion of cloud services will be accessed over the mobile networking infrastructure. The radio segment of a mobile network consumes significantly more energy compared to a wireline network. Here, the energy usage in the radio segment could be optimised by using better cell planning, intelligent antenna systems (e.g. Multiple Input Multiple Output, beam forming, etc.) and state of the art frequency reuse and encoding schemes [38].

C. User Equipment Plane

The traditional user equipment for accessing the cloud service are workstations and laptops. The energy consumed by these products is based on the specification of the hardware and the peripheral devices as well as the load factor. There are international and national bodies that certify the energy ratings of these types of user equipment [15]. Energy optimisation techniques such as clock gating and power gating that are used for servers in the data centre can also be applied here. For compatibility between different equipment manufacturers there are industry compliant standards for switching peripheral devices, CPUs and computer system to different power state based on the level of usage [39]. For cloud intensive processing, thin clients connected to desktop virtualization software in the cloud is an alternative for workstations. They consume up to 80% less energy than workstations and can significantly reduce the energy consumption at the user plane. A mobile or an Internet of Things (IoT) device is another energy efficient option for accessing the cloud. They are inherently energy efficient because of the limited lifespan of the batteries; however they have limited processing capability and storage. These deficiencies can be complemented by similar virtualization services at the cloud based mobile backend.

V. Conclusion and Future Work

The main features of the end-to-end energy consumption model of the cloud discussed above are as follows: firstly, the approach towards energy efficiency is hierarchical from the granularity of the individual hardware components to the end-to-end logical overlay system. At the lowest level, the individual sub-systems of the cloud hardware such as processor, memory, NIC and power supply unit (PSU) have to be energy efficient and the level of energy usage should be measurable and configurable. This information is mapped cross layer to the abstraction of system level efficiency of the hardware devices such as the physical servers and network switches. An end-to-end cloud infrastructure generally has large scale redundancies of equipment and network capacity to deal with varying load and traffic demand. The approach in this model is to aggregate the workload both at the transport level as well as storage and processing. This optimises the energy usage since only a limited subset of the links and servers are run, albeit at a higher utilisation factor. In this context, efficient virtualization is important to virtually partition the network and the servers transparently to ensure that SLAs are not compromised. At the top, the floor space level optimisation is dependent on the optimisation at the underlying layers and the efficiency of the heat distribution, power conversion and the cooling system. In a distributed cloud, further space level optimisation is possible by distributing the workload to the most energy efficient location after factoring the energy overhead of both the network and the cloud server. Secondly, the model has a comprehensive framework for cross-layer communication to signal information on current load, rate of consumption and power budget to assist in the decision making process. Thirdly, at each layer there is an energy efficient robust scheduler to optimally distribute the workload among the competing resources of that layer. In future work, this model can be used as a reference to construct a structured energy management framework for the cloud in which different energy optimization solutions catering to individual planes and layers of the model can be integrated together to provide a unified end-to-end energy efficient cloud solution.

References

[1] "Smart 2020: Enabling the low carbon economy in the information age," The Global e-Sustainability Initiative (GeSI) 2008. [Online]. Available: https://www.theclimategroup.org/sites/default/files/archive/files/Smart2020Report.pdf

[2] A. P. Bianzino, C. Chaudet, D. Rossi and J. Rougier, "A Survey of Green Networking Research," IEEE Communications Surveys & Tutorials, vol. 14, no. 1, pp. 3-20, 2012, doi: 10.1109/SURV.2011.113010.00106.

[3] J. Koomey, "Estimating Total Power Consumption by Servers in the U.S. and the World," Lawrence Berkeley National Laboratory, Stanford University Oakland, CA 2007.

[4] A. Berl, E. Gelenbe, M. D. Girolamo, G. Giuliani, H. d. Meer, M. Q. Dang and K. Pentikousis, "Energy-efficient cloud computing," The Computer Journal, vol. 53, no. 7, pp. 1045–1051, 2010.

[5] J. Hamilton, "Cooperative Expendable Micro-Slice Servers (CEMS): Low Cost, Low Power Servers for Internet-Scale Services," presented at the 4th Biennial Conf. Innovative Data Systems Research (CIDR), Asilomar, CA, USA, 2009.

[6] J. Accenture, "Data Centre Energy Forecast. Final Report," Silicon Valley Leadership Group 2008.

[7] M. Eisa, M. Younas, K. Basu and H. Zhu, "Analysis and Representation of QoS Attributes in Cloud Service Selection", presented at the 32nd IEEE International Conference on Advanced Information Networking and Applications (IEEE AINA-2018), Cracow, Poland, May 16-18, 2018

[8] V. K. M. Raj and R. Shriram, "Power aware provisioning in cloud computing environment," in Proc. International Conference on Computer, Communication and Electrical Technology (ICCCET), India, 2011, pp. 6 - 11.

[10] R. Buyya, A. Beloglazov and J. Abawajy, "Energy efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges," presented at the International Workshop on Middleware for Grids, Clouds and e-Science (MGC 2009), 2009.

[11] "Wireless Intel SpeedStep Power Manager: optimizing power consumption for the Intel PXA27x processor family," Intel whitepaper (30057701) 2004.

[12] B. Paolo, M. Avgerinou, and L. Castellazzi. "Trends in data centre energy consumption under the European Code of Conduct for Data Centre Energy Efficiency," Luxembourg: Publications Office of the European Union, 2017.

[13] "ENERGY STAR* System Implementation, Published by Intel with technical collaboration from the U.S. Environmental Protection Agency," Whitepaper (Document No. 321556-001)," [Online]. Available: https://www.energystar.gov/ia/partners/product_specs/program_reqs/Computers_Intel_Whitepaper_Spec5.pdf

[14] A. Merkel and F. Bellosa, "Memory-Aware Scheduling for Energy Efficiency on Multicore Processors," in Proc. Workshop on Power Aware Computing and Systems (HotPower'08), San Diego, CA, USA, 2008, pp. 123–130.

[15] A. T. AlEnawy and H. Aydin, "Energy-Aware Task Allocation for Rate Monotonic Scheduling," in Proc. 11th IEEE Real Time and Embedded Technology and Applications Symp. (RTAS'05), San Francisco, CA, USA, 2005, pp. 213–223.

[16] J. Zhuo and C. Chakrabarti, "Energy-efficient dynamic task scheduling algorithms for DVS systems," ACM Transactions on Embedded Computing Systems (TECS), vol. 7 Issue 2, Feb 2008.

[17] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy and D. Wetherall, "Reducing network energy consumption via sleeping and rate-adaptation," in Proc. 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI'08), Berkeley, CA, 2008, USENIX Association, pp. 323–336.

[18] M. Armbrust, R. G. A. Fox, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica and M. Zaharia, "Above the clouds: A Berkeley view of cloud computing, Electrical Eng. Computer Sci. Dept., Univ. California, Berkeley, CA, Tech. Rep. UCB/EECS-2009-28," Feb 2009.

[19] "Understanding Full Virtualization, Paravirtualization, and Hardware Assist," VMware White paper. [Online]. Available: https://www.vmware.com/techpapers/2007/understanding-full-virtualization-paravirtualizat-1008.html

[20] J. Baliga, R. Ayre, K. Hinton and R. S. Tucker, "Architectures for energy-efficient IPTV networks," presented at Optical Fiber Communication Conference/National Fiber Optic Engineers Conference, San Diego, CA, Mar. 2009.

[21] M. Al-Fares, A. Loukissas and A. Vahdat, "A scalable, commodity data center network architecture," in Proc. ACM SIGCOMM Conference on Data Communication, New York, 2008, pp. 63–74.

[22] V. K. M. Raj and R. Shriram, "Power Aware Provisioning in Cloud Computing Environment," presented at International Conference on Computer, Communication and Electrical Technology (ICCCET 2011), Tirunelveli, India, Mar 2011.

[23] P. Ranganathan, P. Leech, D. Irwin and J. Chase, "Ensemble-level power management for dense blade servers," presented at 33rd Annual International Symposium on Computer Architecture (ISCA), 2006.

[24] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao and F. Zhao, "Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services," presented at the 5th USENIX Symposium on Networked Systems Design & Implementation (NSDI'08), San Francisco, CA, April 2008.

[25] R. Nathuji, C. Isci, E. Gorbatov and K. Schwan, "Providing platform heterogeneity-awareness for data center power management," Cluster Computing, vol. 11(3), pp. 259–271, 2008.

[26] "The green grid metrics: Describing data center power efficiency.," The Green Grid, Tech. Committee White Paper Feb 2007.

[27] R. Bolla, R. Bruschi, F. Davoli and F. Cucchietti, "Energy Efficiency in the Future Internet: A Survey of Existing Approaches and Trends in Energy-Aware Fixed Network Infrastructures," IEEE Communications Surveys & Tutorials, vol. 13(2), pp. 223 - 244, 2011.

[28] R. S. Tucker, R. Parthiban, J. Baliga, K. Hinton, R. W. A. Ayre and W. V. Sorin, "Evolution of WDM Optical IP Networks: A Cost and Energy Perspective," IEEE Journal of Lightwave Technology, vol. 27(3), pp. 243-252, Feb 2009.

[29] M. Yamada, T. Yazaki, N. Matsuyama and T. Hayashi, "Power Efficient Approach and Performance Control for Routers," presented at Green Communications Workshop in conjunction with IEEE ICC'09 (GreenComm09), Dresden, Germany, June 2009.

[30] J. Baliga, R. Ayre, K. Hinton and R. S. Tucker, . Proc. I, San Francisco, CA, 2007., "Photonic switching and the energy bottleneck," in Proc. International Conference on Photonics in Switching, San Francisco, California, 2007, pp. 125-126.

[31] J. Baliga, R. Ayre, K. Hinton, W. V. Sorin and R. S. Tucker, "Energy Consumption in Optical IP Networks," Journal of Lightwave Technology, vol. 27(13), pp. 2391 - 2403, July 2009.

[32] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy and D. Wetherall, "Reducing network energy consumption via sleeping and rate-adaptation," in Proc. 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI'08), Berkeley, CA, 2008, pp. 323–336.

[33] M. Gupta and S. Singh, "Dynamic Ethernet Link Shutdown for Energy Conservation on Ethernet Links," in Proc. IEEE International Conference on Communications (ICC '07), Glasgow, June 2007, pp. 6156 - 6161.

[34] R. Hays, "Active/Idle Toggling with Low-Power Idle," IEEE 802.3az Task Force Group Meeting 2008. [Online]. Available: http://www.ieee802.org/3/az/public/jan08/hays_01_0108.pdf,

[35] W. Fisher, M. Suchara and J. Rexford, "Greening backbone networks: reducing energy consumption by shutting off cables in bundled links," presented at 1st ACM SIGCOMM workshop on Green Networking (Green Networking '10), New Delhi, India, Aug 2010.

[36] I. Fajjari, M. Ayari, O. Braham, G. Pujolle and H. Zimmermann, "Towards an Autonomic Piloting Virtual Network Architecture," in Proc. 4th IFIP International Conference on New Technologies, Mobility and Security (NTMS 2011), Paris, 2011, pp. 1 - 5.

[37] M. Kasbekar, "On efficient delivery of web content," Akamai Technologies 2010. [Online]. Available: http://www.sigmetrics.org/sigmetrics2010/greenmetrics/MangeshKasbekar.pdf

[38] K. Sinha, B. P. Sinha and D. Datta, "An Energy-Efficient Communication Scheme for Wireless Networks: A Redundant Radix-Based Approach," IEEE Transactions on Wireless Communications, vol. 10(2), pp. 550 – 559

[39] "Advanced configuration and power interface specification 4.0a," Hewlett-Packard, Intel, Microsoft, Phoenix and Toshiba, Jan 2019.