Energy Efficiency and Performance of Cloud Data Centers - Which Role can Modeling Play ?

Paul J. Kuehn
Institute of Communication Networks and Computer Engineering (IKR)
University of Stuttgart
Stuttgart, Germany
paul.j.kuehn@ikr.uni-stuttgart.de

ABSTRACT

Resource Virtualization and Load Balancing are main objectives to reduce the power consumption and to improve the performance of large data centers (DC). . The management of Cloud Data Centers (CDC) requires an accurate planning and an efficient use of system resources in order to save energy consumption ("greening"), to provide Quality of Service (QoS), and to meet negotiated Service Level Agreements (SLA). This contribution addresses the question of modeling and the development of generic queuing models for energy-efficient use of resources for dynamic load balancing in virtualized CDCs. Performance models are developed for energy efficiency through automatic Server Consolidation, Dynamic Voltage and Frequency Scaling (DVFS) under Static Load Balancing; Dynamic Load Balancing can be achieved through Virtual Machine (VM) migrations. The analysis of such models provides quantitative performance figures upon which the system operation can be optimized with respect to guaranteed real-time performance and energy efficiency under prescribed SLAs.

Categories and Subject Descriptor

Systems Organization: Distributed Architectures, Real-time Systems. **Information Systems:** Database Performance Evaluation. **Hardware:** Power and Energy.

Keywords

Cloud data centers, server consolidation, load balancing, process migration, VM migration, service level agreements, queuing systems, performance modeling.

1. INTRODUCTION

Cloud Data Centers are increasingly integrated in our ICT infrastructures for storage of application or network configuration data, for efficient searching, for processing of mass data (“Big Data”), as well as for providing application processing functions (Software-as-a-Service (SaaS), "Apps"). The number of processing resources can be reduced considerably through the method of **Virtualization** which leads generally to a more economic use of processing resources. Energy can be saved either through Server Consolidation by switching-off/sleep mode of servers or by throttling down the processor speed through Dynamic Voltage and Frequency Scaling (DVFS) if the current load does not require all fully activated servers of the configured physical server systems. Temporary overload situations, caused through the volatile user load, can be in principle avoided by rejection of new jobs which causes a temporal shift of load to lower loaded periods which worsens the QoS and energy efficiency. A much better solution provides the method of Load Balancing (LB) through job (or VM) migration when a new job can be transferred to another server system of the local DC or to a server system of a foreign DC when all servers of the local DC are currently occupied and as long as the negotiated Service Level Agreement (SLA) is still met (**"Dynamic" LB**). VM migrations can be initiated either (1) at the instant of arrival or (2) during processing (“VM Life Migration”); the latter one is currently intensively discussed and requires additional pre-copy or post-copy actions. Reactions to sporadic overload situations by re-configuring the physical server systems are unreasonable as they require too much time and energy and are being left to adapt to slowly changing average load variations (**"Static" LB**). Larger periods of overload should be avoided as they may lead to server outages triggered by heat sensors to protect physical device damage. In the past, system managers were conservative with respect to dynamic resource management such as server consolidation; but the situation has changed through technological improvements and governmental regulations (Energy Policy Act) which enforces a higher dynamic system monitoring and management.

DCs have been used in the past mostly for data storage and powerful job processing without strict latency restrictions. Advanced CDC applications require fast access to quickly changing state data, network loads or for fast reactions to breakdowns or emergency events with a strictly bounded reaction time (**Real-time Performance**). Applications of this kind are in Software Defined Networking (SDN) with Open Flow in the Future Internet and in Smart Grid environments. For that reason we will focus our interests on both Energy Efficiency **and** Real-time Performance simultaneously. The complexity of the problem, originating from huge numbers of servers, operating and management functions, load characteristics under consideration of QoS and SLA conditions makes experimentation quite difficult for parametric studies. We attempt, therefore, a modeling approach where we restrict ourselves to generic functional modules as one or several groups of server systems and the most important system and load parameters in order to understand their principal influence on the energy efficiency and on the performance. These models can be analyzed either mathematically or they can be evaluated by simulation techniques. Finally, modeling methods need to be validated against results derived from experimental benchmarks from real systems; these are costly and generally less appropriate for research, design and resource planning.

E2DC, June 21-24, 2016, Waterloo, ON, Canada
© 2016 ACM. ISBN 978-1-4503-4421-0/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2940679.2940688

Data Centers have been around for several decades and have been subjected to energy efficiency and performance studies. New technologies provide powerful multi-core/ multi-processor systems with huge storage capacities where the access is supported through powerful storage area networks. Compute power and storage capacity is still growing exponentially according "Moore's Law". For the operational management DCs are topologically structured into physical server systems which are interconnected through edge and aggregation switching racks forming server farms which are controlled flexibly by operating software. Distant DCs are interconnected through a high-speed packet network (internet). This allows a modular decomposition of Cloud Data Centers into subsystems as physical server clusters ("Server Systems", SS) and interconnection network links between DCs where the arriving jobs ("Virtual Machines", VM) are characterized by their processing and storage demands) which are scheduled for service by VM Monitor ("Hypervisor") and Cluster Control software.

Research on cloud computing has started just about 15 years ago and has increased since then tremendously according to the rapid acceptance of cloud technology. The applied operational strategies follow certainly the wisdom of operating system job scheduling, but the complex interaction with non-IT equipment (as cooling systems and power supply) requires experimentally approved strategies. DCs are today characterized by an aggregated criterion, the so-called "DC Power Usage Effectiveness" (PUE) which expresses the ratio between the "IT based energy consumption" and the "Total DC energy consumption"; the PUE has varied typically between 1.5 and 3, but current values of Google and Facebook DCs direct to values between 1.0 and 1.2 which reflects the influence of greening strategies. Another totally different strategy has been followed by placing huge globally operating DCs into arctic regions where the annual average temperature level is significantly lower than in the lower latitude regions around the equator.

Dynamic load balancing aims at a better use of the current resource configuration or for reasons to meet real-time service conditions with respect to the response time. Load balancing is usually achieved through **static** server system configurations, but this is not adaptable to quick load variations. Virtualization allows for a more economic use of servers through server consolidation, but it may cause severe server "hot spots" of power consumption originating from the dynamic workload offered to the DC which may overload the DC cooling systems and, thus, leading to system down times and a worsened DC PUE. For this reason **dynamic** (automatic) server consolidation methods are required by which the frequency of server activations/deactivations can be reduced which contributes to a much smoother server usage dynamics and smaller hot spot risk, but which should still meet the real-time SLAs. Such methods can be additionally complemented with the Load Balancing strategies between different Server Systems (SS) of the home DC or a foreign DC.

The remaining part of the paper ia as follows: In Section 2 the state of the art in the area of CDCs with server consolidation and load balancing mechanisms will be outlined. Section 3 addresses modeling aspects of CDC with respect to methods of server consolidation and load balancing under the aim of performance with respect to prescribed QoS and SLA requirements. In Section 4 an overview is given on solution approaches for their mathematical analysis. More details on the analysis methods can be found in the literature references. The paper closes with concluding remarks on the use and capabilities of performance modeling for CDCs.

2. STATE OF THE ART

The state of the art in DC technology can be found in standard text books on computer science and computer engineering; more specific information on server technologies and performance can be found in [1-4] and on White Papers of vendor companies, see, e.g. [5-7]. Analytic models for server consolidation are mainly based on multi-server queuing models. Queuing theory, also known as Teletraffic Theory, has more than 100 years of history, starting with the works of A.K. Erlang 1917 on the n-server loss system and the n-server delay system under Markovian traffic conditions, i.e. negative-exponentially ("memoryless") interarrival and service processes. A huge amount of publications has appeared since then on single stage queuing systems, queuing networks and application-specific service models. Fundamental publications have appeared in the Journal of Operations Research and in many other Journals of IEEE, ACM and on international conferences as the Int. Teletraffic Congresses (ITC, since 1955) or of ACM Sigmetrics. Queuing theory has addressed a wide spectrum of applications and is recognized as a powerful methodology to model and anlayze service systems (computer systems, communication networks) operating under stochastically varying arrival and service processes and system operation schedules, see, e.g., [8-9]. Many results and methods have been applied and are part of engineering procedures for resource sizing and network planning. In the following references are only made to publications in the context of DC applications.

Server consolidation without activation overhead and without hysteresis-based activations and deactivations can be modeled by simple n-server delay systems of the type GI/G/n, especially under specific arrival processes of GI and special service processes G (Markovian, Phase-type , Deterministic), see [8-9]. Results are on the average server utilization Y ("carried traffic"), the probabilities of delay W and loss B (in delay-loss systems), the average delay and on the distribution function (DF) W(t) of the waiting times. From these results it is easy to find the energy efficiency for DCs under server consolidation operation without or with server activation times after a switched-off or after a sleeping period. Various papers have appeared on such models under specific model assumptions without server setup times [10] and with server setup times [11]. DVFS can easily be modeled by a state-dependent service rate μ.

Server consolidation models with a hysteresis-based operation are adequate to model memory on the past. The first basic model has been applied in context with overload control strategies for computer controlled switching systems in our own research group [12]; the hysteresis avoids frequent activations of the overload strategy at the critical point between normal load and overload. The principle of hysteresis has been applied to server consolidation in three seminal theoretical contributions by J. Keilson and L. Golubchik [13-15] where exact solutions have been derived by the methods of Green's Function on half-lettice Markov Chains and by Markov Chains with Ergodic Compensation Rates, respectively. The exact solutions are difficult to exploit and have been applied to very small systems only. In [16-20] a systematic method for developing hysteresis-based consolidation algorithms for automatic DC applications has been suggested for queuing systems which are controlled by a Finite State Machine (FSM). This approach allows a fast recursive computation of the probabilities of state and is not only computationally feasible, but considers also given SLAs based on prescribed mean or on percentiles of the response times of arriving VM jobs, hot and cold stand-by server operation and DVFS. The FSM-based algorithm can easily be implemented in the server management control software.

Dynamic LB models are based on the principle of cooperation between different server systems in order to reduce overload situations which result out of the volatile character of the job arrivals and their processing times. Load balancing is achieved through **VM migration** to a currently under-utilized server system of the home DC or even a foreign DC. This principle is by no means a new concept and has been intensively applied for dynamic routing in telecommunication networks, known as "overflow", see references [21-23]. VM migration requires, however, server systems with buffering **and** overflow; finite buffer queuing models with overflow have been analyzed by various methods through a full state analysis or by the characterization of the overflow traffic by higher moments [24-28].

Dynamic LB through **VM Life Migration** is currently a hot topic. This method may be useful for VMs with large processing times during which the overall system states may have caused bottleneck effects; shifting such jobs to another physical server system may help to overcome this situation, but it requires additional overhead through pre-copy or post-copy actions. The pre-copy scenario requires at first a resource reservation at the target server and some pre-copy actions to transfer the "dirty" memory pages before the job is continued at the target server. In the post-copy scenario the minimum register and I/O state is transferred to the target server to continue processing during which the residual pages are copied from the source server to the target server. More details can be found in [4, 29-31]. Analytic performance studies are still in an initial phase.

3. MODELING

3.1 Modeling of Server Consolidation

We consider a queuing model with n servers and a finite buffer space for up to s jobs (or VMs), c.f. Fig.1 We consider this queuing system as being controlled by a Finite State Machine (FSM), whose state is denoted by two parameters (x,z). where x indicates the number of actively processing servers and z the number of arrived and buffered jobs where $0 \leq x \leq n$, $0 \leq z \leq w^{(x)}$, x = 0,1,...,n, with $w^{(0)} = 0$ and $w^{(n)} = s$. s is the total buffer capacity required for meeting the SLA in the worst case of starting to wait from the last buffer position. If a new job arrives at state (n,s) it will be rejected (lost) as its SLA cannot be met. For models without hysteresis no acceptance levels for new jobs in server state x are required as long as x < n; for x = n the acceptance level for a new arriving job is $w^{(n)}$-1 = s-1, i.e., we describe the system simply by the number of jobs in the system. For models with hysteresis the threshold values $w^{(x)}$ indicate the maximum number of jobs which can be in the queue in server state x; they are staggered by constant values w, i.e., $w^{(1)} = w$, $w^{(2)} = 2w$, ..., $w^{(n-1)} = (n-1)w$ and define the maximum number of VMs which can be queued in state x. The value of w is chosen such that a new job which is arriving in the worst case at state (x, $w^{(x)}$ -1) and needing an average service time h = 1/μ will have to wait no longer than the time $t_{W0} = hw$ until its service begin when the mean response time of waiting jobs is used as QoS value under the strict queue disipline FIFO (First-In, First-Out). This time is constituted by the time until the first of the jobs met in service terminates (minimum of all residual service times) and the time that all jobs being met in the queue at the arrival of the considered new job have started service. If t_{W0} is the required mean response time as QoS value, w can be chosen accordingly as $w = t_{W0}/h \geq 1$. If a new job arrives at state (x, $\geq w^{(x)}$) a new server is activated immediately and the new service rate is (x+1)μ, i.e., this job has to wait $(w^{(x)} +1)h/(x+1) \leq t_{W0}$ in average. These results hold exactly only for jobs with negative-exponentially distributed service times under a Markovian job arrival process regime (G = M), i.e., a Poisson process with negative-exponentially distributed interarrival times with mean interarrival time 1/λ, but can be extended to general service times for n =1 and approximately estimated for multi-server systems.

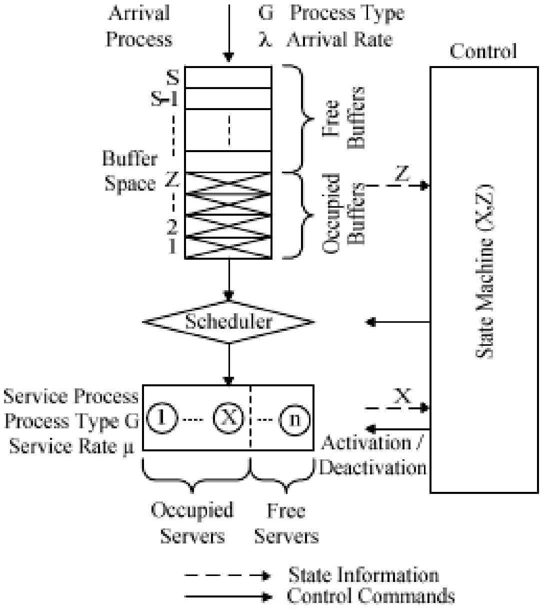

Fig.1. Queuing Model for Server Consolidation

The described model of an n-server system with a staggered hysteresis in the upward direction of the state-transition-diagram (STD) reduces the number of server activations by buffering new arriving jobs as long as their SLA is still met and activates a new server only after crossing the threshold value $w^{(x)}$ of buffered jobs in state x. Deactivations of servers are only applied, when a server becomes idle and the queue is empty, which guarantees the fastest job processing speed in state x. Buffering of jobs arriving in states x < n effects a lower activation/deactivation rate of servers and, thus, smooths the server activities and saves energy for activation/deactivation. The model was introduced first in [16-18] and has been extended later by explicit consideration of finite activation times, hot and cold stand-by of servers and even Dynamic Voltage and Frequency Scaling (DVFS) [20]. Another effect of the consideration of the mean response times as SLA is a rather flat and almost constant average response time over a large range of server occupancies between zero and the capacity limit. The SLA criterion "mean response time" has also been extended to guarantee response time percentiles, c.f. [19].

3.2 Modeling of Load Balancing

A cloud data center is configured by multiple multi-server systems. In the following, two generic models will be introduced for two multi-processor server systems (SS) with a limited buffer capacity to control the SLA which is defined by the response time of accepted jobs. The two SSs serve jobs cooperatively as they can accept jobs of the other SS by job (or VM) migration through mutual overflow (LSSF), or by assigning arriving jobs instantaneously to that SS which guarantees the shortest response time (SRTF), respectively.

3.2.1 Local Server System First (LSSF)

The generic model for the LSSF strategy is shown in Fig. 2.

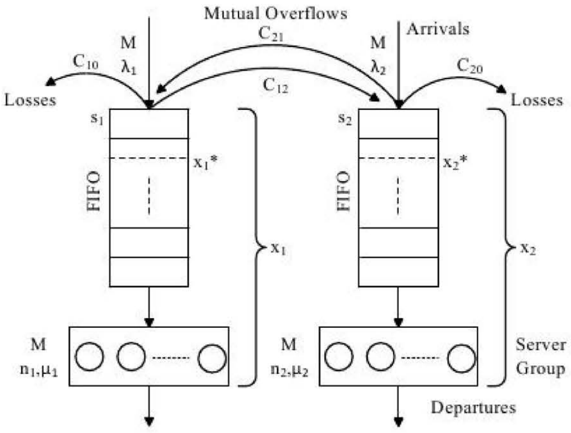

Fig. 2 . System Model for two Server Systems with Mutual Overflow

The operation of this model is as follows:

* Each SS serves its arrivals as long as an idle server is available. If all servers are occupied, arriving requests are assigned to the local queue and buffered there. Buffered jobs will be served in the strict FIFO mode.

* If all buffer places within the queue of the local SS are occupied, an arriving request may overflow (or be migrated) to the complementary SS to be served there if the given SLA can be met.

* The buffer capacities s_1 and s_2 and migration thresholds x_1* and x_2* for the overflow to a complementary DC are chosen such that the SLA of accepted arrivals will always be met.

* SLAs are defined by a threshold x_i* for the mean response time $t_{W,T}$ of a request which has to wait and which is assigned to queue i, i = 1,2. . More specifically, the SLA can also be defined to meet a prescribed percentile ε for the maximum delay an arrival will have to suffer.

The model parameters are defined as follows for the server group i, i = 1,2:

$n_{i,}$, s_i number of servers, buffer capacity

λ_i , μ_i arrival rate of new jobs, service rate of a job

($\mu_i = 1/h_i$, h_i mean service time)

M traffic type for inter-arrivals and services (Markov)

x_i total number of jobs at SS i

x_i* acceptance threshold for migrating jobs

C_{i0} conditions for rejection (loss) of an arrival at SS_i

C_{ij} conditions for migration of a new job at SS_i to SS_j

For strategy LSSF a new arriving job will always be buffered in the queue of the local SS as long as the local queue is not completely filled. If the buffer is filled up and if there is space in the complementary local or foreign SS queue the arriving job overflows to the complementary queue according to logical conditions C_{ij}; if the state at the instant of arrival is such that the SLA cannot be met, the new request will be lost defined by logical conditions C_{i0}, i = 1,2.

The queue capacities s_i are dimensioned according to the SLA condition that the mean response time t_{Wi} in each SS does not exceed a threshold $t_{W,T}$ in the worst case, i.e., for Markovian service times

$$s_i/n_i\mu_i \leq t_{WT}, \quad i = 1,2. \qquad (1)$$

The state thresholds x_i* are dimensioned such that the overhead time t_M for a process migration from one (overloaded) SS to the complementary SS is considered such that SLA is still met for a migration from SS_i to SS_j, i.e.,

$$t_M + (x_j^* - n_j + 1)/n_{j*}\mu_j \leq t_{W,T} \quad \text{or}$$

$$x_j^* \leq (t_{W,T} - t_M) *\cdot n_{j*}\mu_j + n_j - 1, \quad j = 1,2 \qquad (2)$$

For real-time services with a tighter SLA, the dimensioning of the acceptance/migration level can be derived from the complementary distribution function (DF) of accepted jobs $W_i (> t) / W_i$ acc. to

$$W_i (> t_{Th}) / W_i = P\{T_{Wi} > t_{Th} \mid T_{Wi} > 0\} \leq \varepsilon \qquad (3)$$

where T_{Wi} denotes the random variable of the waiting time of an arriving job at SS_i, t_{Th} the response time threshold for an arriving and accepted job and ε the delay percentile, i.e., that the response time T_{Wi} of an arriving and accepted job exceeds the threshold t_{Th} only with a prescribed probability ε. The acceptance level x_i* of an i-job at DC_i follows from the worst-case arrival state $x_i^* \geq n_i$ at SS_i from (3); in that case, the response time distribution is an Erlangian DF of degree $k_i = x_i^* - n_i + 1$, i.e., the sum of k_i exponential phases each with mean h_i/n_i the arriving job has to wait until service begins. If the i-job arrives at state $x_i > x_i$*, it might be migrated to the other data center SS_j, $j \neq i$, if the current state there is $x_j \leq x_j$*. The condition for the value of x_j* follows from the worst-case response time for the migrated job which is composed of the random job migration time T_M and the waiting time at SS_j; the latter one follows again an

Erlangian DF of degree $k_j = x_j^* - n_j + 1$, i.e., from k_j exponential phases each with average h_j/n_j. As the DF of T_M is generally not known, a proper assumption has to be made; in the most optimistic case T_M could be assumed to be constant t_M; in that case, the response time DF is a shifted Erlangian DF. If the conditions at both DCs cannot be met, the arriving job has to be rejected.

3.2.2 Shortest Response Time First (SRTF)

A special case of this strategy has already gained some research interest, better known under the name "Join the Shortest Queue" (JSQ), see, e.g., references [32-37]. By this strategy, several single or multi-queue server systems with *identical* service time distributions and *identical* server numbers are considered. An arriving job is assigned to the currently shortest queue; in case of identical queue lengths, an equal probabilistic assignment is applied. The exact analysis is not completely known even under Markovian traffic assumptions. The attraction of this policy is its guarantee of the absolutely shortest delay and its intrinsic strategy to balance the queue lengths instantaneously.. We will consider a **generalized** strategy of JSQ as an alternative to the LSSF policy above under a modified condition acc. to the propositions of this paper, i.e., to guarantee given SLAs by scheduling the arriving job to that SS which provides the Shortest Response Time (SRTF) and allowing heterogeneous server groups with different numbers of servers, different buffer capacities, and even different server speeds. Fig. 3 shows the server system arrangement.

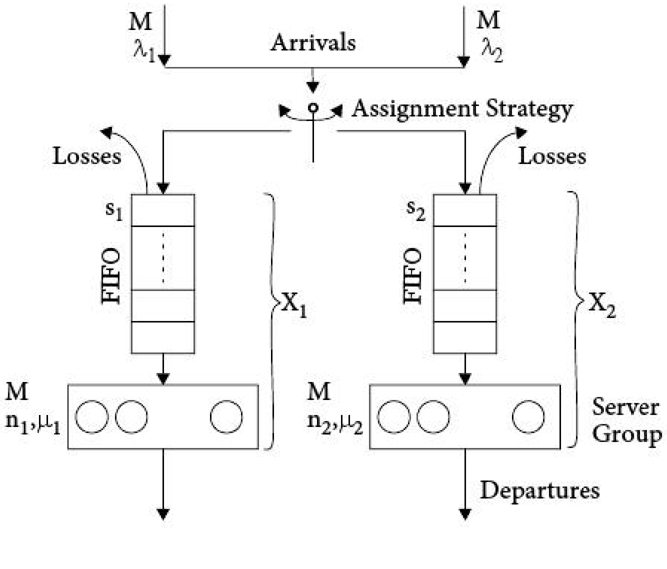

Fig. 3. System Model for two Server Systems under SRTF

The operation of the queuing model is as follows:

* An arriving job (VM), is assigned to that server system with the smallest number of busy servers, if at least one server is still idle in each SS. In principle, the arriving job could be assigned to any SS as the response time is zero in that case, but the assignment to that SS with the smaller number of busy servers enforces a better load balancing.

* If only one SS has fully occupied servers, the arriving job is assigned to the complementary SS.

* If all servers of both SSs are fully occupied the arriving job is assigned to that SS which provides the shortest mean response time, i.e. acc. to the following condition:

$$\min\{(x_1 - n_1 + 1)h_1/n_1\ ,\ (x_2 - n_2 + 1)h_2/n_2\}\ .$$

* If both queues are completely filled, the arriving job will be lost.

The SRTF strategy provides the best load balancing, but requires a higher overhead as each arriving job has to be scheduled according to the condition above. However, this strategy outperforms all other scheduling strategies.

4. MATHEMATICAL ANALYSIS

The mathematical analysis is limited to Markovian assumptions, i.e., to negative-exponentially distributed inter-arrival and service times. The system operation is represented by the method of State-Transition-Diagrams (STD) where states are represented by nodes (vertices) and transitions are represented by arcs (edges) of a directed graph. This representation holds generally, independent of the specific traffic assumptions. Transitions are annotated by transition rates in case of Markovian traffic assumptions where the STD acts as representation of a multi-dimensional Markov Chain. For all models introduced in Section 3.1 exact solutions for the probabilities of state have been derived through a novel recursive algorithm for the hysteresis-based server consolidation models with hot and cold stand-by and DVFS [20]. For the load balancing models introduced in Section 3.2 no closed-form solutions exist; the state probabilities are solved exactly from the equilibrium balance equations through numerical computations; approximate solutions have been developed based on simplifying hypotheses with high accuracy . From the probabilities of state the most important QoS metrics are derived straightforwardly as

* the probabilities of delay, loss, and job migration

* the average server utilizations and average queue lengths

* the average and the distribution function of response times

* energy efficiency metrics dependent on the system load

For models with general distribution types of arrivals and services (G) the performance can be analyzed by computer simulations using the OMNeT simulation tool to study the effects of non-Markovian traffic assumptions.

5. CONCLUSIONS

In this contribution current research activities have been reviewed for the analysis of cloud data centers controlled under operational schedules for energy efficiency and load balancing for real-time performance. Energy efficiency can be achieved through server consolidation schemes through which servers are either switched-off, operated in a low power sleep mode or being throttled down by dynamic voltage and frequency scaling and which need additional energy and time for activation or warm-up, respectively. Load balancing mechanisms aim at defeating sporadic overload situations and can be modeled by VM migration between different server systems. For VM migration two novel models have been suggested, which can be analyzed exactly under Markovian traffic assumptions. Modeling methodology and their results are adequate to describe complex cloud data center configurations and to study their real-time performance quantitatively under QoS and SLA conditions. The results of such studies provide a deeper insight in performance and effectiveness on the energy efficiency. There is no other method available by which these challenges can be met.

6. REFERENCES

[1] A. Beloglazov, R. Buyya, Y. Ch. Lee, A. Zomaya, A Taxonomy and Survey of Energy-efficient Data Centers and Cloud Computing Systems. In: M. Zollkowitz (Ed.), Advances in Computers, Elsevier, San Francisco, 2011 (51p.).

[2] H. Zhang, Research on the Influence of Cloud Computing on the Virtual Operation Performance Management. In: 7th Int. Conf.on Computer Science & Education (ICCSE), Melbourne, 2012, pp. 235 - 238.

[3] K. Ye et al, Virtual Machine Based Energy-efficient Data Center Architecture for Cloud Computing: a Performance Perspective, in: Proc. of 2010 IEEE/ACM Int. Conf. on Green Computing and Communications, pp. 171 - 178.

[4] J. Sekhar, G. Jeba, S. Durga, A Survey on Energy-efficient Server Consolidation through VM Life Migration. Int. J. Adv. Eng. Technol. (IJAT) 5 (1) 2012, pp. 515 - 525.

[5] - Data Center Power: Managing for Energy Efficiency and Cost Savings, White Paper STI-100-013, Server Technology, Nov. 2013,

[6] - Advances in Power and Environmental Monitoring for Increasing Efficiency in the Data Center, White Paper STI-100-014, Server Technology, Oct. 2014.

[7] S. Niles, P. Donovan, Virtualization and Cloud Computing : Optimized Power, Cooling, and Management Maximizes Benefits. White Paper 118 (rev. 4), Schneider Electric UK. http://news.angelbc-mail.com.

[8] L. Kleinrock, Queuing Systems. Volumes I and II, John Wiley and Sons, New York, 1975.

[9] H. Kobayashi, B.L. Marks, System Modeling and Analysis - Foundations of System Performance Evaluation, Pearson Prentice-Hall, Upper Saddle River, N.J., 2009.

[10] H. Khazaei, J. Misic, V.B. Misic, Performance Analysis of Cloud Computing Centers Using M/G/m/m+r Queuing Systems, IEEE Trans. Parallel Distrib. Syst. 23 (5) 2012, pp. 936 - 943.

[11] M. Harchol-Balter, M/G/k with Staggered Setup, Operations Research Letters, Vol. 41, Nr. 4, July 2013, pp. 317 - 320.

[12] P. Tran-Gia, Overload Problems in Stored-Program Controlled Switching Systems - Modeling and Analysis.. Doctoral Dissertation, Univ. of Siegen.36 th Report on Traffic Theory, Institute of Switching and Data Tchnics, Univ.of Stuttgart, 1982.

[13] O.C.Ibe, J. Keilson, "Multi-Server Threshold Queues with Hysteresis", J. Perform. Eval. 21 (1995), pp. 185 - 213.

[14] J.C.S.Liu, L. Golubchik, " Stochastic Complement Analysis of Multi-Server Threshold Queues with Hysteresis", J. Perform. Eval. 35 (1999), pp. 19 - 48.

[15] C.-F.Chou. L.Golubchik, J.C.S.Liu,"Multi-Class, Multi-Server Threshold- based Systems: A Study on Non-instantaneous Server Activation, IEEE Trans. Parallel Distrib. Systems 18(1) 2007, 96-110.

[16] P.J. Kuehn, "Systematic Classification of Self-Adapting Algorithms for Power-Saving Operation Modes of ICT Systems", Proc. 2nd ACM Conf. on Energy-Efficient Computing and Networking (e-Energy 2011), New York, N.Y., USA, May 30- June 1, 2011. http://edas.info/N9577.

[17] P.J. Kuehn, M. Mashaly, "Performance of Self-Adapting Power-Saving Algorithms for ICT Systems", IFIP/IEEE Symp. On Integrated Network and Service Management" (IM 2013), Ghent, Belgium, May 27 – 30, 2013, IEEE Xplore.

[18] M. Mashaly, P.J. Kuehn, "Load-Balancing in Cloud-Based Content Delivery Networks Using Adaptive Server Activation/Deactivation", IEEE Conf. ICET, Cairo, Oct. 10 – 11, 2012.

[19] M. Mashaly, P.J. Kuehn, "Modeling and Analysis of Virtualized Multi-Service Cloud Data Centers with Automatic Server Consolidation and Prescribed Service Level Agreements", 23rd Int. Conf. on Computer Theory and Applications (ICCTA 2013), Alexandria, Egypt, Oct. 29 – 31, 2013.

[20] P.J. Kuehn, M. Mashaly."Automatic Energy-Efficient Management of Data Center Resources by Load-Dependent Server Activation and Sleep Modes", Rev. Version of a Contribution to 2nd Conf. On Energy-Efficient Data Centers (E^2DC 2013), Berkeley, Cal., USA, May 21, 2013. J. Ad Hoc Networks 25(2015), pp. 497 – 504, Elseviers, 2015.

[21] H. Akimaru, K. Kawashima, "Teletraffic – Theory and Applications", Chapter 5 and References there. Telecommunication Networks and Computer Systems Series, M Gerla, A. Lazar, P. Kühn, H. Takagi, (Eds.), Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[22] .R. Krieger, B. Müller-Clostermann, M. Sczittnick, "Modeling and Analysis of Communication Systems Based on Computational Methods for Markov Chains“, IEEE Journal on Selected Areas in Communications, Vol. 8, No. 9, Dec. 1990, pp. 1630-1648.

[23] T.-J. Lu, W. Yue, T. Hasegawa, "A Mutual Overflow System with Simultaneous Occupation of Resources", J. Operations Research Soc. of Japan, Vol. 41, No. 1, March 1998, pp. 81-90.

[24] E. Cinlar, R.L. Disney, "Streams of Overflows from a Finite Queue", Opns. Res. 15, 1967, pp. 131-138.

[25] U. Herzog, P. Kühn, "Comparison of Some Multiqueue Models with Overflow and Load Sharing Strategies for Data Transmission and Computer Systems", Proc. Int. Symposium on Computer Communications Networks and Teletraffic, Polytechnic Institute of Brooklyn, N.Y., 1972, pp. 449-472.

[26] J. Matsumoto, Y. Watanabe, "Theoretical Method for the Analysis of Queuing Systems with Overflow Traffic", Trans. IECE Japan, B-64, 1981, pp. 536-543.

[27] F. Machihara, "On the Overflow Process from the $PH_1 + PH_2/M/S/K$ Queue", Proc. 10th Int. Teletraffic Congress (ITC), Montreal, 1983, Session 4.1, Paper # 5, pp. 1-7.

[28] K.S. Meier-Hellstern, "The Analysis of a Queue Arising in Overflow Models", IEEE Trans. on Comm., Vol. 37, No. 4, 1989, pp. 367-372.

[29] R.W.Ahmad et al, "A Survey on Virtual Machine Migration and Server Consolidation Frameworks for Cloud Data Centers", Int. Journal of Networks and Computer Applications, Vol. 52, 2015, pp..11-25.

[30] V.Shrivastava et al, "Application-aware Virtual Machine Migration in Data Centers", IEEE INFOCOM 2011, Mini-Conference, pp. 66-70.

[31] W. Voorsluys et al, "Cost of Virtual Machine Life Migration in Clouds: A Performance Evaluation", Springer Lecture Notes in Computer Science, LNCS No. 5931, 2011 (12 pages).

[32] Hwa-Chun Lin, C.S. Raghavendra, "An Analysis of the Join the Shortest Queue (JSQ) Policy", Proc. 12 th Int.. Conf. On Distributed Computing Systems, Yokohama, Japan, 1992, pp. 362 – 366.

[33] Hwa-Chun Lin, C.S. Raghavendra, "An Approximate Analysis of the Join the Shortest Queue (JSQ) Policy", IEEE Trans. On Parallel and Distributed Systems, Vol. 7, No.2, pp 301 – 307.

[34] V. Gupta, M.Harchol-Balter, K. Sigman, W. Whitt, Analysis of Join-the-Shortest-Queue Routing for Web Server Farms", J. Performance Evaluation Architecture, Vol. 64, Issue 9 – 12, Oct. 2007, pp. 1062 – 1081.

[35] S.Abimannan, K. Durai, A.V.Jejahumar, S. Krishnaveni, "Join-The-Shortest Queue Policy in Web Server Farms", Global Journal of Computer Science and Technology, Vol. 10, June 2010, pp. 39 - 45.

[36] I.J.B.F. Adan, O.J. Boxma, J.A.C. Resing, "Queuing Models with Multiple Waiting Lines", Queuing Systems No. 37, 2001, pp. 65 - 98.

[37] M. Harchol-Balter, "Scheduling in Server Farms", Talk Slides, Carnegie-Mellon University, harchol@cs.cmu.edu.