Copyright Notice

Institute of Communication Networks and Computer Engineering
University of Stuttgart
Pfaffenwaldring 47, D-70569 Stuttgart, Germany
Phone: ++49-711-685-68026, Fax: ++49-711-685-67983
Email: mail@ikr.uni-stuttgart.de, http://www.ikr.uni-stuttgart.de

DVFS-Power Management and Performance Engineering of Data Center Server Clusters

Paul J. Kuehn
University of Stuttgart
Stuttgart, Germany
paul.j.kuehn@ikr.uni-stuttgart.de

Maggie Mashaly
German University in Cairo
Cairo, Egypt
maggie.ezzat@guc.edu.eg

Abstract*—Dynamic Voltage and Frequency Scaling (DVFS) is a method to save energy consumption of electronic devices and to protect them against overheating by automatic sensing and adaptation of their energy consumption. This can be accomplished either on the program instruction level for electronic devices or on the task or job level for server clusters. This paper models DVFS on the job level and through which Service Levels Objectives can be guaranteed with respect to prescribed mean or quantiles of service delays according to given Service Level Agreements (SLA) between user and service provider. The two parameters V (voltage) and f (frequency) cannot be changed independently of each other; typically only several combinations of V and f values are implemented in hardware for several power states. In this paper a novel analysis of operating DVFS is suggested for Server Clusters of Cloud Data Centers (CDC) under prescribed bounds of service level objectives which are defined by SLAs. The method is based on the theory of queuing models of the type GI/G/n for a server cluster to establish a relationship between SLA parameters and the power consumption and is performed for the example of the Intel Pentium M Processor with Enhanced SpeedStep Power Management. As result of this method precise bounds are provided for the load ranges of service request rates λ for each power mode which guarantee minimum power consumption dependent on given SLA values and job arrival and service statistics. As the instantaneous load in a CDC can be highly volatile the current load level is usually monitored by periodic sensing which may result in a rather high frequency of DVFS range changes and corresponding overhead. For that reason an automated smoothing method is suggested which reduces the frequency of DVFS range changes significantly. This method is based on a Finite State Machine (FSM) with hysteresis levels.

***Keywords*— Cloud Data Centers, Energy Efficiency, Dynamic Voltage and Frequency Scaling, Modeling, Performance Evaluation, Queuing Theory, Optimum Operation Ranges, Automatic Power Management.**

I. INTRODUCTION

Cloud Data Centers (CDC) have become an enabling system component for virtualized network operation of Cyber-Physical Systems. Their energy consumption for computing and cooling is growing fast and amounts already to about 1.5 % of the global energy consumption. The main energy is consumed through computations by servers for processing, data storage, and memory access. Cooling is required to protect the microelectronic parts against overheating and outage. According to experience, the energy consumption of the electronic parts of a data center amounts typically to about one third of the total energy consumption, expressed by the so-called "Power Usage Effectiveness" (PUE), defined as the fraction of the power consumption of the whole data center and of the electronic devices only [1]. By more efficient usage of the electronic parts both, electronic and cooling power consumptions, can be reduced simultaneously. During the last two decades much efforts have been undertaken on the "greening" of CDCs by quite different approaches: 1. Optimization of system parameter settings, 2. Dynamic activations/deactivations and sleep mode operations of CDC servers (Server Consolidation), 3. Dynamic load assignments among different server clusters or DCs (Load Balancing), 4. Scheduling methods or Virtual Machine (VM) Migrations to under-loaded server groups of CDCs, or 5. Dynamic Voltage and Frequency Scaling (DVFS). Methods for analyzing DVFS are theoretical studies on operational strategies, performance modeling and optimization, or experimentally by use of simulation tools or by testbed benchmarks of server clusters or mobile devices. For an overview on main contributions of energy efficiency operation methods we refer to [2-3]. For studies on DVFS specifically we refer to references [4-7]. For device-specific DVFS operations, testbed and simulation experiments we refer to [8-11]. To the best knowledge of the authors, no study has been addressing the theoretical relationship between the load and strict SLA restrictions on the task or job level.

In this paper an analytical performance of a queuing model of the general type GI/G/n is used for the evaluation of a cloud server cluster operating under the DVFS load-dependent strategy, being operated automatically by the cluster operating system to meet either mean values or quantiles of prescribed threshold values on service delays. As the dynamic load of cloud data centers may be highly volatile, the operating system has to monitor the load level. Load variations could result in frequent changes of V and f and corresponding additional processing and time overhead. To reduce such load range changes significantly an automatic control is proposed based on a Finite-State Machine model for the system state (X,Z), where X indicates the actual number of busy cluster servers and Z the actual number of waiting jobs. The FSM is based on a two-dimensional hysteresis state transition diagram with SLA-dependent thresholds.

ISBN 978-3-903176-13-3 © IFIP

The remaining parts of this paper are as follows: In Section II the properties of DVFS will be reminded based on the example of the Intel Pentium M Processor and prescribed Service Level Agreements (SLA) between tenants and the CDC operation management. To connect these two totally independent aspects we will model the DVFS problem in Section III by methods of Queuing Theory using an n-server cluster queuing model of type GI/G/n with general job arrival processes (GI) and general type of service processes (G) which provides the relation between the negotiated SLA requirements on response time, and the server cluster performance. Through this method we establish the optimum operating load ranges for minimum power consumption of the data center server cluster. Section IV addresses the problem how DVFS can be implemented in the Operating System of the Server Cluster for an automated operation with very low frequencies of power state changes in case of highly fluctuating service requests. In Section V the results are discussed and the final Section VI concludes the paper with a summary and an outlook.

II. DVFS AND SLAS

First, the power consumption P of microelectronic devices as CMOS-transistors follows approximately the law $P \sim C \cdot V^2 \cdot f$, where C denotes the capacitance of the transistor, V the supply voltage, and f the clock frequency. For energy saving purposes this relation suggests to reduce the supply voltage as much as possible; this, however, requires more time for charge exchanges, which results in a lower clock operating frequency f. Below, we will consider the special processor Intel Pentium M as a typical example which will be based for the studies in this paper.

According to the original Intel White Paper on the Enhanced SpeedStep Technology for the Pentium M Processor [9] 6 Power States P_i, i = 0,1,...,5, are distinguished together with their clock frequencies f_i and supply voltages V_i , c.f. Table I. Assuming an average processing time of h_0 = 1 s as time unit for one CDC service request during state P_0 we have completed Tables and Charts of [9] for this request by the required average processing times h_i in seconds, the power consumptions P_i of this job in Watts (W), and its energy consumption E_i in Ws, i = 0.1....,5.

TABLE I. POWER STATES OF THE INTEL PENTIUM M PROCESSOR

P-State	*Frequency f_i*	*Voltage V_i*	*Proc.Time h_i*	*Power P_i*	*Energy E_i*
P_0	1.6 GHz	1.484 V	1.00 s	25 W	25 Ws
P_1	1.4 GHz	1.420 V	1.25 s	15 W	17 Ws
P_2	1.2 GHz	1.276 V	1.50 s	10 W	13 Ws
P_3	1.0 GHz	1.164 V	1.75 s	8 W	10 Ws
P_4	0.8 GHz	1.036 V	2.00 s	7 W	8 Ws
P_5	0.6 GHz	0.956 V	2.25 s	6 W	6 Ws

The implementation of DVFS is accomplished by 2 processor registers named IA32_PERF_CTL and IA32_PERF_STATUS through "Get_State" and "Change_State" commands. For the decision on state changes a power manager and a processor driver are responsible. Decisions can be based on various inputs as user power policy, processor utilization, battery level (in portable devices), thermal conditions, or specific events. In our application of CDC server clusters we will assume that the decision is based on the current processor load or utilization **and** a negotiated SLA criterion on the mean **or** quantile of service delays.

Typical Service Level Agreements (SLAs) in a CDC environment refer to the response times of the CDC to service requests. In interactive applications, as in case of Cyber-Physical System (CPS), Smart Grid, Software-Defined Networking (SDN), production automation, traffic or health control, real-time performance criteria are of prime interest. In this paper we will, therefore, define the SLA by the mean of the random waiting time T_W of an arriving service request when it has to wait, $t_W = E[T_W | T_W > 0]$, or the quantile probability $p = P\{T_W > t_{Th} | T_W > 0\}$ by which service requests would have to wait longer than a threshold time t_{Th}. As every state change requires some overhead time (acc. to [9] the Enhanced Intel SpeedStep Technology hardware unavailability is 10 µs instead of 250 µs before), we should try to further reduce the frequency of changes between Power States, c.f., Section IV.

The problem will be solved by establishing a functional relationship between the SLAs, prescribed either by the mean delay t_W or by the quantile p for exceeding a threshold delay t_{Th}, and the job arrival rate λ. These relationships are provided in principal by the performance analysis of the queuing model of type GI/G/n. The problem is, that theoretically exact results are only known for some special cases of the model GI/G/n and, secondly, only in forward direction from a given load figure to performance, where we need the inverse direction from the performance to the load figure. We will solve this problem by a typical traffic engineering approach, c.f. Section III.

III. MODELING OF THE DVFS OPERATION

A. Modeling the Cloud Server Cluster

The best way to model a server cluster is a queuing model with n servers and an unlimited buffer space for arriving service requests which cannot be served immediately at the instant of arrival when all n servers are occupied. Fig. 1 shows this standard queuing model of the type GI/G/n acc. to Kendall's notation, where "GI" denotes the type of arrival process (**G**enerally distributed and **I**ndependent inter-arrival times), "**G**" denotes the type of service time process (**G**enerally distributed service times), and "**n**" denotes the number of servers.

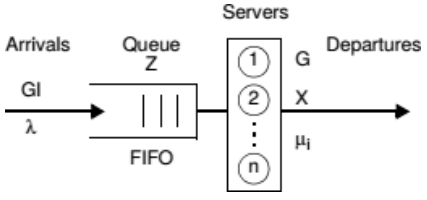

Fig.1. Queuing model GI/G/n for a CDC server cluster

The queuing model is further specified by:

- the arrival rate λ of service requests (mean number of arrivals per second)
- the service rate μ_i of the servers in Power Mode i, where $h_i = 1/\mu_i$ denotes the mean service time of a job in Power Mode i, $i = 0,1,...,5$
- the queuing service discipline FIFO (first-in, first-out)
- the system state random variables (X,Z), where X denotes the number of busy servers and Z the number of buffered service requests.

Queuing Theory has a long tradition of more than 100 years, during which a rich knowledge has been gained through stochastic process analyses in Mathematics, Operations Research, Teletraffic Theory and Engineering, and Computer Performance Modeling. For the general queuing model GI/G/n no exact analysis exists, except for the 3 cases M/M/n, GI/M/n [12] and M/D/n [13], where "M" denotes a Markovian process with negative-exponentially distributed inter-arrival or service times, and "D" denotes a deterministic process with constant service or inter-arrival times. For the general case phase-type approximations exist which are precise enough for practical applications, c.f. standard literature [12,14]. For practical and engineering applications Tables have been provided for a wide range of model types and parameters, c.f. [15,16].

We will apply the GI/G/n queuing model to determine the operating ranges of the request arrival rates λ for the 6 Power Modes under the restriction of the SLA in form of the **mean threshold waiting time t_{WT}** and, alternatively, in form of the **quantile p of waiting times exceeding the threshold t_{Th}** of waiting requests. For the presentation of our method we choose the most simple example of the M/M/n delay system for which exact closed-form solutions exist (below) by the aid of the waiting time performance organized as in the charts of Figs. 2 and 3. The same can be performed for exact results of the two other models GI/M/n and M/D/n [12,13]. For the general model GI/G/n we will use Tables for the mean delays and the coefficient of variation of delay distributions which are based on close approximations, c.f. [14,16]. For the queuing model M/M/n closed-form solutions exist for the probability of delay W of an arbitrarily arriving request, the mean waiting time t_W of a delayed request, the mean queue length Ω, and the complementary distribution function of waiting requests $W^c(t)/W = P\{T_W > t \mid T_W > 0\}$, also known as "Erlang-C formula"

.

$$W = A^n . n/n! \, (n - A) \, / \sum_{i=0}^{n-1} \frac{A^i}{i!} + \frac{A^n . n}{n!(n-A)} \qquad (1)$$

$$t_W = h/(n - A) \qquad (2)$$

$$\Omega = \mathrm{W}.\lambda . t_W \qquad (3)$$

$$W^c(t)/W = \exp(-t/t_W) \qquad (4)$$

where $A = \lambda \cdot h$ denotes the "offered traffic". In a pure delay system A is identical with the average number of occupied servers, i.e., $A = Y = E[X]$.

Exact results are also known for the queuing models of type GI/M/n [12] and M/D/n [13]. The numerical evaluation for the model M/D/n is highly complex, in particular for the delay distribution convergence. The model GI/M/n can be solved by an Embedded Markov Chain, but requires an iterative solution of a characteristic root of a probability generation function. This model observes a special property as that the complementary Cumulative Distribution Function (CDF) of delayed requests is exponential with mean t_W. Both models have been tabled for the mean delay t_W and the complementary CDF of delayed requests $W^c(t)/W$ for queue service disciplines FIFO (First-In, First-Out) and RANDOM [15]. For the general queuing model GI/G/n no exact solution is known. Instead, this model has been solved numerically for phase-type models being fitted to the mean and the coefficient of variation of the inter-arrival and service times, respectively, c.f. [16]. Many other results and validations for the queuing model GI/G/n can be found in [14]. In the following two sub-sections III.B,C we will explain how the load ranges can be determined **exactly** for SLAs on the **mean delay t_{WT}** or on the **delay quantile p**, with the **threshold delay t_{Th}**, for the queuing models M/M/n, M/D/n and GI/M/n based on the above closed-form solution and for GI/D/100 using tabled results of [16]. In sub-section III.D we generalize the method for the model GI/G/n.

B. DVFS for the SLA on Mean Delays of Delayed Jobs

In Fig.2 the mean waiting times t_{Wi} of the 6 Power Modes are plotted dependent on the request arrival rate λ for $n = 100$ servers for the queuing model M/M/100. Note, that the operation range for Power Mode P_0 is up to $A = 100$; at $A = 100$ the system is saturated and the mean waiting time t_W approaches infinity asymptotically; this means that $\lambda_{0,max} = n/h_0 = 100\ 1/s$. We will use the same model for all Power Modes; as the mean service times increase with slower servers (named "processing times" in Table 1) the system saturates for smaller arrival rates λ accordingly, i.e. $\lambda_{i,\ max} = n/h_i$. At the intersection between the curves for the mean waiting times and the SLA threshold value $t_{WT} = 0.05$, we find the upper bounds for the service request arrival rates λ_i for Power Mode i, $i = 0, 1,..., 5$.

TABLE II. OPERATING RANGES λ_i FOR MEAN DELAYS FOR QUEUING MODELS M/M/n, M/D/n, GI/M/n AND GI/D/n FOR MEAN DEALAYS $t_{Wi} = t_{WT} = 0.05\ s$ AS SLA FOR MEAN WAITING TIME

P-States	P_0	P_1	P_2	P_3	P_4	P_5
Type	Upper Thresholds of Arrival Rates λ_i in Power State P_i					
M/M/100	80	60	46	37	29	24 [1/s]
M/D/100	86	66	51	40	32	25 [1/s]
GI/M/100	64	46	35	26	21	16 [1/s]
GI/D/100	66	48	37	27	22	17 [1/s]

The Power Mode operating ranges for the SLA criterion of mean waiting times of delayed service requests t_{WT} are given in Table II for the 4 queuing model examples M/M/n, M/D/n, GI/M/n and GI/D/n for $n = 100$ (with an hyper-exponential-type arrival process with coefficient of variation $c_A = 2$). For the last case we applied approximation results from Tables on

Delay Systems of type GI/G/n which are based in phase-type process approximations [16], see also Section D.

Example: The optimum operating range for Power Mode *i* is $\lambda_{i+1} < \lambda < \lambda_i$, $i = 0,1,...,5$, $\lambda_6 = 0$.

From Table II it can be conjectured that the operating ranges decrease with increasing coefficient of variation of the arrival and service processes.

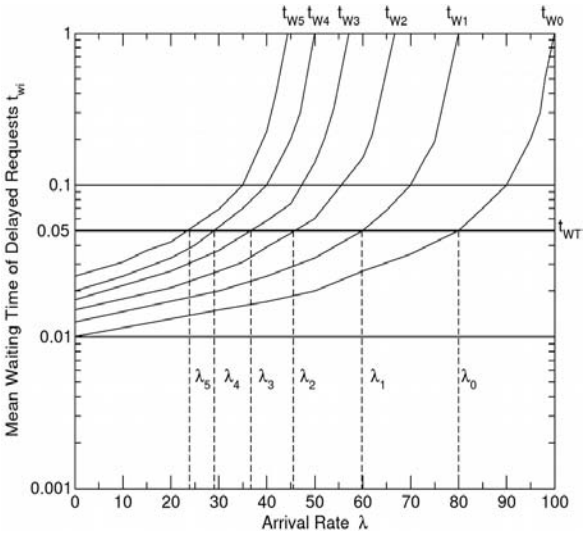

Fig. 2. Mean waiting times of delayed requests versus arrival rate λ for 6 power modes oft the server cluster of Intel Pentium M processors. Example:model M/M/100 for mean waiting time threshold $t_{WT} = 0.05$ *s*

C. DVFS for the SLA on Quantile of Delayed Jobs

In this section we want to show how a more real-time-centric SLA can be guaranteed. For this we choose the quantile *p* of the waiting time of waiting requests $\{T_W \mid T_W > 0\}$ as the SLA criterion, defined by the probability

$$p = P\{T_W > t_{Th} | T_W > 0\} \quad (5)$$

which means that waiting times for delayed requests at their arrival instant exceed a threshold time t_{Th} only with probability p ("quantile"). From the exact analyses of the queuing model GI/M/n (which includes M/M/n) it is known that the complementary distribution function of delays is also exponentially, i.e., we have an explicit formula as in case of the model M/M/n (but with different mean t_{Wi}):

$$W_i^c(t)/W_i = exp(-t / t_{Wi}) = p \quad (6)$$

From this formula we find directly the relation between the mean waiting time t_{Wi} and the SLA tuple (p, t_{Th}):

$$t_{Th} = - t_{Wi} \cdot ln\, p \quad (7)$$

For prescribed SLA tuples (t_{Th},p) we find immediately the corresponding mean waiting time t_{Wi} of a delayed request. With the value of t_{Wi} we find from Fig. 2 $t_{Wi} = f(\lambda,i)$ the upper threshold values (bounds) λ^i for each power mode *i*.

Fig, 3 illustrates the relationship between the complementary delay distribution function $W_i^c(t)/W_i$, the delay quantile *p*, and the time *t*. The delay threshold time $t_{Th,i}$ for delayed requests for Power Mode *i* follows from the intersection point between the two straight lines: The smaller the threshold time t_{Th}, the smaller must be the average delay time t_{Wi} and, thus, the load level λ .

As mentioned before, the mathematically exact analysis of the queuing model M/D/n is known, but its numerical evaluation is more difficult, in particular for the delay distribution which is no longer exponential. We begin with the array of curves tabled for the complementary delay distribution function $W_i^c(t)/W_i$ from the tabled results for the M/D/n queuing system [15] in a chart organized analogously as in Fig.3 for different mean waiting times t_W as parameter. We have to find that curve with the parameter t_W which hits the intersection between the delay curve and the line for the quantile *p* by interpolation. The corresponding load factor λ for this case is found from the array $t_{Wi} = f(\lambda,i)$ for Power Mode *i* as illustrated in Fig.2.

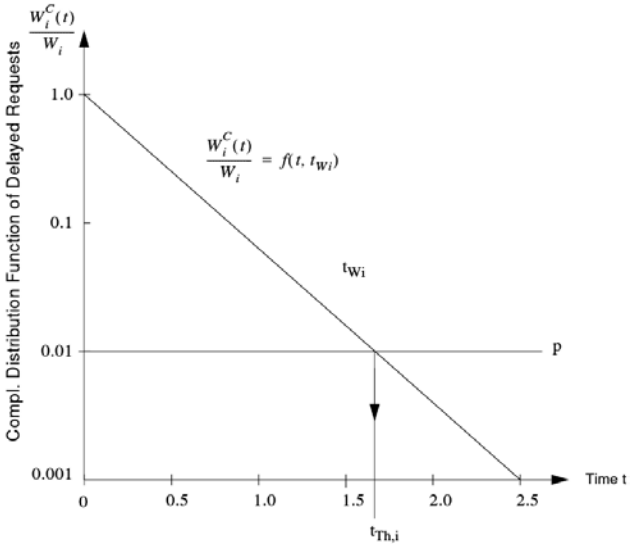

Fig. 3. Complementary distribution function of delayed requests versus time t for the queuing model GI/M/n and delay quantile *p* for the determination of th delay threshold Time $t_{Th,i}$ (principal diagram)

TABLE III. OPERATING RANGES λ_i FOR QUANTILES OF DELAY MODELS M/M/n, M/D/n, GI/M/n AND GI/D/nm FOR QUANTILES OF DELAY $p = 0.05$, THRESHOLD TIME $t_{Th} = 0.10$ *s* AS SLA

P-States	P_0	P_1	P_2	P_3	P_4	P_5
Type	Upper Thresholds λ i/λ i of Arrival Rates in Power State P_i					
M/M/100	70	50	42	27	19	12 [1/s]
M/D/100	72	56	42	30	20	15 [1/s]
GI/M/100	53	35	25	18	12	8 [1/s]
GI/D/100	54	37	28	16	13	8 [1/s]

The Power Mode operating ranges for the SLA criterion of delay quantiles p of delayed service requests are given in Table III for the three queuing models M/M/100, M/D/100 and

GI/M/100 with exact solutions and for GI/D/100 from the tail approximation using results tabled in [16]. The method for non-exponential delay distributions is based on a theorem on the tail behavior of queuing systems of the type Ph/Ph/n under the FIFO queuing discipline, c.f. Section III.D.

D. DVFS for SLAs of General Queuing Models GI/G/n

From Queuing Theory we know, that an arbitrary probability distribution function of a random variable can be approximated by a "Phase-type" distribution function (DF), especially by a graph of exponentially distributed phases [12]. This allows to analyze a more complex model by a multi-dimensional Markov Chain. The problem is, however, to find the parameters for the exponential phases by solving a corresponding approximation problem. From experimental measurements we often settle to meet the first two ordinary moments of the DF of a random variable T, $E[T^i]$, $i = 1,2$, or, equivalently, the mean $E[T]$ and the coefficient of variation (CV) c, where $c^2 = VAR[T]/E[T^2] - 1$. Simple Phase-type models are just a series of phases or a probabilistic choice between a few (especially 2) exponential phases, allowing for the range of coefficients of variation $0 \leq c \leq \infty$. The corresponding queuing system is a special case of a Ph/Ph/n model, where the phase parameters are derived from given values for the mean $E[T]$ and CV c. Queuing models of type Ph/Ph/n have been tabled in [16].

For queuing models of the Type Ph/Ph/n with queue discipline FIFO Yutaka Takahashi has shown in [17] that the tails of the delay distribution function behave asymptotically exponential:

$$W_i^c(t)/W_i \sim a \cdot exp(-bt) \qquad (8)$$

i.e., they are asymptotically linear on the log/linear coordinate plane for $W_i^c(t)/W_i$. This approximation has been proved as of high accuracy. This is exactly that range of the complementary DF which is relevant for our delay threshold as SLA. In the log/linear plane it can be determined easily as a straight line by two points $(0,a)$ and (t_{Th},p). This allows us to find the load region thresholds the same way as explained before. The Tables for Ph/Ph/n [16] are given for

TH	mean service time $E[T_H] = 1$ (Time normalization) i.e., all temporal parameters are multiples of $E[T_H]$
CA	CV of the arrival process $(CA = c_A)$
CS	CV of the service process $(CH = c_H)$
RHO	server utilization $(RHO = A/n)$
PW	probability of delay $(PW = W)$
PB	probability that all servers are busy
TW	mean waiting time of delayed requests $(TW = t_W)$
ELQ	mean queue length $(ELQ = Ω)$
A,B	CDF parameters of delayed requests $(A = a, \quad B = b)$

Table IV provides results for the operating load ranges λ_i (for t_{Wi}) and λ^i (for t_{Th},p) for an optimistic/pessimistic case of hypo-/hyper-exponential arrival/service processes of queuing model types $E_3/E_3/25$ and $H_2/H_2/25$ with $c_A^2 = c_s^2 = 0.33$ and 1.5, respectively. For comparison we have placed the results for the basic queuing model M/M/25 between them $(c_A^2 = c_S^2 = 1)$. For the quantile study we have assumed, that the waiting time should exceed the doubled mean waiting time value only with probability of 1%. Table 4 emphasizes the striking influence of the stochastic arrival and service processes on the optimum operation load ranges: The higher the variability of the arrival/service processes the higher must be the voltage/frequency mode.

TABLE IV. OPERATING RANGES λ_i/λ^i FOR QUANTILE DELAYS FOR QUEUING MODELS E3/E3/25, M/M/25 AND H2/H2/25 FOR MEAN DELAYS t_{Wi} AND FOR QUANTILE p OF DELAYS, MEAN WAITING TIME THRESHOLD $t_{WT} = 0.4$ s, THRESHOLD TIME $t_{Th} = 2t_{WT} = 0.8$ s, QUANTILE p = 0.01

P-States	P_0	P_1	P_2	P_3	P_4	P_5
Type	Upper Thresholds λ_i/λ^i of Arrival Rates in Power State P_i					
E3/E3/25 λ_i	24.2	19.2	15.8	13.2	11.5	10.2 [1/s]
E3/E3/25 λ^i	3.2	18.2	14.8	12.2	10.5	9.0 [1/s]
M/M/25 λ_i	22.5	17.5	14.0	12.5	10.0	9.0 [1/s]
M/M/25 λ^i	19.5	14.5	11.2	9.3	7.5	6.5 [1/s]
H2/H2/25 λ_i	19.0	11.0	8.3	6.2	5.0	4.0 [1/s]
H2/H2/25 λ^i	17.0	12.2	9.2	7.0	5.5	3.8 [1/s]

IV. IMPLEMENTATION ASPECTS OF DVFS

DVFS aims at the adaptation of the power level to the instantaneous traffic load by lowering the power consumption accordingly. This can be accomplished by a permanent monitoring of the current load. As the load is generated by many independent tenants the instantaneous load changes statistically, an effect well-known from massive service systems as the telephone network, or from interactively used computer systems. As already stated above, each change of the power level causes a short time-out which reduces the system capacity and causes extra energy consumption. Besides this effect, load monitoring always lacks behind as statistics are only available over the recent past. In the literature many studies on DVFS were reported using system parameter optimization algorithms where SLA aspects are not considered.

For that reason we advocate for another method of a Finite State Machine (FSM) based on monitoring of the current system state using a staggered hysteresis model for server activations/deactivations which had originally been suggested by the authors for power-saving by server consolidation [18,19]. This principle is based on a strategy to retard new server activations by buffering of arriving service requests as long as possible while still guaranteeing a delay-based SLA. This had the effect that the server activation rate could be lowered considerably which is in particular of interest when each server activation requires additional time and energy [6].

ISBN 978-3-903176-13-3 © IFIP

For these reasons we would like to suggest applying this method also in connection with DVFS.

The server cluster is modeled by a queuing system with an FSM-controlled operation strategy, c.f. Fig. 4. The system state *(X,Z)* indicates the current random number of occupied servers X and the random number Z of queued requests which are served according to the FIFO queuing discipline. The function of the control is indicated by a State Transition Diagram (STD), see Fig. 5. At each number x of busy servers new arrivals are queued until state $(x,w^{(x)}-1)$, where $w^{(x)}$ is chosen such that the SLA is still guaranteed for the mean (or for the quantile) of delay. An arrival at state $(x,w^{(x)})$ will immediately ignite that an idle server has to be activated and starts service, $x = 1,2,...,n-1$, where $w^{(x)} = x \cdot w$, for $x = 1,2,...,n-1$. The value w can be determined easily for the mean waiting time $t_{WT} = w \cdot h$ as SLA in case of exponential service times from the mean residual time until the next server terminates service and x. Similarly, w can also be determined for the case of an SLA for the quantile p and the delay threshold t_{Th} of delays. Server deactivations take only place when the waiting line is empty.

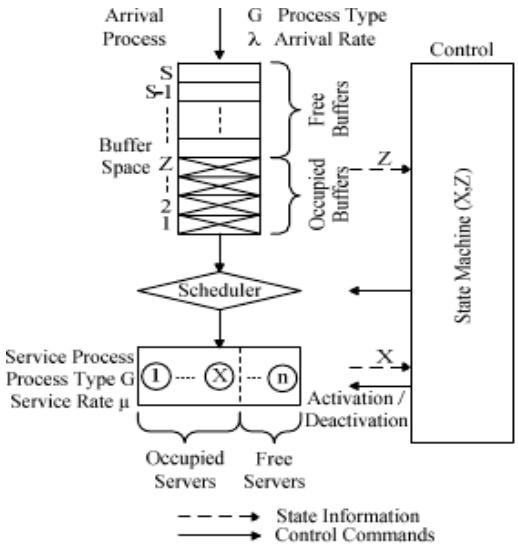

Fig. 4. Model of a server cluster with FSM-control

By this strategy the dynamics of server activations/deactivations is greatly reduced while the SLA can still be guaranteed and the system resides for a much longer time with x servers being busy. When this FSM is applied under a DVFS control regime, the temporal and energy overhead involved with DVFS can be reduced. Note, that server activations and deactivations are indicated in Fig. 5 only at states with bold-faced transition arrows. The implementation for an automatic DVFS control is easy, as the state *(X,Z)* is known by the operating system at any time. **without** any system state sampling. The transition rates in Fig.4 have been used for the analytic performance evaluation without DVFS and are valid only for exponentially distributed inter-arrival and service times to determine the server activation rate R_A [18]. In the context of this paper only the **FSM logic** is of importance to decide **when** a Power State has to be changed up or down, respectively. For the performance analysis within a certain Power State the simpler queuing model of the type GI/G/n without the hysteresis can be applied as outlined in Section 3.

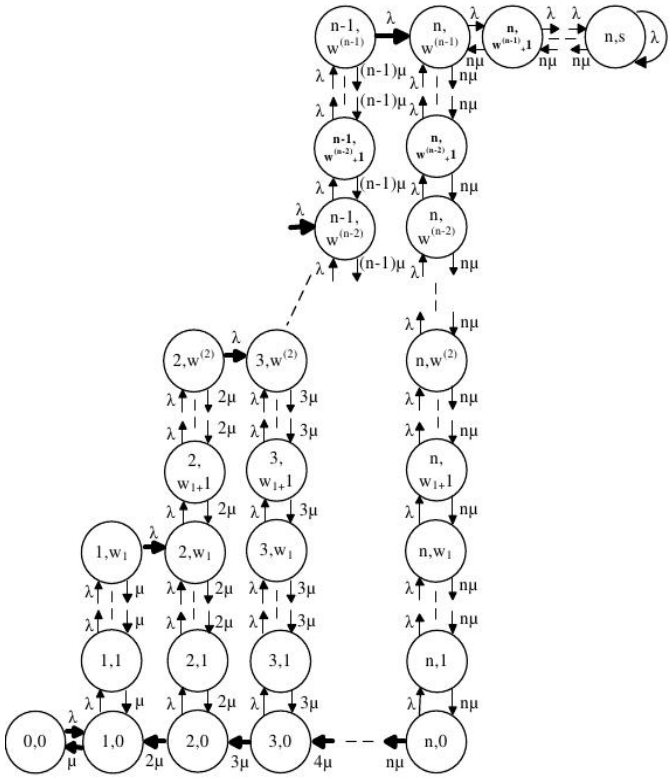

Fig. 5. State-Transition Diagram for a server cluster by FSM control

In Fig. 6 a typical result of the hysteresis model is presented taken from [18], showing the server activation rate R_A versus the job arrival rate λ of events per second, and a constant incremental queue threshold step size $w^{(i)} - w^{(i-1)} = w$, $i = 1,.,n-1$, and $t_{WT} = w \cdot h$ as SLA for the case of the 2-dimensional Markov-Chain Fig.5. The server cluster operating system activates a new server each time when a state $(x,w^{(x)})$ is entered acc. to the transitions indicated in Fig. 5, for $x = 1,...,n-1$. The method is also applicable in the case where an SLA for the quantile p has to be met: In that case the condition for the threshold values $w^{(x-1)}$ have to be derived from $W_i^c(t)/W_i \leq p$ in state $(x,w^{(x-1)})$, where $W_i^c(t)/W$ is Erlang-k distributed with $k = x-1$ and service rate x/h. Fig.7 shows the results for the mean waiting time of buffered service requests for an extended model by the relative idle server activation times h/α, where $1/\alpha$ denotes the mean time for activating a sleeping server. The smoothing effect of the hysteresis control affects an almost constant waiting time for a large range of the job arrival rate in case of zero activation overhead $(\alpha \rightarrow \infty)$; with increasing activation overhead the waiting time naturally increases, but

maintains the relatively small dependence on the load which supports a stable system operation for a wide load range.

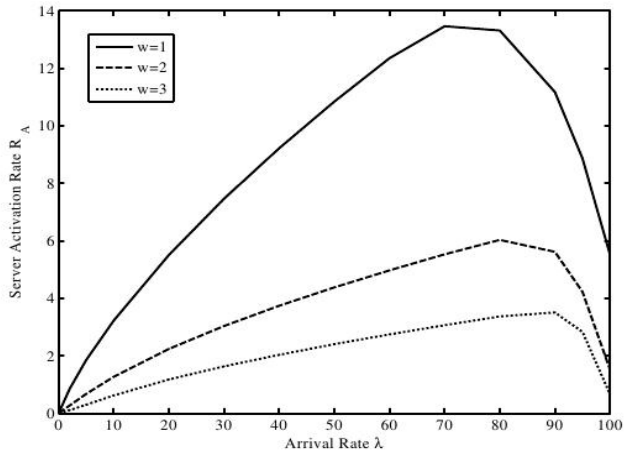

Fig. 6. Server activation rate R_A as function of the job arrival rate.λ for 3 different step parameters w, n = 100 servers, under Markovian assumptions

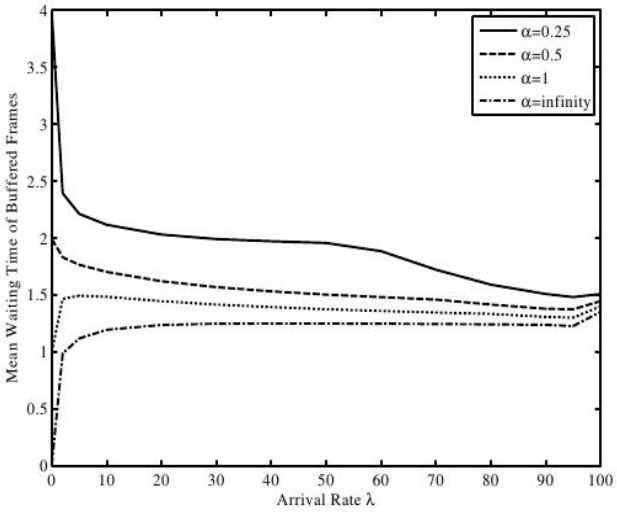

Fig. 7. Mean waiting times of buffered jobs as function of the job arrival rate and various mean server activation overhead times . Step parameter w =2.

V. DISCUSSION

In this Section the relation of our contribution to the options of the Linux Kernel implementation [9] and a recent publication on a queuing analysis for a single server configuration[20] will shortly be discussed.

In [9] the method of the CPUFreq Governor for clock scaling is described, where 6 different policies are distinguished:

1. "Performance": the CPU frequency is set to the highest value within a minimum and a maximum frequency

2. "Powersave": The CPU frequency is set to the lowest value

3. "Userspace": The CPU frequency is user-defined

4. "Ondemand": The CPU frequency depends on the current system load, based on the recent CPU usage statistics

5. "Conservative": similar as Ondemand, but through graceful changes instead of jumping between the maxima of speed levels

6. "Schedutil" The CPU frequency follows from a per-entity load tracking mechanism.

The policy "Ondemand" and "Conservative" are the closest ones to our load-depending method in this paper. The differences are that the Linux Governor depends on periodic load statistics without consideration of any SLA conditions. Both policies allow, however, for options as flexibility of the state sampling rate, for options on "Powersave" bias and frequency steps.

In [20] a single processor is operated under DVFS. The differences of their model to our paper are: A single server model of type M/M/1, i.e., one processor operated under periodical server utilization checks, applied on the AMD Opteron(TM) 4180 processor, and analyzed by a state-based Markov Chain and by simulations. Results are on the mean response time for two operating modes "On demand" and "Conservative" at discrete time instances and server utilization bounds. Numerical examples are reported by a Discrete Time Markov Chain (DTMC) analysis and by a simulation based on the SHARPE modeling tool for correctness verification. The results are not directly comparable with our results, due to the different modeling assumptions, but direct at the same intention to define applicable criteria for power mode changes.

A state-based method for a dynamic adaptation of the service rates had already been used in an earlier approach for server consolidation purposes where the transition rates between system states are **state-dependent**, which had been analyzed under Markovian assumptions to reflect different server speeds, but it was not evaluated specifically for DVFS operation [18,19].

For the case of a single-server model instead of the multi-server cluster as in this paper, our approach can be applied analogously for each Power mode, starting with Power mode P_0, and succeeding to $P_2,...,P_5$ successively for queuing models of the type GI/G/1 (specifically easy for M/G/1 for closed-form solutions) to find the upper arrival rate thresholds λ_i and λ^i for the mean or the quantile of service delays, respectively. Aggregating the 6 power mode models into a one-dimensional FSM and associated description by a STD with corresponding transitions between neighbor states a FSM-based model can be constructed analogously to the two-dimensional STD of Fig. 5 for a state-based decision on power mode changes instead of periodical server utilization checks.

VI. CONCLUSION

DVFS has proved as a powerful method to adapt the power consumption of servers to the application conditions and, in particular, to the actual load. In this contribution a queuing theory-based approach is suggested to adapt the power consumption under Service Level Agreement restrictions with respect to the mean or to the quantiles of service delays. The

method allows to fix the upper load level boundaries exactly which are defined by actual service request rates λ_i or $\boldsymbol{\lambda}^i$ in case of mean or of quantiles of service delays as SLAs to the operative Power Mode levels $P_0,...,P_5$, respectively. The method applies to general application statistics of the request inter-arrival and service times. The method is exemplified for the special case of the Enhanced Intel SpeedStep Technology of the Intel Pentium M processor. There exist various different ways how load control can be applied automatically: Fine-grained on the individual processor level independently, on the server level, or Coarse-grained on the server cluster level of n servers. In this application a server or a whole server cluster is considered where the cluster load situation is controlled by the Operating System. Precise load boundaries are provided for each Power state and two given SLA parameters: mean or quantile of service delays. For the implementation a Finite State Machine control method is suggested which is based only on two actual load state indicators: the number of busy servers X and the number of delayed service requests Z, described by a staggered hysteresis FSM model by which the frequency of Power Mode changes can be reduced drastically which contributes to time and energy savings and an easy implementation **without** specific load level sampling. A simplified application exists also for a single-processor or single-server-system.

As outlook we would like the extend the studies on DVFS to a comparison between the two different operating methods of periodic load sampling and our suggested FSM monitoring without periodic server utilization checks.

REFERENCES

[1] P. 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

[2] A. Ghandi, M.Harchol-Balter, R. Das, C.Lefurgy, "Optimal Power Allocation in Server Farms", ACM SIGMETRICS/Performance '09, June 15 - 19, 2009.

[3] H.H. Kramer, V. Petrucci, A. Subramanian, E. Ochoa, " A Column Generation Approach for Power-aware Optimization of Virtualized Heterogeneous Server Clusters, Publicado em 06/06/2011.

[4] T. Guerout, T. Monteil, G. Da Costa, R.N. Calheiros, B. Buyya, M. Alexandru, "Energy-aware Simulation with DVFS", In: Simulation Modeling Practice and Theory, Elsevier, vol. 39, pp. 70 - 91, 2013.

[5] P. Arroba, J.M. Moya, J.L. Ayala, R. Buyya, "DVFS-Aware Dynamic Consolidation of Virtual Machines for Energy Efficient CloudData Centers", Concurrent and Computation: Practice and Experimence, 2010, Wiley InterScience, www.interscience.wiley.com

[6] V.J. Patel, H.A. Bheda, " Reducing Energy Consumption with DVFS for Real.-Time Services in Cloud Computing", IOSR J. of Computer Engineering (IOSR-JCE), Vol. 16, Issue 3, Ver. II, 120143, pp. 53 - 57.

[7] P. Arroba, J.M. Moya, J.L. Ayala, R. Buyya, "Dynamic Voltage and Frequency Scaling-aware Dynamic Consolidation of Virtual Machines for Energy Efficient Cloud Data Centers", Concurrency and Computation : Practice and Experience, Vol. 29, Issue 10, 31 pages. https://doi.org/10.1002/cpe.4067.

[8] D. Brodowski, N. Golde, "Linux cpufreq governors". Linux Kernel. https://www.kernel.org/doc/Documentation/cpu-freq/governors.txt, 2013. TM) Kernel - Information for Users and Developers https://www.kernel.org/doc/Documentation/cpu-fr

[9] "Enhanced Intel SpeedStep Technology for the Intel Pentium M Processor", Intel White Paper, Order Number 30117ß-001, March 2004.

[10] M.J. Ham, "QoS Handling with DVFS (CPUfreq&DEVfrag), SW Center, Samsung Electronics

[11] E. Calore, A. Gabbana, S.F. Schifano, R. Tripiccione, " Evaluation of DVFS Techniques in Modern HPC Processors and Accelerators for Energy-aware Applications", 2017. https://doi.org/10.1002/cpe.4143.

[12] H. Kobayashi, B.L. Mark, "System Modeling and Analysis - Foundations of System Performance Evaluation", Pearson International Edition, Prentice-Hall Inc., Upper Saddle River, 2009.

[13] C.D. Crommelin, "Delay Probability Formulae when the Holding Times are Constant", POEEJ 25, 1932, pp. 41 - 50.

[14] W. Whitt, "Approximations for the GI/G/m Queue", J. Production and Operations Management, Vol. 2, No. 2, Spring 1993, pp. 114 - 161.

[15] P. Kuehn, " Tables on Delay Systems", Institute of Switching and DataTechnics, University of Stuttgart, Germany, 1976. Online: http://www.ikr.uni-stuttgart.de

[16] L.P. Seelen, H.C. Tijms, M.H. van Hoorn, "Tables for Multi-Server Queues", Elsevier Science Publishers B.V., 1985.

[17] Y. Takahashi, "Asymptotic Exponentiality of the Tail of the Waiting Time Distribution in a Ph/Ph/c Queue", J. Adv. Appl. Probability 13, 1981, pp. 619 - 630.

[18] P.J. Kuehn, M. Ezzat Mashaly, "Automatic Energy Efficiency Management of Data Center Resources by Load-dependent Server Activation and Sleep Modes", Elsevier J.- Ad Hoc Networks, 25, 2015, pp. 497 - 504.

[19] M. Ezzat Mashaly, "Performance of Cloud Data Centers with Service Level Agreement Guarantees", PhD-Dissertation, University of Stuttgart, 2017.

[20] R. Basmadjian, H. de Meer, "Modeling and Analysing Conservative Governor of DVFS-enabled Processors", ACM E²DC Workshop, Karlsruhe, Germany, 2018, ACM Digital Library.A. Beloglazov, J.