Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 337 (2018) 67–86

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Epistemic Uncertainty Propagation in Power Models

Marco Gribaudo[1], Riccardo Pinciroli[2]

Dipartimento di Elettronica, Informatica e Bioingegneria
Politecnico di Milano
Via Ponzio 34/5, Milano 20133, Italy

Kishor Trivedi[3]

Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708, USA

Abstract

Data-centers have recently experienced a fast growth in energy demand, mainly due to cloud computing, a paradigm that lets the users access shared computing resources (e.g., servers, storage, etc.). Several techniques have been proposed in order to alleviate this problem, and numerous power models have been adopted to predict the servers' power consumption. Some of them consider many server resources, some others account for only the CPU, that has proven to be the component responsible for the largest part of a server's power consumption. All these models work with generally inaccurate input parameters. However, none of them takes into account the effects of such inaccuracy on the model outputs. This paper investigates how epistemic (parametric) uncertainty affects a power model. Studying the impact of epistemic uncertainty on power consumption models makes it possible to consider loads with a probability density while investigating the battery depletion time or the amount of energy required for a given task.

Keywords: power consumption, energy consumption, power models, epistemic uncertainty, parametric uncertainty, uncertainty propagation, M/M/c/K

1 Introduction

Servers' energy consumption is now one of the main issues to be considered while developing new applications. For example, the U.S. data-centers energy consumption is estimated to be 73 billion kWh in 2020 [31]. Another energy consumption related problem is about managing those big infrastructures in order to improve

[1] Email: marco.gribaudo@polimi.it
[2] Email: riccardo.pinciroli@polimi.it
[3] Email: ktrivedi@duke.edu

https://doi.org/10.1016/j.entcs.2018.03.034

their efficiency and fault tolerance [19]. Furthermore, uninterruptible power supplies (UPS) are used for emergency power when the main power source fails. To better investigate these problems and to decrease the global amount of energy consumed by data-centers, accurate power consumption models are required. In past years, scientific and business organizations provided several power models to estimate the data-centers power consumption [11,1,14]. Depending on the granularity at which the problem is studied, the power models consider only few resources or the whole data-center.

The power consumption models are generally solved with fixed values for their input parameters. However, since those parameters are estimated from finite number of collected samples, they introduce into the model some uncertainties due to partial information. Thus, the uncertainty needs to be propagated from the inputs to the output. This kind of parametric uncertainty is called epistemic uncertainty [21].

In this paper we wish to apply epistemic uncertainty propagation to power consumption estimation, when the data-center arrivals forms a homogeneous Poisson process and service times are exponentially distributed (i.e., the rates are fixed over time). Although several works in literature take into consideration aleatory uncertainty while analyzing a general model [13,30], just a few consider the epistemic uncertainty. Indeed, while aleatory uncertainty is due to the natural variations of the physical phenomenon modeled by the system, epistemic uncertainty is introduced into the model by a lack of knowledge (i.e., finite number of observations, that in our case makes estimation of Λ and M inaccurate) and needs to be propagated to the output. Albeit in the former case the uncertainty is reduced improving the model itself, in the latter one it may be curtailed by collecting a larger amount of samples for a more accurate input parameter estimation. For this reason, we assume some of the input parameters of the power model are assessed with uncertainty, thus to be stochastic. Those input parameters become input random variables (r.v.) with a probability density function (pdf). Thus, the power model does not return an exact value, but some stochastic results with a confidence interval. Hereinafter, we refer to r.v. with capital letters, and use the small ones to refer to their observed values. In order to propagate the uncertainty from model inputs to its outputs, we adopt a multi-dimensional integration technique [21,23,22] that has been already applied to dependability models. To the best of our knowledge, epistemic uncertainty has never been applied to models that estimate the servers' power consumption.

In order to investigate how epistemic uncertainty propagation affects the power consumption of a machine, we collect several samples while performing some benchmarks, thus considering different resource utilizations. In particular, we observed an ASUS desktop computer with an Intel i7-3770 CPU@3.4GHz, with 4 cores and a Simultaneous multi-threading (SMT) level of 2 (i.e., it can concurrently run 8 parallel threads, in spite of the 4 physical cores), a 16GB memory, a dedicated GeForce GTX 560 GPU, two hard disk drives and a solid state disk, and running an Ubuntu 14.04 OS. Albeit some relationships between CPU utilization and server's power consumptions are given in [11], they are affected by SMT when its level is higher

than 1 [5]. For the sake of simplicity, in this paper we assume the machine we consider has $SMT = 1$, thus we turn off four of the eight available logical CPUs in order to have one thread for each physical core.

As said in [11], CPU utilization is generally the only parameter we need to estimate a server's power consumption. For this reason, the machine analyzed in this paper is modeled only considering its CPU. In particular, we model it as an $M/M/c/K$ queue, where inter-arrival and service times are exponentially distributed, the number of available servers is $c = 4$ (i.e., the 4 physical cores), and K is the maximum capacity of the queue, since data-centers usually have a finite buffer [26]. When dealing with data-centers and cloud resources, new tasks arrival and execution times (e.g., virtual machines and users requests) are often assumed to be exponentially distributed [24,37,2]. Although an empirically measured workload would be more accurate, the exponentially distributed one is typically adopted when analyzing queuing models [20]. Moreover, besides letting us easily estimate the density of inter-arrival and service rates [21], Λ and M, respectively, the $M/M/c/K$ model also allows to consider an always stable system, without requiring further assumptions on the observed values, λ and μ, of the two random variables. Setting the capacity K to high values, we can analyze queues with large capacity, without dealing with the stability condition, (i.e., $\lambda < c \cdot \mu$).

The obtained theoretic results are then applied to a practical case. We consider an uninterruptible power supply that starts working after the main power source fails. In particular, we wish to study the backup battery life when it is strained by a constant load whose characteristics are affected by epistemic uncertainty. Indeed, the load of a UPS may be derived from the power consumption of the server the battery must keep on. Since such a power consumption (i.e., the output of the power model) is uncertain due to the epistemic uncertainty of the input parameters, the UPS load cannot be estimated with certainty.

The remainder of this paper is structured as follows. In Section 2 previous work about existing power models, epistemic uncertainty propagation and battery models is discussed. In Section 3 the adopted power model is presented, and the densities of its input random variable are investigated starting from the samples collected during several measurement campaigns. Section 4 defines the equations used in this paper for uncertainty propagation, describes the obtained results and validates them against the measured data. In Section 5 the UPS case-study is investigated, and Section 6 concludes the paper.

2 Related work

Several power models have been proposed during recent years for studying and developing new techniques to decrease data-centers power consumption. In [29], Priya et al. described the main power consumption models used for data-centers. The linear model proposed by Fan et al. in [11] is the most used one in the literature [6,12,39]. In this case, power consumption is assumed to linearly grow with CPU

utilization, following the equation:

$$P(U) = P_{idle} + (P_{busy} - P_{idle}) \cdot U \tag{1}$$

where P_{idle} and P_{busy} are the power consumptions of an idle server and a fully utilized one, respectively, whereas U is the CPU utilization. This model is adopted in [39] to identify the data-center cooling energy saving potential, and in [12] to model a predictive control that is able to save a significant amount of energy. Cerotti et al. [6] used the linear model to study the power consumption of Pool depletion systems.

Fan et al. also introduced a non-linear model where power consumption increases monotonically with the CPU utilization. The equation that defines the non-linear model is:

$$P(U) = P_{idle} + (P_{busy} - P_{idle}) \cdot (2U - U^r) \tag{2}$$

where r is a calibration parameter that is estimated through experiments.

Buyya et al. [3] proposed a power model similar to the one introduced in [11] in order to deal with the energy-aware resource allocation in cloud environments. Since the ratio between P_{idle} and P_{busy} is generally known (e.g., 70%), they adopted the model:

$$P(U) = c \cdot P_{busy} + (1 - c) \cdot P_{busy} \cdot U \tag{3}$$

where $c = P_{idle}/P_{busy}$. This power model has also been used in [4] to propose an adaptation strategy for managed Cassandra data centers.

Due to the small number of required parameters, these are the simplest available power models. Nonetheless, several other more accurate power models have been proposed in literature. The cost of a greater accuracy is the complexity of the model itself. For example, since dynamic voltage/frequency scaling (DVFS) and hyper-threading have been adopted to decrease server's energy consumption, researchers also proposed some power models that take into consideration these two techniques. It is the case of [5] where CPU utilization in Eq. (1) has been weighted by a factor Δ defined as:

$$\Delta = \left(\frac{C_{th}}{T_{th}} \alpha_{th} + \frac{C_{cr}}{T_{cr}} \alpha_{cr} \right) \cdot \left(\frac{fr}{fr_{max}} \right)^{\eta} \tag{4}$$

where α_{th}, α_{cr} and η are evaluated by a fitting procedure, C_{th} (C_{cr}) and T_{th} (T_{cr}) are the number of used threads (cores) and the total number of available threads (cores), respectively, fr is the average CPU frequency and fr_{max} is the maximum one.

Other power models consider several server components besides the CPU. For example, Bohra et al. [1] used the following model:

$$P = P_{idle} + c_{cpu} \cdot \xi_{cpu} + c_{cache} \cdot \xi_{cache} + c_{mem} \cdot \xi_{mem} + c_{disk} \cdot \xi_{disk} \tag{5}$$

where c_r is the weight of the monitored system events ξ_r for resource r (i.e., CPU, cache, memory and disk).

Other available models take into consideration the power consumption of the whole data-center. For example, Gosh et al. [14] proposed a stochastic model to

evaluate the performability of an IaaS Cloud environment and its cost. In particular, they grouped all the physical machines into three different pools (i.e., hot, warm and cold) based on their current status (i.e., running, turned on and not ready, turned off, respectively). Thus, they gave different costs to performance and dependability parameters in order to evaluate and eventually optimize the cost and capacity of the considered environment. While considering the power consumption of the physical machines and its cost, they also take into account the cooling cost. Indeed, the power consumption for cooling mechanisms is estimated to be between 30% and 50% of the data-center total power consumption [25,10].

For the purpose of this paper, we first adopt the power model in Eq. (2) and, similarly to what has been done by Buyya et al. in [3], we further reduce the number of required input parameters expressing P_{busy} as a function of P_{idle}.

As said in Section 1, in this paper we wish to apply epistemic uncertainty propagation techniques to data-centers power consumption estimation. Epistemic uncertainty – due to a lack of knowledge, and generally ascribed to the finite number of collected samples during the measurement campaign [9,40] – must not be confused with aleatory uncertainty. Indeed, differently from the epistemic uncertainty that is considered in this paper, aleatory uncertainty is caused by variations in the analyzed physical phenomenon [32,34].

Epistemic uncertainty has been considered in dependability models [22,28] and performance related models [36]. To the best of our knowledge, epistemic uncertainty has never been incorporated in the predictions of a power model. For this purpose, we resort to the multiple integral equations given in [21]. In particular, consider a set of input random variables $\{\Theta_i, i = 1, 2, ..., l\}$ and an output measure M defined as a random variable $M = g(\Theta_1, \Theta_2, ..., \Theta_l)$. Computing the output measure at specific parameter values can be seen as computing the conditional measure $M(\Theta_1 = \theta_1, \Theta_2 = \theta_2, ..., \Theta_l = \theta_l)$ (henceforward denoted by $M(\bullet)$) because of the uncertainty introduced by the input parameters. Applying theorem of total probability [35] and using joint epistemic density $f_{\Theta_1,\Theta_2,...,\Theta_l}(\theta_1, \theta_2, ..., \theta_l)$ (hereinafter denoted by $f(\bullet)$), the conditional measure $M(\bullet)$ can be unconditioned. Thus, Cdf of the measure M is computed as follows:

$$F_M(m) = \int ... \int \mathbb{1}(M(\bullet) < m) \cdot f(\bullet)\ d\theta_1\ ...\ d\theta_l \tag{6}$$

where $\mathbb{1}(\zeta)$ is an indicator variable that is 1 when the event ζ is verified, and 0 otherwise. Instead, the expected values of measure M is computed as:

$$E[M] = \int ... \int M(\bullet) \cdot f(\bullet)\ d\theta_1\ ...\ d\theta_l \tag{7}$$

For the sake of simplicity, in this paper, we assume the epistemic random variables to be independent, thus the joint epistemic density is given by the product of the marginals:

$$f(\bullet) = f_{\Theta_1,\Theta_2,...,\Theta_l}(\theta_1, \theta_2, ..., \theta_l) = f_{\Theta_1}(\theta_1) \cdot f_{\Theta_2}(\theta_2) \cdot ... \cdot f_{\Theta_l}(\theta_l)$$

While considering the exploitation case, we investigate the lifetime distribution of a UPS battery when the power model adopted to estimate the system power consumption is using uncertain input parameters. Several battery models have been proposed in the literature: some of them, such as the electro-chemical models [8], are very accurate and need several input parameters to work; some others, like the kinetic battery model (KiBaM) [18,16], need just a few parameters and may be easily used also by inexperienced users. In this paper, we adopt the latter model.

The KiBaM is an accurate and simple model [18,16] that divides the stored charge into two different wells: the *available charge* and the *bound charge*. The two wells are connected by a pipe and the charge can migrate in both directions – based on the level of charge in the two wells – at a given rate ω. At the beginning, a fraction $\phi = [0, 1]$ of the total capacity is in the available charge well, and the remaining $1 - \phi$ fraction is in the bound charge well. When the battery is strained, only the available charge may be used, whereas the bound charge slowly becomes available charge. For the purpose of this paper, we are interested in the lifetime equation. It is derived in [18] by solving the differential equations that models the change of the charge of both the wells that, when the load is assumed to be constant, is:

$$I \cdot t + \frac{(1-\phi)\cdot I}{\phi} \cdot \frac{1 - e^{-\omega' t}}{\omega'} = \hat{\gamma} \tag{8}$$

where I is the load that is straining the battery, t is the lifetime of the battery, $\hat{\gamma}$ is the battery full capacity and $\omega' = \omega/[\phi(1-\phi)]$.

Although [16] has already analyzed how uncertainty on input parameters (i.e., initial capacity and load) affects the KiBaM performance in predicting the GomX-1 [4] satellite state of charge, they expressed the uncertainty on the output measure without considering the number of observations of the input parameters. In this sense, our technique lets the users consider the amount of samples to be collected in order to get a more accurate output measure.

3 Modeling M/M/c/K queues power consumption

For the purpose of this paper, the power model presented in [11] and defined by Eq. (2) is our starting point. *Stress* [5] is the benchmark we used to generate the CPU workload. In order to consider different CPU utilization (i.e., $U = \{0\%,\ 25\%,\ 50\%,\ 75\%,\ 100\%\}$), we made *Stress* work with a different amount of threads (i.e., $Threads = \{0,\ 1,\ 2,\ 3,\ 4\}$). Note that, power consumption samples observed when $U = 0\%$ and $U = 100\%$ are P_{idle} and P_{busy} samples, respectively.

Before adopting Eq. (2) to study the power consumption of a server, samples collection and parameters estimation are required. As previously said, the finite number of observation for the input parameters is the source of epistemic uncertainty. Thus, in this paper, uncertainty is introduced into the system by the stochastic parameters P_{idle}, P_{busy}, r and U.

[4] https://gomspace.com/gomx-1.aspx

[5] https://people.seas.harvard.edu/~apw/stress/

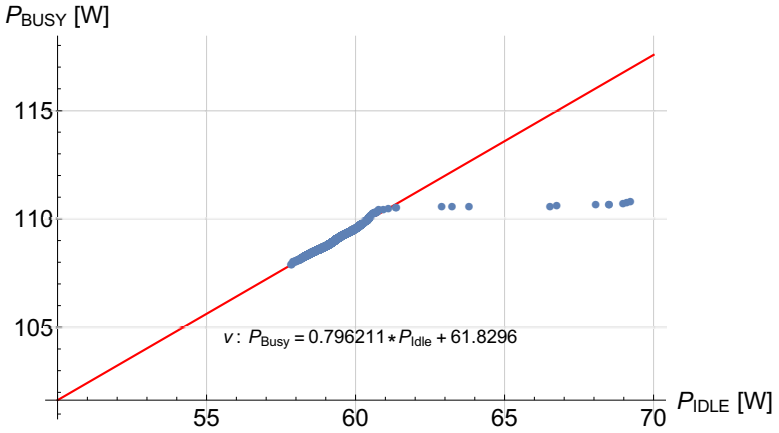


Fig. 1. Q-Q plot representing the linear relationship between P_{idle} and P_{busy} samples.

In order to decrease the number of random variables, we express P_{busy} as a function of P_{idle}. To this end, samples analysis and algebraic manipulation have been performed and a linear relationship between the two parameters has been identified. The relationship is defined by the equation $P_{busy} = m \cdot P_{idle} + q$ where, in the case of the architecture described in Section 1, $m = 0.796211$ and $q = 61.8296$. In Fig. 1 a Q-Q shows the existing linear relationship between P_{busy} and P_{idle}. Some points are not fitted by the linear equation. We argue those outliers are due to background processes that were executed during the measurement campaign, making the system collect some P_{idle} samples when $U \neq 0\%$.

Thus, power model given in Eq. (2) is extended considering the linear relationship, and it becomes:

$$P(U, P_{idle}, r) = P_{idle} + [q + (m-1) \cdot P_{idle}] \cdot (2U - U^r) \tag{9}$$

with $m = 0.796211$ and $q = 61.8296$.

The effectiveness of the power model defined by Eq. (9) is shown in Fig. 2, where a Q-Q plot compares $[P_{Est}|U]$, the distribution of the system power consumption estimated through Eq. (9) with a fixed value U, with $P_{Samp}(U)$, the power consumption distribution estimated from the samples collected during the measurement campaign for the considered utilization level U. In particular, each realization of $[P_{Est}|U]$ is determined starting from the sampled distribution of the system idle, $P_{Samp}(0)$, the considered utilization level U, and the calibration parameter $r = 1.45862$ (where such value has been estimated from the collected samples). In particular we have:

$$[P_{Est}|U] = P_{Samp}(0) + [61.8296 - 0.203789 \cdot P_{Samp}(0)] \cdot (2U - U^{1.45862}) \tag{10}$$

As it can be seen, the obtained curves are closely aligned to the 45° line for all values of U, showing that the proposed procedure is capable of estimating very well the real power consumption distribution for a given utilization level U.

As stated in Section 1, we consider an $M/M/c/K$ queue to avoid stability issues.

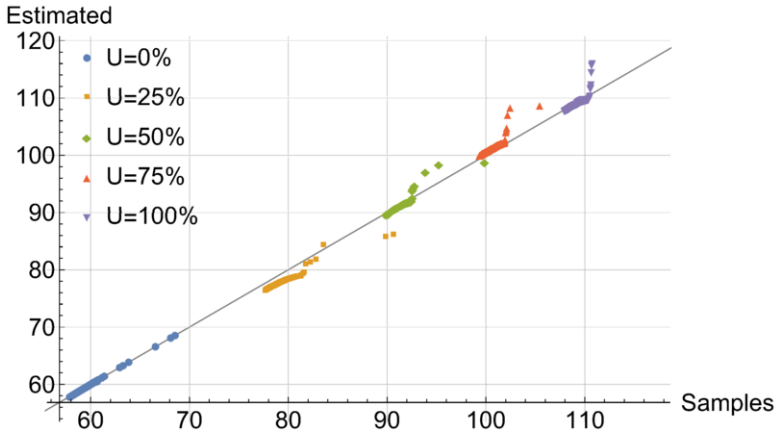


Fig. 2. Q-Q plot comparing the distribution of the estimated power consumption $[P_{Est}|U]$ against the sampled one $P_{Samp}(U)$.

Thus, solving the proper continuous time Markov chain (CTMC), U is computed through equation [33]:

$$U = 1 - \sum_{i=0}^{c} \frac{(c-i)\cdot \pi_i}{c} \tag{11}$$

where the probability to be in state i of the considered CTMC, π_i, is defined as:

$$\pi_i = \begin{cases} \frac{\Lambda^i}{i!M^i}\cdot \pi_0 & \text{for } 0 \le i \le c \\ \frac{\Lambda^i}{c^{i-c}c!M^i}\cdot \pi_0 & \text{for } c \le i \le K \end{cases} \tag{12}$$

and π_0, the probability to be in state 0, is:

$$\pi_0 = \left(\sum_{i=0}^{c-1} \frac{\Lambda^i}{i!M^i} + \sum_{i=c}^{K} \frac{\Lambda^i}{c^{i-c}c!M^i} \right)^{-1} \tag{13}$$

Assuming to know the number of cores of the considered architecture, c, and its buffer capacity, K, the utilization of the resource is a stochastic object on the two r.v. Λ and M, the arrival and service rates, respectively.

Deriving queue utilization, U, starting from Λ and M estimates as in Eq. (11), instead of collecting utilizations samples, has a two-fold advantage: *i)* differently from U, Λ and M estimation does not require to determine an optimal time window; *ii)* it enables *what-if* analysis, thus to study the system under different values of inter-arrival and service rates.

The densities of the input random variables that may affect the output of the power model (i.e., arrival rate Λ, service rate M, P_{idle} and calibration parameter r) are here analyzed.

As said and proved in [21,23,22], the rate of an exponential distribution determined from a finite set of its samples is Erlang distributed. It is the case of the arrival and service rates when dealing with $M/M/c/K$ queues. In particular, since arrival (service) time, X, is exponentially distributed with rate λ (resp. μ), the

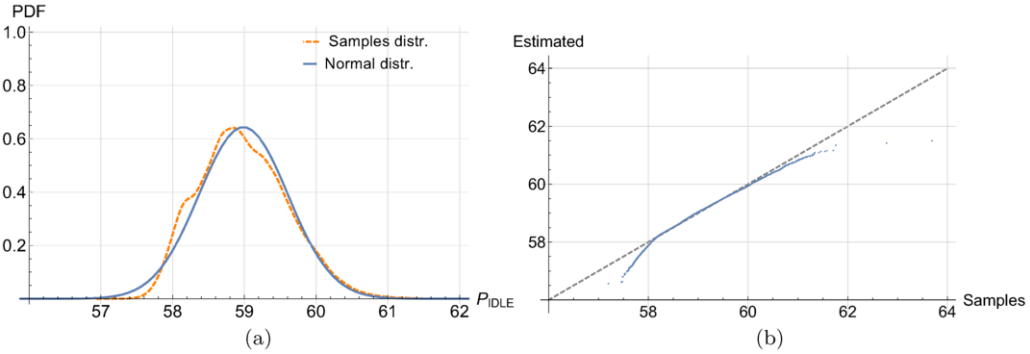


Fig. 3. On the left, the density of the collected P_{idle} samples compared to a $Normal(58.992, 0.619657)$ one. On the right, a Q-Q plot that compares the collected samples to the random variates estimated through the Normal distribution.

random variable $S_X = \sum_{i=1}^{k_x} X_i$ follows a k_X-stage Erlang distribution with rate parameter λ (resp. μ), and its pdf, given $\Lambda = \lambda$ (resp. $M = \mu$), is:

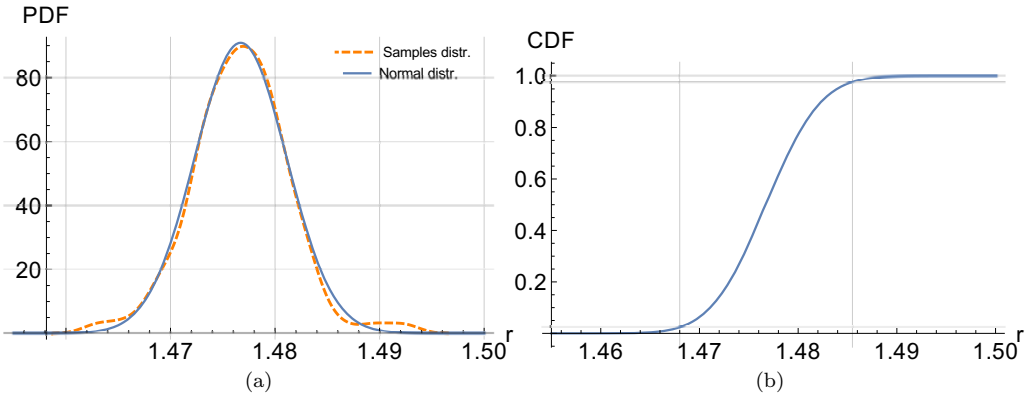
$$f_{S_\Lambda|\Lambda}(s_\Lambda|\lambda) = \frac{\lambda^{k_\Lambda} s_\Lambda^{k_\Lambda - 1} e^{-\lambda s_\Lambda}}{(k_\Lambda - 1)!} \ , \quad f_{S_M|M}(s_M|\mu) = \frac{\mu^{k_M} s_M^{k_M - 1} e^{-\mu s_M}}{(k_M - 1)!} \tag{14}$$

Applying Bayes theorem and using Jeffreys' prior for Λ (resp. M) (i.e., $f_\Lambda(\lambda) = s_\Lambda/\lambda$, and $f_M(\mu) = s_M/\mu$) as done in [22], the pdf of rate parameter Λ (resp. M) is derived as:

$$\begin{aligned} f_{\Lambda|S_\Lambda}(\lambda|s_\Lambda) &= \frac{\lambda^{k_\Lambda - 1} s_\Lambda^{k_\Lambda} e^{-\lambda s_\Lambda}}{(k_\Lambda - 1)!} = Erlang(k_\Lambda, s_\Lambda) \ , \\ f_{M|S_M}(\mu|s_M) &= \frac{\mu^{k_M - 1} s_M^{k} e^{-\mu s_M}}{(k_M - 1)!} = Erlang(k_M, s_M) \end{aligned} \tag{15}$$

With regard to the densities of P_{idle} and calibration parameter r, we use Mathematica [17] to fit the collected samples and find their densities. The results that estimate the P_{idle} density are shown in Fig. 3. They have been derived after collecting 20000 samples. In particular, in Fig. 3a the pdf of the collected samples is compared to the one of the estimated $Normal(58.992, 0.619657)$ distribution. In Fig. 3b a Q-Q plot is used to analyze the two densities. The results confirm that the density of the P_{idle} samples observed during the measurement campaign can be approximated by the estimated $Normal$ distribution. The coefficient of variation (c_v) is computed as the ratio of the standard deviation to the mean, and for $P_{idle} \sim Normal(58.992, 0.619657)$ it is $c_v = 0.0105$. Since it is larger than 1%, the impact of P_{idle} density on the output of the model described by Eq. (9) is further analyzed in the next Section.

In order to evaluate the calibration parameter r density, 3700 samples have been collected for each considered value of utilization, U; for this purpose, we used 3700 samples since it has been found to be a good trade-off between the quality of the measure and the time required for samples collection. Then, 100 sets have been created, each one composed by 185 randomly selected samples (i.e., 37 samples for each value of U). Finally, starting from the 100 sets, Mathematica is used to estimate



(a) (b)

Fig. 4. On the left, the density of the estimated r samples compared to a $Normal(1.45862, 0.00439)$ one. On the right, the Cdf of the normal distribution.

the values of parameter r and its density. The results are shown in Fig. 4, where the pdf and cumulative distribution function (Cdf) of $r \sim Normal(1.45862, 0.00439)$ is plotted. Due to its small coefficient of variation, $c_v = 0.003$, calibration parameter r is not considered as a source of uncertainty for the purpose of this paper and it is assumed to be exactly known.

4 Epistemic uncertainty in power models

In this Section, the techniques used to study uncertainty propagation in power model described by Eq. (9) are presented. In order to identify the parameters that must be considered to obtain more accurate results, we analyze two different cases: *i)* the uncertainty is introduced into the system by P_{idle}, Λ and M, and *ii)* only Λ and M are the sources of uncertainty. Then, the validation of the chosen equation is performed, using Monte Carlo samples, for some different values of resource utilization, U.

4.1 Epistemic uncertainty in M/M/c/K queues

The equations for epistemic uncertainty propagation presented in Section 2 are now extended and adapted to the case of power consumption estimation in $M/M/c/K$ queues.

As said in Section 3, for the purpose of this paper we assume epistemic uncertainty is introduced into the system by input random variables P_{idle}, Λ and M, since P_{busy} is expressed as a function of P_{idle} and the density of the calibration parameter r has a small coefficient of variation.
Considering three input parameters are introducing uncertainty into the power model, a three-dimensional integral (i.e., one for each input random variable) is required to compute the unconditional Cdf and expected value of the output metric. In particular, assuming the epistemic random variables to be independent as

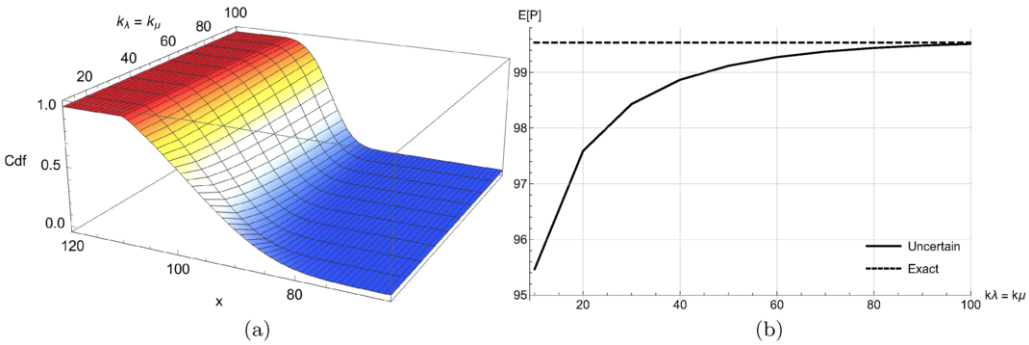


Fig. 5. The power consumption Cdf (a) and expected value (b) plotted as a function of the collected samples for the two input parameters, when $\hat{\lambda} = 6.4$ j/s and $\hat{\mu} = 2.25$ j/s.

said in Section 2, the Cdf described in Eq. (6) becomes:

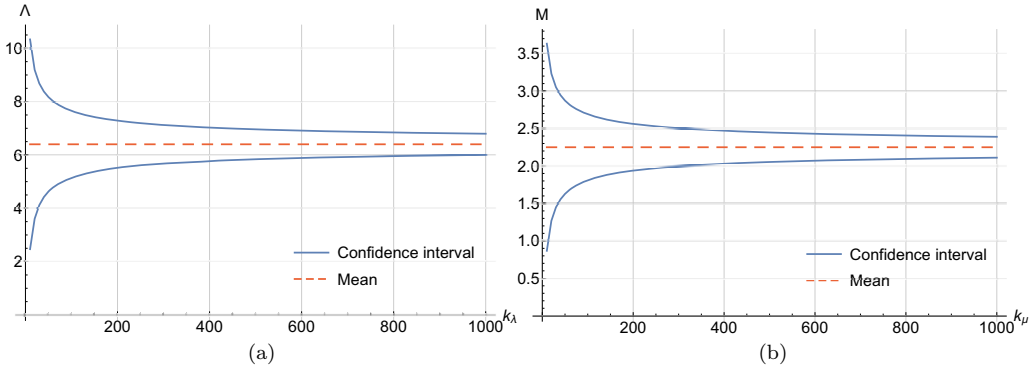
$$F_{P(U|P_{idle}=p,\Lambda=\lambda,M=\mu)}(x) = \int_0^\infty \int_0^\infty \int_0^\infty \mathbb{1}(P(U|P_{idle}=p,\Lambda=\lambda,M=\mu) \leq x) \cdot f_\Lambda(\lambda) \cdot f_M(\mu) \cdot f_{P_{idle}}(p)\ d\lambda\ d\mu\ dp \tag{16}$$

and the expected value in Eq. (7) is computed as:

$$\mathbb{E}[P(U|P_{idle}=p,\Lambda=\lambda,M=\mu)] = \int_0^\infty \int_0^\infty \int_0^\infty P(U|P_{idle}=p,\Lambda=\lambda,M=\mu) \cdot f_\Lambda(\lambda) \cdot f_M(\mu) \cdot f_{P_{idle}}(p)\ d\lambda\ d\mu\ dp \tag{17}$$

The limits of integration of each input parameter vary between 0 and infinity since there are not constraints on the values the input parameters may assume. In particular, since we are considering a finite capacity $M/M/c/K$ queue, the system is always stable and no constraints on the value of Λ are given. The epistemic densities of r.v. Λ and M, i.e., $f_\Lambda(\lambda)$ and $f_M(\mu)$, are given in Eq. (15), whereas $f_{Pidle}(p)$ is the P_{idle} probability density function, that is $P_{idle} \sim Normal(58.992, 0.619657)$ as shown in Section 3.

Let us call respectively k_λ and k_μ the number of samples collected before estimating the values of Λ and M. The Cdf of the power consumption and its expected value derived through Eqs. (16) and (17), respectively, are shown in Fig. 5. They have been solved through numerical integrations – performed by Mathematica – due to the complexity to get closed form solutions. Both the statistical measures are plotted against the number of samples collected for Λ and M estimation. For the sake of simplicity, only the case when $k_\lambda = k_\mu$ (i.e., the same number of samples is collected to estimate arrival and service rates) is represented in Fig. 5, and only $U \simeq 70\%$ is considered (i.e., average arrival rate $\hat{\lambda} = 6.4$ jobs/second, and average service rate $\hat{\mu} = 2.25$ jobs/second). As expected, the results show the importance to collect as many samples as possible for each input parameter to make their con-



(a) (b)

Fig. 6. 95% confidence interval for the power model input parameters Λ and M as a function of the number of observations. On the left, the confidence interval length for $\hat{\lambda} = 6.4$ j/s; on the right, the confidence interval length for $\hat{\mu} = 2.25$ j/s.

fidence interval tighter. Indeed, starting from the number of collected samples, k, it is possible to derive the confidence interval of an exponential distributed random variable as follows [23]:

$$2d = \frac{1}{2s_k}\left\{\chi^2_{2k,\alpha/2} - \chi^2_{2k,1-\alpha/2}\right\} \tag{18}$$

where d is the half-width of the confidence interval of the considered input random variable, s_k is the sum of the k collected samples and $\chi^2_{2k,1-\alpha/2}$ is the critical value of the chi-square distribution with $2k$ degrees of freedom. The confidence interval length of r.v. Λ and M, computed with Eq. (18) for $U \simeq 70\%$, are plotted in Fig. 6 as a function of the number of observations. The straight lines are the lower and upper bounds, while the dashed ones are the average values $\hat{\lambda} = 6.4$ j/s and $\hat{\mu} = 2.25$ j/s. Similar results may be obtained for different resource utilization, varying the average arrival rate, $\hat{\lambda}$.

In order to evaluate the impact of P_{idle} input random variable on the output metric of the considered power model, we adapt Eqs. (16) and (17) to make them take into account only the effect of Λ and M, assuming P_{idle} is estimated with certainty. Thus, they become:

$$F_{P(U|\Lambda=\lambda,M=\mu)}(x) = \int_0^\infty \int_0^\infty \mathbb{1}(P(U|\Lambda=\lambda, M=\mu) \le x) \cdot f_\Lambda(\lambda) \cdot f_M(\mu)\, d\lambda\, d\mu \tag{19}$$

and

$$\mathbb{E}[P(U|\Lambda=\lambda, M=\mu)] = \int_0^\infty \int_0^\infty P(U|\Lambda=\lambda, M=\mu) \cdot f_\Lambda(\lambda) \cdot f_M(\mu)\, d\lambda\, d\mu \tag{20}$$

respectively. Also in this case numerical solutions have been derived. Since the results computed through Eqs. (19) and (20) for $U \simeq 70\%$ are similar to those

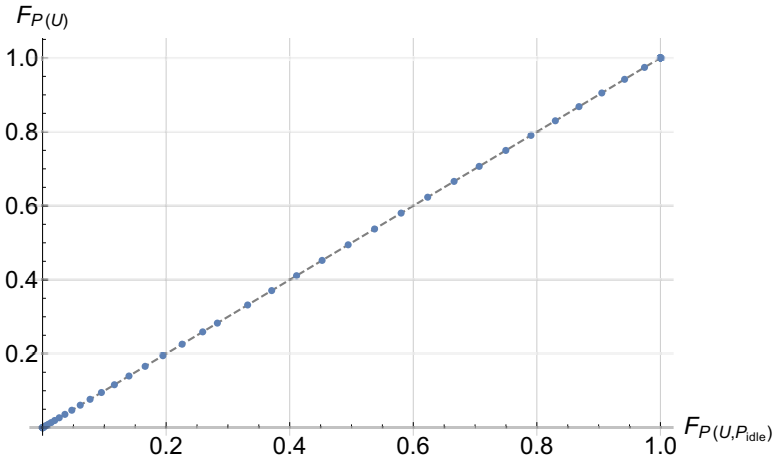


Fig. 7. P-P plot that compares the power consumption Cdfs computed through Eqs. (16) and (19), when $k_\lambda = k_\mu = 10$.

depicted in Fig. 5, they are not plotted. In order to show that the Cdfs and expected values derived in the two different ways are identical, a P-P plot is used for the Cdfs comparison and is depicted in Fig. 7. For the sake of space, only the P-P plot for $k_\lambda = k_\mu = 10$ is shown here. Similar results are obtained considering the Cdfs obtained with larger values of $k_\lambda = k_\mu$.

Based on the presented results, we can assert the epistemic uncertainty introduced by P_{idle} into the considered power model is negligible, and the uncertainty analysis can be performed just studying the densities of Λ and M, thus adopting Eq. (19) to derive the Cdf of the power consumption, and Eq. (20) for its expected value. Note that this assumption has also the effect of reducing the computational complexity, allowing to deal with only a 2-dimensional integral.

Finally, in the limiting case $k_\lambda, k_\mu \to \infty$, the Erlang distributed input r.v. have a zero variance, thus they are deterministic distributed (or degenerate) [15] and they take a single value (i.e., $\lambda = \hat{\lambda}$ and $\mu = \hat{\mu}$). In this case, the considered model is no more affected by epistemic uncertainty, and it may be studied without accounting for epistemic uncertainty propagation.

4.2 Validation

In order to validate Eq. (19) that is used to derive the Cdf of the power model given in Eq. (9) for $M/M/c/K$ queues, we run a further benchmark on the architecture presented in Section 1. The benchmark consists in executing a given number of CPU intensive requests. The arrival and service times for each request are passed by the user at the beginning of the test.

For the purpose of this validation, we generate 1000 exponentially distributed samples for arrival and service times. This step is repeated for each value of resource utilization that has been studied. In particular, the average service rate $\hat{\mu}$ is set to 2.25 jobs/second, and the average arrival rate varies among $\hat{\lambda} = \{0.8, 1.6, 2.4, 3.2, 4.0, 4.8, 5.6, 6.4, 7.2\}$ jobs/second, in order to consider different resource utilizations, that are derived as shown by Eq. (11).

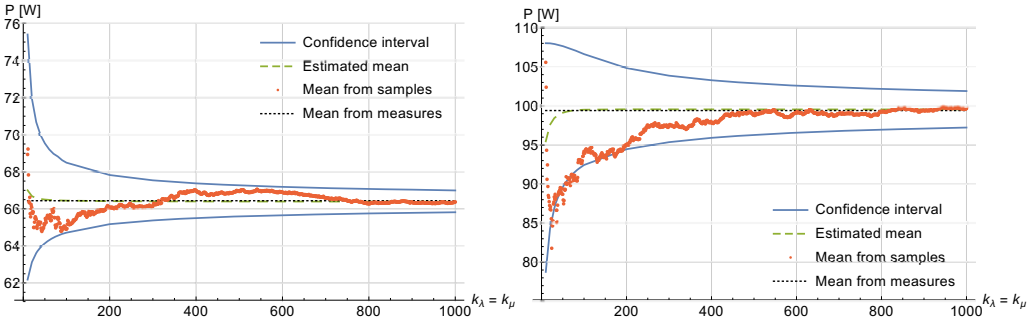


Fig. 8. The results for $U \simeq 9\%$ are depicted on the left, those for $U \simeq 70\%$ are depicted on the right.

In addition to the generated arrival and service times samples, we also collect the power consumption measures recorded by a Yokogawa WT210 power meter [6].

For the sake of clarity, only the results for $U \simeq 9\%$ and $U \simeq 70\%$ (i.e., $\hat{\lambda} = 0.8$ and $\hat{\lambda} = 6.4$, respectively) are depicted in Fig. 8.

They are plotted as a function of the number of samples observed for arrival and service rates estimation, k_λ and k_μ, that in this paper are assumed to be always equal. In both the figures, the lower and upper bounds (i.e., the straight blue lines) have been derived starting from the power consumption Cdf computed through Eq. (19). Then, the 95% confidence interval has been derived for each Cdf, and the two bounds have been plotted varying the number of observations. Eq. (20) is used to compute the expected value in both the figure (i.e., the dashed green line).

The samples of arrival and service times are used to plot the estimated power consumption (the red dots). For this purpose, k_λ and k_μ samples (i.e., the number of considered Monte Carlo samples) are averaged in order to derive the mean arrival and service rates (i.e., $\hat{\lambda}$ and $\hat{\mu}$), respectively. Then, after using these values to compute U as in Eq. (11), it is possible to estimate average power consumption through Eq. (9), with P_{idle} and r set to their mean values. The larger the number of observations of arrival and service times, the more accurate the estimated power consumption.

As expected, the power consumption estimates usually lie between the two bounds and they are closer to the average one when a larger amount of samples is observed.

Finally, the power consumption measured by means of the power meter is depicted in the two graphs as a black dotted line. In both the cases, the measured power consumption and its expected value are close after collecting few samples of arrival and service rates.

[6] http://tmi.yokogawa.com/products/digital-power-analyzers/digital-power-analyzers/

5 Exploitation: life of a UPS battery

In this Section we apply the previous considerations about uncertainty propagation in power models to a real exploitation case. We analyze a UPS battery that starts working after the power source failed. In particular, we wish to estimate the battery lifetime of a UPS that must keep on a server which is processing a constant workload.

Three main UPS architectures exists [38]: *i)* on-line UPS, that provides electrical power from the battery whether there is a failure or not; *ii)* standby UPS, that provides electrical power from the battery only if power source failed; *iii)* line-interactive UPS, that are high-class standby UPS. For the purpose of this paper, distinction between the three architectures is irrelevant since we study the UPS unit when power source has already failed.

Furthermore, batteries are usually not fully drained in order to prevent damages [7]. Thus, a UPS unit is shut down when the battery capacity is lower than a fixed threshold. Nevertheless, in this paper we assume the battery can be completely depleted and the UPS unit does not shut down until the battery is exhausted. Note that, the standard UPS battery configuration may be considered after some adjustments to Eq. (8).

The battery lifetime is derived solving the kinetic battery model for fixed load given in Eq. (8). For that purpose, we set $\phi = 1/2$ and $\omega = 1/100$ as done by Hermanns et al. in [16]. They also assumed the battery initial capacity is uniform distributed between 70% and 90% of full capacity $\hat{\gamma}$, thus assuming to deal with a random environment. Furthermore, in our case, load I is affected by epistemic uncertainty due to the finite number of samples collected for the input parameters Λ and M (i.e., arrival and service rates, respectively). Indeed, we derive load I starting from server utilization in Eq. (11); utilization is used to compute the server's power consumption through Eq. (9), that is divided by voltage ΔV to obtain the load I (i.e., $I = P(U, P_{idle}, r)/\Delta V$). This is a different assumption with respect to [16], where also the load was affected by aleatory uncertainty. Finally, voltage is $\Delta V = 12$ V and full capacity $\hat{\gamma} = 9$ Ah, that are from the specifications of Trust OXXTRON 1000VA UPS [7].

Starting from Eq. (8), we define battery lifetime L as:

$$L(\lambda, \mu, \gamma) = t \mid I(\lambda, \mu) \cdot t + \frac{(1-\phi) \cdot I(\lambda, \mu)}{\phi} \cdot \frac{1 - e^{-\omega' t}}{\omega'} = \gamma$$

For the sake of simplicity, we have dropped from L its dependency on λ, μ and γ. Thus, we propagate the epistemic uncertainty of input parameters Λ and M to the battery lifetime as follows:

$$F_L(l) = \int_0^\infty \int_0^\infty \int_0^\infty \mathbb{1}(L \leq l) \cdot f_\Lambda(\lambda) \cdot f_M(\mu) \cdot f_\Gamma(\gamma)\, d\lambda\, d\mu\, d\gamma \tag{21}$$

[7] http://www.trust.com/en/product/17680-oxxtron-1000va-ups

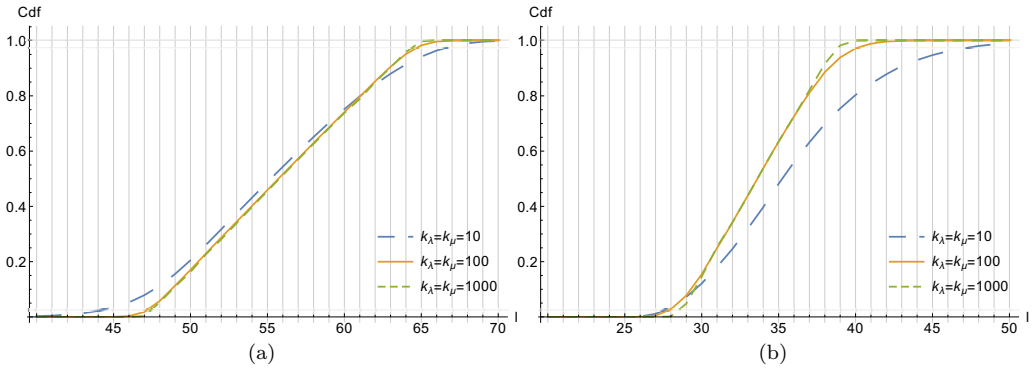


Fig. 9. Battery lifetime Cdfs when the resource utilization is $U \simeq 9\%$ (on the left) and $U \simeq 70\%$ (on the right).

where

$$f_\Gamma(\gamma) \sim Uniform\left(\frac{70\hat{\gamma}}{100}, \frac{90\hat{\gamma}}{100}\right)$$

and $f_\Lambda(\lambda)$ and $f_M(\mu)$ are given in Eq. (15).

The Cdfs obtained solving Eq. (21) for $\lambda = 0.8$ j/s and $\lambda = 6.4$ j/s are depicted in Figs. 9a and 9b, respectively. Although tighter confidence intervals are derived increasing the number of observation for arrival and service rates from 10 to 100, a smaller improvement is obtained when the number of collected samples is further increased. Indeed, in both the cases depicted in Fig. 9, the Cdf of the battery lifetime for $k_\lambda = k_\mu = 1000$ is very similar to the one derived with $k_\lambda = k_\mu = 100$.

In fact, in this case the inaccuracy of the output measure is mainly due to the aleatory uncertainty introduced by the unknown battery capacity γ. That may also be seen considering the limiting case for $k_\lambda, k_\mu \to \infty$ discussed in Section 4, when arrival and service rates, Λ and M, follows a degenerate distribution (i.e., they are known with certainty), while the full capacity γ is still uniformly distributed. The battery lifetime Cdf for $k_\lambda = k_\mu = 1000$ and $k_\lambda, k_\mu \to \infty$, when $\lambda = 6.4$ j/s, are plotted in Fig. 10, where it is shown that the difference between the two Cdfs is negligible. Similar results are obtained for $\lambda = 0.8$ j/s.

6 Conclusion

In this paper we apply epistemic uncertainty propagation to power consumption models. To the best of our knowledge, this is the first time epistemic uncertainty propagation is used to study this kind of problem. Indeed, in previous work the input parameters have always been considered exact, independently on the number of observed samples for their estimations.

The nonlinear power model presented by Fan et al. in [11] is the one we use to analyze our machine that we model as an $M/M/c/K$ queue. We prove the measured power consumption fit the one estimated by the chosen power model. Moreover, we also show that uncertainty of P_{idle} and calibration parameter r are negligible and do not deeply affect the output of the model, since the two input parameters have a

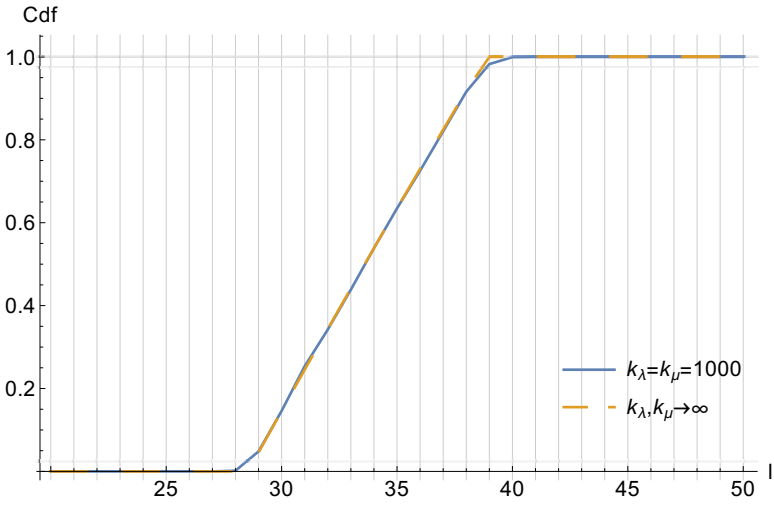


Fig. 10. Battery lifetime Cdfs when the resource utilization is $U \simeq 70\%$ and the number of collected samples is 1000 or tends to infinity.

small coefficient of variation. Thus, we assume only arrival and service rates (Λ and M, respectively) are the input parameters which introduce epistemic uncertainty into the power model. Then, the technique to study the epistemic uncertainty propagation is shown and it is validated using Monte Carlo method and measuring the power consumption of our machine. Finally, the theoretical results are applied to a case-study, where a UPS unit is taken into account. In particular, we estimate how epistemic and aleatory uncertainties impact on the output measure of a model that estimates the lifetime of a UPS battery when it must serve a constant load affected by epistemic uncertainty.

In the future, three main aspects may interestingly extend this paper. First, different power models may be considered in order to analyze different situations; for examples, as said in Section 2, there are power consumption models that let the users consider dynamic voltage/frequency scaling and simultaneous multi-threading, or through which is possible to evaluate the power consumption of a whole data-center. Since a larger number of input parameters is required to use more accurate power models, the multi-dimensional integral technique adopted in this paper could be hardly utilized, thus other uncertainty propagation techniques should be used.

Second, the $M/M/c/K$ queue may be substituted by an $M/M/c/K/Setup$ model; queues with setup have c servers that may be turned on/off based on the incoming workload. However, the off servers require a *Setup* time in order to be properly activated and process new requests. Queues with setup are usually adopted to model data-centers [26,27] and to study their power-performance trade-off. Since the setup time must be estimated from the collected samples, it may be useful to further investigate this kind of models using epistemic uncertainty, in order to better understand how that parameter affects the output measures.

The third direction that may be pursued is about the UPS battery case-study. In particular, it may be interesting to study the UPS system performability, thus considering also the power source failure and repair rates. Indeed, also those two

parameters are introducing some epistemic uncertainties into the battery model, and they should be taken into account with the already studied input random variables in order to obtain more accurate results.

Acknowledgement

Gribaudo and Pinciroli's research was supported in part by the European Commission under the grant ANTAREX H2020 FET-HPC-671623. Trivedi's research was partially supported by US NSF grant number CNS-1523994.

References

[1] Bohra, A. E. H. and V. Chaudhary, *Vmeter: Power modelling for virtualized clouds*, in: *Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on*, Ieee, 2010, pp. 1–8.

[2] Boru, D., D. Kliazovich, F. Granelli, P. Bouvry and A. Y. Zomaya, *Energy-efficient data replication in cloud computing datacenters*, Cluster Computing **18** (2015), pp. 385–402.

[3] Buyya, R., A. Beloglazov and J. Abawajy, *Energy-efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges*, arXiv preprint arXiv:1006.0308 (2010).

[4] Casalicchio, E., L. Lundberg and S. Shirinbab, *Energy-aware adaptation in managed cassandra datacenters*, in: *Cloud and Autonomic Computing (ICCAC), 2016 International Conference on*, IEEE, 2016, pp. 60–71.

[5] Cerotti, D., M. Gribaudo, P. Piazzolla, R. Pinciroli and G. Serazzi, *Modeling power consumption in multicore cpus with multithreading and frequency scaling*, in: *Information Sciences and Systems 2015*, Springer, 2016 pp. 81–90.

[6] Cerotti, D., M. Gribaudo, R. Pinciroli and G. Serazzi, *Stochastic analysis of energy consumption in pool depletion systems*, in: *International GI/ITG Conference on Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance*, Springer, 2016, pp. 25–39.

[7] Daigle, M. and C. S. Kulkarni, *End-of-discharge and end-of-life prediction in lithium-ion batteries with electrochemistry-based aging models*, in: *AIAA Infotech@ Aerospace*, 2016 p. 2132.

[8] Doyle, M., T. F. Fuller and J. Newman, *Modeling of galvanostatic charge and discharge of the lithium/polymer/insertion cell*, Journal of the Electrochemical Society **140** (1993), pp. 1526–1533.

[9] Easterling, R. G., *Approximate confidence limits for system reliability*, Journal of the American Statistical Association **67** (1972), pp. 220–222.

[10] Ebrahimi, K., G. F. Jones and A. S. Fleischer, *A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities*, Renewable and Sustainable Energy Reviews **31** (2014), pp. 622–638.

[11] Fan, X., W.-D. Weber and L. A. Barroso, *Power provisioning for a warehouse-sized computer*, in: *Proceedings of the 34th Annual International Symposium on Computer Architecture*, ISCA '07 (2007), pp. 13–23.

[12] Fang, Q., J. Wang and Q. Gong, *Qos-driven power management of data centers via model predictive control*, IEEE Transactions on Automation Science and Engineering **13** (2016), pp. 1557–1566.

[13] Gelenbe, E. and C. Rosenberg, *Queues with slowly varying arrival and service processes*, Management Science **36** (1990), pp. 928–937.

[14] Ghosh, R., F. Longo, R. Xia, V. K. Naik and K. S. Trivedi, *Stochastic model driven capacity planning for an infrastructure-as-a-service cloud*, IEEE Transactions on Services Computing **7** (2014), pp. 667–680.

[15] Heragu, S. S., "Facilities design," CRC Press, 2008 pp. 539–541.

[16] Hermanns, H., J. Krčál and G. Nies, *How is your satellite doing? battery kinetics with recharging and uncertainty*, Leibniz Transactions on Embedded Systems **4** (2017), pp. 04–1.

[17] Inc., W. R., *Mathematica, Version 11.1*, champaign, IL, 2017.

[18] Jongerden, M. R. and B. R. Haverkort, *Which battery model to use?*, IET software **3** (2009), pp. 445–457.

[19] Kontorinis, V., L. E. Zhang, B. Aksanli, J. Sampson, H. Homayoun, E. Pettis, D. M. Tullsen and T. S. Rosing, *Managing distributed ups energy for effective power capping in data centers*, in: *Computer Architecture (ISCA), 2012 39th Annual International Symposium on*, IEEE, 2012, pp. 488–499.

[20] Meisner, D., J. Wu and T. F. Wenisch, *Bighouse: A simulation infrastructure for data center systems*, in: *Performance Analysis of Systems and Software (ISPASS), 2012 IEEE International Symposium on*, IEEE, 2012, pp. 35–45.

[21] Mishra, K. and K. S. Trivedi, *Uncertainty propagation through software dependability models*, in: *Software Reliability Engineering (ISSRE), 2011 IEEE 22nd International Symposium on*, IEEE, 2011, pp. 80–89.

[22] Mishra, K. and K. S. Trivedi, *Closed-form approach for epistemic uncertainty propagation in analytic models*, in: *Stochastic Reliability and Maintenance Modeling*, Springer, 2013 pp. 315–332.

[23] Mishra, K., K. S. Trivedi and R. R. Some, *Uncertainty analysis of the remote exploration and experimentation system*, Journal of Spacecraft and Rockets **49** (2012), pp. 1032–1042.

[24] Nan, X., Y. He and L. Guan, *Optimal resource allocation for multimedia cloud based on queuing model*, in: *Multimedia signal processing (MMSP), 2011 IEEE 13th international workshop on*, IEEE, 2011, pp. 1–6.

[25] Pakbaznia, E. and M. Pedram, *Minimizing data center cooling and server power costs*, in: *Proceedings of the 2009 ACM/IEEE international symposium on Low power electronics and design*, ACM, 2009, pp. 145–150.

[26] Phung-Duc, T., *Multiserver queues with finite capacity and setup time*, in: *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, Springer, 2015, pp. 173–187.

[27] Phung-Duc, T., *Exact solutions for m/m/c/setup queues*, Telecommunication Systems **64** (2017), pp. 309–324.

[28] Pinciroli, R., K. Trivedi and A. Bobbio, *Parametric sensitivity and uncertainty propagation in dependability models*, in: *10th EAI International Conference on Performance Evaluation Methodologies and Tools*, ACM, 2017, pp. 44–51.

[29] Priya, B., E. S. Pilli and R. C. Joshi, *A survey on energy and power consumption models for greener cloud*, in: *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, IEEE, 2013, pp. 76–82.

[30] Rosenberg, C., R. Mazumdar and L. Kleinrock, *On the analysis of exponential queuing systems with randomly changing arrival rates: stability conditions and finite buffer scheme with a resume level*, Performance Evaluation **11** (1990), pp. 283–292.

[31] Shehabi, A., S. Smith, N. Horner, I. Azevedo, R. Brown, J. Koomey, E. Masanet, D. Sartor, M. Herrlin and W. Lintner, *United states data center energy usage report*, Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775 Page **4** (2016).

[32] Strack, O., R. Leavy and R. M. Brannon, *Aleatory uncertainty and scale effects in computational damage models for failure and fragmentation*, International Journal for Numerical Methods in Engineering **102** (2015), pp. 468–495.

[33] Sztrik, J., *Basic queueing theory*, University of Debrecen, Faculty of Informatics **193** (2012).

[34] Tannenbaum, D., C. R. Fox and G. Ülkümen, *Judgment extremity and accuracy under epistemic vs. aleatory uncertainty*, Management Science (2016).

[35] Trivedi, K. S., "Probability & statistics with reliability, queuing and computer science applications," John Wiley & Sons, 2008.

[36] Trubiani, C., I. Meedeniya, V. Cortellessa, A. Aleti and L. Grunske, *Model-based performance analysis of software architectures under uncertainty*, in: *Proceedings of the 9th international ACM Sigsoft conference on Quality of software architectures*, ACM, 2013, pp. 69–78.

[37] Xia, Y., M. Zhou, X. Luo, S. Pang and Q. Zhu, *A stochastic approach to analysis of energy-aware dvs-enabled cloud datacenters*, IEEE Transactions on Systems, Man, and Cybernetics: Systems **45** (2015), pp. 73–83.

[38] Yin, L., R. M. Fricks and K. S. Trivedi, *Application of semi-markov process and ctmc to evaluation of ups system availability*, in: *Reliability and Maintainability Symposium, 2002. Proceedings. Annual*, IEEE, 2002, pp. 584–591.

[39] Zhang, X., T. Lindberg, K. Svensson, V. Vyatkin and A. Mousavi, *Power consumption modeling of data center it room with distributed air flow*, International Journal of Modeling and Optimization **6** (2016), p. 33.

[40] Zwirglmaier, K. and D. Straub, *A discretization procedure for rare events in bayesian networks*, Reliability Engineering & System Safety **153** (2016), pp. 96–109.