2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)

Multi-Granularity Power Prediction for Data Center Operations via Long Short-Term Memory Network

Ziyu Shen, Xusheng Zhang, Bin Xia, Zheng Liu, Yun Li
Jiangsu Key Laboratory of Big Data Security & Intelligent Processing
Nanjing University of Posts and Telecommunications
Nanjing, China
liyun@njupt.edu.cn

***Abstract*—The increasing numbers of the applications and requirement of cloud computing have made huge power consumption in data centers, which brings the problems of the high cost and resource waste. This problem attracts significant attention from academia and industry. A critical approach to solve this problem is constructing an intelligent energy management system for data centers. Furthermore, an efficient assessment and prediction module of power consumption in data centers is an essential part of the management system. It facilitates cloud service providers to perform workflow scheduling at the minimal cost and energy efficiency management with the requirement of QoS. Since the assessment and prediction of power consumption correlate, this paper presents a multi-granularity approach for power consumption prediction in data centers, which combines multi-task learning with the LSTM network. We first transfer a multi-granularity power prediction problem into a multi-task regression problem to assess and predict the power consumption of data center system maintenance and scheduling operations. Due to the time requirement for workflow and container scheduling, the prediction interval is 30 seconds. Then we propose an efficient long short-term memory network for the multi-granularity prediction. The experimental results show our model outperforms other prediction models on the real datasets.**

***Index Terms*—Data center, Time series prediction, Energy efficiency, Power consumption**

I. Introduction

As the development of cloud computing and increasing numbers of the applications, the data centers have caused large power consumption. In China, the total number of data centers is more than 40,000, of which the annual electricity consumption exceeds 1.5% of the whole social electricity consumption. The yearly expense for power distribution units (PDUs) and cooling devices exceeds 20 billion dollars. The power consumption accounts for half of a data centers total expense [1] as well as generates much carbon dioxide [2], which accelerates global warming. Moreover, the problem of low energy efficiency for data centers cannot be ignored. Generally, about 15% servers are in idle, and CPU usage of 75% servers is less than 20%, which has brought the high cost to the cloud service providers (https://www.zdnet.com/). Therefore, it is a critical challenge to improve energy efficiency and reduce power consumption.

Many technologies are applied in data centers to improve energy efficiency. The most widely used technique is dynamic voltage and frequency scaling (DVFS) [3], which changes the voltage dynamically according to processing speed, to reduce the processor power consumption. Other technologies, such as live migration [4] and task consolidation, control the CPU utilization of the servers to reduce the power consumption. However, these technologies are not enough to improve energy efficiency due to ignoring the relationship among power consumption, cloud applications, and platform architectures.

To effectively reduce power consumption and help cloud service providers to allocate workflows and meet QoS requirements, it is very important to assess and predict power consumption in data centers accurately [5]. Recently, Google is devoted to developing a predictive model of power usage efficiency (PUE) for large-scale data center and use DeepMind to improve its energy efficiency [6] [7]. In this paper, we propose a method to assess and predict the power consumption in data centers simultaneously. Existing works treat the assessment and prediction of power consumption independently but do not explore the correlation. To the best of our knowledge, our work is the first to address multi-granularity energy consumption predictions in data centers from the perspective of multi-task learning. We consider simultaneously obtaining the power consumption at the next second power and the average power consumption at the next period (30 seconds). The next second power prediction can be thought of as a fine-grained prediction, and the next 30 seconds power prediction corresponds to a medium-grained prediction. It is meaningful to assess and predict the power consumption in data centers. We predict the power consumption of the next second to assess the power usage state of the data center, for example, by accurately assessing power consumption to judge whether the data center runs properly. Due to the time constraints of workflow scheduling and task management, we predict the average power consumption of the data center in the next 30 seconds. We use LSTM and multi-task learning to predict multi-granularity power consumption because the two prediction tasks are closely correlated. LSTM fits the problem because it can iteratively fine-tune for the coarse prediction and deal with the time dependence of sequential temporal data well. And multi-task learning is intended to impose shared knowledge when solving multiple correlated tasks simultaneously [8].

Our work contributes in the following two ways.

- We define a multi-granularity prediction problem for power consumption. We set up two power systems where

978-1-7281-4328-6/19/$31.00 ©2019 IEEE
DOI 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00037

different workloads are run to generate power-related data.
- We construct a multi-task LSTM model to assess and predict power consumption and compare our approach with other single task models.

This paper is structured as follows. In Section 2, we introduce the related approaches and background of the power consumption prediction and multi-task learning. Section 3 presents the formulation of the multi-granularity power problem and the framework of the power consumption management system. Section 4 shows the details of our proposed model. To validate the advantage and efficiency of our model, we provide sufficient experimental comparison with other state-of-art methods in Section 5. We conclude the whole paper and discuss future work in Section 6.

II. Related work

Power consumption reduction in data centers is a hot research topic for decades, much effort has been made to improve the energy efficiency of data centers. DVFS, dynamic power management (DPM) [9], power napping [10] and live migration, saving energy technologies, are widely applied in data centers. However, a data center consists of many complex components, such as cooling, power transformation, information and communication technology (ICT) equipment, and management subsystem. These technologies mentioned above are not enough to save much electricity, then researchers like to model the relationship between power consumption and characteristic of data centers to reduce the power consumption effectively.

Some existing study has made efforts in modeling the relationship between power consumption and performance counters. The process of these methods is usually divided into three steps. First, collecting power consumption data and relevant effective factors, such as performance counters and system utilization [11] when running applications. Then, the related parameters are required to be fit to generate the model. Finally, using the model for power prediction. At the initial time, a linear relationship between processor power consumption and multiple performance counters is be found in Bellosa's power model [12], which correlates power consumption with performance counters at the processor level. Joseph,et al [13] and Isci, et al [14] proposed detailed analytical processor power models based on CPU performance counters. However, the approaches are lack of generality and portability because of the limited microarchitectural knowledge of a particular processor. Mantis [15], proposed by Ecomomou, a non-intrusive infrastructure for providing fast and accurate full-system power predictions. Mantis models the linear relationship between power measured from a wall socket and four distinct utilization subsystem counters. However, the simple accumulation models can not describe the power consumption fully and accurately, becasue the models are based on every component power accumulated while not all features are added into them.

As cloud data centers continue to grow in size and power consumption becomes more complex, previous approaches to modeling the relationship between power consumption and data center components based on simple physical measurements failed to achieve all the factors, resulting in power consumption cannot be accurately assessed.

Therefore, the power consumption data has been seen as a time series to assess and predict in recent years. Much effort has been made to develop and improve time series predicting models in many domains, such as electricity demand prediction, financial market prediction, and wind power prediction. Box and Jenkins firstly proposed ARIMA model in 1970. ARIMA is flexible that it can represent autoregressive (AR), moving average (MA) and ARMA models. And the exponential smoothing has many variants: simple exponential smoothing, Holts exponential smoothing, and Holt-Winters. Anwar M Y [16] used ARIMA model to predict future trends in incidence by historical data on the number of endemic malaria infections in Afghanistan. Mazumdar [17] used ARIMA and exponential smoothing models to predict the stability of the data center based on real-time data of batch workload. Maurizio [18] predicted data center power consumption with the Holt-Winter method and experiment on four datasets.

However, these algorithms only take the current feature values into account, not consider the temporal characteristic of the time series data. Recently, many methods based on deep learning are proposed to solve the time series problems. Li [19] proposed two deep learning-based models: a fine-grained model and a coarse-grained model with auto-encoder employed to encode data and also used smoothing to remove the noise in data. Liu [20] adopt the LSTM neural network for workload prediction to allocate VM resource. LSTM is a recurrent neural network architecture, which can be applied for time series prediction. LSTM can eliminate the gradient vanishing problem and capture the long-term dependence in time series [21]. In this paper, we use LSTM based model to assess and predict the power consumption in data centers. We are the first to use LSTM and multi-task learning to assess and predict the power consumption simultaneously.

Multi-task learning [22] utilizes the feature correlations to model the relationships among related tasks. Multi-task learning can boost the performance of the tasks by learning the knowledge sharing [23]. Recent work such as Bayesian [24] and Max-margin [25], mean to find the feature and task correlations. Zhou [26] used temporal group Lasso to capture the intrinsic relatedness among the different tasks and predicted the disease progression. Power consumption is not a simple linear mapping function between the power and components in data centers. Therefore, it is desirable to use non-linear representation and use multi-task learning to find the relationships between the assessment and prediction of power consumption.

In our work, we consider the fine-grained prediction – predicting the next-second power consumption for assessment and the medium-grained prediction – predicting the average power consumption over the next 30 seconds for workflow scheduling and container scheduling.

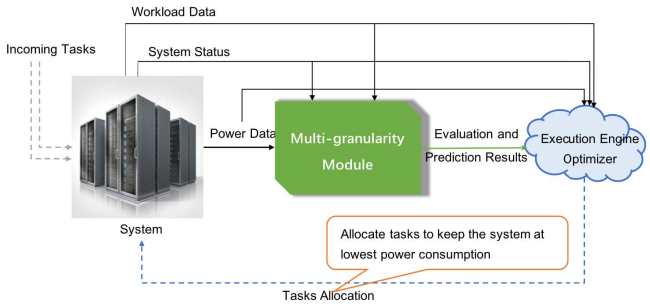

Fig. 1. The overall architecture and workflow of the system under prediction

III. POWER CONSUMPTION SYSTEM PREDICTION AND MANAGEMENT

In this paper, we study the power consumption based on the multi-task learning. We formulate two tasks: a fine-grained problem – power consumption assessment for predicting the next second power consumption, and a medium-grained problem – power consumption prediction for predicting the average power consumption at the period of next 30 seconds. The two tasks correlate with each other. The real-time power assessment can evaluate not only the accuracy of the model but also facilitate the prediction of power consumption. The two tasks correlate with each other. The real-time power assessment can not only evaluate the accuracy of the model we build for the system, but also facilitate the prediction of the power consumption.

In real data centers, a complete framework consists of IT equipment, power supply units, cooling equipment, and other support infrastructures. Cooling equipment, which contains computer room air conditioners (CRACs), cooling towers and chillers, is another large power-consumed part in data centers [27] [28]. In this paper, we consider the power consumption generated by IT equipment. The paper is indicated to the assessment and prediction of power consumption. The multi-granularity module is a vital part of the data center management system, as shown in Fig. 1. The module assesses and predicts the total power consumption of the system, based on the power data, system status, and workload data. Then the predicted values of power consumption are sent to the execution engine optimizer, which monitors the system resources. The execution engine optimizer dynamically deploys a minimum power consumption scheduling scheme based on the continuously predicted results.

In our experiments, we build two power systems and the corresponding datasets to assess and predict power consumption. We run the CPU-intensive workload [29] and the URL-request workloads respectively on the individual system and collect the datasets for training the model. The dataset includes three parts: power consumption data, system status, and workload data. The CPU-intensive power consumption system in Fig. 2(a) consists of a server, a monitor computer, and a PDU. We dynamically adjust the CPU utilization to make sure that the server runs in different states to generate power

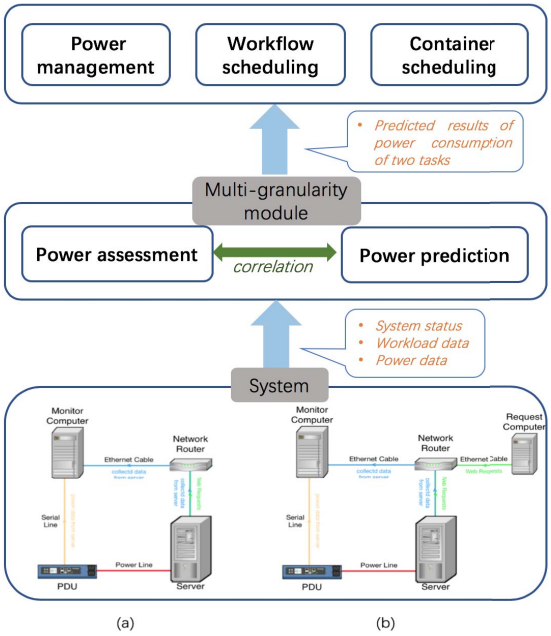

Fig. 2. The structure of the power consumption management systems. We set two different systems: the architecture of CPU-intensive power consumption system (a) and the architecture of URL-request power consumption system (b).

consumption. The URL-request power consumption system in Fig. 2(b) contains one more request computer, which sends HTTP requests to the server. The request is sent at different rates so that the server runs in different states. Monitor computer records the system status of the server and power consumption from PDU per second. We use the two datasets to train the module for power assessment and power prediction. These two tasks correlate and share related features. The execution engine optimizer then selects the best solution from the multiple candidate plans based on the minimal prediction values of power consumption. This multi-granularity module helps cloud service providers perform power management, workflow scheduling, and container scheduling.

IV. MODEL FOR LSTM BASED MULTI-TASK LEARNING

Most work on the power consumption in data centers is a single task about workloads. In this section, we elaborate on our proposed approach for multi-granularity prediction of the power consumption system. The next second power prediction is defined as a fine-grained prediction task, while the prediction of the next 30 seconds average power is used as a medium-grained prediction task. We use LSTM and multi-task learning to predict multi-granularity power consumption. LSTM can capture the long-range dependency of the time series, and multi-task learning is intended to impose shared knowledge when solving multiple correlated tasks simultaneously.

A. Recurrent Neural Network

A recurrent neural network (RNN) is used to deal with a temporal sequence, which is different from the forward neural network in the hidden layers. Since the hidden layers not only receive the input but also receive the output of last hidden layer, the correlations among the data series are generated in the RNN, which significantly improves the ability to analyze data.

However, RNN has an inevitable problem that the components of the gradient vector will grow or decay exponentially when the training sequence is very long [30] [31]. Therefore, the RNN model is not capable of learning well in the long-dependence series for the gradient explosion and gradient vanishing.

Long short-term memory (LSTM) network, the variation of RNN, is proposed to address the difficulty of RNN for learning the long-term dependent sequence [32]. The LSTM has a memory cell in the hidden neutron which consists of three gates: an input gate i_t, a forget gate f_t and an output gate o_t (we define the vectors at the time step t). And the memory cell is c_t, candidate state of the memory cell is $\widetilde{c_t}$ and the hidden value of the memory cell at time t is h_t. The LSTM transition equations are as follows:

$$f_t = \sigma(W_f \cdot [x_t, h_{t-1}] + b_f), \tag{1}$$

$$i_t = \sigma(W_i \cdot [x_t, h_{t-1}] + b_i), \tag{2}$$

$$\widetilde{c_t} = tanh(W_c \cdot [x_t, h_{t-1}] + b_c), \tag{3}$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t}, \tag{4}$$

$$o_t = \sigma(W_o \cdot [x_t, h_{t-1}] + b_o), \tag{5}$$

$$h_t = o_t * tanh(c_t), \tag{6}$$

where x_t is the input at the time step t, W is the weight and b is the bias of each vectors, the entries of three gating vectors is in [0,1], and σ denotes the logistic sigmoid function.

B. Multi-task LSTM model

We propose a multi-task deep neural network model to deal with many related tasks. Fig. 3 shows the structure of our proposed methodology.

We use the shared-layer LSTM model to assess and predict the power consumption for different workloads introduced in Section 3. Each task has its separate LSTM layer, and a bidirectional LSTM layer is set for all the tasks which can capture the shared information.

In Fig. 3, given two tasks (p, q), $h_t^{(p)}$ and $h_t^{(q)}$ represent the input of the two tasks sequence respectively. The hidden shared layer receives the input x_t and the output of the last hidden shared layer $h_{t-1}^{(s)}$. The output of the hidden shared layer $h_t^{(s)}$ at the time step t consists of two parts: the forward output $\overrightarrow{h_t}^{(s)}$ and the backward output $\overleftarrow{h_t}^{(s)}$.

$$h_t^{(s)} = \overrightarrow{h_t}^{(s)} \oplus \overleftarrow{h_t}^{(s)}, \tag{7}$$

where $\oplus$ represents the concatenation operation.

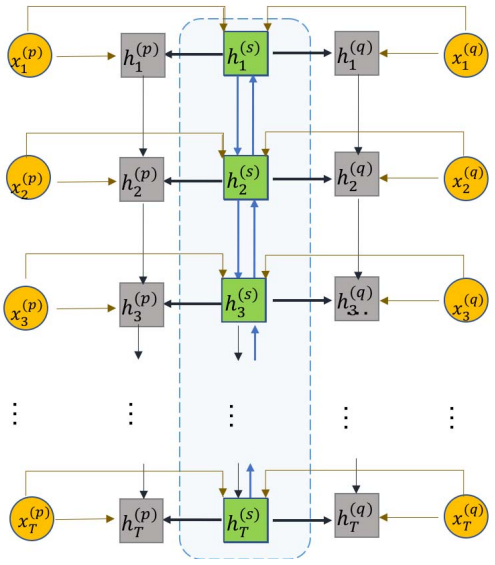

Fig. 3. The structure of the multi-task LSTM model

The shared hidden layer receives x_t and the last output of hidden shared layer $h_{t-1}^{(s)}$ as its input. Moreover, the hidden layer $h_t^{(s)}$ not only sends its output to the next hidden layer but also transforms the output to the task hidden layer h_t. For each task, a gate is introduced between the task separated layer h_t and task shared layer $h_t^{(s)}$ as well as between $h_t^{(s)}$ and $h_{t-1}^{(s)}$ at task hidden layer at the time step t to decide how much information to accept.

Therefore, the Eq. (3) is rewrite in the following, for task p:

$$\widetilde{c_t}^{(p)} = tanh(W_c{}^{(p)} \cdot [x_t, g^{(m)} h_{t-1}^{(s)}] + W_c{}^{(s)} \cdot [g^{(s \to p)} h_t{}^{(s)}] + b_c{}^{(p)}), \tag{8}$$

$g_{(p)}$ is the gate between $h_t^{(s)}$ and $h_{t-1}^{(s)}$ at the hidden layer of task p and $g^{(s \to p)}$ is the gate between the task separated layer $h_t^{(p)}$ and task shared layer $h_t^{(s)}$:

$$g^{(p)} = \sigma(W_g{}^{(p)} \cdot [x_t, h_{t-1}^{(p)}] + b_g^{(p)}), \tag{9}$$

$$g^{(s \to p)} = \sigma(W_g{}^{(s \to p)} \cdot [x_t, h_t^{(s)}] + b_g^{(s \to p)}), \tag{10}$$

The LSTM shared layer can capture the correlated representations for tasks. We use the gate mechanism to decide whether to accept the information from the hidden shared layers, which facilitates the interaction between the task hidden layers and shared layers.

C. Learning

We feed the sequence data to the model and the i-th sample is denoted as ($x_t^{(p)}$, $y_t^{(p)}$) for task p, where the y_t is the real value for the power consumption. And we use $\hat{y}_t^{(p)}$ to represent the output for task p. The parameter set $\Theta = \{W, b\}$ is learned by using the following loss function:

$$\phi = argmin \Sigma || \hat{y}_t^{(p)} - y_t^{(p)} ||_2^2 \tag{11}$$

TABLE I
STATISTICS OF THE DATASETS USED

Dataset	Type	Train size	Test size	Length
D1	CPU-intensive	70%	30%	200000
D2	URL-request	70%	30%	150000

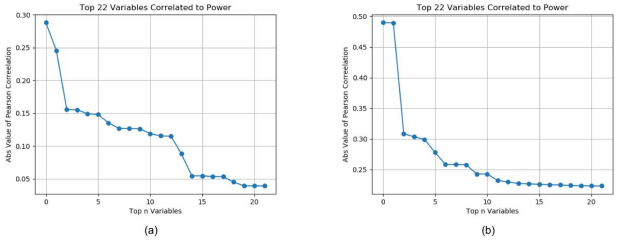

Fig. 4. The correlation between top 22 variables and power consumption on D1(a) and D2(b). The variables are not complete the same because of the different workloads of the two datasets.

V. EXPERIMENT

In this paper, we collect the datasets, including power-time data, system status, and workload data. PDU collects the power consumption of the two systems respectively per second, and we use Collectd, an open-source software, to record system status, such as CPU usage and memory usage of the individual system per second. Note that the information of the two systems is measured independently.

A. Parameters

By establishing two power consumption systems, we get the corresponding datasets. The details of the datasets are listed in TABLE. I.

The final hyper-parameters are as follows:

We use Adam as the optimizer. Learning rate is set to 10^{-3}, and the weight decay rate is 10^{-4}.

We use Pearson correlation to select the most correlated variables as the features for multi-granularity prediction of power consumption. The Pearson correlation measures the linear relationship between the variables and power consumption. The relationship between the two is close when the absolute coefficient of the two variables is close to 1. Excessive variables can cause the model to be over-fitting, and fewer variables can lead to the under-fitting of the model. The variables are arranged in descending order of the Pearson coefficient of the power consumption, and the appropriate first N variables are selected. In our experiment, we choose the top 22 variables to train the model. The remaining variables have little effect on the performance of the model and may even reduce the accuracy of the model. The coefficients of the top 22 most correlated variables are shown in Fig. 4. For the different workloads, the most relevant top N coefficients are not exactly the same. We use these features to train the model, which has a more significant impact on power consumption than other variables and improves the prediction accuracy of the model.

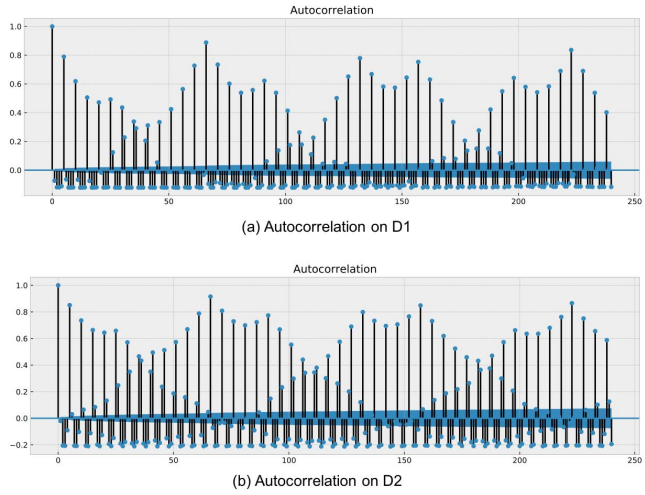

Fig. 5. ACF (Autocorrelation function) diagrams

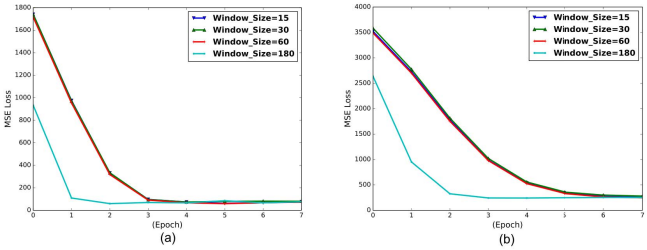

Fig. 6. Convergence speeds of observation window sizes on D1(a) and D2(b)

Then we select the optimized observation window size, investigate the empirical performance of our method on the multi-granularity prediction of power consumption, and compare our model to other baselines.

B. Obeservation Window Size

We performed the multi-granularity prediction tasks on each of our two datasets to compare the performance of our methods and other methods. We consider both predictions for multi-granularity: fine-grained prediction, that is, power consumption for the next second, and medium-grained prediction, that is, average power consumption for the next 30 seconds. The average power prediction at the next 30 seconds corresponds to a medium-grained prediction for workflow scheduling to meet cost savings and the requirement of QoS. In the following figures, task 1 represents the fine-grained prediction, and task 2 represents the medium-grained prediction.

Because the samples of the two multi-granularity prediction tasks are from the same datasets, there is no temporal misalignment between the data. Moreover, the observation windows of the two tasks need to be consistent. First, we compare the size of the observation window for each dataset. The appropriate window size plays an important role in the prediction accuracy of the model. The observation window is the length of each sample used to predict the power consumption by

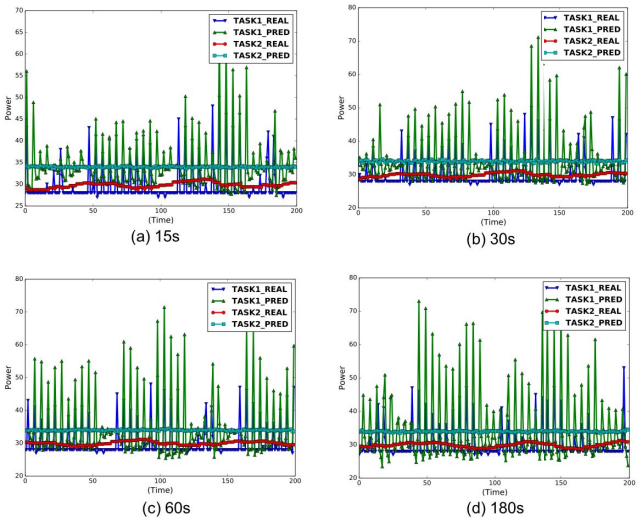

Fig. 7. Predicted results of different window sizes on D1. Task 1 is the fine-grained prediction and task 2 is the medium-grained prediction.

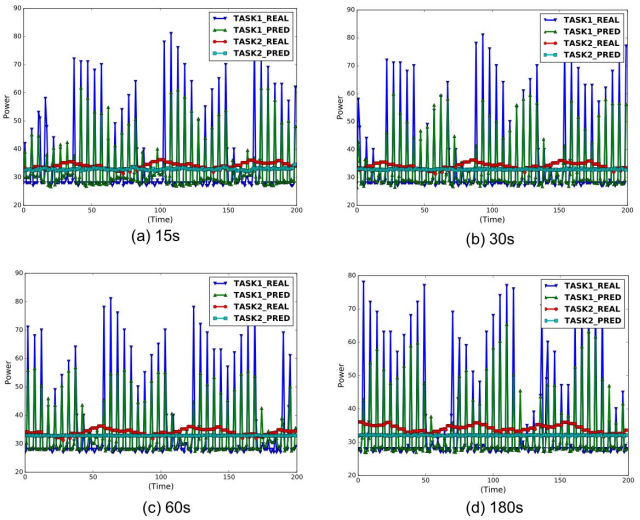

Fig. 8. Predicted results of different window sizes on D2

segmenting the dataset. Figure 5 shows that there are strong autocorrelation and periodicity in the power consumption data. We respectively observe the performance of four candidate window sizes (15s, 30s, 60s, and 180s) on the two datasets. The convergence speeds of the loss functions are shown in Fig. 6. The loss functions for all settings are convergent. When the observation window size is equal to 180s, the loss function converges the fastest because the window of 180s covers the data periods. Fig. 7 and Fig. 8 are the predicted results of different window sizes of our proposed model on the datasets. The predicted errors of the four candidate window sizes are shown in TABLE II. For the fine-grained prediction of power consumption on D1, the best observation window size is equal to 15s while the optimized window size is 60s for the medium-granularity. To prevent the window from being too small to cause under-fitting, we choose the 60s as the size of the observation window on D1. And when the length of window size is equal to 60s, the predicted results of 60s have the lower RMSE errors than the other three sizes of the multi-granularity on D2. This reveals that when the window size is set to 180s, the convergence speed is the fastest, and because the window length is too long, the overall comparison is worse than other candidates. Therefore, we use the past 60 seconds of power values as the optimized observation window size to learn the multi-task at time t on D2. Moreover, we predict the output continuously and set the step of the sliding window is 1s.

TABLE II
RESULT OF DIFFERENT OBSERVATION WINDOW SIZES ON D1 AND D2

	Error	15s	30s	60s	180s
D1	Fine-grained	0.3030	0.3980	0.4058	0.5745
	Medium-grained	0.0453	0.0467	0.0425	0.0470
D2	Fine-grained	0.5133	0.5181	0.4722	0.4752
	Medium-grained	0.0793	0.0772	0.0760	0.0776

TABLE III
PRECTION ACCUARCY COMPARISON OF METHODS

	Error	MGM	LSTM	GBR	ARIMA
D1	Fine-grained	0.4058	0.5010	2.544	3.072
	Medium-grained	0.0425	0.0504	0.4025	0.997
D2	Fine-grained	0.4722	0.5089	4.542	8.957
	Medium-grained	0.0760	0.0783	0.4970	0.9612

C. Comparison

To evaluate the accuracy and the computational complexity, we compare our multi-granularity model (MGM) with the three representative models:

- Standard LSTM network
- Gradient boosting regression (GBR)
- Autoregressive integrated moving average (ARIMA) [33]

The experimental results of the other three prediction methods are shown in Fig. 9. And TABLE III shows the prediction errors for all of the prediction models on the two datasets.

In this paper, the prediction error is defined as the root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{M}\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2} \tag{12}$$

where M is the number of prediction values, Y_t represents the actual power values that the PDU measures at time t and $\hat{Y}_t$ is the prediction values at time t.

Our approach and LSTM have lower errors than the other three models since RNN can deal with the long-range dependency of time series and fit the temporal prediction well. GBR performs better than ARIMA because ensemble learning has a desirable accuracy for prediction. ARIMA performs worst since it only uses power consumption to predict itself.

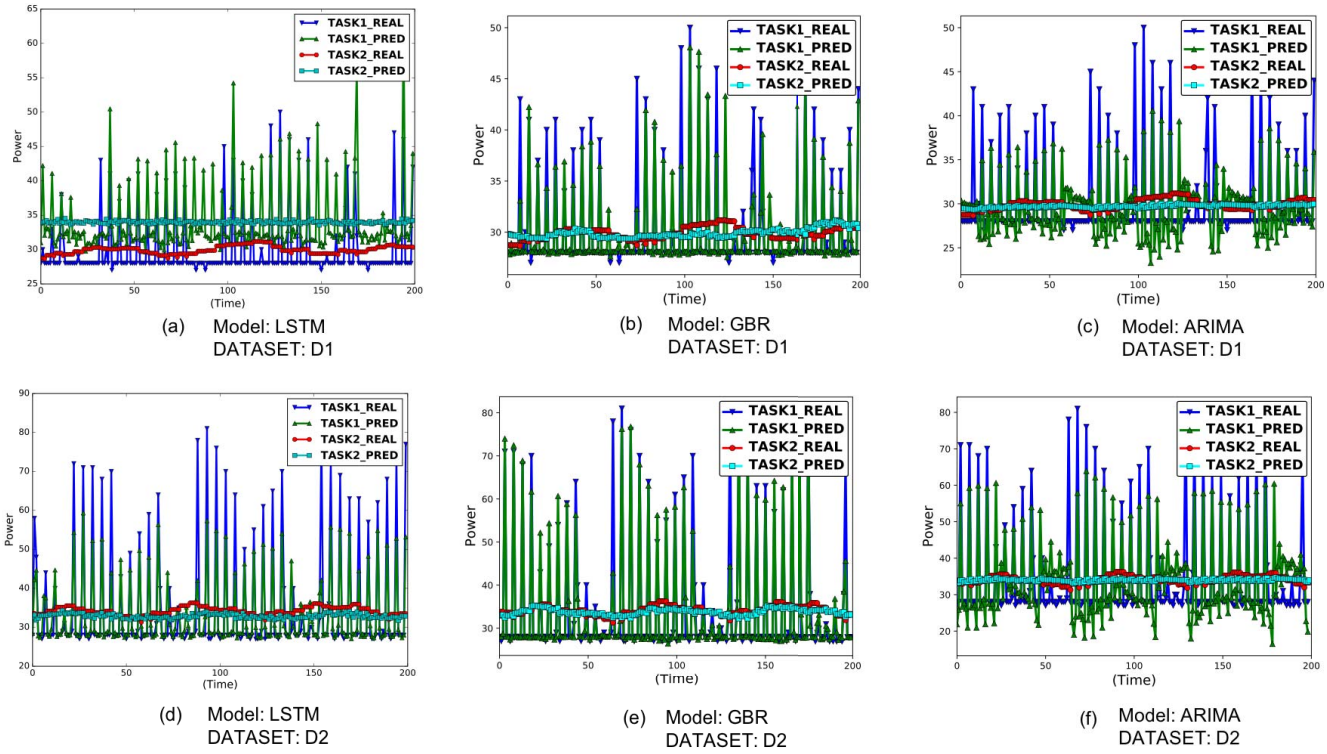

Fig. 9. The predicted results of LSTM, GBR, and ARIMA

Compared to the LSTM, our method improves the accuracy of the tasks, which indicates our method can extract more abstract representations between the correlated tasks. Besides, the medium-grained predictions on the two datasets have significantly lower errors because the accurate power assessment can improve the accuracy of medium-power predictions. Power consumption assessment has a larger error than power consumption prediction for all the models since the prediction for future 30s power fluctuates more softly than the assessment for the next second power consumption.

VI. Conclusion and Discussion

In the data center, efficient power consumption prediction has a great significance for workflow scheduling and power management. In this article, we have developed a multi-granularity prediction for power consumption issues in the data center. Fine-grained prediction is defined as the next second power consumption prediction for data center assessment. The medium-grained prediction predicts the average energy consumption at the next 30s for workflow scheduling and container scheduling. Moreover, because these two tasks have long-term dependence on time and correlation between tasks, we propose an LSTM network based on multi-task learning for this multi-granularity problem. Compared to other predicted models, the multi-task model can capture more abstract features in tasks, thereby improving the performance of correlated tasks. In the future, we will consider more coarse-grained prediction or trend prediction [34] of power consumption into the multi-granularity model, which is different a lot from one second or 30 seconds of power consumption.

Apart from the power consumption caused by IT equipment, there are also some effort on heat prediction and saving in data centers [35] [28]. In this work, we only consider power consumption generated by the IT equipment. We will also pay attention to cooling optimization for power saving in data centers and find more complete features included cooling equipment in the future to improve the multi-granularity model.

VII. Acknowledgement

This work was supported by National Key Research and Development Program of China (2018YFB1003702).

References

[1] J. Hamilton, “Cooperative expendable micro-slice servers (cems): low cost, low power servers for internet-scale services,” in *Conference on Innovative Data Systems Research (CIDR09)(January 2009)*. Citeseer, 2009.

[2] Y. C. Lee and A. Y. Zomaya, “Energy efficient utilization of resources in cloud computing systems,” *The Journal of Supercomputing*, vol. 60, no. 2, pp. 268–280, 2012.

[3] G. V. Laszewski, L. Wang, A. J. Younge, and H. Xi, “Power-aware scheduling of virtual machines in dvfs-enabled clusters,” in *IEEE International Conference on Cluster Computing & Workshops*, 2009.

[4] S. Hacking and B. Hudzia, *Improving the live migration process of large enterprise applications*, 2009.

[5] R. Friedrich and C. Patel, "Data center energy management system," Oct. 16 2003, uS Patent App. 10/122,210.

[6] J. Gao, "Machine learning applications for data center optimization," 2014.

[7] N. Lazic, C. Boutilier, T. Lu, E. Wong, B. Roy, M. Ryu, and G. Imwalle, "Data center cooling using model-predictive control," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 3814–3823. [Online]. Available: http://papers.nips.cc/paper/7638-data-center-cooling-using-model-predictive-control.pdf

[8] R. Caruana, *Multitask Learning*, 1998.

[9] L. Benini, A. Bogliolo, and G. De Micheli, "A survey of design techniques for system-level dynamic power management," *IEEE transactions on very large scale integration (VLSI) systems*, vol. 8, no. 3, pp. 299–316, 2000.

[10] D. Meisner, B. T. Gold, and T. F. Wenisch, "Powernap: eliminating server idle power," in *ACM sigplan notices*, vol. 44, no. 3. ACM, 2009, pp. 205–216.

[11] J. C. McCullough, Y. Agarwal, J. Chandrashekar, S. Kuppuswamy, A. C. Snoeren, and R. K. Gupta, "Evaluating the effectiveness of model-based power characterization," in *USENIX Annual Technical Conf*, vol. 20, 2011.

[12] F. Bellosa, "The benefits of event: driven energy accounting in power-sensitive systems," in *Proceedings of the 9th workshop on ACM SIGOPS European workshop: beyond the PC: new challenges for the operating system*. ACM, 2000, pp. 37–42.

[13] C. Isci and M. Martonosi, "Runtime power monitoring in high-end processors: Methodology and empirical data," in *Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Computer Society, 2003, p. 93.

[14] R. Joseph and M. Martonosi, "Run-time power estimation in high performance microprocessors," in *ISLPED'01: Proceedings of the 2001 International Symposium on Low Power Electronics and Design (IEEE Cat. No. 01TH8581)*. IEEE, 2001, pp. 135–140.

[15] D. Economou, S. Rivoire, C. Kozyrakis, and P. Ranganathan, "Full-system power analysis and modeling for server environments." International Symposium on Computer Architecture-IEEE, 2006.

[16] M. Y. Anwar, J. A. Lewnard, S. Parikh, and V. E. Pitzer, "Time series analysis of malaria in afghanistan: using arima models to predict future trends in incidence," *Malaria journal*, vol. 15, no. 1, p. 566, 2016.

[17] S. Mazumdar and A. S. Kumar, "Forecasting data center resource usage: An experimental comparison with time-series methods," in *International Conference on Soft Computing and Pattern Recognition*. Springer, 2016, pp. 151–165.

[18] M. Rossi and D. Brunelli, "Forecasting data centers power consumption with the holt-winters method," in *2015 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS) Proceedings*. IEEE, 2015, pp. 210–214.

[19] Y. Li, H. Hu, Y. Wen, and J. Zhang, "Learning-based power prediction for data centre operations via deep neural networks," in *Proceedings of the 5th International Workshop on Energy Efficient Data Centres*. ACM, 2016, p. 6.

[20] N. Liu, Z. Li, Z. Xu, J. Xu, S. Lin, Q. Qiu, J. Tang, and Y. Wang, "A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning," 2017.

[21] N. Kalchbrenner, I. Danihelka, and A. Graves, "Grid long short-term memory," *Computer Science*, 2015.

[22] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," 2012.

[23] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[24] M. Yang, Y. Li, and Z. Zhang, "Multi-task learning with gaussian matrix generalized inverse gaussian model," in *International Conference on International Conference on Machine Learning*, 2013.

[25] Z. Yi, "Learning multiple tasks with a sparse matrix-normal penalty," *Advances in Neural Information Processing Systems*, pp. 2550–2558, 2010.

[26] J. Zhou, L. Yuan, J. Liu, and J. Ye, "[acm press the 17th acm sigkdd international conference - san diego, california, usa (2011.08.21-2011.08.24)] proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining - kdd Ï1 - a multi-task learning formulation," in *Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*, 2011.

[27] L. Parolini, B. Sinopoli, B. H. Krogh, and Z. Wang, "A cyber-physical systems approach to data center modeling and control for energy efficiency," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 254–268, 2011.

[28] J. Chen, T. Rui, W. Yu, G. Xing, X. Wang, X. Wang, B. Punch, and D. Colbry, "A high-fidelity temperature distribution forecasting system for data centers," 2012.

[29] B. L. B. X. Z. L. Y. L. Ziyu Shen, Xusheng Zhang, "Pcp-2lstm: Two stacked lstm-based prediction model for power consumption in data centers," 2019.

[30] F. F. Informatik, Y. Bengio, P. Frasconi, and J. Schmidhuber, *Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies*, 2001.

[31] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," 2012.

[32] A. Graves, *Long Short-Term Memory*, 2012.

[33] R. V. Anand, P. B. Sivakumar, and D. V. Sagar, *Forecasting the Stability of the Data Centre Based on Real-Time Data of Batch Workload Using Times Series Models*, 2016.

[34] B. Xia, T. Li, Q.-F. Zhou, Q. Li, and H. Zhang, "An effective classification-based framework for predicting cloud capacity demand in cloud services," *IEEE Transactions on Services Computing*, 2018.

[35] M. Zapater, J. L. Risco-Martn, P. Arroba, J. L. Ayala, J. M. Moya, and R. Hermida, "Runtime data center temperature prediction using grammatical evolution techniques," *Applied Soft Computing*, vol. 49, pp. 94–107, 2016.