Last e-mail outline:*(Note: There is no meeting in last week, this outline is built based on replied e-mail)*
Fulfill the content of next 5 questions:

- What is the problem you are trying to address with the experiment? ✓
- What is the work that has been done in relation to this problem? ✓
- why you focus on this paper? what is special about it?✓
- what is the purpose of the experiment you would like to carry out?✓
- What are you expecting to get which may be different from other people?✓

1 What is the problem that I trying to address

In the environment of 'Industry 4.0' and 'Internet-Of-Things', a large amount of real-time data being generated everyday improves the limit requirement of data processing and management. Cloud services are pay based on need which is a good option for companies to choose. Cloud computing has been regarded as the fifth public resource after water, electricity, gas, and oil[1]. Cloud computing provides Infrastructure as a Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS) which provides a perfect option for companies to choose. When users flood into a data center, how to manage the load balance between physical machines becomes an inevitable problem for the service provider. Virtual machine migration is a cornerstone technology to realize the load balance in the data center[2]. It can release the virtual machines on the underload physical machine and migrate virtual machines on the overload physical machine while providing continuous services for users. However, virtual machine migration costs power as well, and there exist virtual machine interference during migration. Thus building a migration method that can keep load balance with fewer times of migration and lower power cost becomes a hot-spot problem in the research field of the virtual machine migration[3][4][5].
The virtual machine interference problem is caused by unpredictable drastic changes in the resource requested by a virtual machine. This kind of unpredictable drastic changes will cause underload and overload problems on physical machines even after migration. After that, these virtual machines running on underload/overload physical machines will have to immigrate again to another suitable physical machine. Then the total number of the virtual machine migration will be increased, what's more, the energy consumption and bandwidth usage will be affected as well in this circumstance. Hence, building a model that can predict the future usage of CPU, memory, and disk request by a virtual machine is one critical step to handle the problem which is known as virtual machine interference problem[6].

2 What is the work has been done in relation to this problem

Predicting the usage of a virtual machine is one of the core steps of avoiding virtual machine interference. In [7], the authors purpose a deep learning model based on the canonical polyadic decomposition to predict CPU load. In [8], authors use auto-regression filter and neural network to forecast load in a data center. In [9], due to the feature of recurrent neural networks that can record previous information and predict time series problems accurately. The authors apply it to predict CPU utilization. A machine learning-based approach[10] is presented to estimate the CPU burst time. The author chooses 18 attributes of the task as the input vector of the deep learning model.
Apart from machine learning-based approaches, there are some other approaches. Due to time consumption of training and learning hyper-parameters, the authors in [11] apply gaussian progress regression to predict the number of working CPU cores. One cluster-based predicting CPU load approach is proposed in [12]. A dual simplex method is applied to predict the length of the next CPU burst time in [13]. Authors in [14] assume that the frequency of requests in unit-time period apply Poisson distribution, based on which purpose a method to predict probability parameter to estimate the frequency of request in cloud system.
In [15], the authors employ the ARIMA(Autoregressive integrated moving average) model to forecast the CPU load. According to the forecasted CPU load, they select the underload physical machine then match

each of the virtual machines deployed on that physical machine with other physical machines. Based on ‘match-value’, the virtual machine is migrated. In [16], the authors apply a linear regression model to predict CPU and memory usage, then detect the overload period according to the prediction. They built a benchmark to evaluate the contribution of one virtual machine concerning a host physical machine. Based on that to identify the virtual machine for migration and the physical machine as a destination. A scheduling algorithm is proposed in [17] to assign the most suitable physical machine to a virtual machine based on historical memory consumption. The prediction approach in these papers is suffering from shortage since the time series model considers too much previous unrelated information, but the future usage requirement is much more affected by processing tasks ignored in the time series model. As a result, their prediction exists undetected overload/underload period and low accuracy in long-term prediction. Authors in [6] improve the prediction approach, and combine four regression methods such as Weighted moving average (WMA), Exponential smoothing average (ESA), Holt winter’s method (HWM) and Autoregressive (AR) model, then set the result of these methods as input vector of neural network to predict the resource usage. Given the forecast resource usage, they define low/high/normal CPU usage, low/high/normal memory usage, low/high/normal disk usage. Virtual machines on overload and underload physical machines are then migrated to normal-load/underload and normal-load physical machines, respectively. Even though this approach combined four typical time series models, it still cannot avoid the effect of past unrelate information on future resource usage. What’s more, the migration strategy is oversimplified and further improvement may lead to a better solution for the virtual machine interference problem.

3 Why I focus on [6]? what is special about it?

- Although the authors in [6] combine four different time series models to predict resource usage, it still suffers from the same problems as the time series model.
- There exists a huge space that can be improved in the migration phase.
- This paper is published by Elsevier which is a good journal, so it is good to compare with.

4 what is the purpose of the experiment you would like to carry out?

My assumption:

- The same resource rate is requested in a same time period. For example, the requested rates on Mondays are similar.
- More features can be applied to characterized requested resources in each time period, such as expectation and variance of resource utilization, the unpredictability of request, etc.

Based on the assumption, we propose a machine learning method to predict future resource requests.

5 what are you expecting to get which may be different from other people?

- The accuracy of the proposed method is compared to the previous method.
- The number of virtual machine migrations is consumed.
- Load balance between physical machines in the proposed method and the previous method are compared.

References

[1] W. D. Tian and Y. D. Zhao, *Optimized cloud resource management and scheduling: theories and practices*. Morgan Kaufmann, 2014.

[2] F. Zhang, G. Liu, X. Fu, and R. Yahyapour, "A survey on virtual machine migration: Challenges, techniques, and open issues," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 1206–1243, 2018.

[3] J. Liang, J. Cao, J. Wang, and Y. Xu, "Long-term cpu load prediction," in *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, IEEE, 2011, pp. 23–26.

[4] F. Zhang, G. Liu, X. Fu, and R. Yahyapour, "A survey on virtual machine migration: Challenges, techniques, and open issues," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 1206–1243, 2018.

[5] M. Noshy, A. Ibrahim, and H. A. Ali, "Optimization of live virtual machine migration in cloud computing: A survey and future directions," *Journal of Network and Computer Applications*, vol. 110, pp. 1–10, 2018.

[6] G. J. L. Paulraj, S. A. J. Francis, J. D. Peter, and I. J. Jebadurai, "A combined forecast-based virtual machine migration in cloud data centers," *Computers & Electrical Engineering*, vol. 69, pp. 287–300, 2018.

[7] Q. Zhang, L. T. Yang, Z. Yan, Z. Chen, and P. Li, "An efficient deep learning model to predict cloud workload for industry informatics," *IEEE transactions on industrial informatics*, vol. 14, no. 7, pp. 3170–3178, 2018.

[8] J. J. Prevost, K. Nagothu, B. Kelley, and M. Jamshidi, "Prediction of cloud data center networks loads using stochastic and neural models," in *2011 6th International Conference on System of Systems Engineering*, IEEE, 2011, pp. 276–281.

[9] K. Mason, M. Duggan, E. Barrett, J. Duggan, and E. Howley, "Predicting host cpu utilization in the cloud using evolutionary neural networks," *Future Generation Computer Systems*, vol. 86, pp. 162–173, 2018.

[10] T. Helmy, S. Al-Azani, and O. Bin-Obaidellah, "A machine learning-based approach to estimate the cpu-burst time for processes in the computational grids," in *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, IEEE, 2015, pp. 3–8.

[11] D.-M. Bui, H.-Q. Nguyen, Y. Yoon, S. Jun, M. B. Amin, and S. Lee, "Gaussian process for predicting cpu utilization and its application to energy efficiency," *Applied Intelligence*, vol. 43, no. 4, pp. 874–891, 2015.

[12] C. Viswanath and C Valliyammai, "Cpu load prediction using anfis for grid computing," in *IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-2012)*, IEEE, 2012, pp. 343–348.

[13] M. M. Kumar, B. R. Rajendra, C. Niranjan, and M Sreenatha, "Prediction of length of the next cpu burst in sjf scheduling algorithm using dual simplex method," in *Second International Conference on Current Trends In Engineering and Technology-ICCTET 2014*, IEEE, 2014, pp. 248–252.

[14] M. S. Yoon, A. E. Kamal, and Z. Zhu, "Requests prediction in cloud with a cyclic window learning algorithm," in *2016 IEEE Globecom Workshops (GC Wkshps)*, IEEE, 2016, pp. 1–6.

[15] W. Fang, Z. Lu, J. Wu, and Z. Cao, "Rpps: A novel resource prediction and provisioning scheme in cloud data center," in *2012 IEEE Ninth International Conference on Services Computing*, IEEE, 2012, pp. 609–616.

[16] T. H. Nguyen, M. Di Francesco, and A. Yla-Jaaski, "Virtual machine consolidation with multiple usage prediction for energy-efficient cloud data centers," *IEEE Transactions on Services Computing*, 2017.

[17] Z. Tang, Y. Mo, K. Li, and K. Li, "Dynamic forecast scheduling algorithm for virtual machine placement in cloud computing environment," *The Journal of Supercomputing*, vol. 70, no. 3, pp. 1279–1296, 2014.