2016

PROST: Predicting Resource Usages with Spatial and Temporal Dependencies

Ji Xue
William & Mary

Evgenia Smirni
William & Mary

Thomas Scherer

Robert Birke

Lydia Y. Chen

Recommended Citation

Xue, Ji; Smirni, Evgenia; Scherer, Thomas; Birke, Robert; and Chen, Lydia Y., PROST: Predicting Resource Usages with Spatial and Temporal Dependencies (2016). *Proceedings of the 2016 ACM/Spec International Conference on Performance Engineering (Icpe'16)*.
10.1145/2851553.2858678

PROST: Predicting Resource Usages with Spatial and Temporal Dependencies

Ji Xue
College of
William and Mary
xuejimic@cs.wm.edu

Evgenia Smirni
College of
William and Mary
esmirni@cs.wm.edu

Thomas Scherer
IBM Research
Zurich Lab
tsc@zurich.ibm.com

Robert Birke
IBM Research
Zurich Lab
bir@zurich.ibm.com

Lydia Y. Chen
IBM Research
Zurich Lab
yic@zurich.ibm.com

ABSTRACT

We present a tool, PROST, which can achieve scalable and accurate prediction of server workload time series in data centers. As several virtual machines are typically co-located on physical servers, the CPU and RAM show strong temporal and spatial dependencies. PROST is able to leverage the spatial dependency among co-located VMs to improve the scalability of prediction models solely based on temporal features, such as neural network. We show the benefits of PROST in obtaining accurate prediction of resource usage series and designing effective VM sizing strategies for the private data centers.

1. INTRODUCTION

Tools for workload characterization and prediction are key to effective resource allocation in data centers. Based on accurate predictions of upcoming workload within the next timeframe (which can be in the order of minutes, hours, or even weeks, depending on the application), proactive decisions can be made to improve the system's performance. In a cloud data center environment, for example, this information can be used to migrate and consolidate VMs to reduce the number of required physical servers and thus improve the energy efficiency, while at the same time fulfilling the service level agreement with respect to the relevant performance metrics. Depending on the capability of predicting peak load magnitudes and timings, resources can be multiplexed at various degrees across users and time.

Past work has established that resource usage at data centers exhibits strong temporal patterns [2]. Beyond temporal dependencies that are established by usage time series [4], it is common for co-located VMs to simultaneously compete for the limited physical resources, essentially exhibiting strong *spatial* dependency. Indeed, in our past work we have shown that neural networks can be effectively employed for prediction [4], but their effective usage remains prohibitively expensive in practical situations as it suffers by its high training cost. In practice, in a large-scaled data center, with more than tens of thousands of physical boxes and hundreds of thousands of VMs, it is infeasible to rely on neural networks to predict future resource usage.

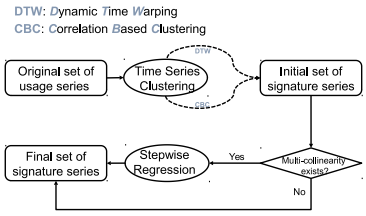


Figure 1: Overview of the PROST in obtaining signature series via the spatial dependency.

We solve this problem by developing a prediction framework, PROST, that discovers spatial dependencies across usage series and exploits them to develop a scalable methodology for predicting a large number of usage series. To this end, we introduce the concept of *signature VM series*, a subset of usage series that are representative of all other usage series. We are able to predict usage series not in the signatures set via a linear combination of signature VM series, which are predicted by the neural networks using their temporal dependency.

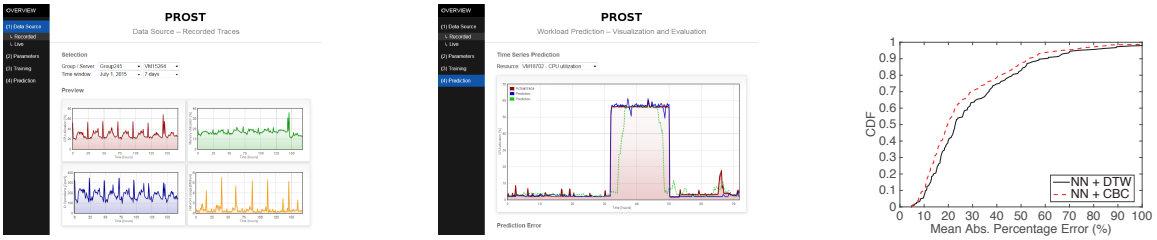
2. TOOL DESCRIPTION

The immediate obstacles of prediction given a large number of usage series of co-located VMs are accuracy, training overhead, and model scalability. Typically, temporal models [3], such as auto regressive and moving average models, are not able to capture well bursty behaviors. More sophisticated temporal models such as neural networks, capture irregular patterns better but at much higher computational overheads. Given such restrictions, it is important to come

ICPE'16 March 12-18, 2016, Delft, Netherlands

ACM ISBN 978-1-4503-4080-9/16/03.

DOI: http://dx.doi.org/10.1145/2851553.2858678



(a) Data source view of the GUI (b) Prediction w/ different spatial methods (c) Case study of PROST

Figure 2: Data source view of the GUI to select the servers of interest, showing a preview of the selected VM's resource utilization traces and prediction of using different spatial methods, and a case study of PROST on predicting usage series at private datacenters.

up with efficient and accurate prediction models that also scale well.

2.1 PROST Framework

Motivated by the strong spatial patterns across VCPU and VRAM, we argue that a small number of signature series as predictors can well represent the entire set of resource usages. To such an end, we propose a prediction methodology, which combines a novel correlation-based time series clustering technique, and stepwise regression. The signature series are predicted via existing time-series models exploring the auto-correlation, which unfortunately incurs very high computational overhead and storage requirement on historical data. Consequently, the linear regression model leveraging the spatial dependency can drastically reduce the computation overhead without sacrificing the accuracy.

Fig. 1 provides an overview of the PROST framework, using a two-step algorithm to identify so-called signature series (1) initial set: initial time series clustering using existing dynamic time warping (DTW) [1] or proposed correlation based clustering (CBC); (2) final set: detecting and removing multicollinearity among signature series using variance inflation factors (VIF) and stepwise regression. The second step is to fix the pitfall that though signature series appear independent combinations of certain series can well present the others in the signature set.

2.2 Implementation and Graphical User Interface

We developed a web based graphical user interface (GUI) to demonstrate the PROST framework. The PROST GUI consists of four main views to select the servers, define the model parameters, train the models, and visualize the predictions obtained via the trained models. Screenshots of the data source and prediction views are shown in Fig. 2(a)-(b).

3. EVALUATION

To test PROST, we use traces containing VCPU and VRAM utilization taken at each 15 minutes. These traces are from production data centers covering for one week 6K physical servers hosting over 80K VMs which serve various industries and use disparate operating systems such as Windows and UNIX. However, due to the computation intensity, especially of the temporal predictions, we evaluate PROST on a subset of 400 randomly selected physical servers and their hosted VMs only. We take the first five days to train the temporal models for the signature series and to train the spatial models for the non-signature series. Then, we use the models to predict the sixth day. First, we predict in one shot the signature series of the sixth day via the temporal models. Afterwards, with the predictions of signature series as input, we predict the sixth day of the non-signature series via the spatial models.

For the temporal models we consider neural networks [4], whereas for the spatial models we consider both the DTW and the CBC clustering techniques. The signature series are predicted by the neural networks, whereas the non-signature series are predicted by the linear models of signature series via DTW and CBC. We evaluate the prediction accuracy in terms of the distribution of the Absolute Percentage Error (APE). Fig. 2(c) presents the CDF of the prediction accuracy with both the DTW- and CBC-based spatial models. One can see that CBC is more precise with lower APE. Indeed, the average APEs of resources usage per physical server for using DTW and CBC to explore the spatial dependency are 31% and 23%, respectively.

Acknowledgment

This work is supported by EU commission FP7 project GENiC (grant agreement no. 608826) and the NSF under grant CCF-1218758.

4. REFERENCES

[1] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD*, 1994.

[2] R. Birke et al. State-of-the-practice in data center virtualization: Toward a better understanding of VM usage. In *IEEE/IFIP DSN*, 2013.

[3] C. Chatfield. *The analysis of time series: an introduction*. CRC press, 2013.

[4] J. Xue et al. Practise: Robust prediction of data center time series. In *IEEE CNSM*, 2015.