Storm Prediction in a Cloud

Ian Davis, Hadi Hemmati, Richard C. Holt, Michael W. Godfrey
David R. Cheriton School of Computer Science
University of Waterloo
{ijdavis, hhemmati, holt, migod}@uwaterloo.ca

Douglas Neuse, Serge Mankovskii
CA Labs, CA Technologies
{Douglas.Neuse, Serge.Mankovskii}@ca.com

***Abstract* — Predicting future behavior reliably and efficiently is key for systems that manage virtual services; such systems must be able to balance loads within a cloud environment to ensure that service level agreements (SLAs) are met at a reasonable expense. In principle accurate predictions can be achieved by mining a variety of data sources, which describe the historic behavior of the services, the requirements of the programs running on them, and the evolving demands placed on the cloud by end users. Of particular importance is accurate prediction of maximal loads likely to be observed in the short term. However, standard approaches to modeling system behavior, by analyzing the totality of the observed data, tend to predict average rather than exceptional system behavior and ignore important patterns of change over time. In this paper, we study the ability of a simple multivariate linear regression for forecasting of peak CPU utilization (storms) in an industrial cloud environment. We also propose several modifications to the standard linear regression to adjust it for storm prediction.**

***Index Terms*— Regression, time-series, prediction, cloud environments**

I. Introduction

Infrastructure as a Service (IaaS) is becoming a norm in large scale IT systems and virtualization in these environments is common. One of the main difficulties of such virtualization is the placing of virtual machines (VMs) and balancing the load. If the demands placed on the infrastructure exceed its capabilities, thrashing will occur, response times will rise, and customer satisfaction will plummet. Therefore it is essential [1] to ensure that the placing and balancing is done properly [2-4].

Proper balancing and capacity planning in such cloud environments requires forecasting of future workload and resource consumptions. Without good forecasts, cloud managers are forced to over-configure their pools of resources to achieve required availability, in order to honor service level agreements (SLAs). This is expensive, and can still fail to consistently satisfy SLAs. Absent good forecasts, cloud managers tend to operate in a reactive mode and can become ineffective and even disruptive.

Several workload forecast techniques based on time series analysis have been introduced over the years [5] that can be applied in the cloud settings as well. The bottom-line of such literature is that there is no “silver bullet” technique for forecasting. Depending on the nature of the data and characteristics of the services and the workload, different statistical techniques and machine learning algorithms may perform better than the others. In some cases even the simplest techniques such as linear regression may perform better than the more complex competitors [6].

To understand the practicality of such prediction techniques on industrial size problems, we set up a series of case studies where we apply different forecasting techniques on data coming from our industrial collaborator, CA Technologies [7]. CA Technologies is a cloud provider for several large scale organizations. They provide IaaS to their clients and monitor their usage. Their cloud manager system basically is responsible for balancing the workload by placing the virtual machines on the physical infrastructure.

In this paper, we report our experience on applying a basic multivariate linear regression (MVLR) technique to predict the CPU utilization of virtual machines, in the context of one of the CA clients. However, unlike many existing prediction techniques, where they minimize the average prediction errors or maximize average likelihoods, we are more interested in predicting extreme cases rather than averages. The motivation comes from the type of workload we are facing in our case study, which is not very uncommon for other cloud-based applications, as well. In our case, the average utilization across all VMs was at most 20%, but the maximum utilization was almost invariably very close to 100%. Applying MVLR in such data (most of the time very low utilization but occasionally reaching to peaks), we realized that though the average predictions are very accurate but the forecast for large values (storms) are drastically poor.

To cope with this problem, we introduce several modifications to the basic MVLR to adjust it for predicting peak values. The results show that subtracting seasonalities extracted by Fourier transform and then using a weighted MVLR provides our best observed results for storm prediction. In the following sections, we describe the details of each modified MVLR and report its results.

II. Subject of Study

We were provided with a substantive body of performance data relating to a single large cloud computing environment running a large number of virtual services over a six month period. In total, there were 2,133 independent entities whose performance was being captured every six minutes. These included 1,572 virtual machines and 495 physical machines. The physical machines provided support for 56 VMware hosts. On average, 53% of the monitored services were active at any time, with a maximum of 85%. The captured data ideally would describe CPU workloads, memory usage, disk I/O and network traffic. However, in most cases only CPU workloads were available. Therefore, we only focused on the CPU workload data. This data was consolidated into average and maximum hourly performance figures.

In terms of the nature of the services, at least 423 services were dedicated to providing virtual desktop environments, while the cloud was also proving support for web-based services, transaction processing, database support, placement of virtual services on hosts, and other services such as performance monitoring and backup.

As is typically the case in desktop environments, individual computer utilization varies dramatically. Much of the time little if any intensive work is being done on a virtual desktop and the virtual service appears almost idle. However, for any virtual desktop there are periods of intense activity, when CPU, memory, disk I/O, and/or network traffic peaks. Similarly, within transaction processing environments, there will be a wide variety of behaviors, depending on the overall demand placed on such systems.

As mentioned, the frequency distribution of the utilizations is highly skewed, with the vast majority of utilizations (83.5%) not exceeding 25%. Therefore, we mostly require a prediction technique that (with a reasonable degree of confidence) indicates when future loads will be high, even if such predictions do not mathematically fit the totality of observed and future data as closely as other statistical approaches.

III. Storm Prediction using Linear Regression

In this section, we apply a basic MVLR and three variations of it to our industrial dataset and report their accuracy in terms of average absolute errors, when predicting peak values.

MVLR: To apply an MVLR on CPU utilization data, we first obtain correlograms from the provided data, by computing the auto-correlation of each time series with each lagged version of the same time series. This indicates the strongest auto-correlation as the hourly (1,676 sources), weekly (247), daily (106) and bi-weekly (41) levels, with these correlations degrades only slowly over longer intervals.

Using the discovered significant lags, multivariate linear regression [5] was then applied using 10 lags of 1 and 2 hours, 1 and 2 days, 1, 2, 3 and 4 weeks, and 1 and 2 months, to identify coefficients which when applied to this strongly correlated lagged data, linearly fit observed data, with minimal least squared residue error. This provided good general predictability across the data sources. The resulting linear equation was then used to predict the next hour's utilization.

To be able to evaluate prediction techniques with respect to peak values, for each data series, the observed utilizations are partitioned into small intervals, in increments of 0.05. For each such partition, the average absolute difference between observed and predicted values is obtained. Plotting these average absolute errors per interval helps understanding the behavior of the predictor algorithm for different input data ranges.

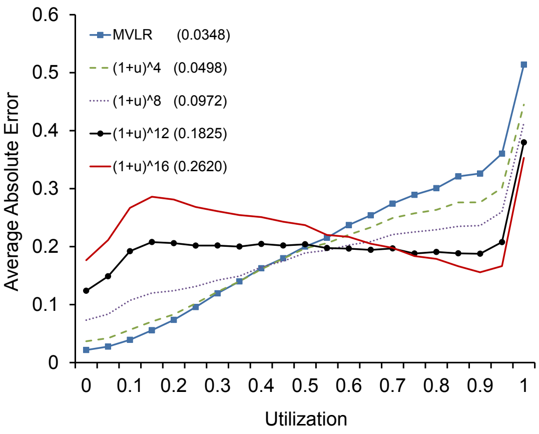

Figure 1. Comparing MVLR with Weighted Regression. The weighting parameter is $c = 4, 8, 12, 16$ which means a data point having utilization u and all lags associated with it are multiplied by $(1+u)^c$. The values in parenthesis are the average absolute errors across all intervals.

A minor problem we encountered that requires special consideration is the *missing values*. In this study, short gaps are approximated by their prior values. However, in our dataset, 769 time series have more missing data than the actual data. In such scenarios we discard the highly missing source of data from our dataset since otherwise they could skew our experimentation results.

Weighted MVLR: To adjust the MVLR to higher values, we first restrict the regression to a 5 week sliding window. Within the regression summations, we then weight [8] each data point. Because the overall distribution of utilizations is observed to be exponential, we employ exponential weighting in which a data point having utilization u as well as all lags associated with this data point were multiplied by $(1+u)^c$. This naturally assigns higher utilizations a significantly greater weight, thus skewing the predictions towards higher values, while simultaneously bounding them by the highest values. As can be seen in Figure 1, increasing c (the weighting parameter), from 4 to 16, reduces the prediction errors for the higher utilization intervals while increases the errors for the lower intervals. Therefore, a more consistent average absolute error across all intervals might be the best choice. For example, c=12 seems to be a good choice for our dataset.

In both MVLR and Weighted MVLR, we only relied on the predefined lags for our predictions. However, one must consider seasonal contributions, as well. Applying Fourier transforms [9] is a typical approach to discover obvious cyclic patterns within the data. The next two approaches employ Fourier transforms.

Scaled Seasonality: Applying a Fourier transformation on our dataset, we fit the summation of the top n (n=10 in this study) sine waves with the largest amplitudes – the terms that describe the most dominant variability within the input data – to the input data. This fits well to the overall seasonality within the provided data, but fails to fit the peaks in the data. To account for the terms not included in the contribution to our

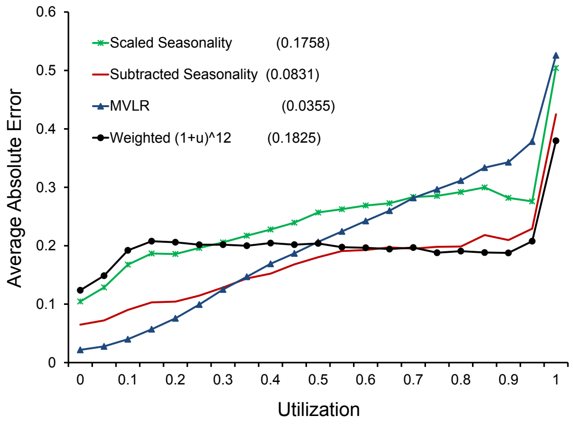

Figure 2. Comparing Fourier-based approaches (scaled and subtracted seasonality) with MVLR and weighted MVLR

prediction, it is reasonable to attempt to scale the Fourier transform to better fit the utilization.

One way of better fitting peaks is to apply a linear transformation to the computed Fourier transform, ignoring all values below some suitable cutoff (e.g. maximum = 0.05). We arrange for the minimum to remain unchanged by subtracting it, and then scale by the mean of observed values divided by the mean of predicted values, before adding the minimum back in.

Figure 2. compares the MVLR, Weighted MVLR and the Scaled Seasonality approahces. MVLR provides better predictive accuracy for low utilizations, and weighted MVLR for high utilizations. Scaled Seasonality is in between MVLR and Weighted MVLR in both cases, however, it also has the potential for longer term predictions (in months).

To improve the accuracy of our predictions, in the next approach, we combine the Fourier and regression analyses.

Subtracted Seasonality: In this approach, we subtract the seasonality from the original data, to remove much of the variability in the data, which makes it more linear, and thus a better fit with linear prediction models. Essentially, we 1) subtract the Scaled Seasonality from the observed utilizations, 2) perform MVLR (as before) on the resulting residue, and 3) add the seasonality back in to the resulting prediction.

The results (Figure 2) obtained are significantly better than using either Fourier transforms, or MVLR/Weighted MVLR alone. Applying this approach on our dataset, roughly, reduced the average absolute error across all inputs for large utilizations by a third, and halved the overall average absolute error.

The most significant drawback of using Fourier transforms is that unlike regression, which could quickly start providing predictions from initially observed results, a substantial amount of prior data must be available, in order to discover seasonality within an input time series. In practice, it is proposed that early predictions are predicated on regression alone, while periodically, as sufficient data becomes available, a Fast Fourier Transform is employed to repeatedly discover seasonality with the input data.

IV. Limitations and Threats to the Validity

The top three limitations of this study, which we currently working on, are 1) having a single dimensional prediction based on the CPU utilization, 2) studying only a linear regression (and its modified versions) prediction approach and 3) evaluating the forecast only based on the prediction accuracy and not the ultimate improvement in terms of impacts on the virtualization and capacity planning process.

As it is common in industrial research, the study is limited to the data which is available for the research team. It is obvious that having knowledge about other performance measures such as memory, disk I/O, and network traffic consumptions would potentially improve the prediction power. In addition, knowledge about workload type and even the business context behind the workload are among variables that may have impact on the future CPU utilization. However, in this study, we only had access to the CPU utilization data from the CA client's systems. The goal, therefore, was to maximize prediction accuracy (specifically with respect to the peak values) using the available data. However, in the future, we are planning to get access to several performance data resources and extend our one dimensional approach to such rich datasets.

While multivariate linear regression can be expected to respond appropriately to changing trends, our presumption (predicated on studying the data) was that no trend would be present within long term seasonality. If trends were present within the observed seasonality, it would be necessary to attempt to scale the seasonality using something more complex than a simple linear equation. Non-linear regression approaches are among the first techniques that we are planning to exercise on our current and future datasets. In addition, machine learning techniques, e.g. neural networks [10], need to be evaluated to find the best forecasting approach.

In this stage of the study, it is difficult to apply the research finding on the company's virtualization and capacity planning process. However, in short term, we are planning to explore more datasets and techniques and increase the supporting evidence around the ideas of storm forecasting, so that the company would be willing to apply them in its virtualization process.

In terms of construct validity, we made a best effort to accommodate missing data, but assumptions as to what missing values might have been, necessarily compromise predictive algorithms. In addition, in terms of external validity, this research was predicated on a single client data, during a comparatively short, six month, interval. Though containing a very large number of physical and virtual services, the behaviour of the system and the data patterns might not be the typical within all cloud computing environments.

V. Related Work

In general, the relevant literature to this work may fall into three categories: 1) workload characterization 2) workload forecasting and 3) prediction techniques. The first category focuses more on the features of the workload that can help analyzing and potentially predicting it [11-13]. The second category explores different data and prediction techniques to predict the future workload [14, 15] but still its focus is more on exploring data than the prediction itself.

In this paper, however, our focus is more on the prediction side, the third category. We use the data that is made available for us by our industrial collaborator and we study possibilities for maximizing the accuracy of the predictions. Therefore, we briefly mention some of the relevant articles in this direction.

Linear regression techniques are among the most popular workload prediction approaches. For example, Andreolini et. al. propose using moving averages to smooth the input time series, and then using linear extrapolation on two of the smoothed values to predict future workload [16].

Exponential smoothing, auto regressive and ARIMA models are the other most used approaches in this area [17]. For instance, Dinda et. al. compared the ability of a variety of ARIMA like models to predict futures [18]. Nathuji et. al. proposed evaluating virtual machines in isolation, and then predicted their behavior when run together using multiple input signals to produce multiple predictive outputs using difference equations (exponential smoothing) [3].

Using machine learning techniques for workload prediction builds up another large category of related literature. For instance, Istin et. al. used neural networks for workload prediction [10] and Khan et. al. applied hidden Markov models to discover correlations between workloads, which can then be used to predict variations in workload patterns [19].

Unlike the existing work, our paper uses basic techniques (linear regression and its modified version combined with Fourier transformation), as a starting point, and applies them on utilization data from a CA technology client with a specific goal of predicting peak utilizations.

VI. Conclusions and Future Work

System utilization can peak both as a consequence of regular seasonality considerations, and as a consequence of a variety of anomalies, that are inherently hard to anticipate. It is not clear that the optimal way of predicting such peak system activity is through approaches such as multivariate linear regression, since such prediction is predicated on the totality of the data observed, and tends to produce smoothed results rather than results that emphasize the likelihood of system usage exceeding capacity.

We have presented a number of modifications to standard multivariate linear regression, and to Fourier transforms, which individually and potentially collectively improve the ability of multivariate linear regression to predict peak utilizations with reasonably small average absolute error.

The best proposed modification subtracts an scaled seasonality, extracted by a Fourier analysis, of the observed utilizations, then performs a weighted multivariate linear regression on the resulting residues, and finally adds the seasonality back in to the resulting predictions.

In the future, we plan to extend this study using more predictive variables such as memory, disk I/O, and network traffic consumptions, as well as workload characteristics and business data. In addition, we plan to evaluate several prediction techniques such as non-linear regression and machine learning techniques, to improve the accuracy of storm prediction.

Acknowledgment

This research is supported by grants from CA Canada Inc., and NSERC. It would not have been possible without the interest, assistance and encouragement of CA. Allan Clarke and Laurence Clay of CA provided valuable input to this paper.

References

[1] X. Meng, V. Pappas, L. Zhang, “Improving the Scalability of Data Center Networks with Traffic-aware Virtual Machine Placement,” International Conference on Computer Communications, 2010.

[2] D. Gmach, J. Rolia, L. Cherkasova, A Kemper, “Workload analysis and demand prediction of enterprise data centre applications,” International Symposium on Workload Characterization, 2007.

[3] R. Nathuji. A. Kansal, A. Ghaffarkhah, “Q-Clouds: Managing performance interference effects for QoS-aware clouds,” European Conference on Computer Systems, 2010.

[4] M. Stokely, A. Mehrabian, C. Albrecht, F. Labelle, A. Merchant, “Projecting disk usage based on historical trends in a cloud environment,” Workshop on Scientific Cloud Computing, 2012.

[5] J. G. D. Gooijer and R. J. Hyndman, “25 Years of time series forecasting”, International Journal of Forecasting, vol. 22, issue 3, 2006, pp. 442–473.

[6] A. Amin, L. Grunske, A. Colman, “An automated approach to forecasting QoS attributes based on linear and non-linear time series modeling,” International Conference on Automated Software Engineering, 2012.

[7] CA Technologies. http://www.ca.com.

[8] N. R. Draper and H. Smith, “Applied regression analysis,” Wiley Series in Probability and Statistics, Third Edition, 1998.

[9] M. Frigo and S. Johnson, “The Fastest Fourier Transform in the West,” MIT-LCS-TR-728, Massachusetts Institute of Technology, 1997.

[10] M. Istin., A. Visan, F. Pop, V. Cristea, “Decomposition based algorithm for state prediction in large scale distributed systems,” International Symposium on Parallel and Distributed Computing, 2010.

[11] A. Williams, M. Arlitt, C. Williamson, K. Barker, “Web Content Delivery, chapter Web Workload Characterization: Ten Years Later,” Springer, 2005.

[12] M. Arlitt and T. Jin, “Workload characterization of the 1998 World Cup Web site,” Technical Report HPL-1999-35R1, HP Labs, 1999.

[13] S. Kavulya, J. Tan, R. Gandhi, P. Narasimhan, “An Analysis of Traces from a Production MapReduce Cluster,” International Symposium on Cluster, Cloud, and Grid Computing, 2010.

[14] D. Gmach , J. Rolia , L. Cherkasova , A. Kemper, “Workload Analysis and Demand Prediction of Enterprise Data Center Applications,” International Symposium on Workload Characterization, 2007.

[15] J. Tan, P. Dube, X. Meng, L. Zhang. Exploiting, “Resource Usage Patterns for Better Utilization Prediction,” International Conference on Distributed Computing Systems Workshops, 2011.

[16] M. Andreolini and S. Casolari, “Load prediction models in web based systems,” International conference on Performance evaluation methodologies and tools, 2006.

[17] T. Zheng, M. Litoiu, M. Woodside, “Integrated Estimation and Tracking of Performance Model Parameters with Autoregressive Trends,” International Conference on Performance Engineering, 2011.

[18] P. A. Dinda and D. R. O’Hallaron, “Host load prediction using linear models,” Journal of Cluster Computing, vol. 3, issue 4, 2000.

[19] A. Khan, X. Yan, S. Tao, N. Anerousis, “Workload characterization and prediction in the cloud: A multiple time series approach,” Network Operations and Management Symposium, 2012.