2013 12th International Conference on Machine Learning and Applications

Future Client's Requests Estimation for Dynamic Resource Allocation in Cloud Data Center using CGPANN

Jawad Ali, Faheem Zafari, Gul Muhammad Khan, S. Ali Mahmud
Department of Electrical Engineering
University of Engineering and Technology
Peshawar, Pakistan 25000
Email:jawad.ali, fahim.zafari, gk502, sahibzada.mahmud@nwfpuet.edu.pk

***Abstract*—Cloud computing is an emerging and rapid growing field of Infrastructure as a Service (IaaS), it has to deal with resource allocation and power management issues. This paper proposes CGPANN to accurately forecast the client's requests for a very short term duration of 1 second. A forecasting accuracy as high as 99.81% has been attained that verifies the accuracy of the proposed model. The experimental results show that the model outperforms all the contemporary models proposed in past.**

***Keywords*—Cartesian Genetic Programming, Cloud Computing, Dynamic Resource Allocation, CGPANN, Data center traffic forecasting**

I. Introduction

The data traffic forecasting helps in efficient utilization of data servers by proper management of the cloud computing infrastructure and energy resources. The dynamic resource allocation schemes, when implemented for a particular data center with realistic forecasting models, facilitates the end user in using the services and assists in fast switching between various subscribed services. Since traffic forecasting in data center can improve the performance of the data center and helps in providing various services efficiently, a range of stochastic forecasting and neural network models [1] are introduced in past to facilitate the resource management. This paper utilizes Cartesian Genetic Programming evolved Artificial Neural Network (CGPANN) for forecasting the number of clients' requests made to NASA web server. CGPANN prove to perform better at obtaining accurate and computationally efficient forecasting models [15]. The model uses historical data of 60, 90 and 100 seconds for forecasting the number of request for a single future instance i.e. one second. Using the sliding windows mechanism, the traffic for future 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 seconds is forecasted as well. A detailed analysis of the network using various architectures and training and testing scenarios is performed to obtain the overall best possible generalized estimation model for the task at hand.

II. Data Center Load Forecasting

Research in the fields of data center load forecasting and Cloud Computing resource allocation has enhanced the performance of data centers in terms of data resource allocation and data processing [1]. [1] uses Neural Network and Stochastic model for forecasting the number of client requests made to the NASA web server. [2] proposes Elastic Load balancer while exploring Dynamic Resource Allocation (DRA) methods based upon the load on Virtual Machines (VMs) concerning IaaS.[3] proposes DRA for Massively Multiplayer Online Gaming (MMOG), which is a concurrent game playing environment, providing Online Role Player Gaming. Users' requests and messages are traced and resources are dynamically allocated on the basis of players supported by each VM in that case. Data flow prediction for network is done in [4] using a hybrid model. The model uses Particle Swarm Optimization (PSO) algorithm in combination with Radial Basis Function (RBF) Neural Network. Data flow rate is estimated for the next instant using PSO-RBFNN while its Mean Percentage error is 5.4% for the given data flow rate. Measurements are made for load prediction and hot spot detection including the raw data of the number of session requests per second at different time scales in [5]. Web workload characterization for autonomic Cloud Computing is employed for Software as a Service (SaaS). A two-step approach is made in [6] for solving the problems like Load Sharing, Admission Control and Overload Control faced by Large Scale Data Networks for Network layer and application layer.

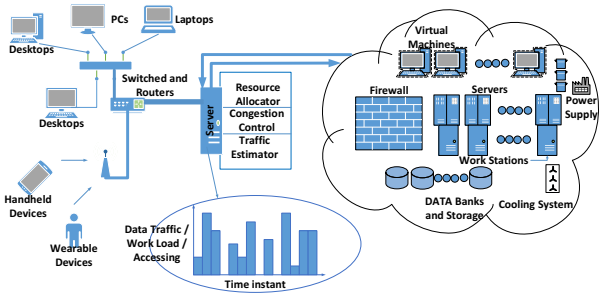

Fig. 1. Cloud Computing Structural Overview

Data centers and cloud computing are based on virtualization techniques that use Dynamic Resource Allocation, as in [7-9], for efficient utilization of power, processing and storage resources including load sharing strategies and equilibrium techniques. The management of such resources need a dynamic framework like Artificial Neural Network, that can learn the trends in the network traffic and give optimized results in

978-0-7695-5144-9/13 $31.00 © 2013 IEEE
DOI 10.1109/ICMLA.2013.189

forecasting loads based on the previous trends experienced by the Data Centers of Cloud architectures.

III. Cloud Computing Overview

Cloud Computing is a Remotely Processing Unit Accessing (RPUA) technique that provides services such as SaaS (Software as a Service) and IaaS on demand [1]. In case of IaaS, it is a virtualization infrastructure technique in which a specific slot of processing unit is allocated to the client, based on the user's specifications and cloud resources. While in SaaS, a client requests specific softwares, that the client does not own either because the client uses it less frequently or does not have the required specifications to use the software.

A virtual workstation is allotted to the client as per request[3]. The number of requests for a resource are noted periodically and main frame servers are turned on, set on sleep state or left idle depending on the requests from the clients [1]. Cloud Computing uses grid computing architecture by employing the fundamental approach of distributed computing while a high bandwidth access between the client (end user) and the processing resource (grids of parallel CPUs) is provided as shown in Fig. 1. Fig. 1 shows the basic architecture of Cloud Computing starting from the client to the work stations, data banks and virtual machines. Various end users of the CCI (Cloud Computing Infrastructure) are using the resources of the cloud while accessing it through different means. A single server is used before the cloud that provides the resource allocation, routing, congestion control, traffic directions and load estimation. The same server makes a history log of traffic data, work load and access history tagged with time stamps for certain management issues. CCI consists of grids of parallel processing units and data banks that needs continuous supply of power and proper ambiance for normal work load.

IV. Data Center Resource Allocation

Several DRA methods including Stochastic models, Neural Networks and Probabilistic approaches [1-2][7-8] are used for managing the clients that access the data centers. Hybrid architectures have also been proposed for multiple connectivity environment [3].

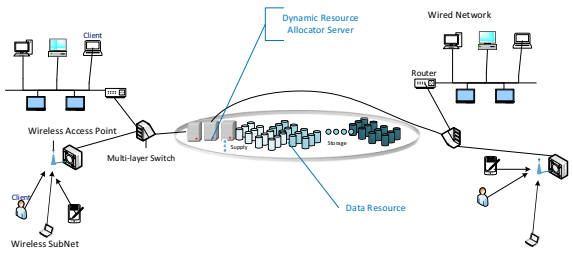

Fig. 2. Centralized Data Center framework; Clients from different network accessing data resource

Fig. 2 shows the data resource allocation for multiple clients. The DRA server uses historical data for forecasting future client requests made to a specific date center. This facilitates the power management and congestion control since the power state of the data banks are adjusted accordingly. The DRA provides resources to a specific client while simultaneously minimizing the wastage of energy in SaaS and IaaS architectures.

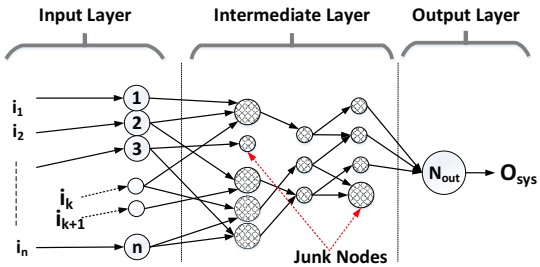

Fig. 3. Layered Architecture of a typical ANN

V. Cartesian Genetic Programming Evolved Artificial Neural Network (CGPANN)

ANN is the mathematical replica of natural neural network in which the synapse is replicated by a random connection amongst the functional nodes with an assigned weight as shown in Fig 3. In order to obtain a specific mathematical model for the task, various parameters of the network are subjected to training, and once it learn the process all the parameters are frozen. There are a range of training methods introduced in literature with CGPANN[10-14] producing the best results both in terms of its accuracy and computational efficiency. CGPANN uses the idea of encoding weights, functions and topologies in a single genotype that upon its evolution results enhanced functioning neural networks with augmented topologies and optimum weights. At the initial stages, the network attributes including connections, their respected weights and node functions are chosen randomly. The evolution process may add new features, replace an existing one or remove some features for the optimum performance of the ANN. Offspring are produced using 1 + λ evolutionary technique where λ is the number of offspring produced in a single generation. The Fittest offspring is then treated as the parent genotype and further mutation takes place using the same evolutionary criteria till the optimum genotype is obtained, that is transformed into ANN.

Fig. 3 shows the typical structure of an ANN, consisting of an input array I (represened by Eq. 1), input nodes N (represented by Eq. 2), intermediate layer of hidden nodes and an output node N_{out}.

$$I = [i_1, i_2, i_3, i_4, ., i_k, i_{k+1}...., i_n] \tag{1}$$

$$N = [1, 2, 3, 4...n] \tag{2}$$

The major characteristics of natural neural network comprises of four entities i.e. node functions, connections, weight assigned to each connection and number of inputs to a node. During the evolutionary process, all these parameters represented in the form of a genotype are modified to obtain the desired phenotypic behaviour.

VI. Application of CGPANN in Network traffic prediction and Dynamic Resource Allocation

The idea of forecasting the clients' requests to the web server, in broad terms, can be applied to a large data centers. Both IaaS and SaaS can use the proposed prediction models for the pre-allocation of their resources, restricting available resources, speeding up the data center services and power management. The same DRA scheme can also be applied to cloud computing environment for its effective and reliable performance. The subsequent sections discuss the application of the model and its evaluation.

A. Experimental Setup

The raw data consisted of data request tags, request instants (time stamped) and IP Addresses. The time stamped request numbers that forms the time series is extracted from the raw data. 30 different CGPANN networks were trained in various scenarios including the data center traffic forecasting for next 1 second based on historical data of past 60, 90 and 100 seconds. The network architecture was modified by varying the number of nodes being increased from 50 to 500 with an increment of 50. The choice of selecting node number between 50 and 500 enables fast learning and ANN reshaping. Note that CGP eliminates the junk nodes from the final prototype thus the phenotype might consist of nodes less than 50.

Historical data of a single day was used to train the network while it was tested on 28 days data. Using sliding windows mechanism, the data center traffic for 5 ,10, 20, 30, 40, 50, 60, 70, 80 ,90 and 100 seconds was also forecasted. Fitness is used as a performance evaluation criteria as shown in Eq. 3.

$$FITNESS = 100\% - MAPE \tag{3}$$

where MAPE is

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|RA_{if} - RA_{iA}|}{RA_{iA}} \times 100\% \tag{4}$$

In Eq. 7, RA_{if} is the forecasted value and RA_{iA} is the actual value at instant i. Here N is the prediction duration across which the error is being calculated. Mean Absolute Percentage Error (MAPE), given in Eq. 7 was used as the performance criterion. The mutation rate (μ_r) was kept 10% due to its better outcomes as proposed in [12]. 1+λ evolutionary technique is used to produce offspring during evolutionary with λ set to 9. All the networks are evolved for one million generations.

B. Performance and Evaluation

The training results for the CGPANN network that uses the historical data of past 60, 90 and 100 seconds for forecasting the data center traffic for next 1 second is tabulated in table I. The number of nodes were increased from 50 to 500 with an increment of 50 nodes per step. The results show that the model with 50 nodes provide the optimum results. Also the model using the historical traffic data of 100 seconds has the optimum training results. The testing results are given in Table II. It is evident from the table that the networks with 50 nodes provides the best forecasting results in terms of MAPE. Also the network that uses historical data of 100 seconds for forecasting the next one second provides better results compared

TABLE I. MAPE values for the training session

Nodes	60sec input CGPANN	90sec input CGPANN	100sec input CGPANN
50	**0.64647**	**0.66233**	**0.620905**
100	0.72526	0.66658	0.702408
150	0.77917	1.12083	0.947033
200	1.03882	0.98727	0.769213
250	0.95979	1.01773	0.977573
300	1.04223	1.11239	0.917853
350	0.90869	0.74537	1.258149
400	1.13937	0.82540	0.772861
450	0.99462	1.05678	1.018454
500	0.98618	1.08210	1.084759

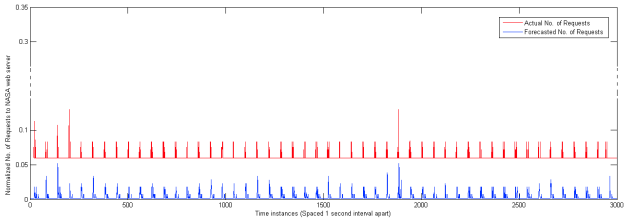

Fig. 4. Comparison of Real data set vs. forecasted values for first 3000 seconds, Aug, 4

to other networks. This is because the network has a longer data series to learn from. Table III provides the Root Mean Square Error (RMSE) values for various instances, whereas Eq. 5 defines RMSE. The optimum results are again obtained for a network that uses historical traffic data of past 100 seconds for forecasting the load for next one second. The significance of the proposed model is that as the forecasting horizon increases, the performance of the network becomes even better i.e. RMSE reduces with the increase in the forecasted duration.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(RA_{if} - RA_{ia})^2} \tag{5}$$

TABLE II. MAPE values for the testing session spanning 28 x 86400sec points

Nodes	60sec trained CGPANN	90sec trained CGPANN	100sec trained CGPANN
50	**0.4648**	**0.4644**	**0.4411**
100	0.4931	0.4708	0.4901
150	0.5424	0.8980	0.6994
200	0.7954	0.7804	0.5420
250	0.7170	0.7605	0.7223
300	0.7852	0.8495	0.6717
350	0.6773	0.5259	0.9951
400	0.8784	0.7967	0.9951
450	0.7473	0.5753	0.7753
500	0.7311	0.8419	0.8570

Figure 4 shows the comparison of the actual and forecasted values of number of clients' requests with a CGPANN network of 50 nodes that forecasts the data traffic for 1 second using the historical traffic data of past 100 seconds. The proximity of the two curves highlight the accuracy of the model. The forecast curve is leveled by +0.06 above the actual data curve on the normalized axis for better visualization. Table IV presents a comparison of the proposed CGPANN model with some of the other machine learning techniques proposed in literature to date. It is evident that CGPANN performs far better than any other model proposed to date. A MAPE value of 0.198%

TABLE III. NORMALIZED RMSE VALUES FOR 10 INSTANCES, FOR 100SEC UP TO 900SEC TIME HORIZONS AT DIFFERENT STEP SIZES

	Time Horizon(Sec)								
Time Instant	900sec	800sec	700sec	600sec	500sec	400sec	300sec	200sec	100sec
2065	0.022941591	0.021663521	0.019559343	0.017243589	0.015397123	0.014176548	0.011515427	0.008325455	**0.0019810**
2066	0.022941591	0.021786490	0.019559344	0.017243589	0.015397123	0.014176551	0.011515426	0.008325455	0.0072278
2067	0.022941591	0.021786491	0.019559345	0.017243590	0.015413535	0.014176551	0.011515427	0.009231078	0.0072441
2068	0.022941591	0.021798092	0.019559354	0.017243588	0.015690203	0.014176565	0.011515431	0.009231077	**0.0019810**
2069	0.022941590	0.021798092	0.019559370	0.017243588	0.015690203	0.014176568	0.011515432	0.009231079	0.0177704
2070	0.022941595	0.021798095	0.019559374	0.017243591	0.015690207	0.014176568	0.011515436	0.009258432	0.0092507
2071	0.022941596	0.021809689	0.019559385	0.017243594	0.015690207	0.014176568	0.011515435	0.009258432	**0.0019810**
2072	0.022941594	0.021809688	0.019559385	0.017243591	0.015706308	0.014176568	0.011515430	0.009285693	**0.0019810**
2073	0.022941592	0.021821276	0.019559384	0.017243591	0.015722396	0.014176566	0.011515430	0.009285693	**0.0019810**
2074	0.022941583	0.021832850	0.019572289	0.017243579	0.015722384	0.014194422	0.011515411	0.009285674	0.0019745

TABLE IV. COMPARISON OF DIFFERENT ALGORITHMS ON THE BASIS OF 100 POINTS FORECASTING HORIZON RMSE/MAPE

Model/Algorothm	RMSE/MAPE	Reference
BP Neural Network	1.05% MAPE	[18]
RBF-PSO	0.6% MAPE	[18]
Radial Basis Funtion (RBF)	3% MAPE	[18]
Levenberg Marquardt (LM) Algorithm	0.019 RMSE	[19]
Gradient Descent (GD)	0.0142 RMSE	[19]
Resilient back Propagation (RP)	0.0031 RMSE	[19]
Scaled Conjugate Gradient (SCG)	0.0128 RMSE	[19]
Conjugate Gradient with Flitcher-Reeves updates (CGF)	0.0128 RMSE	[19]
Conjugate Gradient with Polak-Ribiere updates (CGP)	0.0118 RMSE	[19]
Conjugate Gradient with Powell-Beal restarts (CGB)	0.0128 RMSE	[19]
One Step Secant (OSS)	0.0128 RMSE	[19]
CGPANN	**0.001981** RMSE, 0.198% MAPE	Proposed Model

and RMSE value of 0.001981 is far superior to the forecasting results of other models.

VII. CONCLUSION

This paper proposed CGPANN for accurately forecasting the client request's sent to the NASA web server. Historical traffic data of a single day was used to train the network while it was tested on the data of 28 days. The model forecasts the number of client's request for one second using the historical data of 60, 90 and 100 seconds. The obtained results highlights the accuracy of the model promising that the proposed CGPANN model is ready to encode the past trends from the provided data, performing better than other models elucidated in the paper. The proposed CGPANN can also be used as a forecasting tool in other time series including river flow, atmospheric pressure, temperature, solar irradiate, humidity, data traffic, instantaneous channel bandwidth, wind power generation, wind speed etc.

REFERENCES

[1] John J. Prevost, Kranthi, Manoj Nagothu, Brian Kelley and Mo Jamshidi, " Prediction of Cloud Data Center Load Using Stochastic and Neural Models", *Proc. Of 6th international conference on Systems Engineering Albuquerque, New Mexico, USA* - June 27-30, 2011.

[2] A. Inomata, T. Morikawa, M. Ikebe, Y. Okamoto, S. Noguchi, K. Hujikawa and H. Sunahara, "Proposal and Evaluation of a Dynamic Resource Allocation Method based on the Load of VMs on IaaS", *4th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pp. 1-6, 2011.

[3] Ginhung Wang and Kuochen Wang, "An efficient hybrid P2P MMOG cloud architecture for dynamic load management," *International Conference on Information Networking (ICOIN)*, Vol. 199 Issue. 204, pp. 1-3 Feb. 2012.

[4] Zhang Yu Bin, Lin Li Zhong and Zhang Ya Ming, "Study on network flow prediction model based on particle swarm optimization algorithm and RBF neural network," *3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, Vol.2, pp. 302-306, July 2010.

[5] P. Saripalli, G. V. R. Kiran, R. Shankar, H. Narware and N. Bindal, "Load Prediction and Hot Spot Detection Models for Autonomic Cloud Computing," *Fourth IEEE International Conference on Utility and Cloud Computing (UCC)*, pp. 397-402, Dec. 2011.

[6] Riccardo Lancellotti, Mauro Andreolini, Claudia Canali and Michele Colajanni, "Dynamic request management algorithms for Web-based services in cloud computing", *Proc. of the IEEE Computer Software and Application Conference (COMPSAC 11)*, Munich Germany, July 2011.

[7] M. A. Arfeen, K. Pawlikowski and A. Willig, "A Framework for Resource Allocation Strategies in Cloud Computing Environment," *2011 IEEE 35th Annual Computer Software and Applications Conference Workshops (COMPSACW)*, pp. 261-266, 18-22 July 2011.

[8] Fei Teng and F. Magoules, "Resource Pricing and Equilibrium Allocation Policy in Cloud Computing," *Computer and Information Technology (CIT)*, pp. 195-202, July 2010.

[9] N. R. R. Mohan and E. B. Raj, "Resource Allocation Techniques in Cloud Computing – Research Challenges for Applications," *Fourth International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 556-560, Nov. 2012.

[10] X. Yao. "Evolving artificial neural networks." *In Proceedings of the IEEE*, volume 87(9), pages 1423-1447, 1999.

[11] J. Peralta, G. Gutierrez and A. Sanchis, "Time series forecasting by evolving artificial neural networks using genetic algorithms and estimation of distribution algorithms," *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 18-23 July 2010.

[12] G. M. Khan, S. Khan and F. Ullah, "Short-term daily peak load forecasting using fast learning neural network," *Intelligent Systems Design and Applications (ISDA)*, pp. 843-848, 22-24 Nov. 2011.

[13] Zhang Yu Bin, Lin Li Zhong and Zhang Ya Ming, "Study on network flow prediction model based on particle swarm optimization algorithm and RBF neural network," *3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, vol.2, pp.302-306, 9-11 July 2010.

[14] Samira Chabaa, Abdelouhab Zeroual and Jilali Antari, "Identification and Prediction of Internet Traffic Using Artificial Neural Networks, *J. Intelligent Learning Systems & Applications (JILSA)*, Vol.2, pp. 147-155, 2010.