Data Mining in Adaptive Control of Distributed Computing System Performance

Ravi Kumar Gullapalli[#1], Dr.Chelliah Muthusamy[*2], Dr.A.Vinaya Babu[#3]

[1#]Technical Expert, Hewlett-Packard
Bangalore, India

*[*2]Academic Relations - Head*
Yahoo, Bangalore, India

[#3]Principal, JNTUH College of Engineering
JNTUH, Hyderabad, India

***Abstract*— The Distributed Computing Systems are significant building blocks in design and implementation of the business critical applications. It is always important for such environments to provide high performance for all kinds of workload variations. There are different solutions identified to deal with the performance problems in Distributed environments. There are recent investigations in exploring the feedback control based solutions in various computing environments. The self-managing capabilities are inherent qualities of Adaptive Control and such mechanisms are investigated to address performance problems in computing including distributed computing systems. The objective of Adaptive Controllers is to provide the intelligence in tuning the control parameters that would maintain the system in a desired state. In this paper we explore and identify the applicability of Data Mining based approaches to build Adaptive Controllers in self-managing the performance. We implemented a simple Adaptive Control using Time-Series Analysis. It predicts the frequency of occurrence of statements in a database so that the database driver can cache them according to the predicted values which demonstrates the feasibility of building Data Mining based Adaptive Controller**

***Keywords*— Data Mining, Adaptive Control, Distributed Computing Systems**

I. Introduction

With the advent of internet there is a significant change in the software applications and services development and delivery, particularly in distributed computing systems [1]. The non-functional requirements such as performance and availability [2] have become very important. The demand for making such applications with self-managing and self-healing capabilities [3] is an important behaviour to be exhibited. To support such abilities there are various mechanisms and solutions are experimented and explored, In this direction the feedback control systems are heavily investigated in computing for more than a decade [4]. There is a lot of research and investigations done in applying feedback control in different areas of computing. The majority of the computing areas where the feedback control is applied is in Distributed Enterprise and Cloud environments [5] such as workload regulation of Web and Application Servers [6] , improve the web caching [7][8], database driver caching [9] database servers [10], Computer Networks TCP Congestion control [11], CORBA [12], power management in data centres [13]. In these investigations there are few attempts in applying the Adaptive Control mechanisms [14] such that software applications and services can provide the self-managing capabilities. We have observed that Adaptive Control mechanisms are majorly applied in Application Servers [15], Web Servers [16] through Fuzzy Control. The recent trends in the area of Cloud Computing, there are Service Delivery Platforms (SDPs) that provide rapid cloud service development and deployment. These services run in cloud environments and it is very important for them to provide high performance and scalability irrespective of the changing workload dynamics. This demands the need to build autonomic SDPs [17] and we see usage of adaptive control is a huge opportunity.

Adaptive Control generally means to adapt to the change in the environment [18] and ensure that the system being controlled exhibits the desired behaviour. When adaptive Control is applied in computing, the controller tunes the control parameters such that the system performs as desired at any instant of time.

This demands the accurate decision capability as one of the important characteristic of any adaptive control scheme. There are already a set of Adaptive Control Schemes that are introduced for the last many decades. But in the case of Distributed Systems the decision making abilities in the adaptive control are heavily data oriented and intelligent decisions are required to be taken based on an effective data analysis. There are investigations in applying some of the Data Mining techniques in pro-active management in large clusters to handle faults to provide availability[19], there are experiments in applying Intelligent controllers [20] such as Artificial Neural Networks [21][22][23] and Fuzzy Logic based controllers [24] in Distributed Systems. Additionally there is a study to apply Data Mining in Distributed Enterprise Systems performance management [25], building reliable Distributed Systems [26]. We clearly observe the importance of using Data Mining in building high performance

Distributed Computing Systems. This motivated us to explore further and investigate in analysing the Data Mining [27] algorithms and approaches to build Adaptive Control techniques that can provide accurate decisions in tuning the control parameters that can self-regulate the systems. In this paper we did a comprehensive study of Data Mining, analysed the different algorithms and approaches of Data Mining. We identified the some of the Data Mining schemes that are suitable to build adaptive controllers for different problem areas of Distributed Computing Systems. In order to demonstrate this we have implemented a simple adaptive control using Time-Series Analysis. In this paper we review the Adaptive Control mechanisms, brief the Data Mining primitives that are relevant for adaptive control and discuss the analysis identifying the most suitable Data Mining algorithms and how they can be used to design and implement Adaptive controllers in Distributed Computing Systems.

II. ADAPTIVE CONTROL PRIMITIVES [28]

In this section we very briefly review the definition of Adaptive Control and various adaptive control schemes. All the figures in this section are redrawn based on diagrams present in [28]

A. DEFINITION AND PROPERTIES

An Adaptive Controller is defined as a controller with adjustable parameters and mechanism to adjust those parameters. The design and implementation of the controller will have the following steps:

- Characterize the desired behaviour of the closed-loop system
- Determine a suitable control law with adjustable parameters
- Identify a parameter adjustment mechanism
- Implement the control law

The Fig 1 below shows a sample Adaptive Control system

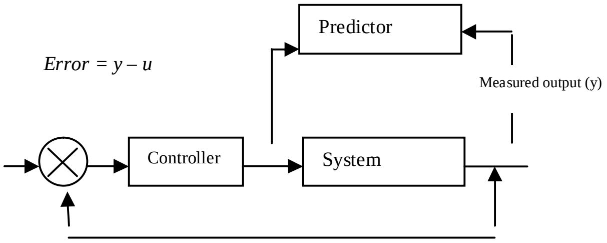

Fig. 1 Adaptive Control System [28]

B. MODELING

The computing system to be controlled needs to be modelled first before adjusting those using adaptive controllers. There are different mechanisms to model such as difference equations [29], ARMA modelling [30] both being very popular. It is required to estimate the model parameters before designing the controller. Typically these model parameters are estimated using a training data set.

C. ADPATIVE CONTROL SCHEMES

In this section we present the different Adaptive Control schemes briefly.

1. GAIN SCHEDULING

The Fig 2 below shows Gain Scheduling which is an adaptive control scheme to tune the controller gain for different operating points of the system. It is used in cases where the measurable variables of the system that correlate well with the changes in the dynamics of the system. It has two loops. The inner loop contains the controller and the system to be controlled. The outer loop contains the gain scheduling algorithm that tunes the controller based on the measured variables. The Gain Scheduling is regarded as a mapping from the system measured parameters to the controller parameters.

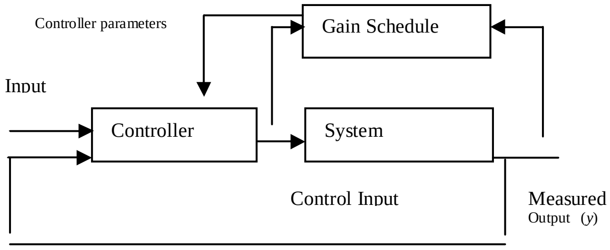

Fig. 2 Gain Scheduling [28]

2. MODEL REERENCE ADAPTIVE SYSTEMS

The Model Reference Adaptive Systems (MRAS) is an indirect adaptive control scheme, in which the performance requirements are specified in terms of a reference model. The Fig 3 below shows the block diagram of which has two loops. The inner loop consists of the feedback controller and the system to be controlled. The outer loop tunes the controller parameters such that the error between the actual measured output of the system and the model output is small. The hallenge in MRAS is to keep the error as small as possible hich is non-trivial. The MRAS uses the parameter adjustment rule called as MIT Rule [31].

The equation (1) represents the MIT rule

$$d\theta/dt = -\gamma e(\partial\theta/\partial t) \qquad (1)$$

Where
e = y - y_m is the model error and
θ = controller parameter
$\partial\theta/\partial t$ = sensitivity derivative of the error
γ = adaptation rate

The MIT rule is regarded as a gradient scheme to minimize the squared error e^2.

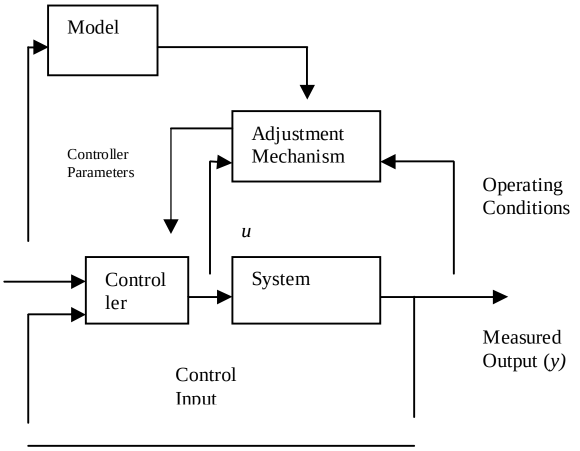

Fig. 3 Model Reference Adaptive System [28]

3. *SELF-TUNING REGULATORS*

Self-Tuning Regulators (STR) is an indirect adaptive control scheme. In this scheme the system parameters are estimated and updated. The controller parameters (or gain) are obtained using the estimated parameters. There are two loops in STRs; the inner loop consists of the controller and the computing system. The outer loop consists of a recursive parameter estimator and a design calculation to determine the controller gain value. The advantage of the STRs is that the dynamics of the system are considered and the model parameters are also updated regularly and hence the controller gain. The STRs block diagram is shown in the Fig 4 where there is a recursive parameter estimator, controller calculations. The parameter uncertainties are not considered even in STRs.

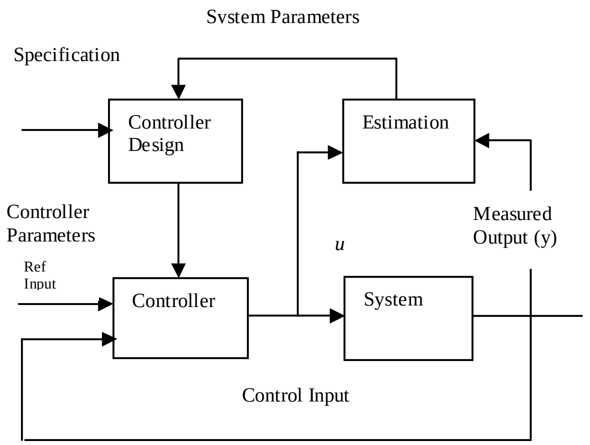

Fig. 4 Self Tuning Regulator [28]

4. *DUAL CONTROL*

The schemes described above are indirect and heuristic in nature, have the bottleneck of handling the uncertainties. In Dual control the uncertainty factor is taken into account. It means that the uncertainties in the estimated parameters are considered and controller knows to take special actions when it has poor knowledge about the system.

The Fig 5 shows the block diagram of a Dual Control where the uncertainties are handled by the nonlinear control law.

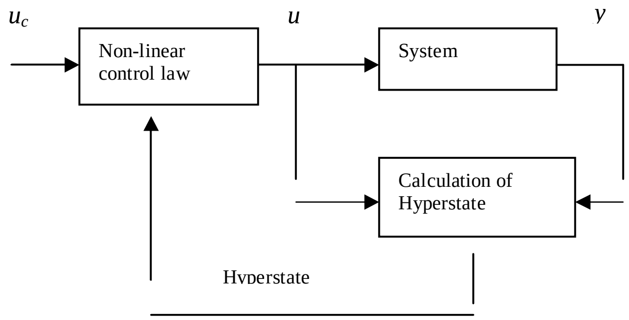

Fig. 5 Dual Control [28]

The dual control has Non-linear estimator which generates the conditional probability distribution of the state which is called the hyperstate of the problem. The feedback controller of the dual control is a non-linear function that maps the hyperstate into the space of control variables. The hyperstate is computed online. The dual control is from non-linear stochastic control theory where the system, its parameters, and the environment are described using a stochastic model. The criterion is to minimize the loss function and it is difficult identify an optimal controller that minimizes this loss function. An optimal solution can be found using dynamic programming, which is called the Bellman equation. The optimal control though drives the output to the desired value, but introduces disturbance when the estimates are uncertain, which will improve the future control. The optimal controller achieves a correct balance between good control and small estimation errors, which is dual control.

III. DATA MINING PRIMITIVES [32]

Data Mining is a process to identify, valid and useful patterns. These kinds of patterns are useful in interpreting and identifying the behaviour of the system under study.

A. *CONCEPT AND CLASS DESCRIPTION*

Data is associated with concepts or classes. Concepts are useful to explain the data in a concise and precise manner. Such descriptions are called Class/Concept descriptions.

These descriptions can be derived either by Characterization or Discrimination

- Characterization: It is a summarization of the general characteristics of target class of data set which is under study. Attribute Oriented Induction is an approach to describe the concepts.
- Discrimination: It is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. The discrimination descriptions are expressed in rule form is called as discriminant rules.

B. MINING FREQUENT PATTERNS

Frequent Patterns are patterns that frequently occur in the data. There are different frequent patterns called itemsets, subsequences, and substructures. Mining frequent patterns help in performing Association Analysis determining the support and confidence factors. Association rules help in determining how interesting the data is by defining the thresholds for support and confidence. Frequent pattern mining will be a potential Data Mining approach in designing adaptive controllers that predict the computing system behaviour in future.

The Correlation Analysis helps in excluding the uninteresting rules that have been generated by the Association Rules.

The Constraint based Association rule mining will be helpful further to filter between interesting and uninteresting rules. The different constraints include Knowledge type, Data, rule constraints.

C. CLASSIFICATION AND PREDICTION

Classification is a process to determine the model based on a data collection. This model classifies the entire data into different classes or concepts. This model can be used to predict the future data sets. The models can also be derived using training data set.

The derived models are represented as

- IF-THEN rules called Decision trees
- Neural Networks with the neurons connected using the weighted connections,
- Bayesian rules,
- Support vector machines
- k-nearest neighbour classification

The classification is used to predict discrete, unordered data sets based and analyse the class labelled data sets.

Prediction is used to predict continuous valued functions. In Data Mining the Prediction is used for numeric prediction. The statistical methods such as Regression Analysis and its variants can predict the future data set based on the a priori data set.

D. CLUSTER, OUTLIER ANALYSIS

Cluster Analysis classifies the data objects without consulting any reference class label. It generates the class labels on its own. The interclass similarity is used to identify the data objects which are most similar such that they all can be grouped together. Clustering also helps in taxonomy formation that organizes the observations into hierarchy of classes that group similar events together.

The Outlier Analysis identifies the data sets that do not comply with the model derived. Such data objects are called Outliers. These are noise and exceptions in the data sets. This can be very important.

E. EVOLUTION AND TIME-SERIES ANALYSIS

The behaviour of the computing systems under study will have varying behaviour over a period of time. There are different reasons such as workload variations, business environment changes, socio-economic dynamics. In order to study model and predict the computing system it will be important to analyse their behaviour over a period of time. There are different statistics based analysis that are incorporated in Data Mining for these kinds of periodicity pattern matching, similarity based data analysis. The ARMA models are one of such examples that can predict under these conditions. Exponential Smoothing techniques are some of the effective algorithms that can do prediction with higher accuracies.

IV. DATA MINING APPROACHES FOR ADAPTIVE CONTROL

In this section we discuss the possible Data Mining approaches that are most suitable to design and implement Adaptive Controllers. In the context of Distributed Computing Systems performance, the following are very important requirements. We explain for each of this category, the possible and relevant Data Mining schemes that can be applied. There are other possible applications of Data Mining in different areas of Distributed Computing Systems, but our discussion is limited to the performance management related aspects.

A. WORKLOAD PREDICTIONS

There are different types of servers that are part of Distributed Computing Systems such as Enterprise Messaging Servers, Web Servers, and Application Servers. As discussed it is important for these servers to exhibit high performance all the time. The variations of the workloads in either internet or enterprise environments will be very dynamic and it is important for these servers to have the ability to predict the future load so that the entire distributed environment is ready to meet these future loads particularly when there are sudden and huge increase in the workload.

Data Mining provides a big set of schemes using which the workload predictions can be done with a higher accuracy. The following are some of such Data Mining algorithms that provide the ability to predict the workloads in advance.

- Time-Series Analysis: The Time-Series Analysis is a very useful algorithm that can be used to predict the workloads. There are different algorithms based on Time-Series Analysis such as Moving Average, Single, Double and Triple Exponential Smoothing techniques [33].
- Artificial Neural Networks (ANNs): ANNs have the proven ability to do pattern matching, estimation. Before using the ANNs for any workload predictions, they must be trained so that the workload predictions are more accurate.
- Fuzzy Control: The Fuzzy rules can be defined from the a priori data set. These fuzzy rules defined will capture the patterns of the workload variations. These rules will be helpful to predict similar work load variations. There are earlier investigations done in applying Fuzzy controls in Web environments [22].
- Event Prediction: Event Prediction is attempted for availability prediction [34], we propose this as a useful approach for performance management. During the working conditions of the servers, if the important events when captured along with their previous states. This combination of information will be useful to perform predictions of such events. In order to make use of the Event Prediction, the knowledge base of the following has to be created.
  - Important events that occur, the external noise and disturbance that cause the events to occur
  - Behaviour of the servers before, during the event and after the event occurrence

  If the event is related to sudden spikes in the load causing performance degradation, if this information is captured then certainly Event Prediction can be a useful mechanism to design Adaptive controllers.
- Episode Discovery: Data Mining has algorithms to detect the frequent episodes. This will be useful to predict any such frequent episodes that are going to occur in the future. Episode Discovery process allows mining the occurrence episodes within a given window. Such two adjacent windows will be similar to each other.

B. WEB AND DATABASE CACHING IMPROVEMENTS

Caching is proved to be an improved technique in performance improvements in Web and Database environments. There are techniques such as LFU, LRU and many improvements over them are identified and implemented to improve the cache hit ratio.

- Frequent Pattern Mining: Many of caching techniques have adopted Pattern matching based mechanisms [35]. We see a potential opportunity in using the Frequent Pattern Mining based techniques to implement caching algorithms.
- Clustering: As Clustering provides the ability to classify the data objects, it is a potential mechanism in caching. In the case of web servers or database drivers when there are different classes of users or applications accessing the servers, the web pages or statements that get cached can be classified based on the users or applications providing a fair share of the caching resources.
- Time-Series Analysis: Though it may not appear to be a natural relevance for the Time-Series Analysis application in improving cache ratios, we have investigated and proved that it provides a significant improvement of cache ratio in database drivers [9]. This triggers to explore the possibility of applying Fuzzy mechanisms, Decision Trees and Artificial Neural Networks for predicting the access patterns of the cache. As these kinds of techniques are resource intensive methods that consume the CPU and memory resources for their processing, we recommend them for use in enterprise class massive applications.
- Outlier Analysis: It is very important to identify radically different patterns present in the data set. Based on the a priori knowledge, implementations of outlier analysis in the adaptive controllers can predict such patterns much in advance such that the servers are ready with required resources.

V. IMPLEMENTATION

In [9] we have implemented a Time-Series Controller to predict the database statements that would be accessed in future so that at any point of time the database driver cache retains the statements such that the cache ratio is high. The following Fig 6 shows the cache hit ratio using this Controller. We can observe the improved cache hit ratio with the Triple Exponential Smoothing Controller.

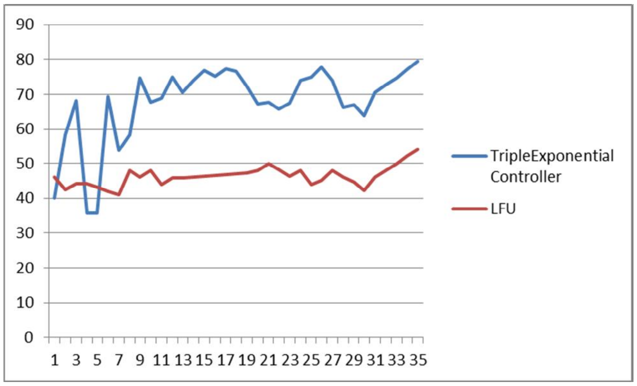

Fig. 6 The Cache Hit ratio using a Triple Exponential Controller

We have implemented a Java based controller based on the Triple Exponential Smoothing with its coefficients as α=0.25, β=0.50, γ =0.75 using the OpenForecast API[36].

VI. CONCLUSION AND FUTURE WORK

In this paper we have identified the different problem areas in Distributed Computing Systems and the possible Data Mining algorithms that can be implemented in Adaptive Control. We intend to implement Adaptive controllers using the discussed Data Mining approaches in the areas of workload prediction of Java based Enterprise Servers (JEE Servers). We want to evaluate the performance improvement using data mining based adaptive controllers with conventional controllers. Additionally we want to extend this to the web and database driver caching environments, and identify the most optimal data mining based controller for Distributed Computing Systems.

ACKNOWLEDGMENT

We thank Dr.AVN Krishna, Principal PJMSCET Hyderabad and M.Upendra Kumar, Associate Professor MGIT, Hyderabad for their encouragement and constant support.

REFERENCES

[1] S. Abdelwahed, N. Kandasamy and S. Neema, "A Control-Based Framework for Self-Managing Distributed Computing Systems",Workshop on Self-Managed Systems (WOSS'04), Newport Beach,CA USA, 2004.

[2] Joseph L Hellerstien, Kaan Katircioglu, and Maheswaran Surendra, “A Business-Oriented Optimization of Performance and Availability for Utility-Based Computing”, Journal on Selected Areas of Communications, Oct., 2005.

[3] Chris Toft et al, www.hpl.hp.com/techreports/2004/HPL-2004-49.pdf Self Managed Systems - A Control Theory Perspective HPL, 2004

[4] Joseph L. Hellerstien, Yixin Diao, Sujay Parekh, and Dawn Tilbury Feedback Control of Computing Systems, John Wiley 2004

[5] Ravi Kumar Gullapalli, Dr.Chelliah Muthusamy, Dr.A.Vinaya Babu, “Control Theory Applications in Java based Web and Enterprise Environments – A Survey”, IJACSA, Vol 2, No 8, 2011

[6] Wei Xu, Zhangxi Tan, Armando Fox, David Patterson, “Regulating Workload in J2EE Application Servers”, http://www.controlofsystems.org/febid2006/files/16225_Wei.pdf

[7] Ying Lu, Tarek Abdelzaher and Gang Tao, “Direct Adaptive Control of A Web Cache System”, Proceedings of the American Control Conference, Denver, Colorado, 2003

[8] Ying Lu, Avneesh Saxena and Tarek E Abdelzaher Differentiated Caching Services; A Control-Theoretical Approach, IEEE International Conference on Distributed Sysytems, 2001

[9] Ravi Kumar Gullapalli, Dr.Chelliah Muthusamy, Dr.A.Vinaya Babu, Raj N. Marndi, “A Feedback Control Solution in improving database driver caching”, IJEST, Vol 3, No 7, Jul 2011

[10] Tarek Abdelzaher, Yixin Diao, Joseph L Hellerstein, Chenyang Lu, and Xiaoyun Zhu., “Introduction to Control Theory and Its Applications to Computing Systems”, International Conference on Measurement and Modeling of Computer Systems SIGMETRICS□08

[11] Seungwan Ryu, Chulhyoe Cho,”PI-PD-controller for robust and adaptive queue management for supporting TCP congestion control”, 132 - 139 18-22 April 2004

[12] Xiaorui Wang et al, “FC-ORB: A Robust Distributed Real-time Embedded Middleware with End-to-End Utilization Control”, ACM Journal of Systems and Software, Vol 80, Issue 7, 2007

[13] What Does Control Theory Bring to Systems Research? Xiaoyun Zhu, Mustafa Uysal, Zhikui, Wang , Sharad Singhal, Arif MerchantPradeep Padala, Kang Shin, ACM SIGOPS Operating Systems Review, Volume 43 Issue 1, January 2009

[14] Adaptive control: “http://en.wikipedia.org/wiki/Adaptive_control”

[15] Giovanna Ferrari, Santosh Shrivastava,Paul Ezhilchelvan, “An Approach to Adaptive Performance Tuning of Application Servers”, IEEE International Workshop on QoS in Application Servers, 2004

[16] N. Gandhi and D. M. Tilbury, Y. Diao, J. Hellerstein, and S. Parekh “MIMO Control of an Apache Web Server, Modeling and Controller Design”, IEEE American Control Conference, 2002

[17] Robert D. Callaway, Michael Devetsikiotis, Yannis Viniotis, Adolfo Rodriguez, “An Autonomic Service Delivery Platform for Service-Oriented Network Environments”, vol. 3 no. 2, pp. 104-115, April-June 2010

[18] Giovanna Ferrari, Santosh Shrivastava,Paul Ezhilchelvan, “An Approach to Adaptive Performance Tuning of Application Servers”, IEEE International Workshop on QoS in Application Servers, 2004

[19] http://www.research.ibm.com/PM/

[20] Ravi Kumar Gullapalli, Dr.Chelliah Muthusamy, Dr.A.Vinaya Babu, “A Study of Intelligent Controllers application Distributed Systems”, IJCSE, Vol 2, No 4, Aug-Sep 2011

[21] Fuquan Tian , Wenbo Xu , Jun Sun, (2010),Web QoS Control Using Fuzzy Adaptive PI Controller, Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science

[22] Palden Lama , Xiaobo Zhou (2010), Autonomic Provisioning with Self-Adaptive Neural Fuzzy Control for End-to-end Delay Guarantee, IEEE International Symposium on Modeling, Analysis and Simulation of Computer Telecommunication Systems

[23] ANN: http://en.wikipedia.org/wiki/Artificial_neural_network

[24] Fuzzy Logic : “http://en.wikipedia.org/wiki/Fuzzy_logic”

[25] David A. Cieslak, Douglas Thain, and Nitesh V. Chawla, “Short Paper: Troubleshooting Distributed Systems via Data Mining”, High Performance Distributed Computing, 2006

[26] Michael Mock and Dennis Wegener, “A data mining based approach to reliable distributed systems”, SRDS 2009

[27] Data Mining, “http://en.wikipedia.org/wiki/Data_mining”

[28] Karl J.Astrom and Bjorn Wittenmark, “Adaptive Control”, *Pearson Education, 2009*

[29] Erwin Kreyzig, “Advanced Engineering Mathematics”, *John Wiley and Sons*

[30] ARMA: http://en.wikipedia.org/wiki/Autoregressive_moving_average_model

[31] Simon H. Fu, C.C. Cheng, C.Y. Yin, “Direct Adaptive Control for a Class of Linear Discrete-time Systems”, ASCC 2004

[32] Jiawei Han and Michline Kamber, “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006

[33] http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc435.htm

[34] AK Pujari, “Data Mining Techniques”, Universities press, 2007

[35] A. Radhika Sarma and R. Govindarajan : An EfficientWeb Cache Replacement Policy, , In the Proc. of the 9th Intl. Symp. on High Performance Computing (HiPC-03), Hyderabad, India, 2003

[36] Open Forecast API : http://www.stevengould.org/software/openforecast/index.shtml