IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. I (May-Jun. 2014), PP 50-55
www.iosrjournals.org

Analyzing Process Behavior to Predict Resource Allocation in Distributed Environment by Using Time Series and Online Predictive Approach Algorithm

Rucha Ravindra Galgali[1], Prof. S. G. Vaidya[2], Prof. S. M. Tidke[3]
[1]*Student, Computer Science and Engineering, Shreeyash Engineering college, Aurangabad, India*
[2] *Assistant Professor, Computer Science and Engineering, Shreeyash Engineering college, Aurangabad, India*
[3]*Assistant Professor, Computer Science and Engineering, Shreeyash Engineering college, Aurangabad, India*

Abstract:*A distributed system is a collection of different computers to handle large amount of data, the connected computers can share and coordinate their data on network. It is very difficult to for server to handle, analyze & process such a large amount of data, so that the performance of operations performed by these systems is reduced and they will produce inefficient data as a result. To handle this problem this paper has implemented the online predictive approach algorithm which uses time series to monitor process behavior, analyze it and predict the future observations to determine the resources required by the process in future.The resource prediction is used to optimize the data access operations like read, write, uploading and downloading the file to and from distributed system & which will improve the performance of distributed system. Time series is generated for every operation so that client will get performance chart.*
Index terms *Application analysis, Application prediction, Distributed system, Time series ,Time series analysis*

I. Introduction

A Distributed system is a software system in which components located on networked computers can communicate and coordinate their actions by passing messages. While handling large amount of data we will face some improper problems such as the time for executing the process will be more, the efficient data is not provided by the operation (Read/Write).

There are many scientific applications which produces large amount of data. It is very difficult to handle, analyze and process such data. There are many existing systems which causes processing, handling and analyzing such a large data. For example clusters and grids [1].

The aim of this paper is to improve scheduling decisions in large scale environments by avoiding historical processes. This paper uses time series to predict process resources by avoiding historical data. Here we are forming every process as time series. If we formed as time series means we can easily understand how much time taking for each process. At a same time we can reduce application execution time. by using time series we can easily analyze the operations such as Read/Write.

The purpose of this system is to improve the performance of the distributed system. This can be achieved by predicting the resources and reducing the data access to database. There are many techniques available to optimize data access like data replication, migration, distribution and access parallelism but these techniques doesn't consider the dynamic behavior of process [1].

This paper supports a strategy which evaluates the efficient reducing system memory locations and predicts the application files and maintain every process that is read and write operations time series for each operations , so we can able to easily analyze the operation. This can also avoid historical data duplication.

This paper contains data optimization approach organizes application behaviors as time series and, then, analyzes and classifies those series according to their properties. By knowing properties, the approach selects modeling techniques to represent series and perform predictions, which are, later on, used to optimize data access operations [1]. Advantages of this system are this system improves resource allocation and also characterizes, predicts application workloads in distributed environment.

II. Literature Study

This section gives the works related to analyze application behavior and then predict application behavior to reduce data access. The Literature study gives the studied done by some authors on existing system, their advantages and disadvantages.Following are the few papers studied for analyzing behavior of various processes and allocation of resources for their execution.

Renato and Mello [1] propose the strategy which supports the online prediction of application behavior in order to optimize data access operations on distributed systems without requiring any information on past execution. In this approach the first step is to monitories the process behavior. After that these processes are

converted into time series, according to properties time series get analyzed and classified. Then modeling technique is selected to model these time series to get some future observations. In last step these predictions are used to optimize the data access. Rahman and Barker [2],propose a framework that predicts the sites transfer times which are hosting replicas, the data from various sources is used for that purpose. They also used the neural network to predict the transfer time of different sites that currently hold file replicas. Chervenak [3,1],propose a framework called Replica Location Service (RLS) which maintains and provides information on physical locations of replicas.

AL-Mistarihi and Yong [4,1], implemented an approach to select the best replica which presents the lowest access cost, for applications.Devarakonda and Iyer[5, 1] propose a statistical approach to predict the consumption of CPU, file system I/O, memory. This study is verifying the process behavior with automaton stored in database. When a new process arrives at the system it is verified with the automaton, if any automaton from the database is capable to represent that process then that automaton is used to estimate the resource requirements for the process.

Kim and Chandra [6, 1] presented accessibility aware resource selection techniques to choose nodes which can efficiently access data from remote data sources. They showed that the accessibility of node is depending on the local data access observations collected from the nodes neighbors. They also proposed the heuristic to reduce execution time of data intensive applications. Vazhkudai and Foster[7], designed and implemented high level replica selection service which uses information regarding replica location and user preferences to guide selection from storage replica alternatives. They presented a dynamic information collection using Globus information service capabilities concerned to storage system properties and how this information can help to improve and optimize the selection process.Faerman and Wolski [8, 1], propose the adaptive regression modeling (AdRM) to determine file transfer times for network bound distributed data intensive applications.AdRM method accurately predicts data transfer times in wide area multiuser environments.

Wang [9,1],employ machine learning techniques to model requests to storage devices, based on Classification And Regression Trees (CART). This work proposes two prediction approaches, the first considers one prediction model for every data access request and the second models and predicts observations based on an average behavior of requests. Oldfield and Kotz [10, 1] propose Armada framework to monitor, control and execute the applications. Armada represents process and data flows by using Graph structure. It also improves network throughput. Oldfield and Kotz [11,1], also described the design of the flexible parallel system that allows the application to control the behavior and functionality of file system aspects.

Senger [12, 1] propose an online approach to acquire, classify and extract process behavior. By using this approach one can simply monitor the user application without any need of recompilation or modification. This approach is capable of automatically modeling process behavior. Senger [13, 1] also propose an approach to predict execution times of parallel applications. This approach has improved scheduling decisions in large scale environments and it has also provided the knowledge of application to system scheduler which makes resource allocation.

Ning Wang, Xian-Yao Meng [14], propose a novel online self constructing fuzzy neural network for time series prediction that speed up the learning process and build a more parsimonious fuzzy neural network. It also achieves accuracy on the characteristics if growing and pruning.O.B.Yaik, Haron F., Chan Huah Yong [15], implemented a model to perform time series prediction using adaptive association rules. This model uses the idea that if a segment of repeatable time series pattern has occurred, it has the possibility that the following segments of the repeatable pattern appear. Techniques applied to mine time series pattern are data mining and pattern matching. This model also has the ability to provide confident level for each prediction it made and perform continuous adaptation.

Draghisi A., A. Costan, V. Cristea [16], presented a presentation architecture developed within the MONALISA monitoring framework which also provides methods for estimating future values for different parameters on various periods of time. These predictions enhances the self adaptive behavior of several data intensive applications. This paper presents the research that focuses on machine learning algorithm correlated with statistical techniques for data mining purposes in order to perform n step ahead time series predictions and to dynamically evaluate their performances.

III. System Design

Renato and Mello propose the strategy which supports the online prediction of application behavior in order to optimize data access operations on distributed systems without requiring any information on past execution [1]. Our paper uses online predictive approach algorithm to develop the system.

This paper implements the proposed algorithm that is online predictive approach algorithm. The data flow for development of the system is shown in the following figure.

Data flow diagram:

Level 0:

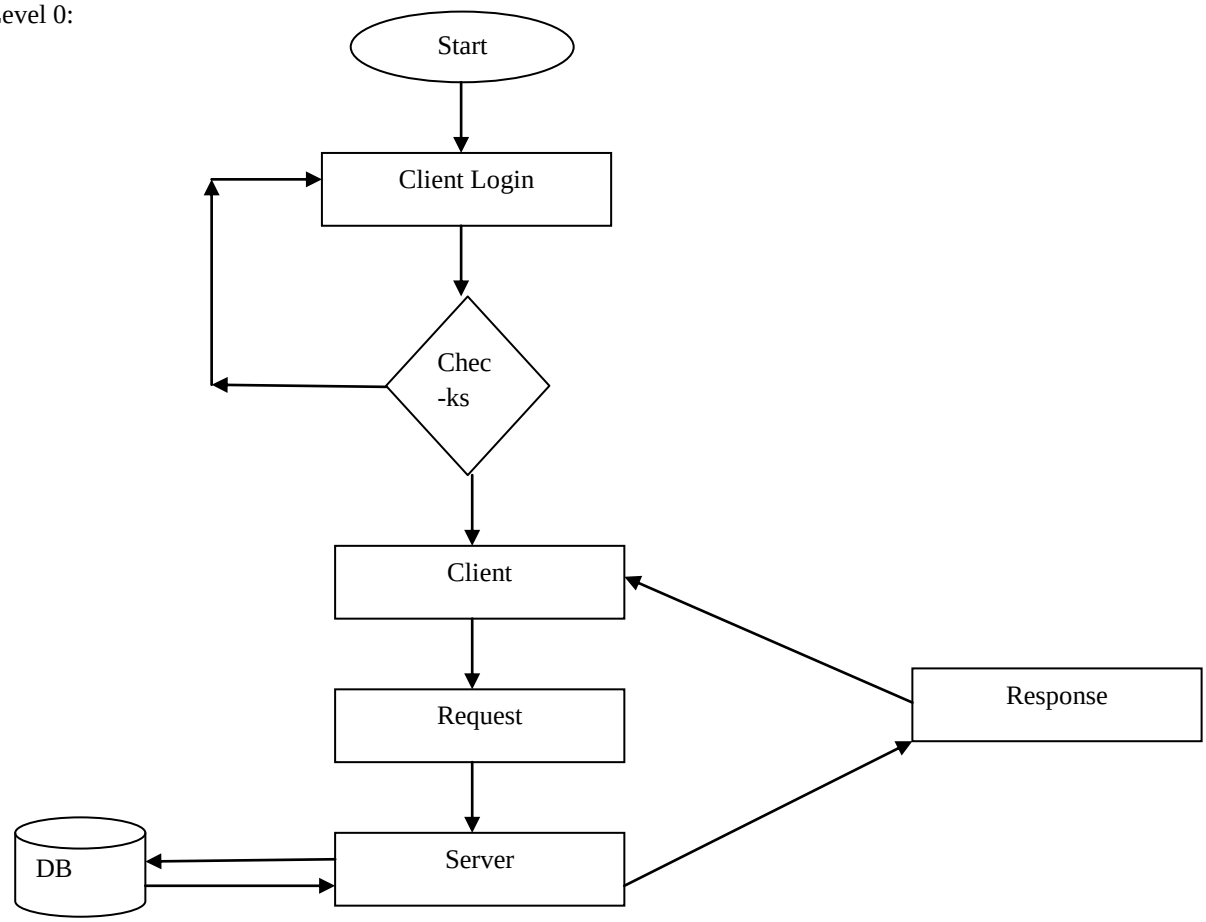

Fig. 1 Level 0 Data flow diagram

Explanation

Thisdata flow diagram shows how the operation flow is happening first step the execution start from if the new client the register process will be first otherwise already register client means directly process with login process. Client will process with user name and password. After that the user name password will check with database if exist means the client page will open. Otherwise again it will process login page itself. After that client give the request to server and server will responds to client.

Level 1:

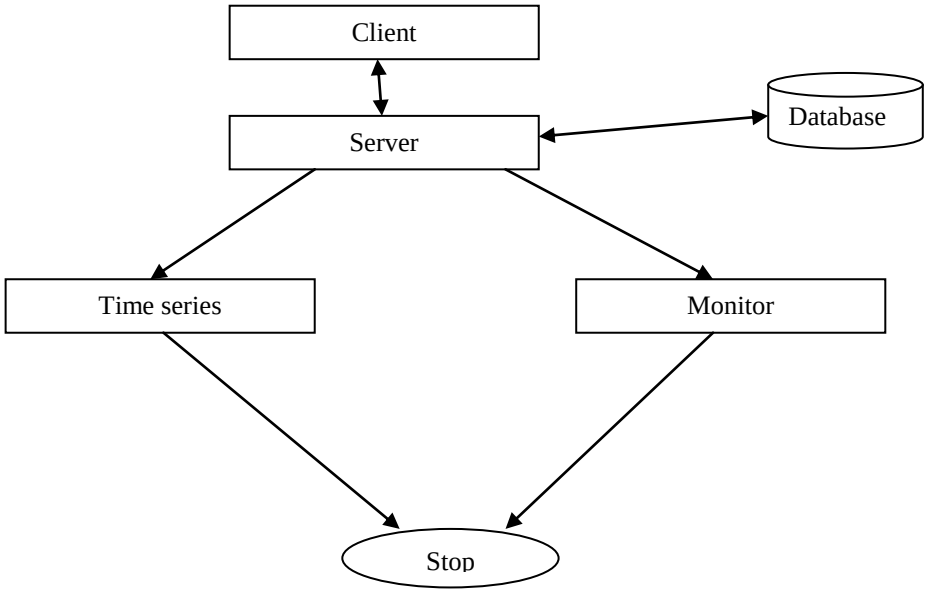

Fig. 2Level - 1 Data flow diagram

Explanation
Data flow diagram shows how the operation flow is flowing between client and server in this case the first step shows the client request and server will receive the client request then server will responds to client. After finishing all operations between client and server time series will calculate for each client request and server responds. Finally the graph will generate for each process. Then how many client are running that will be monitor.

3.1 Diagram for Online predictive approach algorithm steps

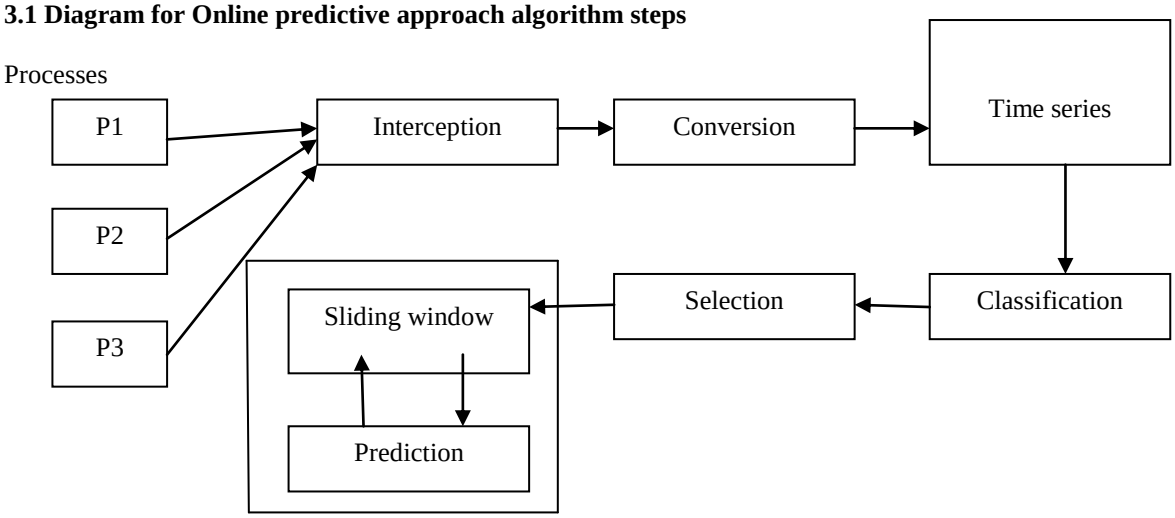

Fig. 3 Steps for Online predictive approach algorithm [1]

Explanation of Online Predictive Approach Algorithm

3.1.1 Interception: The first step, i.e., application knowledge acquisition, is responsible for monitoring process behavior by using event interception. The interception mechanism is associated with the process under execution.When a program calls a function, DLSym intercepts the call and injects any code instead.

3.1.2 Conversion: After extracting the application behavior, it transform the sequence of read-and-write events in a multidimensional time series.

3.1.3 Time Series : The third step evaluates the generation process of time series TR according to specific properties: stochasticity, linearity, and stationarity

3.1.4 Selection:Based on the evaluation of the time series generation process, we select an adequate modeling technique.e.g. when the series is deterministic, a reconstruction is conducted by considering the Takens' immersion theorem which relates series observations over time.

3.1.5 Adaptive Sliding Window:In this step we consider the adaptive sliding window (ASW) mechanism proposed to estimate the number of time series observations to be predicted, based on process behavior changes.

3.1.6 Prediction:After the previous steps, the prediction is performed on the time series, which represents process behaviors.

3.2 Modules:

The following modules can be used to develop the proposed algorithm:

3.2.1 User interface

To connect with server user must give their username and password then only they can able to connect the server. If the user already exits directly can login into the server else user must register their details such as username, password, Email id, City and Country into the server. Server will create the account for the entire user to maintain upload and download rate. Name will be set as user id. . Logging in is usually used to enter a specific page, which trespassers cannot see. Once the user is logged in, the login token may be used to track what actions the user has taken while connected to the site. Logging out may be performed explicitly by the user taking some action, such as entering the appropriate command, or clicking a website link labeled as such. It can also be done implicitly, such as by the user powering off his or her workstation, closing a web browser window, leaving a website, or not refreshing a webpage within a defined period. In the case of web sites that use cookies to track sessions, when the user logs out, session-only cookies from that site will usually be deleted from the user's computer. In addition, the server invalidates any associations with the session, making any session-handle in the user's cookie store useless.

3.2.2 Client

In this module after enter into client profile page the client can perform many operations if the client want to view the particular applications in server means, client can easily view the application file using application prediction. After that the list will display on the screen. If the client wants to show the content from the particular file means it is easy to view the file first client should select the file after that using request button client can view the data from the server. Then client can upload the files to server using upload button the server can upload the file. While clicking upload button new window button will open. After that which file client want upload just select that file then simply give click open button then the file will upload into server. If client want to download some file means it's possible. First client should select that particular file after selecting particular file just click download button then file will successfully download. Suppose client want to know how the performance is implemented in this case just we introduced time series regarding your operation time series will generate.

3.2.3 Server

Network servers make use of communication ports to assign client users to devices. Connections are made through a local area network (LAN). Different assignment tasks are handled by specific network servers. In this case the server handles the most operations using of database if the server gets the request from the client(example login). Regarding to client request the server will match the request with database. depends on the request server will periodically give response to client. The monitor will monitoring how many clients are in online and which operations they performing. if the clients need to download file. After getting request from the client the server will get the file from the database. if the client gave invalid file means the server will responds to client. While client doing registration

3.2.4 Predict application execution

In this case using online prediction it will split all those applications into chunk file it means all splinted application are stored in database. Depending on client request the sever will get the data from database. Here for retrieving purpose data we are using selection technique. If client requesting data to server means server will get the exact data from the database depending on the cline request and here we predicted all the files regarding applications. So server can easily get the data from the database no need to read all application files. At the same time we predicted all application behavior

3.2.5 Time series

In this case the depending on the client operations the time series will evaluate. If the client performing the read operation means client giving request to sever then time will calculate from the request time until reaching to server this time called as request time. After reaching the request to server then the server will response regarding to client request then after reaching response to client that time will calculate. Between these two time periods the request and response time will calculate then result time period will store into data base. If client want to perform more operations like read, write download ,upload, etc all those time series will store into database. Using these time series client will get performance chart. So depend on the time series client can analyze the process behavior and application behavior. Using the time series client will get predicted application execution time.

IV. Conclusion

4.1 Conclusion

In this case system minimizes the application execution time by optimizing data accesses and, therefore, improving decisions on replication, migration, and consistency. From that, data access operations are transformed into time series. By modeling those series, we can understand the behavior of applications and, therefore, predict future observations. By predicting future observations we can optimize data access and improve the performance of distributed system. This system also avoids historical data duplications.

4.2 Future Enhancement

The future work improves to avoid the storage of long historical and efficiently adapts according to variations on the behavior of applications. Performance analysis will be carry out in our project to verify the works. The performance of processes will be analyzed as well as graphically deployed. This graphical view will be used to present this system effectively and easy to understand.

References

[1] Renato Porfirio Ishii and Rodrigo Fernades de Mello, "An Online Data Access Prediction and Optimization Approach for Distributed Systems", vol.23, no. 06, June 2012.

[2] R. M. Rahman,, K. Barker and R. Alhajj, "APredictiveTechniqueforReplicaSelectioninGridEnvironment",Proc. IEEE Seventh Int'l Symp. Cluster Computing and Grid, pp. 163-170, May 2007.

[3] A.L. Chervenak, R. Schuler, M. Ripeanu, M.A. Amer, S. Bharathi, I. Foster, A. Iamnitchi, and C. Kesselman, "The Globus Replica Location Service: Design and Experience," IEEE Trans. Parallel Distributed Systems, vol. 20, no. 9, pp. 1260-1272, Sept. 2009.

[4] H.H.E. AL-Mistarihi and C.H. Yong, "On Fairness, Optimizing Replica Selection in Data Grids," IEEE Trans. Parallel Distributed Systems, vol. 20, no. 8, pp. 1102-1111, Aug. 2009.

[5] M. Devarakonda and R. Iyer, "Predictability of Process Resource Usage: A Measurement- Based Study on Unix," IEEE Trans. Software Eng., vol. 15, no. 2, pp. 1579-1586, http://dx.doi.org/ 10.1109/32.58769, Dec. 1989.

[6] Jinoh Kim, A. Chandra and J. B.Weissman, "UsingDataAccessibilityforResourceSelectioninLarge-ScaleDistributed Systems ",Parallel and Distributed Systems, IEEE Trans. , vol. 20 , pp. 788 – 801, 2009.

[7] S.Vazhkudai. , S. Tuecke and I. Foster, "ReplicaselectionintheGlobusDataGrid", Cluster Computing and theGrid, 2001. Proceedings. First IEEE/ACM International Symp., pp. 106 - 113, 2001.

[8] M. Faerman, A. Su, R. Wolski, and F. Berman, "Adaptive Performance Prediction for Distributed Data-intensive Applications," Proc. ACM/IEEE Conf. Supercomputing (Supercomputing '99), p. 36, 1999.

[9] M. Wang, K. Au, A. Ailamaki, A. Brockwell, C. Faloutsos, and G.R. Ganger, "Storage Device Performance Prediction with Cart Models," Proc. IEEE CS 12th Ann. Int'l Symp. Modeling, Analysis, and Simulation of Computer and Telecomm. Systems (MASCOTS '04), pp. 588-595, 2004.

[10] R. Oldfield and D. Kotz, "Improving Data Access for Computational Grid Applications," Cluster Computing, vol. 9, no. 1, pp. 79-99, Jan. 2006.

[11] R. Oldfield and D. Kotz, " Armada: a parallel file system for computational grids ", Cluster Computing and the Grid, 2001. Proceedings. First IEEE/ACM International Symp., pp. 194 - 201 , 2001.

[12] L. Senger, R.F. Mello, M.J. Santana, and R.H.C. Santana, "An On-Line Approach for Classifying and Extracting Application Behavior on Linux," High Performance Computing: Paradigm and Infrastructure, pp. 381-401, John Wiley and Sons Inc., 2005.

[13] L.J. Senger, M. Santana, and R. Santana, "An Instance-based Learning Approach for Predicting Parallel Applications Execution Times," Proc. Third Int'l Information and Telecomm. Technologies Symp., pp. 9-15, Dec. 2005.

[14] Ning Wang & Xian-Yao Meng, "Time Series Prediction Using Self Organizing Fuzzy Neural Networks", Information, Computing and Telecommunication, 2009. IEEE Youth Conference, pp. 367-370, Sept.2009.

[15] O.B.Yaik, Chan Huah Yong, Haron F., Time Series Prediction Using Adaptive Association Rules" , Distributed Frameworks for Multimedia Applications, 2005. First International Conference, pp. 310-314, Feb 2005.

[16] A. Draghisi., A. Costan., V. Cristea, "Prediction of Distributed Systems State Based on Monitoring Data", Parallel and Distributed Computing (ISPDC) symp., pp. 173-180, July 2010 .