Network Traffic Prediction for Load Balancing in Cloud Access Point Controller

Zhifei Zhang, Shilei Cheng
School of Computer and Information Technology
Beijing Jiaotong University
Beijing, China
zhfzhang@bjtu.edu.cn

Jingpeng Tang, Abraham Teng
Department of Computer Science
Utah Valley University
Orem, Utah, USA
jtang@uvu.edu

Damian Lampl, Kendall Nygard
Department of Computer Science
North Dakota State University
Fargo, ND, USA
kendall.nygard@ndsu.edu

***Abstract*— In cluster cloud access controller (AC) solutions, load balancing algorithms typically consider the number of access points (APs), the number of users, the network traffic at the ACs, as well as central processing unit (CPU) and memory usage. However, because the network traffic has bursts and the user traffic on APs is unbalanced, it is not enough to consider only these factors. We report on the development of new traffic prediction models and their use in load balancing algorithms. The methods are evaluated with simulation experiments using MATLAB and CLOUDSIM. The methods utilize phase space reconstruction sequencing of the user network traffic. The result is improved load balancing efficiency when compared with alternative existing approaches.**

Keywords- cloud controller; wireless access point; load balancing; network traffic prediction.

I. INTRODUCTION

Traditional load balancing algorithms are usually divided into static load balancing algorithms [1] and dynamic load balancing algorithms [2] according to their strategy. Static load balancing algorithms do not consider the runtime operating state of each node in the cluster, but allocates the load of each node with a predetermined load balancing strategy based on its processing capacity. Dynamic load balancing algorithms monitor and collect information on the load of each node, such as CPU utilization, storage, memory and bandwidth utilization, in order to calculate the load balancing weight of each node in real time, and then distribute the traffic to corresponding nodes. Popular static load balancing algorithms include random balancing, polling, hash target address, and source address hash [3]. Dynamic load balancing algorithms include least connection, weighted least connection scheduling, weighted polling and minimum response time [4]. In addition to the classical algorithms, heuristic optimization algorithms, such as genetic algorithms [5], ant colony optimization [6][7], simulated annealing [8] and particle swarm optimization algorithms [9], are also used for scheduling problems.

In the conventional AC-AP architecture, AC is the most important equipment since it manages the AP's configuration, controls the station's access authentication and even forwards the station's packets centrally. With the progress of cloud technology, many companies have released Control and Provisioning of Wireless Access Points (CAPWAP) protocol-based cloud AC management systems. Examples include Ruckus, Relay2 and Google, and some domestic enterprises have also released cloud AC management systems. With the dramatic increase of AP, clustered cloud ACs (called AC pools) have become necessary to obtain performance requirements and unified resource management. Therefore, delivering an incoming AP to its proper cloud AC has become increasingly important in clustered cloud AC systems.

Our proposed load balancing strategy works as follows. First, the user traffic prediction model is set up and the user traffic is predicted based on its partially similar characteristics. Then, the incoming AP's load is predicted based on the user's traffic predictions. Finally, the incoming AP is distributed to a target cloud AC based on the AP's load predictions by the load balancer.

The rest of paper is organized as follows. Section II presents the wireless user traffic characteristics. Section III specifies the user traffic prediction model and its features. Section IV proposes the load balancing algorithm. Section V simulates and evaluates the performance of our proposed algorithm. The paper concludes with section VI.

II. RELATED WORK

Research on wireless user traffic typically involves obtaining its regular pattern through analysis of the user traffic data. It is known that the user traffic has self-similarity, periodicity [10] and burst characteristics [11]. Most studies on the characteristics of wireless user traffic take the statistical data for analyzing during a certain period

 ISBN: 978-1-61208-514-2

of time. In [12][13], based on the statistical result of the user number and user network traffic, it is proved that the user number and user network traffic of an AP has partial similarity with a period measured in days.

According to the CAPWAP protocol used in centralized forwarding mode, AP control packets and data packets are forwarded to the AC through a CAPWAP tunnel. In [14], it is proved that the CPU utilization and memory usage increases rapidly in centralized forwarding mode, while remaining almost constant in local forwarding mode when the number of AP and the number of users remains unchanged, and the user network traffic is increasing. Therefore, it indicates that the user traffic has no impact on network control messages between APs and ACs. Furthermore, the CPU utilization and memory usage increases in both the local forwarding mode and the centralized forwarding mode when the number of AP and user network traffic remains unchanged and the number of users is increased. As we know, there are many more data packets than management packets in centralized forwarding mode; load on cloud AC mainly depends on the user network traffic in centralized forwarding mode.

As shown in Figure (1), the traffic of 140 APs during 7 days while sampling 1440 time points per day illustrates that the traffic varies rapidly during a day but implies a period of day similarity.

Because network traffic has been shown to include attributes of self-similarity, periodicity and burst, a Poisson model is not suitable to describe the characteristics of the network traffic. Moving average models predict a result based on the historical average and require smaller storage space as well as less calculation compared to other models. Weighted moving average methods give different weights to history data, so the predicted value obtained may go awry since the user traffic is partial-similar data. An exponential smoothing model is a time-series forecasting method based on the moving average method and includes the single exponential smoothing, second-order exponential smoothing, and cubic exponential smoothing models. Exponential smoothing models are simple and practical with the potential to reach a high predictive accuracy in some cases.

Artificial neural networks are suitable for large-scale data because of their memorizing, calculating, learning and other intelligent features. Therefore, the neural network prediction method can be used to describe the non-linear characteristics of network traffic and exhibits better performance than autoregressive (AR), autoregressive moving average (ARMA) [15] and other linear network prediction methods. Specifically, since the neural network can remember the variation of network traffic during the training process, which results in less effect on the forecast value of the number of prediction steps, it is suitable for long-term prediction. The autoregressive integrated moving average (ARIMA) model for network traffic prediction gives a bigger error prediction deviation because of its multiple difference on non-stationary time series, which makes the characteristics of network traffic disappear. Furthermore, due to the long time required for its prediction algorithm, the ARIMA model cannot guarantee the required performance necessary for real-time network traffic forecasting. In this paper, we focus on wireless user network traffic prediction by the moving average model, exponential smoothing model [16] and back propagation (BP) neural network model [17], and select the suitable AC network traffic prediction model by comparing the prediction results of these three models.

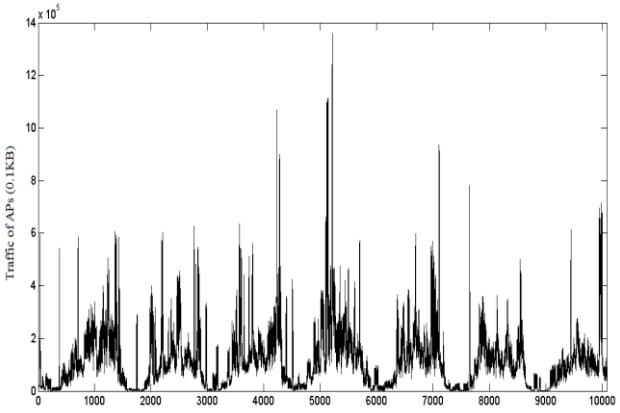

Figure 1. Traffic of 140 Aps for 7 days

In the centralized forwarding mode, the traffic will be forwarded to the AC after the AP successfully accesses the AC, which requires a load balancing strategy to not only consider the current load of the AC, but also take into account the long period load prediction. In consequence, we take the prediction as the AC's load after a long time prediction for user network traffic.

III. USER TRAFFIC PREDICTION MODEL

In this paper, due to the user network traffic periodic partial similar characteristics, the user network traffic time series phase space is reconstructed according to Takens' embedding theorem [18], and the reconstructed series are used to predict the user network traffic of the next day.

Given the user network traffic time series as:

$$X(i) = \{X(1), X(2), X(3) \cdots X(N * T)\}, T = 1440, \quad (1)$$

with the total sequence length being $N*T$, The reconstructed sequences of the same time each day:

$$Y_n(k) = \{X(n), X(n+T) \cdots X(n+k*T)\}, \\ k = 1,2 \cdots N-1, N > 1, \quad (2)$$

forms T reconstruction sequences set:

$$Y_n(k), n = 1,2, \cdots T. \quad (3)$$

In consequence, modeling the network traffic by secondary exponential smoothing as following:

$$S_n^{(1)}(N+1) = \alpha Y_n(N) + (1-\alpha) S_n^{(1)}(N), \quad (4)$$

$$S_n^{(2)}(N+1) = \alpha S_n^{(1)}(N+1) + (1-\alpha) S_n^{(2)}(N), \quad (5)$$

TABLE I. SQUARED ERROR OF DIFFERENT MODEL

Average AP Number per Day	Moving Average Model SSE(10^{15})	Second Exponential Model SSE(10^{15})	BP neural network model SSE(10^{15})
1	0.027	0.028	0.278
2	0.114	0.118	0.492
3	0.145	0.123	0.614
4	0.252	0.234	0.797
5	0.348	0.370	1.080
6	2.215	1.637	2.623
7	2.226	1.657	3.754
8	2.418	1.728	8.305
9	2.454	1.722	12.32
10	2.492	1.795	14.19

where n is the time index of a day and N is the day index, $S_n^{(1)}(N)$ is the single exponential smoothing value of N^{th} day n^{th} time slot, $S_n^{(2)}(N)$ is the second exponential smoothing value of N^{th} day n^{th} time, α is the Smoothness index where $0<\alpha<1$.

$$a_n(N+1) = 2S_n^{(1)}(N+1) - S_n^{(2)}(N+1), \quad (6)$$

$$b_n(N+1) = \frac{\alpha}{1-\alpha}\left[S_n^{(1)}(N+1) - S_n^{(2)}(N+1)\right], \quad (7)$$

where $a_n(N+1)$, $b_n(N+1)$ are intermediate variables. And the network traffic prediction of AC for the next day is:

$$F_n(N+1) = a_n(N+1) + b_n(N+1). \quad (8)$$

The user network traffic sequence $Y_n(k)$ based on the moving average, second exponential smoothing, and BP neural network models are modeled and predicted respectively. We take $n_{max} = 1440$ sample data per day and calculate the squared error for different models. Given the prediction value to be $F_n(k)$ and the real value to be $Y_n(k)$, the sum of squared error on the k^{th} day is [19]

$$SSE = \sum_{n=1}^{n_{max}} [F_n(k) - Y_n(k)]^2 . \quad (9)$$

The prediction results in Table (1) show that the moving average and second exponential smoothing models have similar sum of squared errors when the AP number is small.

TABLE II. SECOND EXPONENTIAL SMOOTHING MODEL WITH CONSTANT AND DYNAMIC EXPONENT

	Fixed Exponent	Dynamic Exponent
SSE(10^{13})	2.5674	1.0156

The second exponential smoothing model has the smallest sum of squared error among the three models and is therefore more suitable for wireless user network traffic prediction, although all of their sum of squared errors increase when the AP number increases. On the other hand, the exponential smoothing model takes less execution time than the BP neural network model, and the moving average model has the shortest execution time. After taking into account the practical application, the exponential smoothing models and moving average models both meet the time requirements for our AC-AP architecture. In conclusion, the second exponential smoothing model is used to predict the user network traffic when correcting the load balancing algorithm.

For the second exponential smoothing models, the smoothness index α is an important factor in algorithm prediction precision. In addition, the greater value of α results in a faster model prediction process. Usually, the smoothness index α is a constant determined by experience when the network traffic is predicted by second exponential smoothing models. However, it is very hard to set α by experience on the case of AC-AP architecture because the connection time and order is unknown, and the network traffic for each user on the AP is different. Therefore, in this paper, a dynamic exponential smoothing [20] model is used to dynamically adjust the smoothness index α in order to decrease prediction deviation.

The user network traffic in time nth of one day may vary sharply compared to its historical data because of the traffic's periodical and burst characteristics. In addition, the value of the exponent α is constantly adjusted based on the smallest network traffic deviation. Experiments for fixed index and dynamic index exponential smoothing models are shown in Table (2).

IV. THE PROPOSED ALGORITHM

As shown in Figure (2), there are m cloud ACs $AC_1, AC_2 \cdots AC_m$ and k APs $AP_1, AP_2 \cdots AP_k$. The AC load balancer relays an AP's packet to a cloud AC and vice versa. To simplify the algorithm, it is assumed that the each AP-AC tunnel has the same quality and the same configuration for every AP.

For cluster cloud ACs set:

$$AC = \{AC_1, AC_2 \cdots AC_m\}, \ m > 1. \quad (10)$$

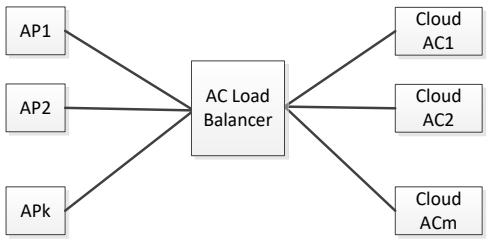

Figure 2. AC Load Balancing Model

Let AC nodes have the same configuration, $L(AC_i)$ represents the mean user network traffic of cloud AC_i. The sum of all cloud AC's mean user network traffic gives as:

$$L_{SUM} = \sum_{i=1}^{m} L(AC_i) \quad , \tag{11}$$

and $C_N(AC_i)$ represents the current AP number managed by cloud AC_i, therefore, the sum of AP managed by all cloud AC's gives as:

$$C_{NSUM} = \sum_{i=1}^{m} C_N(AC_i) \quad . \tag{12}$$

Let the weight coefficient be β, where 0<β<1. Given β=0.5 defines the weighted traffic load $T(AC_i)$ of AC_i as:

$$T(AC_i) = \beta \frac{C_N(AC_i)}{C_{NSUM}} + (1-\beta) \frac{L(AC_i)}{L_{SUM}} . \tag{13}$$

Finally, the minimum traffic load $T_{min}(AC)$ among cluster cloud AC set is:

$$T_{min}(AC) = \min\{T(AC_i), i = 1,2 \cdots m\} \tag{14}$$

V. SIMULATION AND EVALUATIONS

This section presents the simulation setup and evaluations using MATLAB and CLOUDSIM simulation tools. The wireless user network traffic prediction and the process for APs accessing ACs is simulated by MATLAB, while the process time of cloud AC is simulated by CLOUDSIM after APs access ACs successfully.

Euclidean distance is widely used in the sequence similarity research [21]. The Euclidean distance between sequence $X(k) = \{X_1, X_2 \cdots X_n\}$ and sequence $Y(k) = \{Y_1, Y_2 \cdots Y_n\}$ is given as:

$$D(X, Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \quad . \tag{15}$$

A. MATLAB Simulation

Based on 140 APs 14 days of experimental data, the process for AP accessing 3 ACs is simulated through the traffic prediction & AP number algorithm, AP number only algorithm, current traffic load algorithm, and current traffic & AP number algorithm respectively.

After all the AP access the 3 cloud ACs system successfully, the traffic sequence is taken from each AC for 1440 sample data points per day. The similarity of traffic sequence is indicated by their Euclidean distance $D(X, Y)$ which reflects the load balancing among cluster cloud ACs.

TABLE III. LOAD EUCLIDEAN DISTANCE FOR ALTERNATIVE ALGORITHMS

Euclidean Distance	Traffic Prediction & AP Number Algorithm	AP Number Algorithm	Current Traffic Load Algorithm	Current Traffic & AP Number Algorithm
$D(L_1, L_2)$ (10^6)	2.7477	2.8207	2.8409	2.8083
$D(L_2, L_3)$ (10^6)	2.8292	2.7969	2.7960	2.8272
$D(L_1, L_3)$ (10^6)	2.8603	2.8997	2.8251	2.8269
Summary (10^6)	8.4372	8.5173	8.4620	8.4625

Let L1, L2 and L3 be the load sequence of AC1, AC2 and AC3 respectively, the average Euclidean distance for a 100 times simulations is shown in Table (3).

A smaller Euclidean distance indicates a higher similarity. The sum Euclidean distance of the traffic prediction & AP number algorithm is 0.94%, 0.30% and 0.30% smaller compared to the AP number algorithm, current traffic load algorithm, and current traffic & AP number algorithm respectively. Therefore, the traffic prediction & AP number algorithm is more efficient for cluster cloud ACs load balancing algorithm.

B. CLOUDSIM Simulation

After all the AP access the 3 cloud ACs system successfully with the load balancing algorithm through the traffic prediction & AP number algorithm, AP number only algorithm, current traffic load algorithm, and current traffic & AP number algorithm respectively, the traffic sequence is taken from each AC for 1440 sample data points per day. The process time is simulated through CLOUDSIM and the Euclidean distance for the process time sequence is calculated to evaluate the different load balancing algorithms' efficiencies.

In this simulation, one broker and three virtual hosts with node id 1, 2 and 3 are created with simulation parameters as MIPS = 250, RAM = 512MB, bandwidth = 1000Mbps and the image space size =10000MB. The broker takes the load of AC obtained from the MATLAB simulation and delivers the traffic to different virtual host which represents the cloud AC.

Let T1, T2 and T3 be the process time sequence of AC1, AC2 and AC3 respectively. The average Euclidean distance for a 5 times simulations is illustrated in Table (4).

As before, a smaller Euclidean distance means higher similarity. The sum Euclidean distance of the traffic prediction & AP number algorithm is 1.8%、0.32%、0.17% smaller compared to the AP number algorithm, current traffic load algorithm, and current traffic & AP number algorithm respectively. Therefore, the process time simulation by CLOUDSIM gives a result similar to the load simulation of MATLAB, and the traffic prediction & AP

 ISBN: 978-1-61208-514-2

TABLE IV. PROCESS TIME EUCLIDEAN DISTANCE FOR DIFFERENT ALGORITHMS

Process Time Euclidean Distance	Traffic Prediction & AP Number Algorithm	AP Number Algorithm	Current Traffic Load Algorithm	Current Traffic & AP Number Algorithm
$D(T_1, T_2)$ (10^6)	2.0236	2.0856	2.0741	2.0473
$D(T_2, T_3)$ (10^6)	2.0547	2.0652	2.0434	2.0591
$D(T_1, T_3)$ (10^6)	2.0774	2.1206	2.0582	2.0597
Total (106)	6.1557	6.2714	6.1757	6.1661

number algorithm is more efficient for cluster cloud ACs load balancing.

To sum up, in the centralized forwarding mode, since the user network traffic in AC has similarity with a period of a day, it is feasible to predict the user traffic and take the prediction into account for cluster cloud ACs load balancing. The resultant traffic prediction & AP number algorithm gives a better result among cluster cloud ACs load balancing algorithms.

VI. CONCLUSIONS

In this work, the load balancing strategy is studied for APs accessing cluster cloud ACs. The similarity of wireless user network traffic is researched and a prediction algorithm is proposed to forecast the network traffic. The algorithm efficiency is compared with their load and process time Euclidean distance using simulation with MATLAB and CLOUDSIM. The simulations show that the second exponential smoothing model is more suitable for wireless user traffic prediction and the traffic prediction & AP number algorithm is more efficient among cluster cloud ACs load balancing algorithms.

REFERENCES

[1] Chonggun Kim, Kameda H. An algorithm for optimal static load balancing in distributed computer systems [J].Computers, IEEE Transactions on, 1995(41):381-384.

[2] Whang K Y, Kim SW, Wiederhold G. Dynamic Maintenance of Data Distribution for Selectivity Estimation [J].VLDB Journal, 1994, 3(1):29-51.

[3] Ben-Asher Y, Cohen A, Schuster A. The impact of task-length parameters on the performance of the random load-balancing algorithm[C]. Parallel Processing Symposium, 1992 Proceedings, Sixth International, 1992:82-85.

[4] Haddad E. Optimal dynamic redistribution of divisible load in distributed real-time systems[C]. Real-Time Applications, 1994 Proceedings of the IEEE Workshop on, 1994.21-26.

[5] J Gu, J Hu, T Zhao, and G Sun. A new resource scheduling strategy based on genetic algorithm in cloud computing environment [J] Journal of Computers, 2012, 7(1):42-52.

[6] Gupta E, Deshpande V. A Technique Based on Ant Colony Optimization for Load Balancing in Cloud Data Center[C]. Information Technology (ICIT), 2014 International Conference on, 2014:12-17.

[7] Xu Zhihong, Hou Xiangdan, Sun Jizhou. Ant algorithm-based task scheduling in gridcomputing[C]. Proceedings of the IEEE Canadian conference on electrical and computer engineering, 2003:1107-1110.

[8] Paletta M, Herrero P. An Awareness-Based Simulated Annealing Method to Cover Dynamic Load-Balancing in Collaborative Distributed Environments[C]. Web Intelligence and Intelligent Agent Technologies, 2009 WI-IAT 09 IEEE/WIC/ACM International Joint Conferences on, 2009:371-374.

[9] Zhao Yongyi, Xia Shengxian. Research on Load Balancing for Multidimensional Network Services Based on Particle Swarm Optimization Algorithm[C]. Intelligent Networks and Intelligent Systems (ICINIS), 2010 3rd International Conference on, 2010:411-414.

[10] Groschwitz N K, Polyzos G C. A time series model of long-term NSFNET backbone traffic[C], Communications, 1994. ICC'94, SUPERCOMM/ICC'94, Conference Record, 'Serving Humanity Through Communications.' IEEE International Conference on. IEEE, 1994: 1400-1404.

[11] Erramilli A, Singh R P, Pruthi P. Chaotic maps as models of packet traffic[C], Proc. 14th Int. Teletraffic Cong. 1994, 1: 329-338.

[12] McNett M, Voelker G M. Access and mobility of wireless PDA users [J]. ACM SIGMOBILE Mobile Computing and Communications Review, 2005, 9(2): 40-55.

[13] Balachandran A, Voelker G M, Bahl P. Characterizing user behavior and network performance in a public wireless LAN[C], ACM SIGMETRICS Performance Evaluation Review. ACM, 2002, 30(1): 195-205.

[14] Songqun Huo, Research on the networking and forwarding technologies in WLAN [D], Beijing University of Posts and Telecommunications, 2010:45-48.

[15] Xue Ke, Li Zengzhi, Liu Liu and Song Chengquan, Network traffic prediction based on ARIMA model [J]. Microelectronics and Computer, 2004, 21(7):84-87.

[16] LIU Y, JIN X. Research on Network Traffic Prediction-based Dynamic Exponential Smoothing Model [J]. Fire Control and Command Control, 2008, 3: 029.

[17] Li Z, Qin L, Xue K, et al. A Novel BP Neural Network Model for Traffic Prediction of Next Generation Network[C], Natural Computation, 2009. ICNC '09. Fifth International Conference on. IEEE, 2009:32-38.

[18] Lu Jinjun, Wang Zhiquan, Internet traffic data follow forecast by RBF neural network based on phase space reconstruction, Transactions of Nanjing University of Aeronautics & Astronautics [J], 2006, 23(4):316-322.

[19] Hecht-Nielson R. Theory of the back propagation neural networks [M]. Washington D. C. Proceedings of IEEE international Joint conference on Neural Networks. 1989.

[20] Bonsdorff H. A comparison of the ordinary and a varying parameter exponential smoothing [J]. Journal of Applied Probability, 1989.784-792.

[21] Keogh E. Fast similarity search in the presence of longitudinal scaling in time series databases[C], Tools with Artificial Intelligence, 1997. Proceedings, Ninth IEEE International Conference on. IEEE, 1997: 578-58.

 ISBN: 978-1-61208-514-2