RESEARCH ARTICLE OPEN ACCESS

Enhancement of Single Moving Average Time Series Model Using Rough k-Means for Prediction of Network Traffic.

Theyazn H. H. Aldhyani *, Manish R. Joshi**
* *School of Computer Sciences North Maharashtra University, Jalgaon , India*
Email: *th0ha0@yahoo.com*
** (*School of Computer Sciences North Maharashtra University, Jalgaon , India*
Email: *joshmanish@gmail.com*

ABSTRACT
In the last decade, real–time audio and video services have gained much popularity, and now occupying a large portion of the total network traffic in the Internet. As the real-time services are becoming mainstream the demand for Quality of Service (QoS) is greater than ever before. It is necessary to use the network resources to the fullest, to satisfy the increasing demand for QoS. To solve this issue, we need to apply a prediction model for network traffic, on the basis of network management such as congestion control and bandwidth location.
In this paper, we propose an integrated model that combines Rough K-Means (RKM) clustering with Single Moving Average (SMA) time series model to improve prediction loading packets of network traffic. The single moving average time series prediction model is used to predict loading of packets volume in real network traffic. Further, clustering granules obtained by using rough k-means is used to analyze the network data of each year separately. The proposed model is an integration of the prediction results that were obtained from conventional single moving average prediction model with centriods of clusters that obtained from rough k-means clustering. The model is evaluated using on line network traffic data that has been collected from WIDE backbone Network MSE, RMSE and MAPE metrics are used to examine the results of the integrated model. The experimental results show that the integrated model can be an effective way to improve prediction accuracy achieved with the help of rough k-means clustering. A Comparative result between conventional prediction model and our integrated model is presented.
***Keywords*:** Time series model, prediction loading packets of network traffic, RKM, SMA

I. INTRODUCTION

With growing scientific data, e-commerce, banking and business to unprecedented volumes and the requirement to share such huge amounts of data by increasing numbers of geographically distributed collaborators, the Quality of Service (QoS) is necessitated for efficient data access. Network traffic prediction and forecasting are the significant issues for a given time stamp with same estimated tolerance enables better network data flow routing and transfers that are chiefly important for big data movements, which can be found in Internet of Things (IoT). Only one way to improve the impact of traffic growth is prediction m network traffic, for planning and designing stable and flexible networks. Network administrators, network engineers, and operators require accurate, reliable, and rapid network modeling tools to assist model traffic trends, to visualize traffic growth, and to decide when optimization and upgrades networks are necessary.

Furthermore, performance monitoring and predicting is an active area of research. In the rising world of networking, more stress is being given on speed, connectivity and reliability.

Furthermore, with the increase in the use of the Internet and more wide networks are expanded. And therefore, the network architecture is gradually getting complicated and network congestion and emergence are increasing. Due to the heterogeneity and the constantly varying nature of the network traffic, there are only a few works are available to provide prediction of the network performance. Network modeling and prediction has a fundamental part for plan and designing the network for increasing the performance of network.

In this paper single moving average time series prediction model is applied. Furthermore, we propose to use rough k-means technique to enhance conventional single moving average time series prediction model for network traffic prediction. The real network traffic is used to test the existing time series model and our integrated model. Evaluation and comparison between conventional prediction model against our proposed model is presented.

The rest of paper is organized as follows. Section 2 discusses related work. The description of time series model is presented in section 3. Section 4 describes integrated model. The experimental

analysis followed by results analysis is shown in section 5. Finally, section 6 concludes the paper.

II. RELATED WORK

Network traffic prediction is an important issue that has received much interest recently from computer network community. The network traffic of prediction is one of the typical issues useful for monitoring network, network security, avoid congestion and increase speed of networks. Different techniques are used by researchers for network traffic prediction. In the following sections we will discuss some of the works done so far, which has motivated this work.

Jian et al. [7] proposed hybrid model two dimensional corrections and Single Exponential Smoothing (SES) for prediction the mobile network. They collected data from Xinjiang mobile company on different time interval days and hours. Their experiments shows that their model has performed well, a comparative prediction between their model and tradition model is presented. They observed that their mode is giving prediction efficiency and the accuracy. Daniel et al. [8] proposed Simple Moving Average (SMA) and Exponential Moving Average (EMA)
for prediction network traffic. They collected real network traffic from web services to test their model. Their attention was test if these models can help to forecast long-term prediction. From the experiments, they observed these models are not appropriated for long-term prediction. Cortez et al. [9] proposed Neural Network Ensemble (NNE) ARIMA and Holt- Winter approaches with time series model for predicting Internet traffic. Authors collected data from Internet service provider based on TCP/IP protocol. They applied their techniques to data obtained at different intervals of time such as five minute, one hour and one day. They used MAPE metric to measure prediction performance of their techniques. A comparative prediction result between NNE, ARIMA and Holt-Winter methods is presented. They shared that the Holt-Winter approach is better to predict network at one day time scale but the ENN and ARIMA approaches achieved better prediction at five minute and one hour time interval but ARMA is very complex than NNE. And also they noted that the NNE is more accurate overall.

Ng Kar et al. [10] introduced ARMA time series model for predicting large file transfer of network traffic. Authors experimented with File Transfer Protocol (FTP) data set. The network protocol analyzer (wireshark) is applied to capture and collect packets at different time intervals. They applied necessary transformation (log return) for computing the number of packet while capturing data. They transferred 1GB file and 10 MB file through network. Then applied ARMA approach for forecasting different files at different time interval. They concluded that transfer of small file is better than large file. And also, they recommended breaking a large file into several small size files. Transmitting smaller files lead
to use of little bandwidth. They suggested that the most appropriate model to be used is the ARMA different size of file. Poo et al. [11] proposed a time series model with ARMA approach to predict network traffic pattern of bit torrent application. They evaluated ARMA technique, collected 6 sets of real data at a period of six days from bit Torrent Point to Point Network (P2PN) by using wireshark software. They developed a new simulation tool called as Simple Network Prediction Studio (SNPS). This simulation used ARMA model to predict cyclical and seasonal bit torrent network traffic patterns. The Mean Square Errors (MSE) metric applied to measure and evaluates performance of ARMA model on cyclical and seasonal patterns. A comparative prediction result of ARMA model at cyclical and seasonal patterns of bit torrent is presented. They shared that the ARMA approach achieved better prediction in cyclical pattern.

III. TIME SERIES MODEL

A time series is a sequence of set data points, measured typically over successive times and occurring in time intervals. The time series is mathematically defined as a set of vectors $x(t)$, $t = 0, 1, 2, ...$etc where t represents the time periods beyond. The $x(t)$ referred as a random variable. Moreover, the measurements taken through an event in time series are in order of a proper sequential order. The univariate time series contains the records of a single variable. However, where the records contains more than one variable is termed as multivariate. As well as a time series can be continuous or discrete [12]. Last few decades, the time series models are a one of most important research area to which researchers are paying attention for developing new model or enhancing existing time series models. Collecting the past observations of a time series to develop an suitable model which illustrate the inherent structure of the series is major aim of the time series models. Generating future values from the past observation is time series forecasting. Thus, time series forecasting is a subset of prediction for obtaining future observation. The time series is applied in much important demand such as business, science and engineering, etc. Obviously, that a successful time series forecasting future values depends on suitable model that fitting. A lot of attempts have been done by researchers and experts through many years for development of efficient models to enhance the prediction accuracy. There are numerous significant

time series prediction models, we will show in literature.

IV. INTEGRATED MODEL

Fig 1 displays the integrated model of SMA model with rough k-means clustering. Firstly, it is necessary to get the extraction of data traffic, available at WIDE data repository. The logarithm method is used for normalizing network data. The SMA model is a very simple model, SMA model depends on the average size. It is observed that the average size of 2 has less prediction error hence, we decided to use it. The focus is in the centriods of clustering numbers that are obtained from clustering approach for improving conventional model. The prediction which is obtained from the SMA model is integrated with centriods of cluster number that are obtained from RKM clustering. Performance measures are employed to test results of conventional and our integrated model. The enhanced prediction is a function of existing prediction value and appropriate cluster centriod value as follows. $EP_i = f(P_i, C_j)$ where P_i is a prediction value that is obtained by conventional SMA model. C_j is the centriod of the cluster to which the i^{th} sample belongs to. Finally, it is observed that the results of integrated model by using performance measures. The results show that the integrated model improves the existing prediction model.

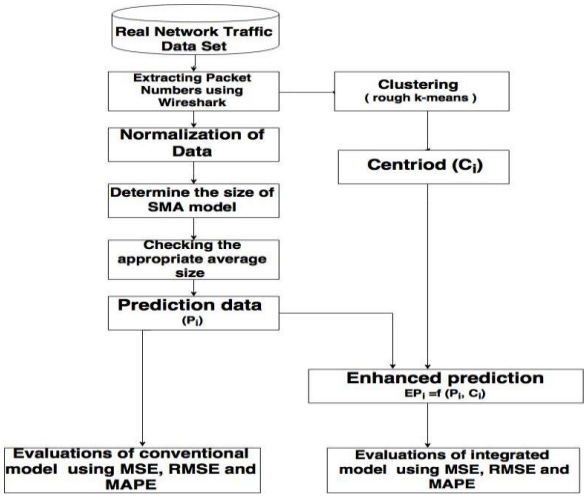

Fig1: Integrated model

A. Data Set

The traffic data to be predicted is from a backbone network with 150 Mbps link. The Measurement and Analysis on the WIDE Internet (MAWI) traffic repository archives have the traffic data collected from the WIDE backbone. The WIDE backbone network repository is maintained by the MAWI working group networks. The WIDE network (AS2500) is a Japanese academic network connecting universities and research institutes. These data daily trace at the transit link of WIDE (150 Mbps) to the upstream ISP. For our analysis, two years 2013 and 2014 data has been used, which gets aggregated every one hour. Packets loading from network traffic have been captured. We extracted the numbers of packets using wireshark tool [6].

B. Normalization

Most data sets are scaled through the data preprocessing phase. The two main advantages of scaling are to avoid instances in greater numeric ranges from dominating those in smaller numeric ranges, and to prevent numerical difficulties during the prediction model. Natural logarithm is employed for scaling our data. This method transformed data in the range of 14 to 21 scales.. Fig2 displays two years original data after normalization.

C. Performance Measures

To evaluate the prediction model three performance measures were used. Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are applied as performance indices. These standard performance measures methods are defined as follows.

$$MSE = \frac{1}{N} \sum_{k=1}^{n} (x_t - \overline{x_t})^2 \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{k=1}^{n} (x_t - \overline{x_t})^2} \quad (2)$$

$$MAPE = \frac{\sum |(x_t - \overline{x_t})/x_t|}{n} * 100 \quad (3)$$

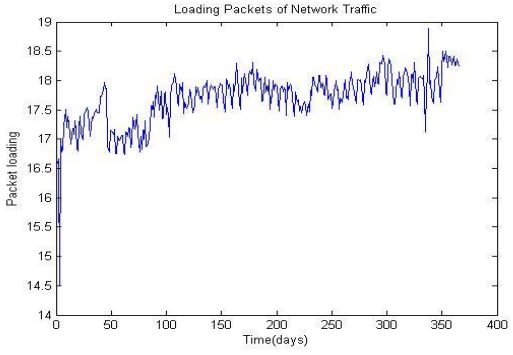

Fig2: Normalization data of year (2013)

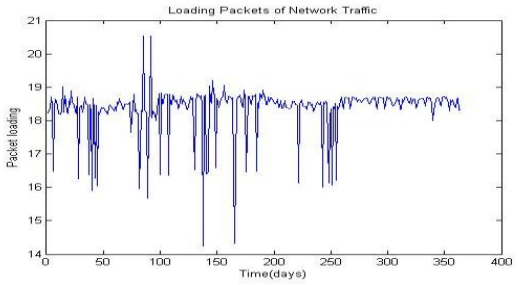

Fig3: Normalization data of year (2014)

D. Single Moving Average Model

The Single Moving Average model (SMA) is a simple model for predicting network traffic. It requires little computation. It acts as computing an average of the most recent *n* data values for the series and using this average for forecasting the value of the time series for the next period. The SMA model is the mean of the previous *n* data sets. The formula of Single Moving Average is as follows:

$$M_t = \frac{X_t + X_{t-1} + \cdots + X_{t-N+1}}{N}$$

E. Rough K-means Approach (RKM)

The notion of rough set was presented by Pawlak [13], [14]. The rough k-means is an updated version of k-means algorithm. Rough k-means is mathematical tool employed to deal ambiguity objects. The rough k-means use to represent the cluster as lower approximation and upper approximation. The rough k-means is an updated version of k-means algorithm. Rough k-means is mathematical tool employed to deal ambiguity objects. The rough k-means use to represent the cluster as lower approximation and upper approximation. We represent each cluster c_i, $1 \le i \le k$, using its lower A (c_i) and upper A (c_i) approximations. All objects that are clustered using the algorithm follow basic properties of rough set theory such as:(P1) An object $\vec{x}$ can be part of a lower approximation of at most one cluster

(P2) $\vec{x} \in A(\vec{c}_i) \Rightarrow \vec{c} \in A(\vec{c}_i)$

(P3) An object $\vec{c}$ is not part of any lower approximation

$\Updownarrow$

$\vec{x}$ belongs to upper approximation of two or more clusters. An object is assigned to lower and/or upper approximation of one or more clusters. For each object vector, $\vec{v}$ let d ($\vec{v}$, $\vec{c}_j$) be the distance between itself and the centroid of cluster $\vec{c}_j$. Let d ($\vec{v}$, $\vec{c}_i$) = min $1 \le j \le k$ d ($\vec{v}$, $\vec{c}_j$). The ratios d ($\vec{v}$, $\vec{c}_j$)/d($\vec{v}$, $\vec{c}_j$), $1 \le I, j \le k$, are used to determine the membership of $\vec{v}$. Let T = { j : d ($\vec{v}$, $\vec{c}_j$)/d ($\vec{v}$, $\vec{c}_j$) $\ge$ threshold and $i \ne j$ }

1. If T = ϕ , $\vec{v} \in$ A ($\vec{c}_j$) and $\vec{v} \in$ A ($\vec{c}_j$), $\forall j \in$ T. Furthermore, $\vec{v}$ is not part of any lower approximation? The above criterion guarantees that property (P3) is satisfied.
2. Otherwise, if T = ϕ , $\vec{v} \in$ A ($\vec{c}_j$). In addition, by property (P2), $\vec{v} \in$ A ($\vec{c}_j$).

V. EXPERIMENTAL ANALYSIS

We evaluated the integrated model on the real network data for predicting loading packets of network traffic. The proposed model is implemented in Matlab. MSE, RMSE and MAPE performance measures are used to measures the prediction performance of conventional single moving average time series model and the integrated model. The results obtained from conventional time series model and our integrated model is as follows:

A. Results of Single Moving Average Model

After the process of normalizing data, single moving average is implemented. SMA model, an alternative way to the simple average (which takes all past samples into account), is to compute the mean of smaller sets of past data. However, SMA model averaging process is continued by advancing one period and calculating the next average. In other words, it is essential to determine the size of average so that, it can help to obtain prediction for specifying data point. We calculate moving averages with different sizes 2,3,4,5 and 7. The moving averages with different sizes of the average size are examined using MSE performance measure. We observed that the size of moving average 2 gave fewer errors. Table I shows the results obtained from Single Moving Average model when the average size is 2. The results show that the SMA model gives good results. Fig 4 and 5 demonstrates the performance of single moving average model.

Table I: Results of SMA model

Performance measures	2013	2014
MSE	0.0346	0.2970
RMSE	0.1859	0.5450
MAPE	0.7584	1.6350

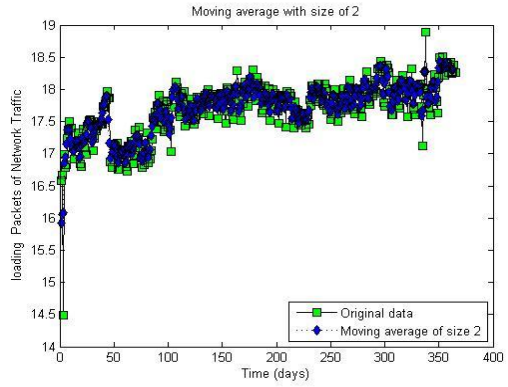

Fig4. Prediction performance of SMA model of year (2013)

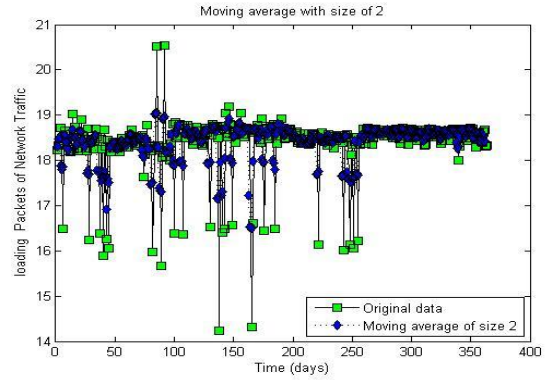

Fig5. Prediction performance of SMA model of year (2014)

B. Results of Integrated Model with Rough K-Means Clustering

The non-crisp rough k-means approach is applied for enhancing conventional time series models. Rough k-means approach clusters the objects into 5 clusters some objects belong to upper approximation and some belong to lower approximation. The centriods of lower approximation are selected. However, the objects that belong to upper approximation are judged to be ambiguous objects that belong to two or more cluster numbers. These ambiguous objects are addressed by taking average of centriods of those clusters to which objects belongs. The centriods are obtained from rough k-means clustering integrated with the prediction results that are obtained from the conventional SMA time series model. This method is based on the computing of the centriods of clustering with prediction that are obtained from prediction model. Tables II clearly show results that are obtained from integrating the SMA conventional model with rough k-means approach. The Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) performance measures have been selected to evaluate the performance of integrated model. The predicted results show that integrated model can achieve higher accuracy than those obtained by existing SMA time series model. Network data has a lot of variation and fluctuation and it is complex in nature. It is examined that the rough k-means can handle ambiguity that reduces the accuracy of network traffic prediction. Fig6 and 7 illustrate graphically prediction performance of integrating SMA conventional time series model and our integrated model.

Table II: Results of integrated model

Performance measures	**2013**	**2014**	**Average improvement (%)**
MSE	0.0192	0.0841	**58.09**
Improvement (%)	44.50	**71.68**	
RMSE	0.1385	0.2899	**36.14**
Improvement (%)	25.49	46.80	
MAPE	0.4545	0.9332	**41.49**
Improvement (%)	40.07	42.92	

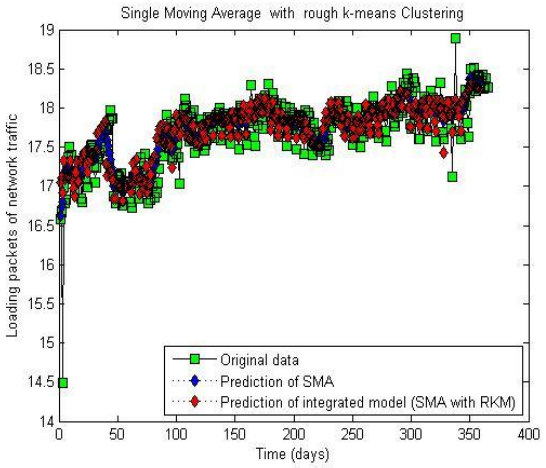

Fig6. Prediction performance of integrated model of year (2013)

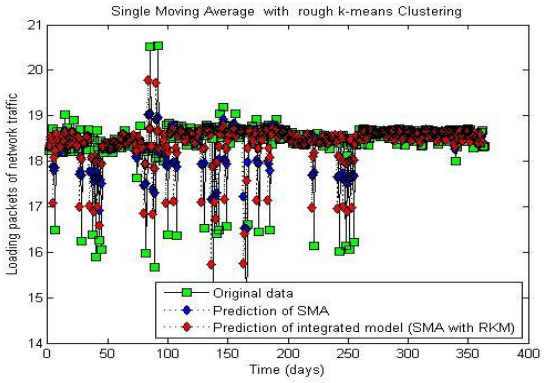

Fig7. Prediction performance of integrated model of year (2014)

VI. CONCLUSION

The ability to estimate bandwidth requirement is important for efficient service supplying and intelligent decision making to avoid growing traffic data and changing traffic patterns. Thus, in order to enhance time series prediction models for predicting number of users, the number of hosts and total Internet traffic in future , has become attention of many of the network organizations, companies and academics. We proposed to integrate clustering with existing SMA time series model. The proposed model is an integrating conventional SMA model with non crisp RKM clustering approach is elaborated. The evaluation as well as the comparison of conventional model with proposed model is presented.

The novelty of the integrated model is the use of applied rough k-means approach for enhancing existing SMA time series model in network traffic prediction. The RKM clustering is proposed for improving the prediction of loading packets in network traffic. The average improvement is 58.09, 36.14, 41.49 % according to MSE, RMSE and MAPE measurements and graphical representation of the performance improvement as results of the proposed model can be observed.

We concluded that the integrated model can help the network engineers for planning and designing network with respect to the Quality of Services (QoS) of network traffic.

REFERENCES

[1] T. H.H.Aldhyani and M. R.Joshi "A Review of Network Traffic Analysis and Prediction Techniques" arxiv.org, 2015,pp 1-24.

[2] T.H.H.Aldhyani, M. R.Joshi "Clustering to Enhance Network Traffic Forecasting" Springer, 2016.

[3] T. H. Hadi and M. Joshi, "Handling ambiguous packets in intrusion detection," in Signal Processing, Communication and Networking (ICSCN), 2015 3rd International Conference. IEEE, 2015, pp. 1–7.

[4] T. H.H.Aldhyani, M. R.Joshi, "An integrated model for predictionof loading packets in network traffic," in Second International Conference on Information and Communication Technology for Competitive Strategies. ACM, 2016.

[5] Packet traces from wide backbone," in http://mawi.wide.ad.jp/mawi/.

[6] Jian Kuang, Dongwei Zhai, Xinyu Wu, and Yanwen Wang. A network traffic prediction method using two-dimensional correlation and single exponential smoothing. In *Communication Technology (ICCT), International Conference*, pages 403–406. IEEE, November 2013.

[7] Daniel W. Yoas and Greg Simco. Using long-term prediction for web service network traffic loads. In *International Conference on Information Technology*, pages 21–26. IEEE, 2014.

[8] Cortez P., Rio M., Rocha M., and Sousa P. Internet traffic forecasting using neural networks. In *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, pages 2635–2642. IEEE, 2006.

[9] Ng Kar Hoong, Poo Kuan Hoong, Tan I. K. T, and Muthuvelu. Impact of utilizing forecasted network traffic for data transfers. In *Advanced Communication Technology (ICACT), 13th International Conference*, pages 1199–1204. IEEE, February 2011.

[10] Poo KuanHoong, Ian K. T., and Tan CheeYikKeong. Bittorrent network traffic forecasting with arima. *International Journal of Computer Networks and Communications (IJCNC)*, (4):143–146, 2012.

[11] Shen Fu Ke, Shanghai China, Zhang wei, and Chang Pan. An engineering approach to prediction of network traffic based on time-series model. In *Artificial Intelligence, International Joint Conference on*, pages 432–435. IEEE, 2009.

[12] Zdzis law Pawlak. Rough sets. *International Journal of Computer and Information Sciences*, 11(5):341–356, October 1982.

[13] Manish R. Joshi, Pawan Lingras, and C. Raghavendra Rao. Analysis of rough and fuzzy clustering. In *Rough Sets and Knowledge Technology: 7th International Conference,*, page 679–686. Springer, December 2010.

[14] Yimu J., Yonggea Y., Chuanxine Z., Chenchena J., and RuChuana W. Research of a novel flash p2p network traffic prediction algorithm. Page 1293–1301. Elsevier, 2015.

[15] B. Yang and M. Jiang. Ensemble of flexible neural tree and ordinary differential equations for small-time scale network traffic prediction.
International Journal of Computer, vol. 8, no. 12, pp. 3039–3046, Dec. 2013.

[16] P. J., R. S. M., T. P., and K. D. S .Network traffic prediction model based on training data. in *Computational Science and Its Applications,*
International Conference. Springer, 2015, pp. 117–127.

[17] Yonglin Liang and Lirong Qiu. Network traffic prediction based on SVR improved by chaos theory and ant colony optimization. *International Journal of Future Generation*

Communication and Networking, 8(1):69–78, 2015.

[18] T. H. Hadi and M. Joshi, "Handling ambiguous packets in intrusion detection," in *Signal Processing, Communication and Networking (ICSCN), 2015 3rd International Conference*. IEEE, 2015, pp. 1–7.

[19] M.-S. Han, "Dynamic bandwidth allocation with high utilization for xg-pon," in *International Conference on Advanced Communication Technology, 16th Conference on*. IEEE, 2014, p. 994997.