<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>A Forecasting Model for Data Centre Bandwidth Utilisation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Authors(s)</strong></td>
<td>Talpur, Samar Raza; Kechadi, Tahar</td>
</tr>
<tr>
<td><strong>Publication date</strong></td>
<td>2016-09-22</td>
</tr>
<tr>
<td><strong>Conference details</strong></td>
<td>The SAI Intelligent Systems Conference (IntelliSys 2016), CentrEd at ExCel London, London, United Kingdom, 21-22 September 2016</td>
</tr>
<tr>
<td><strong>Link to online version</strong></td>
<td><a href="http://saiconference.com/Conferences/IntelliSys2016">http://saiconference.com/Conferences/IntelliSys2016</a></td>
</tr>
<tr>
<td><strong>Item record/more information</strong></td>
<td><a href="http://hdl.handle.net/10197/8368">http://hdl.handle.net/10197/8368</a></td>
</tr>
</tbody>
</table>

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)

Some rights reserved. For more information, please see the item record link above.
A Forecasting Model for Data Center Bandwidth Utilization

Samar Raza Talpur
School of Computer Science
University College Dublin
Belfield Dublin, Ireland
Email: samar.talpur@ucdconnect.ie

Prof. Tahar Kechadi
School of Computer Science
University College Dublin
Belfield Dublin, Ireland
Email: tahar.kechadi@ucd.ie

Abstract—Bandwidth optimization and its efficient utilization is more challenging in operating data centers. Our model can assist for proper usage of resource utilization and accommodate large scale of bursty data. In this paper we propose forecast model for Data Center Bandwidth Utilization system; a forecast model for data centers to predict and estimate proper bandwidth utilization in real-world situations. Based on self-learning procedures, the proposed forecasting model will optimize the traffic and predict bandwidth more efficiently. Our approach is based on Time Series and Vector Autoregression (VAR-Model) models, it optimizes the bandwidth traffic detecting and diagnosing the future based on historical data.

Keywords—Bandwidth Utilization, Network Forecasting, Time Series, ARIMA Models, Exponential Smoothing, VAR Model.

I. INTRODUCTION

While Data Centers are growing in use and popularity, the business communities are also more concerned about speed, reliability and better performance of networks. A typical Data Center uses the shared network resources. The network performance is affected when the bandwidth consumption increases. When the demand is above a certain limit of its total capacity, the whole network becomes unstable. Network bandwidth importance is very valuable when it comes to business use, especially for Business-to-Business (B2B) community, as they heavily rely on healthier internet connectivity; it is a prerequisite for all users that are connected with each other and need a good speed level of data exchange. Certainly they need fast internet and higher bandwidth and more importantly when using data center resources involving significant data volume. Global network traffic evolved with an impressive usage, in relation to the foregoing observations of the last two decades. In 1992 the total Internet traffic per day was 100 GB, soon after one decade later in 2002 it jumped at 100 GB per second, even as in 2014 it arrived at 16,144 GB per second, whereas the prediction for 2019 is expected to be more than 51,974 GB per second [21], [5], [29]. Motivated by system modeling, Network Bandwidth Predictor (NBP) [7] used a neural-network based approach to predict the bandwidth usage and network performance. NBP combined with NWS (Network Weather Service) is developed for observing subsystems and measuring the network behavior. GLMM [15] presented for high speed in data network analysis, a new predictive model which analyzes the traffic patterns and conditions of several networks. The simulation based study claims high accuracy in MSPE (Mean Squared Predication Error). Moussas [28] also used ARIMA model for traffic monitoring to predict the future traffic utilization, by using four different components (a) Campus networks (b) dialup connections (c) Local Area Network and (d) backbone. Network Bandwidth Utilization Forecast Model on High Bandwidth Network [38] uses time series model and claims efficient resources of network utilization. The results showed that it can reduce computational time by 83.2% compared to other traditional approaches. This model can help scheduling data movements on high bandwidth networks. However, if the planning about network usage is known in advance, the performance can be improved significantly. The challenge, here is how to build a forecast model for real-world traffic in data centers. This will help not only improving the performance of both data centers and the networks but also help managing, planning and upgrading both the network and data center resources. Considering all the above mentioned models neither are specifically designed for data centers nor they all have real network traffic of data centers. Our model is entirely different and exclusively designed for data centers, we used absolute genuine traffic to experiment with and apply the same for monitoring and prediction. In order to improve traffic accuracy in data centers, we developed Data Center Bandwidth Utilization forecasting model. Statistical models observed high volume of real traffic and can predict accurately network performance which can be helpful for a good network administrator. With the growth of data centers, the big number of data center service providers have thousands of servers and other equipment installed, doubling all these equipment every 18 months and proving the prediction of Moore’s law [26]. Once we look at the running cost, millions of dollars were spent on diversified hardware, complex workload and thousands of various applications. Despite of all the conventional data centers neither provide proper access for public trace nor real time monitoring system for researchers. Linear regression, using Granger causality and VAR model (Eviews tools) forecasting is also a nicer option for predicting the bandwidth utilization [21], [11].

II. METHODOLOGY

A. Design of Data Center

Among many data center architectures, we opt for a typical three-tier architecture [24] which contains layer 2 to 4 devices such as core routers, aggregation and access switches. The hierarchies of networks are connected with each other from higher layers to lower layers. The distribution and aggregation
layer switches at higher layers are directly connected with the outside traffic (the Internet), while edge layer switches are connected with servers and end-user computers.

We also fixed a dedicated database server, which operated constantly (24 x 7) as a traffic capturing server for collecting network traffic and network behavior. As shown in figure, we collected different types of data from both networks i.e. public and private networks. The secured network, treated as private network, are connected with firewall via a number of restricted policies for In-Out traffic. On the other-hand unrestricted networks bypass the firewall and open for all users, these networks are public.

![Fig. 1: Standard Topology of Data Center](image)

**B. Traffic Generation**

More than 150 computers are running absolutely (24x7) to generate diversified traffic within a particular interval of time. The captured traffic (In-Out) from different LANs and WANs (public and private) is monitored and stored in database server. Database server constantly running and storing the traffic without any interruption and downtime.

![Fig. 2: Captured Traffic of 60 Minutes](image)

![Fig. 3: Captured Traffic of 24 Hours](image)

![Fig. 4: Captured Traffic of 30 Days](image)

All generated network traffic of LANs and WANs is divided into different categories of time and intervals i.e; minutes, hours and days. For example, Figure-2 shows 60 minutes of generated traffic with 30 second interval, and Figure-3 shows 24 hours generated traffic and captured with average of 5 minutes’ intervals, while the Figure-4 displays 30 days of traffic captured at time intervals of 60 minutes. The purpose of this exercise is to demonstrate different granularities of time slots with standard refresh time of intervals (i.e. minutes, hours and days) so that selected sample of small and big units to update the real time traffic and to cope with more accuracy and better performance. In future we plan to utilize these data sets to deal with latency and jitter of traffic.

**C. Data Collection**

The fundamental reason for capturing the real-time traffic is to monitor the whole network, and fulfill the upcoming demand of bandwidth and installation of network devices. We captured the traffic on routers’ ports with the following factors. (Note that the real-time traffic was taken from June 2014 to June 2015). However routers can have multiple types of ports and connectors, but we connect fast Ethernet port of routers and collect the traffic in raw format. So far we clean the data (raw to csv) for popular format for slicing and visualization, by going through all the process of data mining and modeling. Subsequently the traffic presents in shape with In-Out in kilo bits per second and Kilobytes per second on initial stage.
D. Traffic Observation

During the experiments, we try to monitor the In-Out traffic on on Router’s port1 (FastEthernet0/0) (Bandwidth in both KBytes/s and kb/s). All the traffic data is captured and stored in database server and it is directly connected to a router on fast Ethernet port. Using Wireshark software tool we captured all the incoming and outgoing traffic. The traffic was recorded for various time intervals: minutes, hours and days format from networks to the Internet. The reasons for capturing diversified traffic of various variations to obtain more precise results.

III. Experimental Results

ARIMA(1,0,1): This is the combination of autoregressive (AR) model and Moving Average (MA) models with integration of order zero. Let T(In) be considered for In traffic which measured in kb/s and ε can be uses normalisation of data. It can be jotted down as follows:

\[ T(In)_t = \mu + \alpha T(In)_{t-1} + \gamma \epsilon_{t-1} + \epsilon_t \]  

Where \( \alpha \), non-zero is the mean or drift component and \( \epsilon_t \) is the disturbance term that is normally distributed. The \( \alpha \) and \( \gamma \) are the coefficients of the lag terms T(In) and Errors respectively.

In the similar fashion, model can be constructed for the Tout traffic follows:

\[ T(Out)_t = \mu + \alpha T(Out)_{t-1} + \gamma \epsilon_{t-1} + \epsilon_t \]  

Focusing on various results was an additional motivation to work with different models, based on multiple sources and previous data we got the deliverable outcomes. Our productive outcomes help the analytical results for decision making. Subsequently endeavoring couple of models and compare the difference with one another to optimize the process and precise the network bandwidth. To predict the futuristic movement the time series (ARIMA) and VAR models were best options respectively [40].

R-Language has built-in facilities for analyzing time series data, using "auto.arima" function and passing multiple arguments easily produce some basic forecasting results. Linear filtering of time series \((X, T, S \text{ and } \epsilon)\) are major factors, where \(X=\) time series, \(T=\text{trend}, \ S=\text{seasonal component and} \ \epsilon=\text{remainder components.} \ "ts" \ function converts numeric vector in to R, first we convert one variable at a time into time series object by using "ts" function. The typical format of "ts" has four standard object for observations i.e. vector, start, end and frequency, example given here;

\[ \text{port.1.365.days.TIN.ts} = \text{ts(port.1.365.daysTraffic}_1 N, \text{start} = c(2014, \text{as.POSIXlt}("2014 - 05 - 28")yday), \text{frequency} = 365) \text{plot.ts(port.1.365.days.TIN.ts)} \]

However available filtering analysis used in weeks, months and quarters i.e. (a=2, a=12, and a=40) respectively [40].

![Fig. 5: correlation between In and Out traffic](image)

We can see in above figure a visual representation of the original data, red lines represent In data and green lines Out data. As shown in the same figure the quarterly visual representations between two variables (In-Out). Logarithm presents in four equal quarters (July 2014-April 2015), we assume maximum variation in third quarter jumps much higher, and hitting the maximum amount of bandwidth. It demonstrates the rush hour of time when maximum Internet users’ are on-line, and network requests are much higher than its total capacity. During that time congestion may occurs and users may face problems while browsing.

During the time of investigation we had many choices to select the software tools, after investigation R-language was better option to decide on. R has more natural and steep learning, major reasons (a.) free distribution and widely used for analysis. Immense on-line support with millions of blogs and lectures.
(b.) however R is not built for professional developers but more close to scientists and maintain by scientists, (c.) No need to write larger code, R studio provides better interface, this environment gives free hand to observe the output lively.

The visual representations of results are shown as under;

Fig. 6: Forecast ARIMA, ETS(M,Ad,N) and ETS(M,N,N)

B. Exponential Smoothing Functions

ETS (M, Ad, N) Exponential Smoothing Damped Trend Method with multiplicative Errors

If

\[ \epsilon_t = \frac{T(In)_t - T(In)_{t-1}|t}{T(In)_{t-1}|t} \]

then

\[ T(In)_t = (l_{t-1} + \theta c_{t-1})(1 + \epsilon_t) \]

\[ l_t = (l_{t-1} + \theta c_{t-1})(1 + \alpha \epsilon_t) \]

\[ b_t = \theta b_{t-1} + \beta(l_{t-1} + \theta c_{t-1})\epsilon_t \]

ETS(M, N, N): Simple Exponential Method with Multiple Errors

If

\[ \epsilon_t = \frac{T(Out)_t - T(Out)_{t-1}|t}{T(Out)_{t-1}|t} \]

then

\[ T(Out)_t = (l_{t-1} + \theta c_{t-1})(1 + \epsilon_t) \]

\[ l_t = (l_{t-1} + \theta c_{t-1})(1 + \alpha \epsilon_t) \]

\[ b_t = \theta b_{t-1} + \beta(l_{t-1} + \theta c_{t-1})\epsilon_t \]

ETS(M, N, N): Simple Exponential Method with Multiple Errors

If

\[ \epsilon_t = \frac{T(In)_t - T(In)_{t-1}|t}{T(In)_{t-1}|t} \]

then

\[ T(In)_t = l_{t-1}(1 + \epsilon_t) \]

\[ l_t = \alpha T(In)_t + (1 - \alpha)l_{t-1} \]

\[ l_t = l_{t-1}(1 + \alpha \epsilon_t) \]

In the similar fashion model can be constructed for T(Out) as follows:

If

\[ \epsilon_t = \frac{T(Out)_t - T(Out)_{t-1}|t}{T(Out)_{t-1}|t} \]

then

\[ T(Out)_t = l_{t-1}(1 + \epsilon_t) \]

\[ l_t = \alpha T(Out)_t + (1 - \alpha)l_{t-1} \]

\[ l_t = l_{t-1}(1 + \alpha \epsilon_t) \]

While using time series in R, ts-package offers Alpha (\(\alpha\)), Beta (\(\beta\)) and gamma (\(\gamma\)) functions, as HoltWinters generalized the procedure to compact with trend and seasonal variation. Given parameters Alpha describes “level, Beta deal with “trend” and Gamma deal with “seasonal variation”. However data-set for exponential smoothing functions are described in above code. [19], [17], [40]

However for Exponential Smoothing function we use the following equation

\[ F_{t+1} = Ft + \alpha (At-Ft) ORF_{t+1} = \alpha At + (1-\alpha)Ft \]
The forecast values are converted in the shape of two models i.e. Exponential smoothing and ARIMA [9], [20].

1. Forecast from ETS (M,Ad,N) ETS=Exponential Trial Smoothing (Error, Tend, Seasonal) Where M=Multiplicative, Ad=additive damped, N=None
2. Forecast from ETS(M,N,N) ETS=Exponential Trial Smoothing (Error, Tend, Seasonal) Where M=Multiplicative, N=None

The table shows fifteen exponential smoothing methods of trend components and seasonal components of ARIMA model using R-language [18].

<table>
<thead>
<tr>
<th>Trend Component</th>
<th>Seasonal Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (None)</td>
</tr>
<tr>
<td></td>
<td>A (Additive)</td>
</tr>
<tr>
<td></td>
<td>A (Additive damped)</td>
</tr>
<tr>
<td></td>
<td>M (Multiplicative)</td>
</tr>
<tr>
<td></td>
<td>M (Multiplicative damped)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A, N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A, A</td>
</tr>
<tr>
<td></td>
<td>A, A, M</td>
</tr>
<tr>
<td></td>
<td>M, M</td>
</tr>
<tr>
<td></td>
<td>M, M, N</td>
</tr>
</tbody>
</table>

Fig. 7: Exponential Smoothing Methods

C. Vector Autoregression (VAR) (p) model

However predicting and obtaining the futurist data in my research study, We used “Vector Autoregression” (VAR) (p) model to predict the In and Out bandwidth traffic (kbps) and have used the AIC (Akaike information criterion) and LR test (Likelihood-ratio test) for lag selection criterion. The equations for Vector Autoregression (VAR) can be represented are as under:

\[
In_t = c + a1(In)_{t-1} + a2(In)_{t-2} + a3Out(t) - 1 + a4Out(t) - 2 + e1t
\]

\[
Out(t) = c + b1(Out)_t - 1 + b2(Out)_{t-2} + b3(In)_{t-1} + a4(In)_{t-2} + e2t
\]

We estimated two equations and results are given in Figure 7 and 8, as results show for equation-1 all coefficients are statistically significant except the coefficient \( a3 \). There are two coefficients statistically insignificant \( b2 \) and \( b4 \), it means Lag 2 of In and Out variables are not significant. However, we applied granger causality test to check for In and Out bandwidth traffic affected by the past values of In and Out bandwidth. The result of Granger causality test is in Figure-8. Both Hypothesis are rejected so we can conclude that the IN and Out bandwidth can be predicted by the past values of IN and Out bandwidth traffics.

Fig. 8: Results of Vector Autoregression (VAR) (p) Model

We have used VAR(1) to predict the bandwidth traffic Out (kbps) the estimated VAR(1) is given by

\[
Xt = c + aXt - 1 + e
\]

\[
Xt = 606.2201 + 0.5148Xt - 1 + e
\]

Estimated results of the above equation is given in Figure-7, results shows that \( c= 606.2201 \) with \( p = 0000 \) which is highly significant and \( a = 0.514803 \) with \( p-value = 0000 \) which is also highly significant. F-value is 256.1357 with \( p value = 0.0000 \) which shows that the our VAR(1) is highly significant. Durbin-Watson stat shows that there is no problem of auto-correlation (as value is close to 2). Residuals are very large therefore \( R2=0.265114 \) which is very small. Residuals are shown in the graph (i.e. Figure 4: Residuals Graph)

This shows that variation in the period more than 400 is larger as compare to the variation less than 400, which may cause the small \( R2 \) and large standard error.

D. Implementation tools and Specifications

Following programming tools and applications are used for above research:
IV. BACKGROUND AND RELATED WORK

Our proposed forecasting model is suitable for proper resource utilization and also accommodates large scale of bursty data in data center networks. A number of other forecasting models are also proposed to measure the network and bandwidth traffic, most of them used to collect the network traffic on short term basis (minutes to hours), while some of other models used for full-term basis (hour to days) [6], [25]. But our forecasting model used for both from short time as well as full time to capture the traffic i.e minutes, hour and days. The reason for both short term and full term to acquire the detailed traffic traffic and obtain the better results with more precise outputs.

Numerous researches have presented, examined, and compared for forecasting tool to predict bandwidth and its proper utilization for example:

Linear regression by using Granger causality and forecasting by using VAR model (by Eviews tools) is also nicer option for predicating the futurist data. [2], [11], [13].

UANM (Unified architecture for network measurement) offered an end to end measurement tool for bandwidth estimation. Using UANM tool, authors illustrated to achieve synchronized dimensions and avoid interferences, also increase the accuracy and reliability of any measurement of bandwidth from end to end [11].

Traffic-prediction-assisted dynamic bandwidth assignment for hybrid optical wireless networks proposes performance based extensive simulation based on extensive simulation and architectural tool for scheduling. The mechanism propose for dynamic bandwidth assignment (DBA) and optical network unit (ONU) which predict the incoming traffic and manage the network scheduler for better performance on network [27].

Dynamic bandwidth allocation with high utilization (DBAHU) Algorithm proposes to utilize the unused bandwidth of a service class. The exemplified procedure uses simple techniques for dynamic bandwidth allocation (SFDBA) algorithm. Both SFDBA and DBAHU practice a collective accessible byte counter and a mutual counter for multiple queues of a service class. Moreover it addresses the problem of un-utilized bandwidth with service class, and unoccupied bandwidth used by available byte counter [12].

The SNMP data pattern used in Network Traffic Characteristics of Data Centers in the Wild uses paper, author studied numerous network patterns [12]. A flexible reservation algorithm for advance network provisioning proposed a framework for network reservation and claimed to deliver the guaranteed bandwidth [3].

While An empirical study of the multi-scale predictability of network traffic illustrated an experimental study to find out the errors in forecast on diversified time-scales. Study tries to present that the forecasting inaccuracies do not reflect monotonically reduction with ironing on big platform [32].

The estimation of bandwidth can be calculated by sending the probe packets on network and measured with list of easily available applications and tools i.e. IGI, Pathload, pathChrip, Spruce and IGI [16], [22], [33], [35].

In the paper “The Nature of Data center Traffic: Measurements Analysis” [23], the experiment was simulated to create the traffic pattern intentionally to observe for mining the huge data-sets for analyzing the performance.

One we see the network architecture of data centers, the geographically distributed data centers have three layered connectivity (a) layer-2 and over fiber, (b) Layer-3 with WAN over dark fiber and (c) storage extension. However the distributed data centers connectivity expect the compatibility and fasten support within all network vendors and all technologies. At present DCI uses layer 2 - 3 Virtual Private Networks with Multi-Label Switching (VPN-MPLS), Secure Socket Layer with Virtual Private Networks (SSL-VPN), and some other bundles of secure protocols. Various other protocols i.e. IPSEC-VPNs and Vx (virtual private networks and bundle of virtual s) for secure connections are used.

Multipathing in Data Center Ethernet (DLBMP) is an alternative solution of STP (Spanning Tree Protocol), DLBMP propose the solution to overcome the proper bandwidth utilization on data link layer (L2) by using Dijkstra algorithm. Since STP has problem of unexpected blockage for links and ports. In DLBMP redundant physical links have deployed to overcome the failure of physical links, it has more performance and can handle 300% more bandwidth capacity while compare with STP. The communication between nodes and traffic are dynamically adjustable, the load balancing are feasible and ease to achieve their efficient link with proper bandwidth utilization [39].

Portland and VL2 [10], [30] uses the architectural model of Switch-Centric routing structure, which controls the communication by using network switches for routing, the same anatomy used in three-tier (i. access ii. aggregate iii. core) and fat-tree. This type of architectures are largely used in conventional data centers’ physical topology. But the three-tier topology schemes are very large, complicated and heavy looking for price and power [4]. The Helios (Hybrid Electrical and Optical structure) [8] combines pod switches with core switches, the architecture propose the reductions of switching elements, cabling, cost, and power consumption. While cThrough [37] architecture by combining the optical and electrical technology, the optical segment routing performs one hop exchange of communication while the electrical segment works like routing in tree, although optical solution has better performance in power saving but rarely used in data centers due to high-priced cost of switches and complex configuration.

V. CONCLUSION AND FUTURE WORK

This research paper presented an innovative idea and approach to support the research communities of networking especially data center professionals. However it’s challenging to build forecasting models for real world traffic of Data
Centers, more important to make predication of futuristic traffic when designing, managing and upgrading the complex network of data centers. In this paper we will focus on two main objectives; (1) proposing simple and yet scalable techniques for analysis and forecasting, (2) Implementing and evaluating these techniques on real-world data centers. Our next project for research to work on complex protocols used in data centers; especially take counter measures for the security of data centers.

ACKNOWLEDGEMENT

I take this opportunity to express gratitude to all unknown reviewers for their feedback and make me able to participate for this conference. This research was supported by Sukkur Institute of Business Administration, this prestigious institute allowed me to mentioned the name to acknowledge. I would like to express my sincere gratitude to my supervisor Prof. M-Tahar Kechadi, who is second author of this paper; this study is nothing with the exception of his continuous support and motivation. My sincere thanks to my ex-colleague Mr. Fahad Rahim Qasmi for providing the partial data and excess of data center.

REFERENCES

[33] V Ribeiro and Riedi R Baraniuk Retal Pathchimp, Efficient available bandwidth estimation tools, Proceedings of the 3rd ACM


