

Effective Transmission Capacity Management Technique Using Z-Score Outlier Normalization Algorithm

Lutfur Rahman¹, Md Abdur Rahman², Dr Risala Tasin Khan³

¹MSc in IT, Institute of Information Technology, Jahangirnagar University

²Technical Engineer, Optical Transport Network, ZTE Corporation

³Professor, Institute of Information Technology, Jahangirnagar University

Abstract

In this surging tide of information world, it becomes the great challenge for the mobile telecommunication operators and data service providers to maintain their data transmission network with high reliability and accessibility from everywhere considering implementation and operational expenses. Hence, network and capacity expansion could play a vital role in gathering more and more customer to enrich the organizational portfolio and revenue. In the contrary, wrong or error data analysis in deciding network capacity expansion may lead to the loss or meaningless investment which directly impact on CAPEX & OPEX. So, the capacity analysis and management is a vital factor for data service providers to outreach the business prospect as well as revenue. In the process of transmission capacity management, analyzing and reporting of transmission link utilization is inevitable building blocks. Any error data or outlier could result in over or underutilization which may lead the engineers to take wrong decision thus affecting the CAPEX & OPEX of the company. In this paper, we will implement a python based automation tool using Z-Score outlier detection algorithm to calculate the average transmission link utilization.

Keywords: Data Science, Data Analysis, Transmission Capacity, Data Visualization, Python for EDA, IP-MPLS, Utilization, Performance Analysis, Delay, Jitter & Packet Loss.

1. Introduction

To connect the whole country, The Mobile Telecommunication Service Providers (MTSP), International Internet Gateway (IIG), International Gateway (IGW), Internet Exchange (IX), Interconnection Exchange (ICX), International Long Distance Telecommunications Services (ILDTS), Internet Service Providers (ISP), International Private Leased Circuit (IPLC) operators and a lots of group of companies have built their IP networks based on IP-MPLS architecture. To analyze the utilization of the transmission links, there are no low cost, user friendly and error free tool. Although, some telecommunication equipment manufacturer supply their own tool for analyzing capacity utilization but those are very costly and vendor specific. So, small IIG, ISP and organizations cannot afford that expense. For this reason, a light weight, error free and open source tool is need to be developed for data utilization analysis for IP-MPLS networks. This paper proposes a design and analysis method based on Python using Z-Score algorithm for outlier detection and the calculation of average utilization.

1.1 Overview of Transmission Systems

1.1.1 MW Transmission Network

Microwave Transmission is a kind of wireless transmission system where signals are traversed through the air. MW transmission uses different frequency bands which are shown in the following figure.

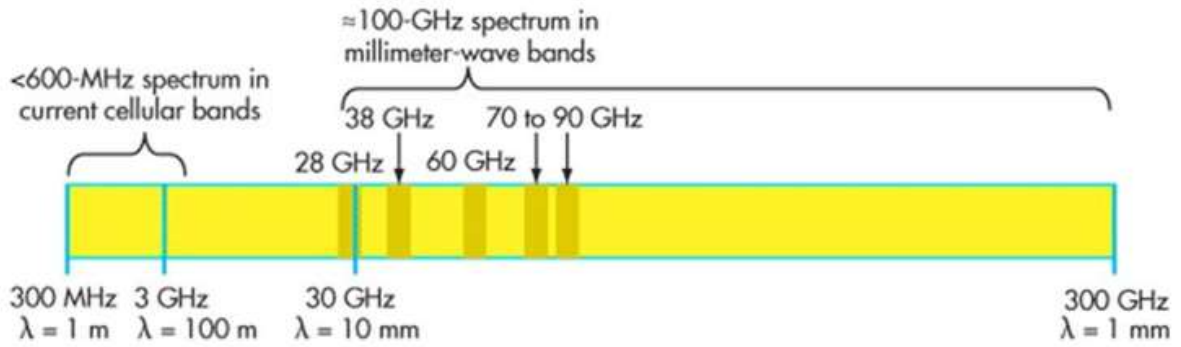


Figure: Frequency Range of MW Transmission Network

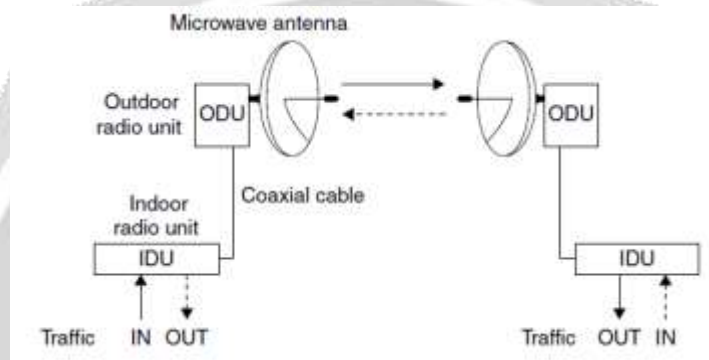


Figure: Point-Point MW Transmission Link

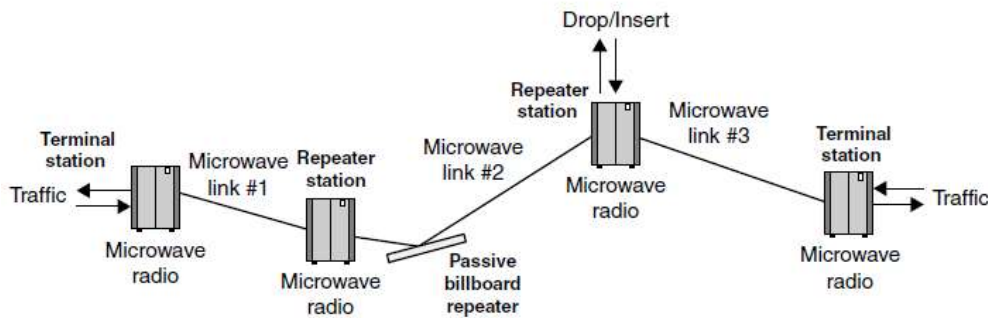


Figure: MW Transmission Traffic Flow

In MW transmission system, each range of Frequency has impact on throughput, signal impairment and BW. Based on the distance of traffic throughput requirement one can choose specific MW devices like Microwave antenna, ODU (Outdoor Unit) and IDU (Indoor Unit). The IP devices could be connected with MW devices to make IP-MPLS network to serve the customer. Huawei, ZTE, Ericsson and NEC are major vendor here in Bangladesh for MW network equipment.

1.1.2 SDH Transmission Network

Synchronous Digital Hierarchy (SDH) is a data transmission standard to transmit data from one point to another abiding by a set of standard protocols. SDH technology enables low bit rate data into higher rate streams. The highest capacity could be achieved is STM-64. Through this transmission standard different types of services could be transported like E1 STM, FE, GE, ATM, SONET [12] etc. So that, this is also known as Multi Service Transport Platform (MSTP). SDH device carries traffic through

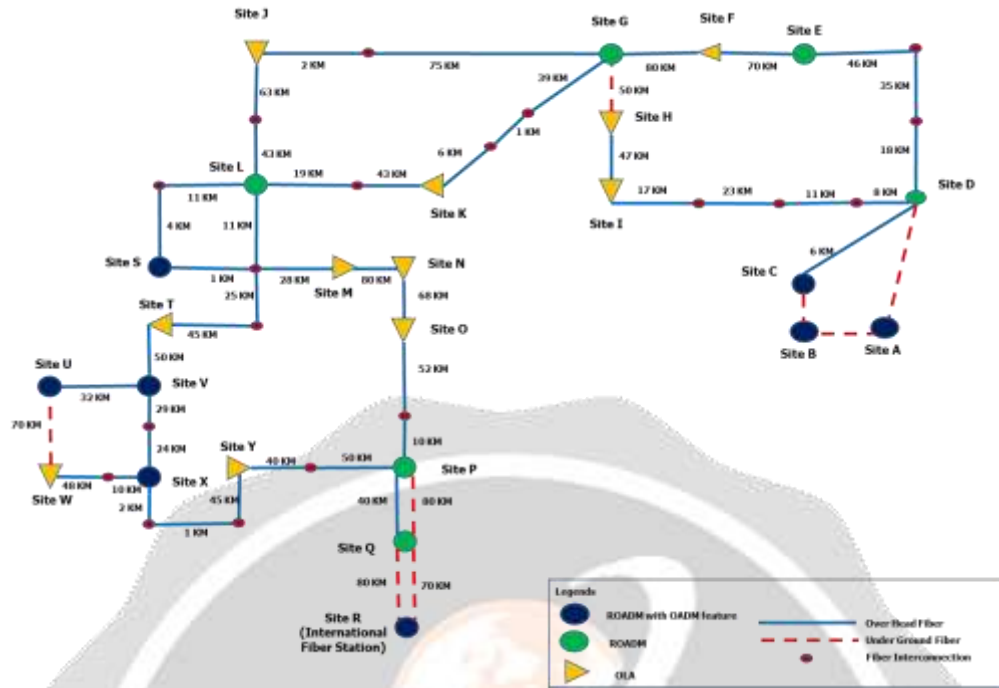


Figure: A Commercial DWDM network

1.1.4 IP Transmission Network

IP Transmission network is basically a logical network where physical network is orchestrated through either MW SDH or DWDM or hybrid transmission media. IP-Transmission network is actually a Layer 3 network [12] which implements different protocols like OSPF, IS-IS, MPLS and BGP etc. Routers, Switches & Firewalls are mainly 3 types of IP transmission devices used in IP-CORE network, IP-RAN, WAN, MAN, C-RAN, SD-RAN and LAN networks. In Bangladesh, CISCO, Juniper, HUAWEI, ZTE are major vendors for implementing IP networks. Following figure shows a typical IP transmission network based on hybrid data transmission network.

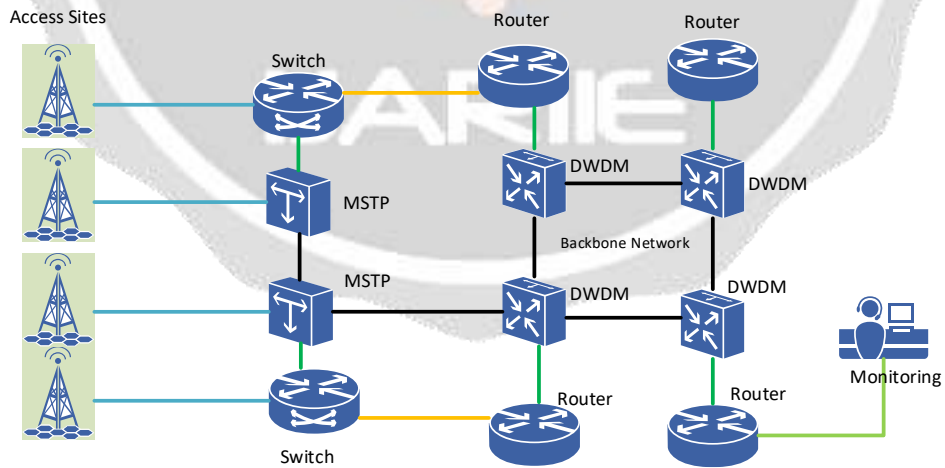


Figure: IP transmission network based on hybrid transmission

2. Capacity Management of Transmission Link

2.1 Capacity Management Process

Capacity management for transmission link is vital for an organization as it can directly impact on company’s existence. Capacity management could be described as a wide range of planning deals to ensure that, the network infrastructure is bearing sufficient capacity and resources to serve the customer demands and growth as per computed forecast and continue business operations without any interruption.

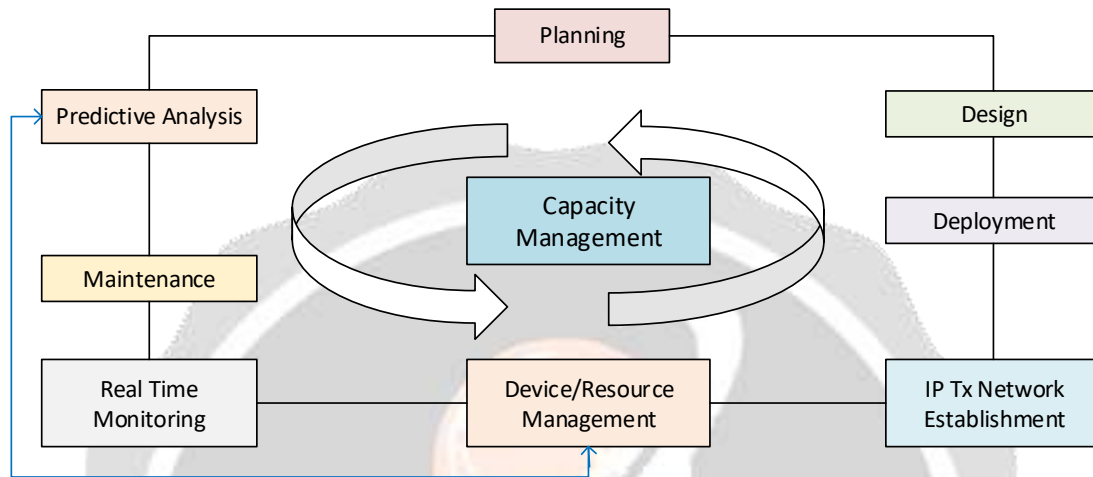


Figure: Capacity Management Process

2.2 Strategies for Capacity Management

There are various strategies for transmission network capacity management which are used, rely on the requirements of the existing network traffic and its tolerance for traffic growth as per the predictive analysis. The most popular and effective strategies for operating capacity management are as follows.

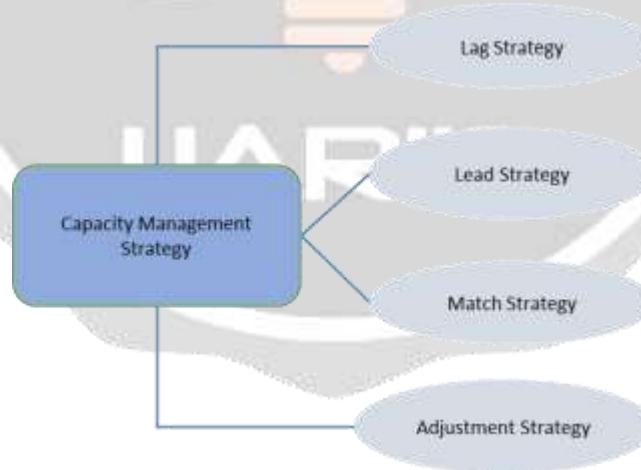


Figure: Capacity Management Strategies

2.2.1 LAG Strategy

In this type of strategy, one needs to feed the demand based on the requirements when it reveals itself. For example, In IP-transmission network there are currently several zones and each zone have two transmission links of 100Gbps (1+1 protection) and current TX (transmission) link BW (Bandwidth) utilization is running between 80% - 90% as per weekly average peak link utilization report [9]. So, organization will increase link capacity only when link utilization reach near to 100% or when specific traffic growth demand raised by some events. So, on that case need to initiate transmission link expansion process which involve

DWDM physical link expansion, DWDM 100G tributary card procurement & expansion, vendor service procurement & Lambda (λ) configuration and acceptance, Router & card procurement and expansion etc.

In capacity management lag strategy is the most primitive in the sense that, it pursue to skip over-all amount of resources like DWDM, Router, deployment service etc. This method creates the hazardous & dangerous situation and that lead the company spend un-necessary and excessive amount of money, time and materials which in-term rising the CAPEX and OPEX that are not at all a satisfactory wish.

The peril of unwanted and excessive expenditure on devices, materials, and accessories must be a trade-off with the results when requirements for expansion surpass budget, project time, and urgency [15]. As an example, a 100G TX link expansion is urgently triggered due to holiday event otherwise spike in utilization may incur congestion which in terms will cause all service interruption under those links and create bad customer experience and may impact company reputation and revenue but on the other hand if company do jump for expansion then need extra employee engagement and chance to burnout, consume available 100G cards of DWDM and routers, procure vendor services which incur more CAPEX cost.

Nevertheless, this strategy must take it into account that to act reactively it may need to tradeoff between latency to manage resource and services, and accomplishment of the new raised requirement so that there prevails ultimate balance between under-allocation and over-allocation.

2.2.2 Lead Strategy

This strategy looks up to predict resource (device, materials, accessories, human resources etc.) demands and proactively achieve them well ahead of time they are needed. Let's assume, an organization wishes to expand its transmission link and target to cover all traffic during the upcoming festival, it might need carefully procure hardware, materials, and accessories, local/foreign vendor service and human resources in contemplation of the upcoming need.

Forecasting upcoming required resources can be a very challenging method, filled with prediction, market operation analysis, customer segments, traffic trend analysis and prediction. Companies are looking up to skip the aftermath results that can arise from being run out resources, but the other threat is an increase in expenditure on CAPEX/OPEX that are not expected. After all, the company may not be able to anticipate factors like market fall, rise of competitors, or an unenthusiastic customer retort to its growth strategy.

Those who get involved in a lead strategy for capacity management must, therefore, be prepared to retort to events where the required resources are not wanted. This often apparent in the form of cutoff and tuning to the forecasted need. The business will also face opportunity costs, such as innovation projects, that could have been engaged with had they not over-anticipated the need for resources.

2.2.3 Match Strategy

This strategy always looks for adjusting the quantity of available resources to significantly reflect existing and near-future need. This type of method is the "market counterbalance" approach to exactly meet supply with demand, as stated above.

While on paper having an exact match of resource supply to demand may sound ideal, there are cons to the strategy worth considering. Firstly, repeatedly estimating the need can be a resource-intensive process. It is also dependent on prediction [6]. These forecasting accuracy may increase day by day over time, but they may result in an organization to exaggerate to catalysts that may later turn out to be not-so-important. Moreover, it might be tough for some companies to prepare LRP and strategy if resources are constantly flapping.

In a summary a match strategy is well-suited for companies that have advanced resource analysis, prediction and planning capabilities. They must also be wishing to trade off immediate capacity availability (found in lead strategies) or overall resource cost savings (as often found in lag strategies) for an ability to achieve their resource demands exactly in the middle.

2.2.4 Adjustment Strategy

An adjustment method is one of the popular and common technique to capacity management because it responds to need but not in accurate real-time window. The company may follow a lag strategy for a specific time window and a lead strategy for another time window. They may look for achieving an exact match during times when balancing resource availability with budget constraints is absolutely paramount.

In opposition to match strategy, where activity put into repeatedly calculating the existing and near-future expansion need, an adjustment strategy responds to factors on a less-frequent basis [7]. The timeline for tuning the strategy could be quarterly or mid yearly, monthly, or in some cases even weekly. Again, the key is that the company looking for using the perfect strategy needed given the lagging and leading indicators in their own particular industry.

An adjustment strategy could be considered as the most-balanced and appropriate capacity management method, but it also does forego the strongest advantages of the strategies above. By seeking to be neither conservative nor consistently proactive with resource and materials procurement, the organization may face opportunity costs compared to choosing one of the strategies above.

An adjustment strategy take the strengths received in being both responsive and reactive, depending on the facts and situation, without the level of effort needed to get involved with an exact match strategy.

3. Outlier Detection Algorithm

A lone data point that resides far from the average value of a set of data point is called Outlier. Outliers might be far different that present outside samples of a group of data as well. Specifically, an outlier is an entity that is noticeably dissimilar from the usual in some extent.

Outliers are very important factor in statistics as they can have a catastrophic effect on overall scenario. Especially, in case of small amount of data sets, a single outlier may ridiculously impact on averages and bias the final results. Following figure presents outlier basics.

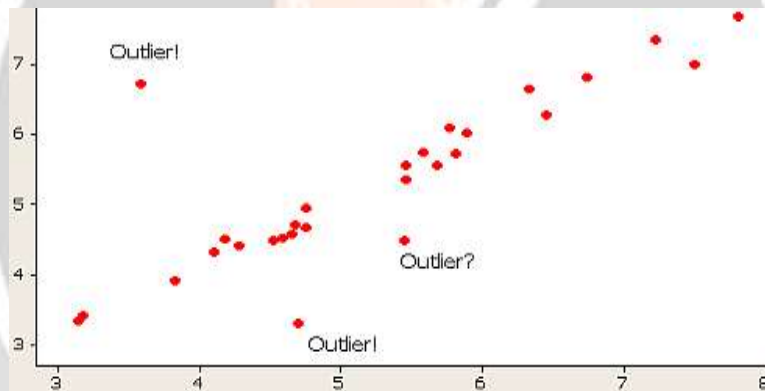


Figure: Example of Outlier

In above picture, the different data are plotted on the graph and the similar group of data are located in the closest area but the items or things which are differ from particular group of data are located far from the groups.

As like other statistical systems, the outliers are generated and impact on transmission link utilization. Some major reasons are, NMS server processing issue, FC utilization due to Fiber cuts, Network Device Faults, Unusual Traffic spikes, Event triggered traffic.

Following are some popular outlier detection methods used in data science.

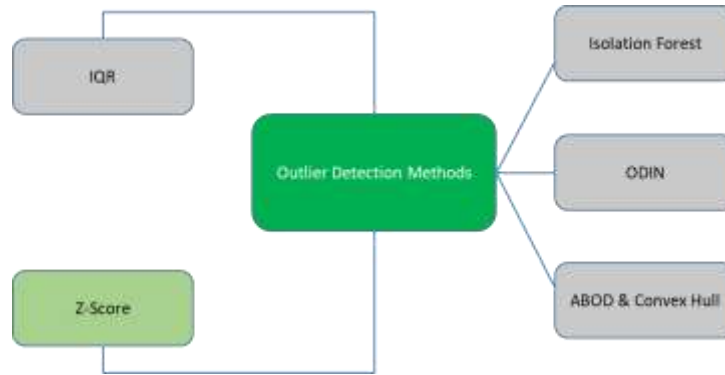


Figure: popular outlier detection methods

3.1 Z-Score outlier Detection Algorithm

The Z-Score or standard-score is a popular technique in statistics for finding outpaced individual data point which is how far from the average. By put in Z-transformation we switch the distribution and make it 0 average with unit **Stdev**. As an example, a Z-Score of 2 would indicate the data point is 2 **Stdev** away from the average of the data set.

Also, Z-Score of any data point can be computed with following equation.

$$Z_i = \frac{X_i - \mu_x}{\sigma_x}$$

Here,

X_i = Particular Data Point

μ_x = Mean of data set of X

σ_x = Stdev of Data set X

Example:

It is considered that the data is normally distributed and the percentage of data sets that lie between $-/+1$ standard deviation is $\sim 68\%$, $-/+2$ standard-deviation is $\sim 95\%$ and $-/+3$ standard-deviation is $\sim 99.7\%$ [8]. So therefore, if the Z-Score is >3 then we can surely identify that point as an outlier point.

Below Figure shows the Z-Score method.

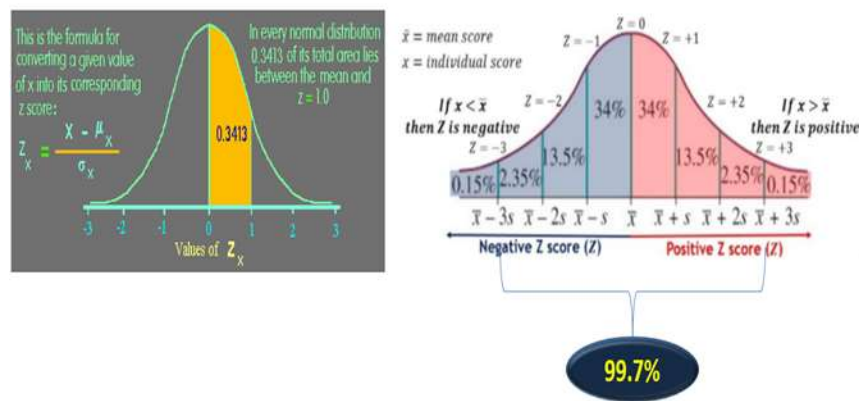


Figure: Z-Score method [10]

In this paper, we have applied Z-Score algorithm to detect outlier and corrected it to compute average of the link utilization. Following algorithm is prepared in python [1, 2, 3] for this automation tool creation in computing transmission link utilization.

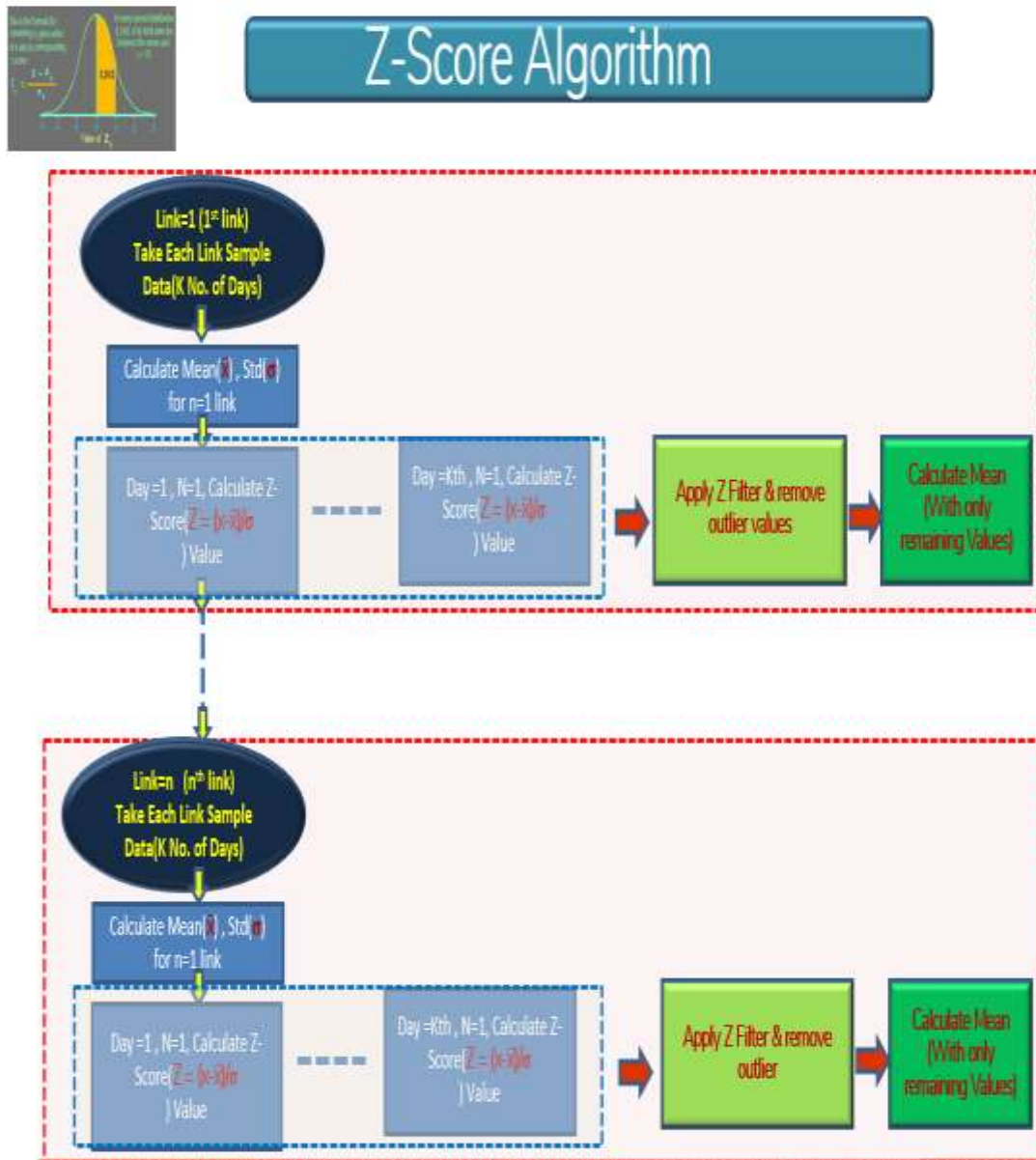


Figure: Z-score Algorithm for Transmission Link Average Utilization Calculation

4. Transmission Link Raw Data Collection and Processing

The utilization dump files has been collected from a running IP-MPLS network of second largest mobile telecommunication service provider in Bangladesh. For data collection and parsing has been done in accordance with the powerful algorithm which is used as the raw file input for the python program. In this program, renowned python libraries are used for data processing and TKinter library [11] is used for Graphical User Interface (GUI) creation.



Figure: GUI Window developed in python TKinter

4.1 Output of file collection and log parsing code

```
IP Link Traffic Report_Daily_20221211000000_20221219131314.xls
IP Link Traffic Report_Daily_20221212000000_20221219131323.xls
IP Link Traffic Report_Daily_20221213000000_20221219131332.xls
IP Link Traffic Report_Daily_20221214000000_20221219131340.xls
IP Link Traffic Report_Daily_20221215000000_20221219131348.xls
IP Link Traffic Report_Daily_20221216000000_20221219131359.xls
IP Link Traffic Report_Daily_20221217000000_20221219131410.xls
```

```
['E:/2022/BH/PMT report/Dec-2022/18 Dec_2022/IP Link\\IP Link Traffic Report_Daily_20221211000000_20221219131314.xls', 'E:/2022/BH/PMT report/Dec-2022/18 Dec_2022/IP Link\\IP Link Traffic Report_Daily_20221212000000_20221219131323.xls', 'E:/2022/BH/PMT report/Dec-2022/18 Dec_2022/IP Link\\IP Link Traffic Report_Daily_20221213000000_20221219131332.xls', 'E:/2022/BH/PMT report/Dec-2022/18 Dec_2022/IP Link\\IP Link Traffic Report_Daily_20221214000000_20221219131340.xls', 'E:/2022/BH/PMT report/Dec-2022/18 Dec_2022/IP Link\\IP Link Traffic Report_Daily_20221215000000_20221219131348.xls', 'E:/2022/BH/PMT report/Dec-2022/18 Dec_2022/IP Link\\IP Link Traffic Report_Daily_20221216000000_20221219131359.xls', 'E:/2022/BH/PMT report/Dec-2022/18 Dec_2022/IP Link\\IP Link Traffic Report_Daily_20221217000000_20221219131410.xls']
]
```

```
number of row is 7
Number of column is 1
```

Figure: Output of Log collection

	Link Name	LINK_BW	12/11/22	12/12/22	12/13/22	12/14/22	12/15/22	12/16/22	12/17/22
0	CG_AL02_NE40E_RS02(Eth-Trunk2)--T_CG_AL02_	300000	136839.63	139225.78	138890.21	140647.92	135992.19	145241.51	146197.21
1	DH_PB02_NE40E_X0A_PE2(Eth-Trunk2)--T_DH_PB_	600000	183663.20	180941.90	183626.08	177581.11	163968.30	166934.70	175893.82
2	T_CG_AK09_NE40E_X0A_PE2(Eth-Trunk2)--CG_AK09_	300000	168510.52	163909.21	150565.82	154670.02	147626.85	155320.92	168441.94
3	CG_BD01_NE40E_X0A_PE1(Eth-Trunk5)--CG_BD01_	400000	141589.55	141840.90	148205.35	140323.54	149071.56	148128.29	152539.48
4	CG_AK09_NE40E_IB02(Eth-Trunk2)--T_CG_AK09_NE4_	200000	106605.27	106780.82	105308.49	104362.76	92148.65	92446.66	104809.39
1108	DH_PB02_PS_S9312_MCE2(Eth-Trunk3.70)--DH_PB02_	20000	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1109	DH_PB02_PS_S9312_MCE2(Eth-Trunk3.100)--DH_PB0_	20000	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1110	DH_KL07_NE40E_PE1_Eth-Trunk3_dot1q4091--DH_KL_	1000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1111	DH_KL07_NE40E_PE1_GE40/3--DH_B002_NE40E_PE1_	10000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1112	DH_KL07_NE40E_PE1_GE40/2--DH_B002_NE40E_PE1_	10000	NaN	NaN	NaN	NaN	NaN	NaN	NaN

1113 rows * 9 columns

Figure: Output after TX link utilization log file parsing

4.2 Output of Z-Score algorithm Code

Below figure shows the output of Z-score code after replacing the outlier data by NaN.

	Link Name	LINK_BW	12/11/22	12/12/22	12/13/22	12/14/22	12/15/22	12/16/22	12/17/22
0	CG_AL02_NE40E_RS02(Eth-Trunk2)--T.CG_AL02_	300000	136839.63	139225.78	138090.21	140647.92	NaN	NaN	NaN
1	DH_PB02_NE40E_X0A_PE2(Eth-Trunk2)--T_DH_PB_	600000	183663.20	NaN	183626.08	177593.11	NaN	NaN	175893.82
2	T.CG_AK09_NE40E_X0A_PE3(Eth-Trunk2)--CG_AK09_	300000	158510.52	153809.21	150565.82	154670.02	NaN	155320.92	NaN
3	CG_BD01_NE40E_X0A_PE1(Eth-Trunk5)--CG_BD01_	400000	141589.55	141840.90	148205.35	140323.54	140071.56	148128.29	NaN
4	CG_AK09_NE40E_IBG2(Eth-Trunk2)--T.CG_AK09_NE4_	200000	106605.27	106780.82	105308.49	184362.76	NaN	NaN	104809.39
1108	DH_PB02_PS_S8312_MCE2(Eth-Trunk3.78)--DH_PB02_	20000	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1109	DH_PB02_PS_S8312_MCE2(Eth-Trunk3.100)--DH_PB0_	20000	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1110	DH_KL07_NE40E_PE1_Eth-Trunk3_dot1q4091--DH_KL_	1000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1111	DH_KL07_NE40E_PE1.GE4/0/3--DH_BG02_NE40E_PE1_	10000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1112	DH_KL07_NE40E_PE1.GE4/0/2--DH_BG02_NE40E_PE1_	10000	NaN	NaN	NaN	NaN	NaN	NaN	NaN

1113 rows x 9 columns

Figure: Output of Data Frame after replacing the outlier by NaN

Average utilization calculation after normalized data by Z-Score algorithm which is shown in the following figure.

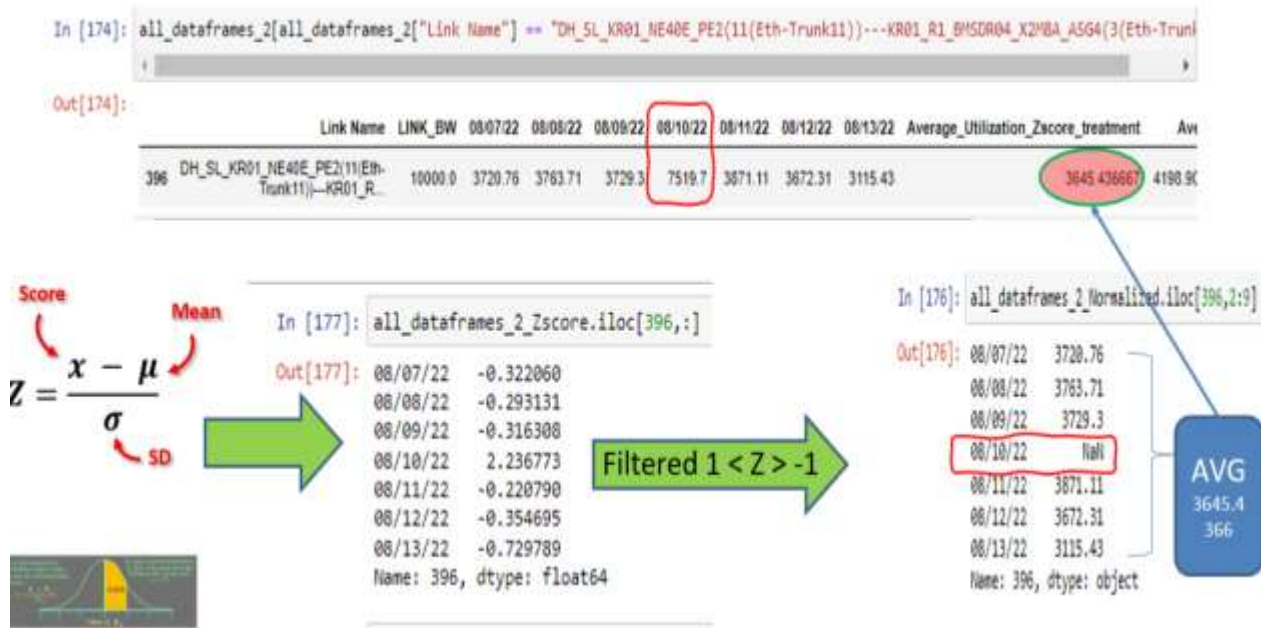


Figure: Average Calculation after outlier normalized by Z-Score

We observed that, in normal calculation, the average is 4198.90 where outliers are exists and 3645.44 after outliers are normalized by Z-Score algorithm. The average with outliers is more than 15% higher than the average value of the Z-Score normalized data. So, this 15% would surely impact on the decision making for the network and capacity expansion process and may invest the company resources which will have no revenue return.



5. Summary

This Project is a practical, pragmatic & cost optimized solution for IP-MPLS transmission links quality & performance log data analysis for identifying link utilization. It shows real log processing, parsing, cleaning garbage, feature data extracting & data frame structuring and give insight through visualization for IP-MPLS network optimization. As like other applications and tools there is always open door for future development based on user requirement. In future, Machine Learning codes could be integrated with any EMS/NMS server to predict the anomaly with one click rather than making separate GUI windows. Moreover, with ML codes, notification management system could be developed for over and underutilization.

6. Conflicts of interest

Authors have no conflicts of interest

Authors' Biography

	<p>Lutfor Rahman is currently working as a specialist in Robi Axiata Limited since 2011. He has more than 16 years of in-depth telecom experience in MW, Fiber and DWDM Transmission Operation, MW Planning & Implementation field for Multination Telecom Industries including Ericsson and Motorola. He started his career as telecom system engineer in 2004. He received his Master's in Information Technology, Jahangirnagar University in 2023 and B.Sc. (Engg.) Degree in Electrical and Electronics Engineering from RUET in 2004. He also participated different professional training programs in Huawei Technologies in Malaysia, Ericsson Academy in Malaysia, India and Sweden, Alcatel Lucent (ShangHai) and achieved certification with updated technologies in MW operation, WDM operation and Fiber operation. His research interest includes Applications of ML in practical field of Telecommunication.</p>
	<p>Md Abdur Rahman received the B.Sc. in Engineering in Electronics and Telecommunication Engineering from the Rajshahi University of Engineering and Technology at the examination of 2011. During his Bachelor study, he worked on automated water level controller, speed control of DC motors- application to the power electronics. He started his career as Assistant Engineer in Planning and Development department of a NTTN operator and closely worked with SDH & DWDM systems. Later, he worked for Huawei Technologies Ltd. as Optical Transmission Engineer and as Specialist in Transport Operations department in Robi Axiata Ltd. Currently, he is working as Technical Engineer (Optical Transport Network) in ZTE Corporation. He has expertise in MSTP, DWDM, FOADM, OTN, OTDR, ASON and ROADM systems. His research interest includes, ML applications in Optical Transmission and IP network optimization.</p>



Dr Risala Tasin Khan is currently working as a Professor at the Institute of Information Technology of Jahangirnagar University from 2020 where she has been working since 2009. She completed her B.Sc. from Computer Science & Engineering Department of Jahangirnagar University in 2003, M.Sc. in 2005 and Ph.D. in 2019 from the same University and department. Her Ph.D research work was on the performance evaluation of CRN over fading channel incorporating space diversity. Her research interests span both computer networking and wireless communication. Recently she has also started doing research on resource allocation of wireless networks using machine learning and the security aspect of IoT. Dr Risala Tasin Khan authored more than 30 research papers on peer reviewed Journals and Conference Proceedings supervised more than 70 students in different fields of wireless communication. She is a senior member of IEEE and also acted as a counselor of IEEE WIE Affinity Group JU SB and EXCOM member of IEEE CS Bangladesh Chapter.

References:

1. M. Summerfield, Programming in Python 3: A Complete Introduction to the Python Language, 2nd Edition, Developer's Library. Addison-Wesley Professional, 2009
2. J. Vanderplas, Python Data Science Handbook-Essential tools for working with Data, Cha-Data Manipulation with Pandas, pp. 160–200. O'Reilly Media, November 2016.
3. V. Arora, "Exploratory Data Analysis (EDA) – A step by step guide" <https://www.analyticsvidhya.com/blog/2021/05/exploratory-data-analysis-eda-a-step-by-step-guide/>
4. Mubarakah, N. and Fadhilah, D.D., 2020, September. Point to point communication link design by using optical DWDM network. In 2020 4rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM) (pp. 265-268). IEEE.
5. Antil, R., Pinki, S.B. and Beniwal, S., 2012. An overview of DWDM technology & network. Int.J. Sci. Technol. Res, 1(11), pp.43-46.
6. W. McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy and IPython (Paperback)-2nd Edition". O'Reilly Media, November 2014.
7. R. Obe and L. Hsu, "PostgreSQL: Up and Running". O'Reilly Media, Inc., July 2012.
8. Q. Nguyen, "Mastering Concurrency in Python: Create faster programs using concurrency, asynchronous, multithreading, and parallel programming". Packet Publishing, November 2018.
9. E. Forbes, "Learning Concurrency in Python". Packet Publishing, August 2017.
10. Dr. Saul McLeod "Z-Score: Definition, Formula, Calculation & Interpretation" <https://www.simplypsychology.org/z-score.html>
11. David Amos "Python GUI Programming with Tkinter" <https://realpython.com/python-gui-tkinter/>
12. Ma, Weidong. (2000). Analysis for IP transmission network technologies. 1. 627 vol.1. 10.1109/ICCT.2000.889281