

The Development of the Education Related Multimedia Whitelist Filter using Cache Proxy Log Analysis

B Indrawan* and Y Kerlooza

Department of Magister Information System, Faculty of Post Graduate, Universitas Komputer Indonesia, Bandung, 102-116 Dipatiukur Street, Indonesia

*berny@labkom.unikom.ac.id

Abstract. The purpose of this study was to show the development of the education related multimedia whitelist filter using cache proxy log analysis. Regulation of bandwidth utilization should be used according to needs and main goals or process business of the organization. For Higher Education Institutions, access to education and science-related contents must be prioritized. Large size content access needs a right regulation, especially for multimedia contents. This research focuses on the development of whitelist filter based on an analysis of internet files access logs taken from squid proxy logs from April to November 2017. This technique can recognize 1,98% of the logs are direct links related to education and 0 % are indirect links to the education-related multimedia files and 98,02 % are links to non-education related multimedia files.

1. Introduction

Universitas Komputer Indonesia (UNIKOM) as a private higher education based on Information Technology, the higher education provides a network for ease of searching information to support of teaching and learning process and research of the Students, Lecturers, Employees, or the academic community in higher education environment. However, the ease and freedom of using internet network facilities in higher education environment presents a new problem such as freedom in the use of large-scale bandwidth to access and download applications and sites other than applications and educational sites such as applications or entertainment sites (movies, YouTube videos, music, social media, etc.) so that, the quality of the Internet network speed performance is reduced. Efforts are needed to resolve and minimize the problems by applying the bandwidth management model. Bandwidth management is the easiest way for universities to improve the Quality and Quantity of their bandwidth for educational purposes [1].

Umeh Innocent Ikechukwu discusses that to manage bandwidth in order to meet the needs of the organization and users of internet network then a rule is required to be able to regulate internet usage. The rules should be able to control the bandwidth usage to manage the use of application access or critical content such as applications or content that are related to business processes. In higher education network environment, applications and content are allowed if related to education or research and non-critical applications or content not related to education and research when bandwidth availability is possible. This rule can reduce the use of applications or content that is not appropriate on the network [2]. In research was conducted by Devajitet et al., Ronget et al. and Cao et al, wherein their research stated that the bandwidth management will be more efficient if conducted monitoring, reporting,



filtering access to the internet and see the use of traffic has been in accordance with business processes run by the Organization or not, by doing the process of monitoring, reporting and checking the use of traffic bandwidth can increase the productivity of bandwidth usage in accordance with the organization's business processes and reduce unsuitable traffic bandwidth usage [3].

Another research conducted by Kartik Bommepally and friends: An administrator has to ensure that all the users get a fair share of the bandwidth. An administrator in a campus may wish to ensure controls on Internet usage. In general, most campuses end up restricting usage over specific periods of the day. This has the disadvantage that a genuine user who needs access to information is also denied. To solve the problem, an analysis of Internet traffic data from proxy log is required to study "user access patterns", where the analysis can help the network management system in traffic shaping and monitoring [4]. To know the user access patterns, then the data collection process about user's actions should be done, and for collect data we can use the approach of an HTTP proxy which is inserted between the client and server. It intercepts all traffic and outputs log data with details about any requests sent to servers as well as the replies that a server sends back to the client [5].

In order to make traffic bandwidth usage not to be used to access content or applications that are not accordance with the business processes run by the Organization, then the development of whitelist needs to be done. Wherein the research conducted by Emyana Ruth Eritha Sirait said that whitelist can filter the content or applications that are in accordance with an organization's business processes and overcome the bandwidth usage to access content that is not related to business processes [6]. An application whitelist is a list of applications and application components that are authorized for use in an organization [7]. Whitelist can also be used to prevent unlicensed or undesired programs from being installed, such as gaming or file sharing programs that would reduce worker productivity or inappropriately use network bandwidth in a workplace environment [8]. Whitelisting is recommended to be used for environments which are relatively static and perform same functions routinely i.e. fewer changes on the system are expected to be made on daily basis [9].

Based on the research above, the authors will try to manage bandwidth usage to be more effective using the Education Related Multimedia Whitelist Filtering technique based on the analysis of Internet Access Log data contained in the Squid Proxy server logs to learn the habits user access patterns. The filters that will be developed in this research is based on Internet Access Log data from Squid Proxy server log of higher education from April to November 2017. So by using Education Related Multimedia Whitelist Filtering technique, authors can recognize, know and can ensure the content or multimedia applications accessed by the user has been in accordance with business processes in higher education and the development of the whitelist can manage the use of bandwidth to make it more efficient.

The purpose of this study was to show the development of the education related multimedia whitelist filter using cache proxy log analysis.

2. Research methods

The research method used to analyze data in this research is a method of Network Development Life Cycle (NDLC). The NDLC method used is cyclical to support continuous development [10]. The NDLC method consists of six stages of the mechanism required to design, build and develop a system, the six stages of the NDLC method i.e. analysis, design, simulation prototyping, implementation, monitoring and management. However, the method to be used in this research only focuses on network analysis and design analysis.

2.1. Network analysis

Analysis proxy logs can help network administrators in understanding the browsing behavior of computer users and in providing them with an overview of Internet usage in an organization. Furthermore, it can assist them in gathering evidence (data) left behind. Such evidence may involve excessive Internet usage for non-work purposes or access to websites that are incompatible with business processes and other illegal activities [11]. Processing log entries can produce useful summary statistics

about workload volume, document types and sizes, popularity of documents and proxy cache performance [12].

In this research, data analytics on the network or network analysis is done by collecting and analyzing data from Squid Proxy server logs of higher education where from the log data authors can view data reports of Bandwidth usage and Network Traffic for eight months starting from April to November 2017. (See Figure 1).

```
1494262802.687 560 194.158.47.145 TCP_MISS/200 327 POST http://differentia.ru/diff.php - ORIGINAL_DST/144.217.254.3 text/html1494262813.131 506 194.158.47.145 TCP_MISS/200 488 POST http://w.api.mega.co.nz/JnarVvVEckuf0hDvgYpERJ3LttqY - ORIGINAL_DST/31.216.147.147 text/plain1494262816.516 2 194.158.47.145 TCP_MISS/404 1527 GET http://www.unikom.ac.id/images/slider.png - ORIGINAL_DST/222.124.203.55 text/html1494262870.004 1367 194.158.47.145 TCP_MISS/200 601 GET http://detectportal.firefox.com/success.txt - ORIGINAL_DST/52.222.232.136 text/plain1494262899.388 4231 194.158.47.145 TCP_MISS/200 488 POST http://w.aq 931.505 756 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n114.userstorage.mega.co.nz/ul/P9ZcAIix8TpwgQJ35tBNFzrnq5Q7929AczNYY37VwISg3v2EuinM1zn_I1ni3vZCPNGMw56; 55/200 389 POST http://gfs270n088.userstorage.mega.co.nz/ul/YDh4TD4mMdxQAKXjixp2G91yLV3zoKJ2f3WuW1MwjmbCbcfmmHoeo70IXg8ypf783eRjKY15ac1f3lmmKzdg/0? - ORIGINAL_DST/31.216.147.145 application/octet-stream1494262951.549 1193 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n051.userstorage.mega.co.nz/ul/serstorage.mega.co.nz/ul/_ZCZ-wdvbuuFYgLEEXJ4v0YwZhuZk8s3ordH4ppMlx-eaIAp0n87V8opUGA5N3T5yy8rdei0YtecuX2hgKXnmQ/0? - ORIGINAL_DST/31.216.144.76 application/octet-stream1494262952.383 1916 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n144.userstorage.mega.co.nz/ul/KjHx88zvPS4vnnQ1L20dXbFAuXs8LEZct0ZjNB7x1217_yc4VnAqaDdNswLTXGBGhnc1xNm; 62954.534 1158 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n144.userstorage.mega.co.nz/ul/KjHx88zvPS4vnnQ1L20dXbFAuXs8LEZct0ZjNB7x1217_yc4VnAqaDdNswLTXGBGhnc1xNm; 62954.534 1158 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n024.userstorage.mega.co.nz/ul/GdSPxtY1Tm3AfWkPKC1cR5n0CjTc4D4yAWfuQxdt80vZg1lHwK_BIbCwYrWh0cfl3QjwuiQMj8x7YRxcldX6g/0? - ORIGINAL_DST/31.216.144.34 44.48 application/octet-stream1494262972.738 2254 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n076.userstorage.mega.co.nz/ul/MzUhcBbnBz2hgkPgM tet-stream1494262972.738 2254 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n014.userstorage.mega.co.nz/ul/T_M_sEeNT_3fAJVTaxHLGCN0Vz8W-GkRzbahfhZdsiuRvX1G1Tcl kxPKnKlMpfFztouY0iAGFyrQzERlrvjvgCtathGfKzpbZJYb4HM-Vq-Mm2qB4gC115GQ/0? - ORIGINAL_DST/31.216.145.107 application/octet-stream1494262974.867 1666 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n031. 144.43 application/octet-stream1494262982.229 1245 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n139.userstorage.mega.co.nz/ul/5_VxxR_tkLsZHQlsGPCRHGs-6KdS8W8uSf 262983.580 1637 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n116.userstorage.mega.co.nz/ul/19WLM9Yt1-1VQJQFismoQpAsRy1IPcgDbrBjP2IYYLdn5pExN0om4V0MsZXXhVYEsNj 144.61 application/octet-stream1494262988.898 924 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n112.userstorage.mega.co.nz/ul/MbeDYFR5EFfDuAJ1T36-Juuo642rxKCu6 262989.569 1165 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n140.userstorage.mega.co.nz/ul/qT672KkyrHye3AJp8W8o8xbcc3PbPB16VULW81f2XwuzfEokGays0k1XeFmVgjuYCh otjolisYk5W8oEpokU0v8s5kh052H1StYnkw/0? - ORIGINAL_DST/31.216.145.50 application/octet-stream1494262991.681 1127 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n099.userstorage.mega.co.nz/ul/JUJZLk w/0? - ORIGINAL_DST/31.216.145.22 application/octet-stream1494262993.158 4846 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n074.userstorage.mega.co.nz/ul/qFrXtmk_LZ3D2wLZtLxyzoYfUH5cp-kcdHGT9IXCW6X6EQXZmh40Q1Wb6Qag; 145 194.158.47.145 TCP_MISS/200 389 POST http://gfs270n071.userstorage.mega.co.nz/ul/yhDXLSLpmH-TxOgIwT2K41zQ4f5ibE1VVeYdk96xu0nziWgK0fA5S0Bqet3HQcAvIRG1gKg09TbdsKzDY8hW/0? - ORIGI s270n122.userstorage.mega.co.nz/ul/_y2oKFTFT_7tQLW40u7s_xVtgrGQD7Km9xAKoj3GUGx0mX0m08B13_kwBnk8ydlly9qX0RgmtZmrtPgqL_w/0? - ORIGINAL_DST/31.216.145.76 application/or g8EVbyXjy8teg3aJhwpPeTXGWIbqCh4Yr_wLhdf70-VQfIMATURP6p4LR1jM4f0bvZ8bo5ZM1M4S1W9KF_edQ/0? - ORIGINAL_DST/31.216.145.52 application/octet-stream1494263000.922 2553 194. hzLIoSfGQACZCzcko75YBL42Tolzc8bZgk1fFwRdInP1wk7Bg/0? - ORIGINAL_DST/31.216.144.63 application/octet-stream1494263002.855 1143 194.158.47.145 TCP_REFRESH_MODIFIED/200 80PhCk4M_t76QZw1cvtflbzfxm3Mmk8Q6Xu8_g/0? - ORIGINAL_DST/31.216.144.57 application/octet-stream1494263005.821 87 194.158.47.145 TCP_REFRESH_MODIFIED/200 4699 GET ht
```

Figure 1. Display data log server squid proxy of higher education.

However, the result of data from Squid Proxy server logs is still difficult to read, so to facilitate the process of data analysis, the author uses Sawmill as a tool for analytics solution. Sawmill is universal log file analysis and reporting. From the data analysis, the author can get statistical data in the form of bandwidth usage which can be viewed based on time, URL and file type that has been used by the user. Here is the result of data analysis by using the Sawmill [13]: (See Figure 2)

Mime types 02/Apr/2017 – 13/Oct/2017, 195 days

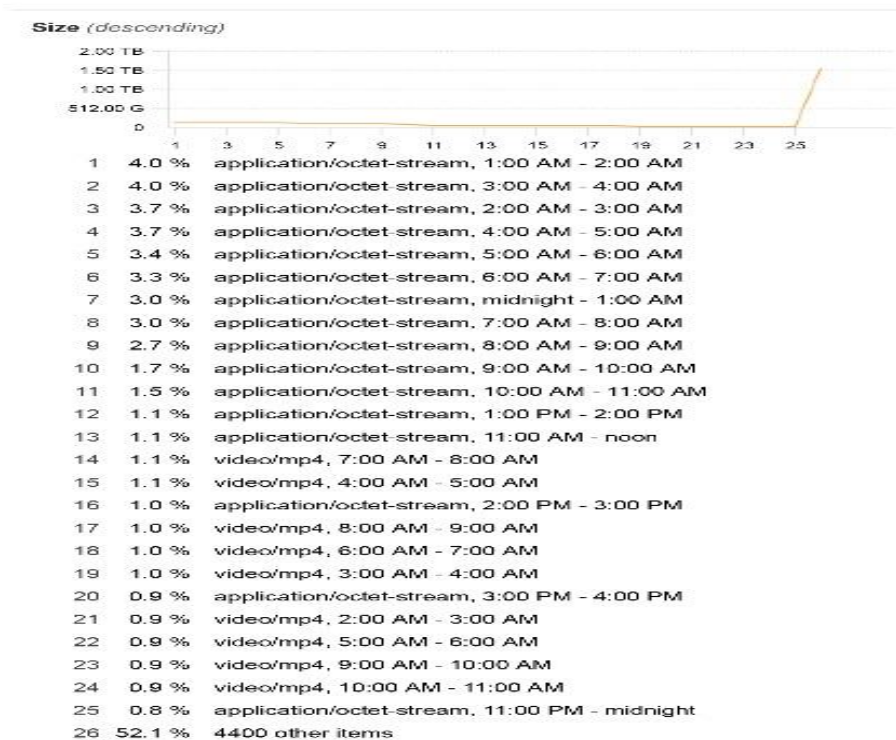


Figure 2. Display data squid proxy server logs by using Sawmill.

From the results of Squid Proxy server logs, authors can create a list of bandwidth traffic usage that has been used by the academic community such as officers, lecturers, employees, and students in the higher education environment. Here is the list of bandwidth usage of an academic community of higher education from April to November 2017: (See Figure 3).

Date	Lecturers Traffic (GB)	Employees Traffic (GB)	Top Officers Traffic (GB)	Students Traffic (GB)	Academic Activities
January-2017	462	150	481	938	Silent Week & Final Semester Test (UAS) in Odd Semester Academic Year 2016/2017
February-2017	404	188	725	703	Holiday Schedule after UAS in Odd Semester Academic Year 2016/2017 and the beginning of the trusteeship period
March-2017	403	197	949	1220	The beginning of class in Even Semester and Closing period of trusteeship and class selection
April-2017	481	236	333	1098	Silent Week and Mid Semester Test (UTS) in Even Semester Academic Year 2016/2017
May-2017	329	220	572	884	Normal class
June-2017	231	209	794	544	Normal class
July-2017	347	308	942	529	The last Class, Eid Al –Fitr Holiday, and UAS in Even Semester Academic Year 2016/2017
August-2017	322	302	514	451	Holiday Schedule after UAS in Even Semester Academic Year 2016/2017 and the beginning of the trusteeship period
September-2017	318	351	1044	608	The beginning of class in Odd Semester and Closing period of trusteeship and class selection – stage one Academic Year 2017/2018
October-2017	414	565	770	432	Normal class and Closing period of trusteeship – stage 2 for the student who fails the Autodebet process
November-2017	381	640	555	342	Normal class
Total	4,1 TB	3,4 TB	7,7 TB	7,7 TB	

Figure 3. Traffic bandwidth usage by academic community from April to November 2017.

To make the traffic bandwidth usage data in figure 3 easy to read, the authors make the graph of traffic bandwidth usage shown in figure 4.

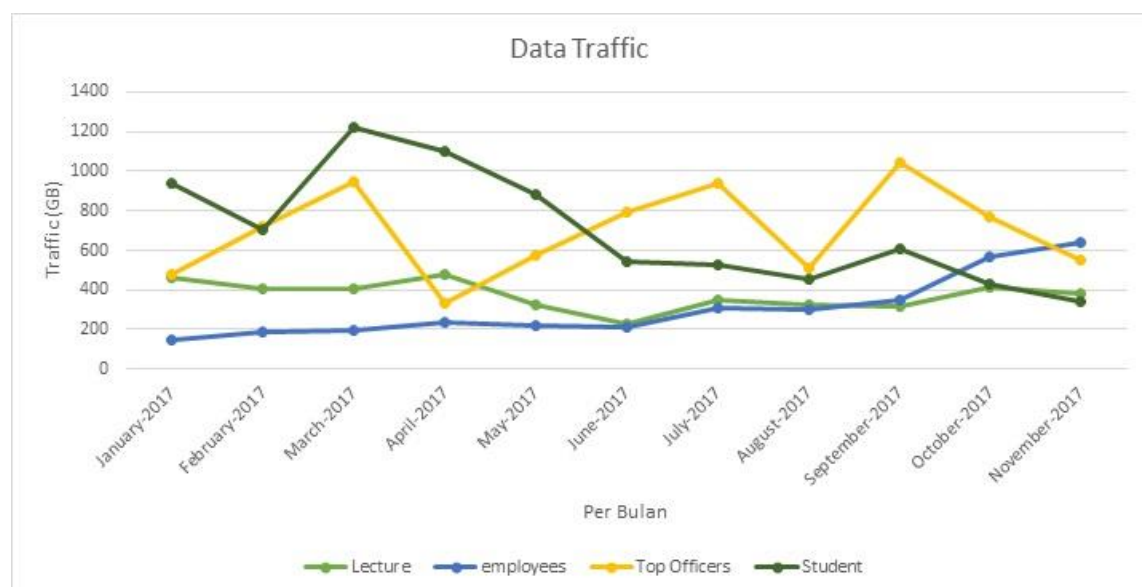


Figure 4. The graphic of traffic bandwidth usage.

After doing network analysis by collecting data of traffic bandwidth usage from Squid Proxy server logs, then the authors try to analysis learning patterns of user habit based on the research of Devajitet

al., Ronget al. and Cao et al using the result of graphic analysis on figure 4. Learn the patterns of user habits is important to do in order to maximize the traffic bandwidth usage and ensure the traffic bandwidth usage has been in accordance with business processes run by the organization, wherein this research the business processes must be in accordance with the rules run by higher education to access content or applications related to education or content that can support the research. To find out and learn the patterns of habits of traffic bandwidth usage, the authors make data details of traffic bandwidth usage that has been used by the academic community based on the content or applications that have been accessed. Here are the details of the pattern of traffic bandwidth usage: (See Figure 5,6)

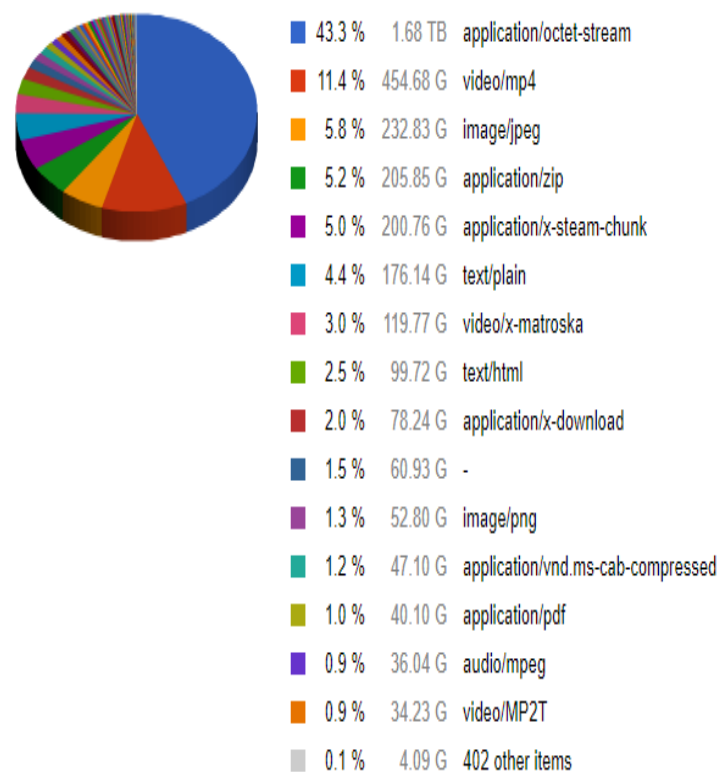


Figure 5. Traffic bandwidth usage: content, applications, and multimedia.

	URL	Mime type	Size	
1	http://www.uho.ac.id/file/	video/mp4	581.69 K	0.1 %
2	http://download.sbmpn.ac.id/media/	video/mp4	72.72 M	14.5 %
3	http://senit2017.poltektegal.ac.id/assets/	video/mp4	1.22 M	0.2 %
4	http://arsip.ipb.ac.id/videoarsip/	video/mp4	1.65 M	0.3 %
5	http://www.2025.unsw.edu.au/apply/	video/mp4	17.62 M	3.5 %
6	http://admisi.ipb.ac.id/pdss/	video/mp4	489.81 K	0.1 %
7	http://penmaba.unj.ac.id/assets/	video/mp4	5.63 M	1.1 %
8	http://penmaba.unj.ac.id/assets/	video/webm	841.87 K	0.2 %
9	http://edulms.chimpgroup.com/wp-content/	video/mp4	639.17 K	0.1 %
10	http://edulms.chimpgroup.com/wp-content/	video/webm	4.86 M	1.0 %
11	http://itsc.unikom.ac.id/new/	video/mp4	360.13 M	72.0 %
12	http://usm.itb.ac.id/Profil.mp4?	video/mp4	26.45 M	5.3 %
13	http://fisip.unpas.ac.id/wp-content/	video/mp4	3.11 M	0.6 %
14	http://uppi.stainkudus.ac.id/files/	video/mp4	4.18 M	0.8 %
Total			500.05 M	100.0 %

Figure 6. Traffic bandwidth usage: multimedia content related to education.

From the results of figure 5 and figure 6, the authors make an analysis and percentage of the results of traffic bandwidth usage. Where from the stage of analysis, the authors find the results that only about 1.98% of the total of traffic bandwidth usage is used to access links content, applications and multimedia files that are related to educational and research and 0% links that do not directly lead to content, applications and multimedia files which is related to education and 98.02% is used to access content, application and multimedia files unrelated to education or research. Here is the Percentage of Traffic Bandwidth Usage: (See Figure 7)

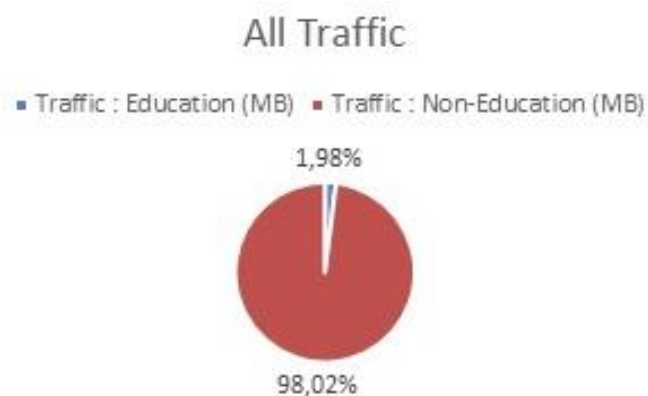


Figure 7. Percentage of traffic bandwidth usage.

2.2. Design analysis

The next stage is to make the process of design analysis where the design before applying the whitelist shown in this block diagram: (See Figure 8).

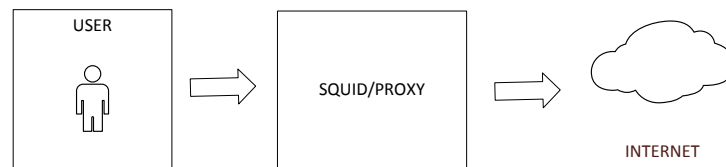


Figure 8. Schema design before applying whitelist.

Before applying the whitelist filters any user who will be accessing the Internet then the data will only be recorded in the Squid Proxy server logs and then the user can use the internet facility to access all content or applications. With such conditions, the bandwidth usage in higher education environment is not in accordance with the business processes, this can be seen in figure 7 where traffic usage for non-educational content reaches up to 98.02%.

Based on the results of data analysis and design analysis above then the Whitelist filter needs to be develop, so that the use of traffic bandwidth to access content or applications, especially multimedia files that require large-scale of traffic bandwidth can be controlled and maintained, so the traffic bandwidth usage can be maximized in accordance with business processes run by the organization.

3. Results and discussion

By using whitelist filter, it is expected that the traffic bandwidth usage to access educational content is higher than non-educational content, especially for educational content related to multimedia.

To develop a whitelist filter, authors make the design of network topology system or internet network scheme using whitelist filter. Here is the design scheme of network topology system that will be used in the development stage of whitelist. (See Figure 9).

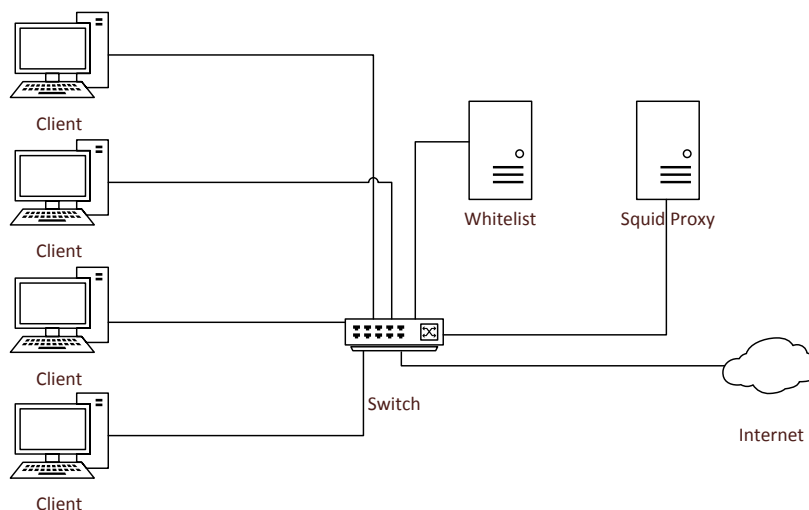


Figure 9. Design of network topology system.

After making the design of Network topology system that will be used, then the next stage is to make schema design using whitelist filter. Here is the scheme design using whitelist filter: (See Figure 10).

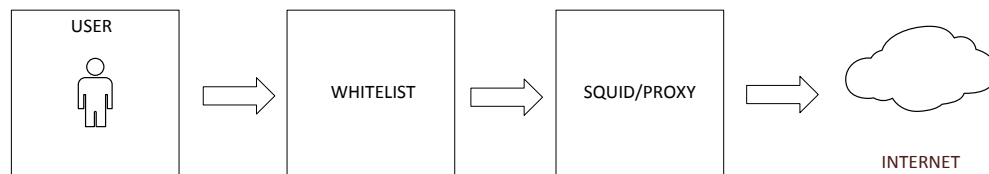


Figure 10. Schema design after using whitelist filter.

With the scheme design shown in Figure 10, every link that will be accessed by the user will be checked by the whitelist filter to view the content or applications has been in accordance with business processes in higher education or not, if the user has access the appropriate content then the user connected to the internet or link destination.

Based on the block diagram of the scheme design using whitelist filter, then the author makes engineering development of whitelist. This whitelist filter will be applied to the internet network at the higher education to maximize the use of traffic bandwidth especially for the content or multimedia application according to the business process run by the higher education, and here is a flowchart of whitelist that will be developed in this research: (See Figure 11).

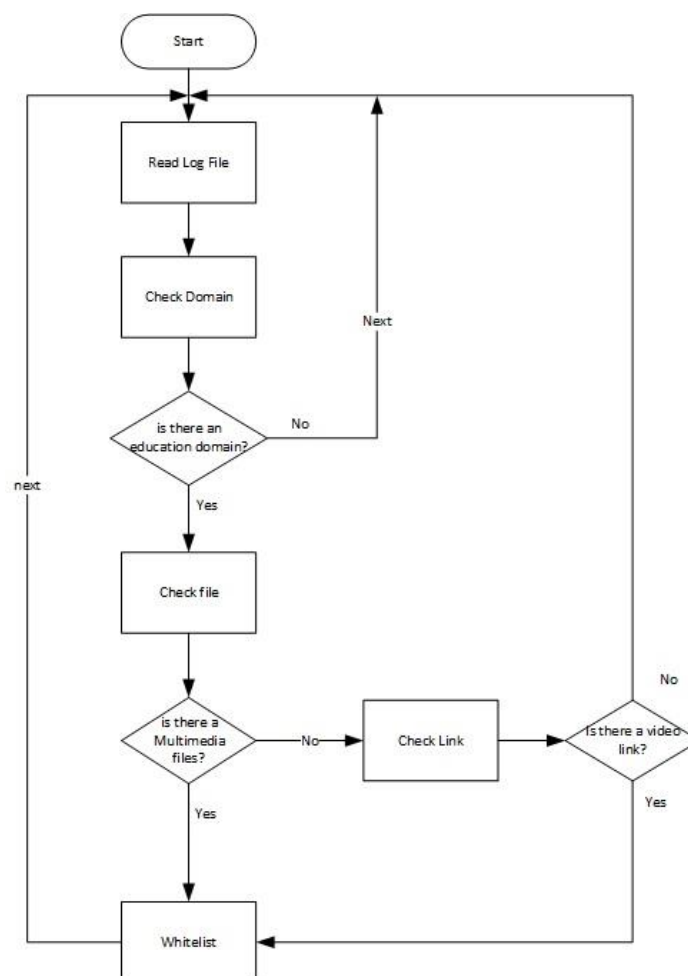


Figure 11. Flowchart of whitelist development.

Based on the flowchart of whitelist development in figure 11, the system will record and check the log file data and then check the domain to ensure that the domain has been accessed in accordance with business processes or related to education, if the domain is related to education then the system will do further checking to check if that domain has a multimedia file related to education, if the domain has a multimedia file then the domain will be added to the whitelist, but if the domain only contains video content link related to education then the link will also be added to the whitelist.

So, with the development of whitelist then any user can only access the domain that contains the file or multimedia link in accordance with the business process run by the higher education i.e. the domain that contains the file or a multimedia link related to education and research.

4. Conclusion

With the whitelist filter is expected to increase the number of access URL related to education and research and can detect the URL of multimedia related to education and research, so the percentage of the traffic bandwidth usage to access content, applications, and multimedia related to education and research that originally only 1, 98% is expected increase up to 99,9%. So, by using a whitelist filter can maximize the use of traffic bandwidth in accordance with business processes run by the higher education.

Acknowledgements

I would like to express my sincere gratitude to Universitas Komputer Indonesia (Unikom) and all staffs of Pusat Software Komputer, Unikom for the good cooperation in data collecting and finishing this research. I also would like to thank AP2SC of Unikom for the supportive guidance during the research, and thanks to the anonymous reviewers for suggestions and comments.

References

- [1] Gakio K 2006 African Tertiary Institutions Connectivity Survey (ATICS)-2006 Report.
- [2] Ikechukwu U I 2017 "A Survey on Bandwidth Management Techniques Via the OSI Model Network and Application Layers," *Global Journal of Computer Science and Technology*.
- [3] Mahanta D, Ahmed M and Bora U J 2013 "A study of Bandwidth Management in Computer Networks," *International Journal of Innovative Technology and Exploring Engineering* **2** (2).
- [4] Bommepally K, Glisa T K, Prakash J J, Singh S R and Murthy H A 2010 "Internet activity analysis through proxy log," *In Communications (NCC), 2010 National Conference on* (pp. 1-5). IEEE.
- [5] Atterer R, Wnuk M and Schmidt A 2006 "Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction," *In Proceedings of the 15th international conference on World Wide Web* (pp. 203-212). ACM.
- [6] Sirait E R E 2016 "Public demand to whitelisting: Another alternative for more secure internet access," *In Information Technology Systems and Innovation (ICITSI), 2016 International Conference on* (pp. 1-7). IEEE.
- [7] Sedgewick A, Souppaya M and Scarfone K 2015 "Guide to application whitelisting," *NIST Special Publication* **800** 167.
- [8] Lo J 2011 *Whitelisting for Cyber Security: What It Means for Consumers* (Public Interest Advocacy Centre).
- [9] McIntyre A, Lindqvist U, Peterson B and Tudor Z 2012 "Host protection strategies for industrial control systems," *In Homeland Security (HST), 2012 IEEE Conference on Technologies for* (pp. 87-92). IEEE.
- [10] Rianafirin K and Kurniawan M T 2017 "Design network security infrastructure cabling using network development life cycle methodology and ISO/IEC 27000 series in Yayasan Kesehatan (Yakes) Telkom Bandung," *In Computer Applications and Information Processing Technology (CAIPT), 2017 4th International Conference on* (pp. 1-6). IEEE.
- [11] Fei B, Eloff J, Olivier M and Venter H 2006 "Analysis of web proxy logs," *In Advances in Digital*

Forensics II (pp. 247-258). Springer, Boston, MA.

- [12] Mahanti A, Williamson C and Eager D 2000 "Traffic analysis of a web proxy caching hierarchy," *IEEE Network* **14** (3) 16-23.
- [13] Sawmill 2017 Sawmill - Universal Log File Analysis and Reporting [available at: <https://www.sawmill.net/>]