Recording and Storage Traffic Management in Storage Systems

Tatyana Tatarnikova[a], Ekaterina Poymanova[a] and Ekaterina Kraeva[a]

[a] *Russian State Hydrometeorological University, ul. Voronezhskaya, 79, 192007 St. Petersburg, Russia*

Abstract

The article discusses a complex solution for managing traffic recording and storage in data storage systems. In the conditions of modern legislation, the issue of storing a large amount of data becomes acute. Physical storage management avoids the unnecessary costs of scaling storage systems. The article proposes the structure of a hardware and software complex for managing physical data storage for storage systems that can be used by owners of technological communication networks to store traffic. Control mechanisms are considered, such as the distribution of data over various media using Kohonen neural networks and forecasting capacity extension using a statistical model and machine learning methods.

Keywords 1

traffic, data storage system, data distribution, physical data storage, machine learning, neural network, forecasting

1. Introduction

The requirements of modern legislation in the field of citizen security pose serious challenges to various organizations, including data storage. The anti-terrorist amendments adopted in 2016 (the so-called "Yarovaya law") obliged telecom operators to store traffic metadata for three years, and the traffic itself for six months. In addition, in June 2020, the Ministry of Digital Development, Communications and Mass Media of the Russian Federation proposed a bill, according to which the owners of technological communication networks are required to store traffic for three years [1].

There is also a legislative norm obligating to increase the capacity of traffic storages by 15% annually. Even though the government has postponed the introduction of this norm for 1 year, the problem of using and extending the physical resources of the storage is very acute.

The volume of Internet traffic over the past 4 years ranged from 32470.782391 PB to 61,226.217838 PB (Fig.1) [2]. That is, in just four years, the volume of traffic has almost doubled. Consequently, the above norm on the annual increase in storage capacity is insufficient, while its implementation requires significant financial costs.

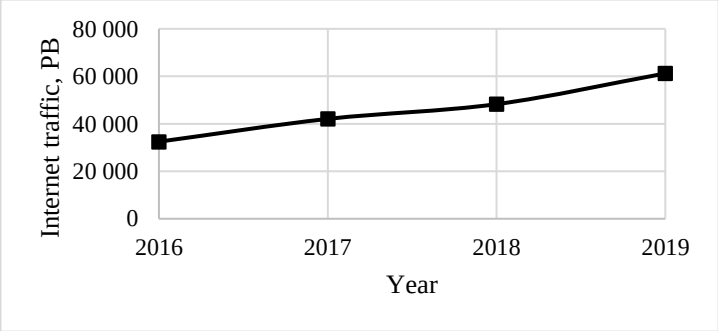

Figure 1: Increasing in the amount of traffic in 2016-2019

Proceedings of the 12th Majorov International Conference on Software Engineering and Computer Systems, December 10-11, 2020, Online & Saint Petersburg, Russia

EMAIL: tm-tatarn@yandex.ru; e.d.poymanova@gmail.com; kate.smitt.by@mail.ru

ORCID: 0000-0002-6419-0072; 0000-0001-9318-6454; 0000-0002-6938-1775

In October 2020, only “Rostelecom” spent 7.8 billion rubles on data storage equipment. Other operators also purchase various storage systems (table 1) [3].

Table 1
Costs of telecom operators for data storage systems

№	Month, Year	Company	Cost, $
1	May 2018	MegaPhone	12,64 million
2	March 2019	MTS	191,66 million
3	October 2020	Rostelecom	106,78 million
4	planned	Tele2	58,87 million

As can be seen from Table 1, telecom operators of the Russian Federation suffer serious financial costs for the purchase of equipment for storage systems. On the other hand, modern information technologies make it possible to manage the resources of data storage systems, use them efficiently and, therefore, avoid unnecessary costs.

The data storage system can manage the recording of the incoming data stream and, firstly, distribute it among different types of media, and secondly, monitor the state of the storage and make a forecast of capacity growth for its timely extension.

2. Physical Data Storage Managing During Recording and Storing Traffic

A research a study has been carried out in which a data storage system is considered as a storage management system that performs the following functions:

- Distribution of data files on various types of media, depending on the file size and storage time
- Monitoring the storage state based on snapshots of each media state
- Forecast of storage capacity extension. [4,5].

The storage management system diagram is shown in Figure 2.

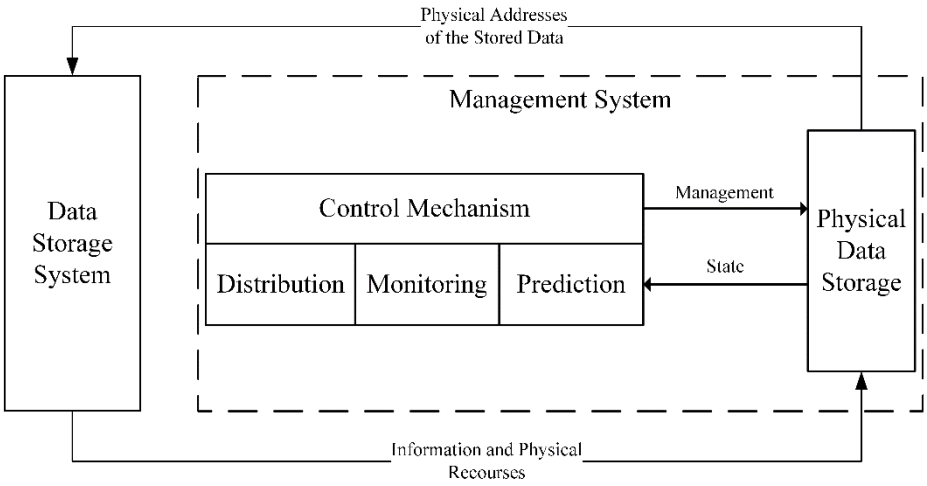

Figure 2: Data Storage Management System

Obviously, for the implementation of such a storage management system, a soft-ware-hardware system is needed that performs the above functions.

It is proposed to include a programmable logic controller (PLC) in this system, which distributes files to media and software that monitors the state of the physical storage and builds a forecast for its extension (Figure 3).

The controller receives an incoming data stream (for example, internet traffic). The controller performs clustering of incoming traffic using Kohonen's neural networks and, in accordance with the resulting topological map, distributes data files to media in the physical data storage.

Physical storage can be organized depending on the information being recorded. In paper [6] there was considered a 3x3 matrix storage and assumed the distribution of files first by one of the storage levels, depending on the storage time, and then - the distribution among the level volumes depending on the file size.

This solution can be easily adapted to the needs of the owners of technological communication networks [7]. Since the storage time of data files, as well as metadata, in accordance with the existing legislation and the upcoming amendments is limited to three years, data files can be distributed across various types of media de-pending on the type of data (text, sound, video) and size.

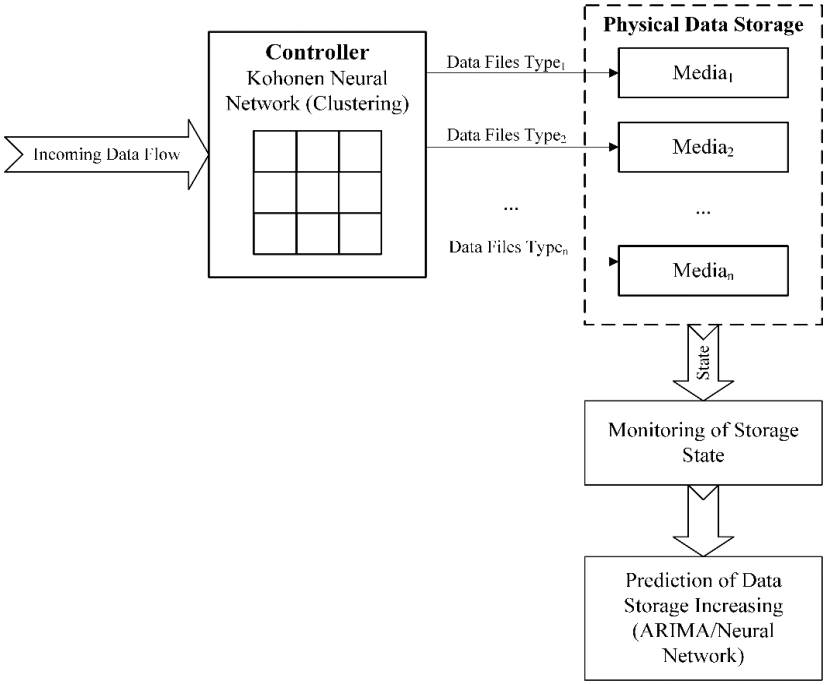

Figure 3: Structure of Hardware and Software System

The structure of physical data storage is determined by the storage system administrator and can contain, for example, RAID arrays for text files, streamers for audio and video files. In addition, volumes inside a RAID array can have different operating systems with different sizes of the logical data block, which will avoid the "under-filling" of files during writing (Figures 4, 5).

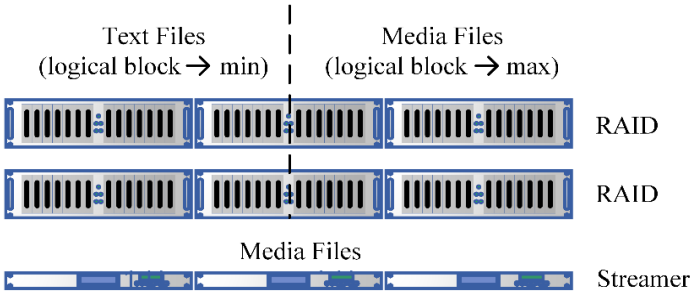

Figure 4: Structure of Physical Storage

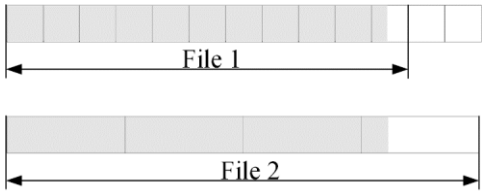

Figure 5: Comparing files with the same amount of data and logical blocks of different sizes

This partitioning helps conserve disk space on RAID arrays.

The state of the storage can be monitored based on the state snapshots coming from the physical data storage. These state snapshots should show the fullness of each media (media volume) of data in physical storage.

It is planned to build a forecast based on the monitoring data for the capacity ex-tension of physical storage. Wherein each storage tier containing a specific type of media is considered.

The capacity growth forecast is based on the model presented in fig.6 [8].

Use only styles embedded in the document. For paragraph, use Normal. Paragraph text. Paragraph text. Paragraph text. Paragraph text. Paragraph text. Paragraph text. Paragraph text. Paragraph text.

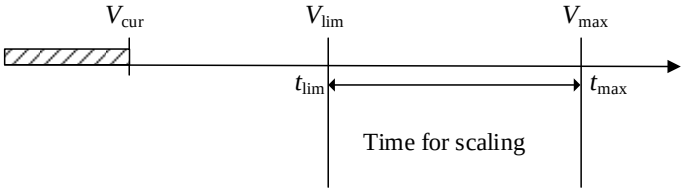

Figure 6: Characteristics of Data Media

$$V_{lim} = \int_{1}^{t_{lim}} f(t)dt = T\sum_{1}^{t_{lim}} f(t), \quad (1)$$

$$V_{max} = \int_{1}^{t_{max}} f(t)dt = T\sum_{1}^{t_{max}} f(t), \quad (2)$$

where t_{limmn} – time to reach limited media capacity;
t_{maxmn} – time to reach maximum media capacity;
$f(t)$ – incoming data function;
T – partition step equal to the unit of the minimum selected time scale.

The forecasting task is to find the timeline point at which the limited capacity and the maximum capacity of each media are reached [5].

To solve this problem, it is necessary to predict the amount of incoming traffic in the storage system.

The forecast can be made by various methods, while it is necessary to consider the peculiarities of the data stream entering the recording. Due to uneven user activity associated with weekends and working days, vacation periods, etc. the incoming data stream is heterogeneous and has a seasonal structure (Fig. 7)

In [6,9], a comparison was made between different forecasting methods: statistical forecasting using an autoregressive model and an integrated moving average (ARIMA) and machine learning methods. The results showed that the ARIMA model is the most suitable for short-term forecasts (Fig. 8), and for mid-term forecasts, machine learning methods (Fig. 9).

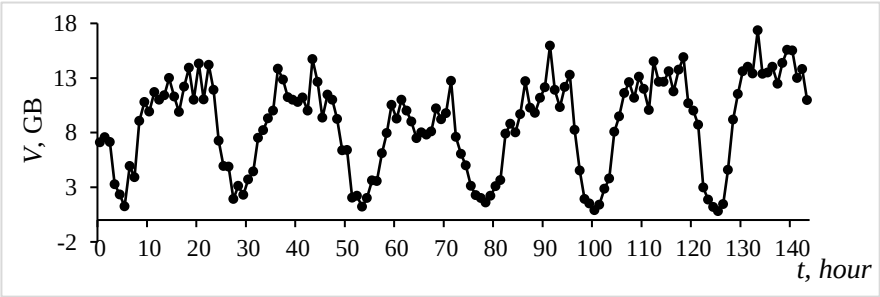

Figure 7: Incoming stream LTE traffic data by MTS

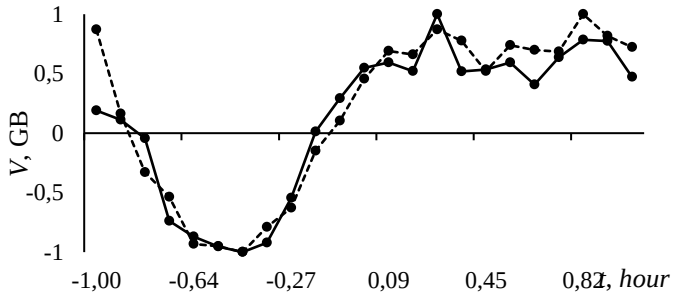

Figure 8: Demonstration of the difference between real traffic and predicted obtained using the ARIMA model (mid-term forecast): MSE=0,04

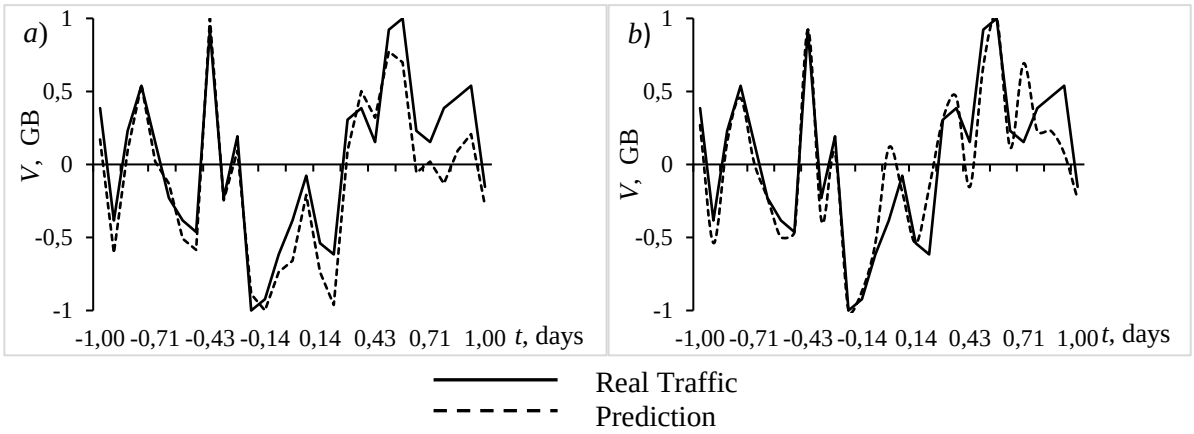

Figure 9: Demonstration of the difference between real traffic and predicted, obtained using the machine learning method: a – decision tree, MSE = 0,047; b – random forest, MSE = 0,047 (mid-term forecast)

To implement the forecast mechanism, an application was developed. This application helps automate the process of predicting the capacity extension of each cell of the storage matrix [10].

Thus, for the further implementation of the hardware-software system, it is necessary to develop a programmable logic controller that distributes files inside the physical data storage.

Programmable logic controllers are widely used in automatic control systems. The performance of modern controllers allows them to use the most efficient control algorithms, such as, for example, neural networks.

3. Conclusion

The norms of modern legislation oblige the owners of technological communication networks to store a large amount of data using their own data storage systems. This leads to serious costs, which, in the end, fall on the end user of communication services.

At the same time, modern technologies make it possible to create systems for managing physical data storage that can efficiently consume physical storage resources. Existing virtualization technologies make it possible to create structures containing various types of storage media and distribute the saved traffic files over them depending on certain characteristics of the files.

Since there is a need for regular scaling of the data storage, it is necessary to monitor its status and scale only those media whose capacity limits tend to be maximized. Predicting capacity extension allows for timely scaling.

Thus, dividing the total incoming data stream by media, predicting capacity consumption, and monitoring the state of physical data storage allow owners of technological communication networks to rationally use physical storage resources and avoid unnecessary costs when increasing storage.

4. References

[1] The Ministry of Economic Development supported the draft law on three-year storage of technological networks traffic [MER podderzhalo zakonoproyekt o trekhletnem khranenii trafika tekhnologicheskikh setey] https://tass.ru/ekonomika/9574021

[2] Communication networks exchange statistics https://digital.gov.ru/ru/pages/statistika-otrasli/

[3] Yarovaya's law has been strengthened in hardware. Newspaper "Kommersant" №185 (09.10.2020), p. 10 https://www.kommersant.ru/doc/4522028 [Zakon Yarovoy usililsya apparatno. Gazeta "Kommersant" №185 ot 09.10.2020, str. 10] https://www.kommersant.ru/doc/4522028

[4] Tatiana M. Tatarnikova, Ekaterina D. Poymanova. Algorithms for Placing Files in Tiered Storage Using Kohonen Map//Selected Papers of the IV All-Russian scientific and practical conference with international participation "Information Systems and Technologies in Modeling and Control" (ISTMC'2019) Yalta, Crimea, May 21-23, 2019. Pp. 193-202

[5] Tatarnikova T. M., Poymanova E. D. Differentiated Capacity Extension Method for System of Data Storage with Multilevel Structure// Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2020. T. 1. No 1. P. 66–73. doi:10.17586/2226-1494-2020-20-1-66-73

[6] Sovetov B. Ya., Tatarnikova T. M., Poymanova E. D. Organization of multi-level data storage. Informatsionno-Upravliaiushchie Sistemy [Information and Control Systems], 2019, no. 2, pp. 68–75 (In Russian). doi:10.31799/1684-8853-2019-2-68-75

[7] Bogatyrev V.A., Bogatyrev S.V., Derkach A.N. Timeliness of the Reserved Maintenance by Duplicated Computers of Heterogeneous Delay-Critical Stream//CEUR Workshop Proceedings, 2019, Vol. 2522, pp. 26-36

[8] Sovetov B. Ya., Tatarnikova T. M., Poymanova E. D. Storage scaling management model. Informatsionno-Upravliaiushchie Sistemy [Information and Control Systems], 2020, no. 5, pp. 43–49. doi:10.31799/1684-8853-2020-5-43-49

[9] Poymanova, E.D., Tatarnikova, T.M. Applying machine learning methods for forecasting In 2020 Wave Electronics and its Application in Information and Telecommunication Systems, WECONF 2020

[10] The Forecast Application for Capacity Extension of Data Storage Systems. Poymanova E.D., Tatarnikova T.M., Yagotintseva N.V. Computer Registration Certificate RU 2019661945, 12.09.2019. Application for registration № 2019619010 22.07.2019.