

THE TASK OF CLEANING UP RETROSPECTIVE DATA ON RESOURCE USAGE IN THE WORKLOAD FORECASTING SYSTEM OF A DISTRIBUTED INFORMATION SYSTEM COMPONENT

O. Ja. Kravets (1), Yu. V. Moiseeva (1), E. I. Mutina (2), O. Yu. Zaslavskaya (3),
A. V. Gorshkov (4), Yu. V. Redkin (5), Y. V. Atlasov (6)*

⁽¹⁾ Voronezh State Technical University, Voronezh; ⁽²⁾ Moscow State University of Technology "STANKIN", Moscow; ⁽³⁾ Moscow City University, Moscow; ⁽⁴⁾ *MAKS Research and Production Centre, Moscow*; ⁽⁵⁾ Admiral Ushakov Maritime State University, Novorossiysk; ⁽⁶⁾ Moscow University of Ministry of Internal Affairs of Russian Federation named by V. Ja. Kikot, Moscow;
Russian Federation

* Corresponding Author, e-mail: csit@bk.ru

Abstract: The article considers an approach to solving the problem of noise reduction of retrospective data on the level of utilization of computing resources of physical machines (using the example of processor cores) used as a training sample for a workload forecasting system of a virtualized data centre. The problems of noise reduction of retrospective data, such as the problem of "Noisy Neighbours" and Software Aging, are considered. A generalization of the subject area of data noise reduction for preprocessing time series of retrospective data based on signal processing methods, in particular, variants of mode decomposition methods, has been performed. Approaches to reducing the noise level of data based on the method of decomposition into variational modes (DVM) and variants of the method of decomposition into empirical modes (DEM) are considered. A hybrid scheme of the module for preprocessing a time series of retrospective data on the level of processor core utilization is proposed based on a comparative evaluation of the results of its variational mode decomposition and complementary multiple decomposition into empirical modes with adaptive noise in order to improve the quality of neural network forecasting of the workload of a virtualized data centre.

Key words: Virtualized data centre, load forecasting, use of computing resources, mode decomposition of signals, oscillatory mode functions.

1. INTRODUCTION

The development and improvement of information services, both general-purpose and with specialized functionality, has become the basis for the development and widespread use of virtualized data centres (VDC), which are the computing basis of

these services. VCCS are based on the paradigm of virtualization of computing resources and their provision as part of virtual machines and/or containers.

Effective implementation of such a solution is impossible without an appropriate management system for computing resources of physical machines. The workload of a VDC can be either periodic, based on linking its growth or decrease to the time of day or other time markers, or spontaneous, associated with sudden spikes in service requests to computing resources. Solving the problem of predicting the occurrence of such template workload values increases the efficiency of VDC management. The solution to this problem refers to proactive workload management methods based on workload forecasting by using the retrospective data on the level of computing resource usage stored by the VDC administration system.

The main goal of this paper is to present a hybrid scheme of the module for preprocessing a time series of retrospective data on the level of processor core utilization based on a comparative evaluation of the results of its variational mode.

2. PREVIOUS RESEARCH

The issues of proactive VDC workload prediction are quite widely presented in [1, 2]. Thus, in [1] an attempt is made to formalize the task of predicting the workload of the VDC, and in [2] an extended analysis and generalization of methods and algorithms for the implementation of this task is given. The solution implemented in the software of the VDC workload management system is based on statistical methods for processing retrospective data on the use of computing resources (Figure 1).

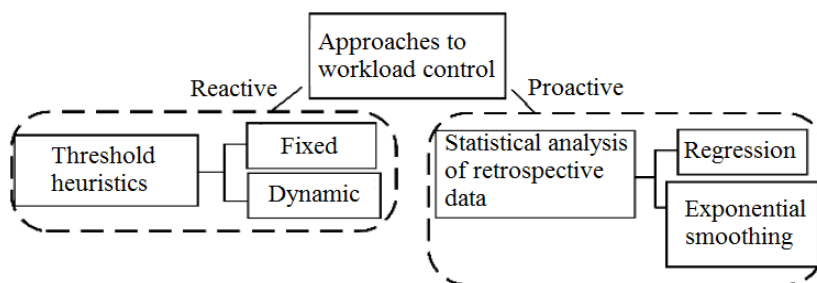


Figure 1. Overview of existing approaches to managing VDC workload

Retrospective data on the use of computing resources is the result of a resource monitoring system that tracks the values of their metrics with a specified frequency. Examples of such a system in modern large-scale VDC are [3]:

- AWS CloudWatch VDC Amazon Web Services;
- Azure Monitor VDC Microsoft Azure;
- Google Resource Manager VDC Google Cloud Platform.

One of the important tasks is to eliminate the noise component in the resulting implementation of the time series of values of the retrospective data in order to obtain an objective picture of the use of the selected type of computing resource. In [4], such a function of the workload management system is called "data clearing".

One of the most promising approaches to implementing the data cleaning function

is to represent the implementation of a time series of values of retrospective data in the form of a time domain of a signal, which allows using methods of its analysis to separate the useful and noise components of values.

3. THE PROBLEM OF NOISE IN THE TIME SERIES OF RETROSPECTIVE DATA ON THE LEVEL OF PROCESSOR CORE USAGE

The analysis of sources [4, 5] devoted to the problems of noise values of time series of data on the use of computing resources revealed two of the most significant factors associated with this process:

- noise associated with the problem of "Noisy Neighbors";
- the noise associated with the problem of "aging" of the software (Software Aging).

At the same time, from the point of view of the impact on the noise level of the values of the time series of data in the VDC, the problem of "noisy neighbors" is crucial, since it is related to the dynamics of the functioning of the VDC resource management subsystem.

Figure 2 provides a generalized representation of the occurrence of the "noisy neighbors" problem. The moment of time t is represented when three multithreaded processes P_1 , P_2 and P_j are assigned to processors (processor cores) 1, 2, n .

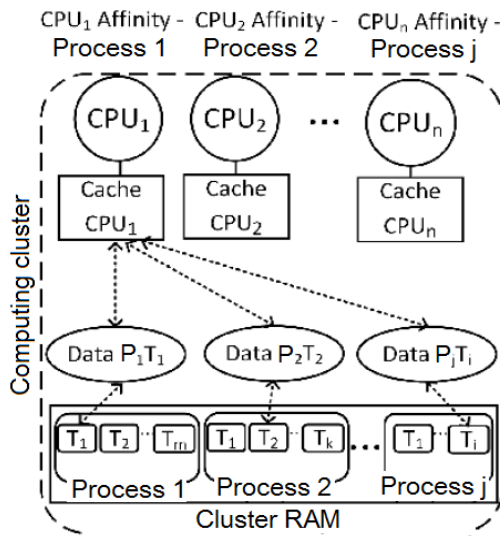


Figure 2. Generalized representation of the occurrence of the "noisy neighbors" problem in a multiprocessor VDC cluster

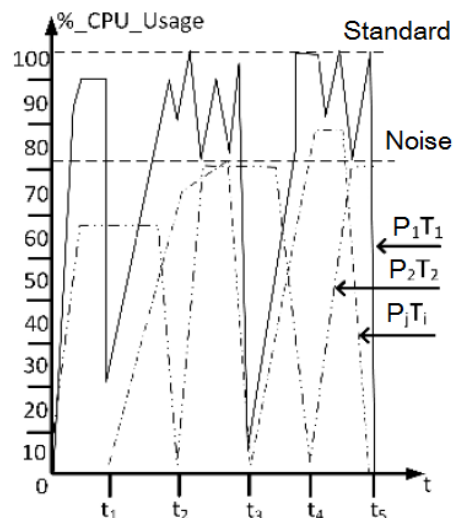


Figure 3. Visualization of the noise of the time series of stream 1 of process 1 by accessing the cache of "noisy neighbors": stream 2 of process 2 and stream i of process j

In this case, the thread T_1 places its data $\text{Data}(P_1T_1)$ in the cache of processor 1. However, at previous time points $t-1$ and $t-2$, the VDC load management system assigned processor 1 to processes P_2 and P_j , respectively, whose threads have placed

their data $Data(P_2T_2)$ and $Data(P_jT_j)$ in the cache of processor 1. Being re-anchored at time t behind processors 2 and n , processes P_2 and P_j , however, require their data to be retrieved from processor P_1 cache, which leads to the need to suspend process P_1 . Depending on the frequency of such cache access, the performance of the process will decrease, which will affect the time series of values of the "% CPU usage" metric of the monitoring system. In this case, processes P_2 and P_j act as "noisy neighbors" of process P_1 , "making noise" with their values over the time series of its values. A visualization of the process of such noise is shown in Figure 3.

4. PRELIMINARY PROCESSING OF RETROSPECTIVE DATA ON THE LEVEL OF UTILIZATION OF COMPUTING PROCESSING OF NOISY RESOURCES BASED ON THE MODE DECOMPOSITION OF SIGNALS

The ability of the resource monitoring system to collect the values of performance metrics of VDC computing nodes in the form of time series of their values and save them in the retrospective data storage system allows the resource manager to solve the problem of workload forecasting in specified time periods.

One of the approaches to solving the problem of clearing noisy data is to use methods of mode decomposition of a signal represented by a time series in a particular area. The methods of mode decomposition of the signal underline the Hilbert-Huang transformation [6]. Their basis is the decomposition of a complex signal into many simpler functions, called oscillatory mode functions OMF. OMFs are functions of a more general form, unlike the Fourier method, in which the signal decomposition is performed in the basis of trigonometric functions with a given frequency and phase. Each element of the OMF set is actually an amplitude-frequency modulation in a given narrow frequency band, which allows it to be associated with a specific process.

Thus, the retrospective data of the values of the time series of metrics for the use of processor cores can be represented by a set of OMFs in which the noise component will be separated or minimized relative to the base values of the series.

The most well-known variants of mode decomposition include the empirical mode decomposition (DEM) method and its variants and decomposition into variational modes (DVM) [7]. The DVM method, working in the frequency domain of a signal, is a computational procedure that solves a variant of a variational problem.

Let us consider the application of the DVM method to the decomposition of a time series of data from retrospective values of processor core usage metrics that determine the workload of the computing process $w(t)$. In fact, the method decomposes the time series $w(t)$ into k OMF functions, the OMF_k function has a limited bandwidth:

$$OMF_k = A_k(t) \cos(\varphi(t)) \quad (1)$$

where $A_k(t)$ is the value of the instantaneous amplitude of the $OMF_k(t)$ function, followed by $f_k(t)$. The value of the bandwidth $f_k(t)$ can be estimated using Gaussian numerical integration smoothing.

The DVM method is based on a variational model of the form:

$$\min_{(f_k), (c_k)} \begin{cases} \sum_{k=1}^K \left\| \partial_t \left[\sigma(t) + \frac{i}{\pi t} f_k(t) \right] e^{-j c_k t} \right\|^2, \\ \text{on condition } \sum_{k=1}^K f_k(t) = w(t) \end{cases}, \quad (2)$$

where ∂_t is the partial derivative of t , and the elements of the set $c_K = c_1, c_2, \dots, c_k$ are the values of the central frequencies of each element of the set of OMF functions.

To obtain an optimal solution to this variational problem, the Lagrange multiplier method with quadratic multiplication of the penalty function is used:

$$L(f_k, c_k, \lambda) = \alpha \sum_{k=1}^K \left\| \partial_t \left[\sigma(t) + \frac{i}{\pi t} f_k(t) \right] e^{-j c_k t} \right\|_2^2 + \left\| w(t) - \sum_{k=1}^K f_k(t) \right\|_2^2 + \left\langle \lambda(t), w(t) - \sum_{k=1}^K f_k(t) \right\rangle, \quad (3)$$

where λ is the Lagrange multiplier, α is the coefficient of the penalty function.

We obtain an extreme solution for the frequencies of each element of the OMF set and their central frequencies c_k :

$$\hat{f}_k^{n+1}(c) = \frac{\hat{w}(c) - \sum_{k=1, k \neq K}^k \hat{f}_k(c) + \frac{\hat{\lambda}(c)}{2}}{1+2} \alpha (c - c_k)^2 \quad (4)$$

$$c_k^{n+1} = \frac{\int_0^\infty c |\hat{f}_k(c)|^2 dc}{\int_0^\infty |\hat{f}_k(c)|^2 dc} \quad (5)$$

To obtain an optimal solution to expression (4) for all elements of the OMF set, the well-known Alternating Direction Method of Multipliers ADMM [8] is used, which decomposes a complex variational problem into particular subtasks. Its stages are:

1. Setting zero parameter values $f_k(t)$, c_k , λ_k^1 , n .
2. Updating the values of $f_k^{n+1}(t)$ and c_k^{n+1} (expressions (4) and (5)).
3. Updating the value of the Lagrange multiplier operator λ^{n+1} :

$$\hat{\lambda}^{n+1}(c) = \hat{\lambda}^{n+1}(c) + \tau \left(\hat{f}(c) - \sum_{k=1}^K \hat{f}_k^{n+1}(c) \right) \quad (6)$$

4. Steps 2-3 are repeated until a value $\hat{f}_k^{n+1}(c)$ is obtained if the following

condition is met: $\sum_{k=1}^K \frac{\left\| \hat{f}_k^{n+1} - \hat{f}_k^n \right\|_2^2}{\left\| \hat{f}_k^n \right\|_2^2} < \varepsilon$.

The DVM method, being a computational procedure, will demonstrate more accurate results (the type of elements of the OMF set), which significantly affects the quality of further recognition of workload patterns. However, this method is computationally complex and requires a preliminary determination of the power of many OMF functions.

5. A HYBRID APPROACH TO PREPROCESSING NOISY RETROSPECTIVE DATA ON THE LEVEL OF COMPUTING RESOURCE UTILIZATION

Let us consider the proposed hybrid approach to preprocessing a time series of retrospective data on the level of utilization of computing resources of VDC physical machines, which is a component of the VDC neural network workload forecasting system. In general, the block diagram of this approach is shown in Figure 4.

The $w(t)$ VDC workload forecasting system consists of two modules:

- preprocessing of the time series $w(t)$;
- the forecasting module $w(t+1)$.

The input data of the system is a time series $w(t)$, represented in particular by the values of the metric "% CPU Usage" in time ranges: t_0-t_1 ($w_1(t)$), t_1-t_2 ($w_2(t)$), t_2-t_3 ($w_3(t)$), stored in the VDC load retrospective data storage system. It is assumed that the values of the specified time series are noisy.

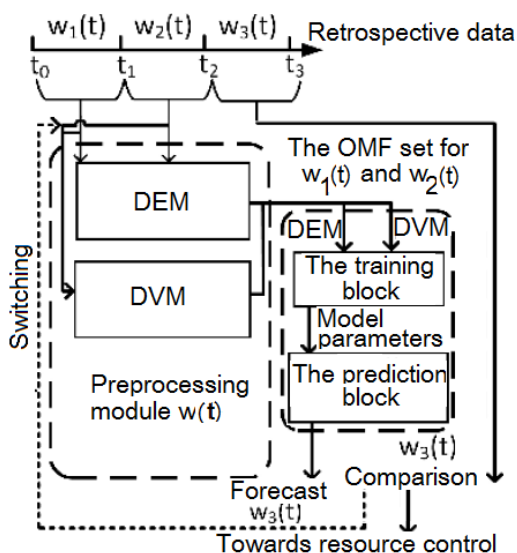


Figure 4. Block diagram of the neural network workload forecasting system $w(t)$ with a hybrid module for preprocessing its time series

In this case, for the forecasting module, the values $w_1(t)$, $w_2(t)$ represent the training sample (forecast $w_2(t)$ based on analysis $w_1(t)$), and the values $w_3(t)$ represent the control sample.

To reduce the degree of noise of the training sample data ($w_1(t)$, $w_2(t)$), their values are fed to the input of the data preprocessing module, in particular to the DEM block implementing the hybrid method. As a result of iterations, the DEM block sequentially generates two sets of OMF functions for $w_1(t)$ and $w_2(t)$, which are input data for the training block of the forecasting module. The result of its operation is the parameters of the neural network model, which are transmitted to the prediction block.

The prediction block analyses the OMF data values of the time series $w_1(t)$ and $w_2(t)$ generates forecast values of the time series $w_3(t)$, which are compared with control values $w_3(t)$ from the VDC load retrospective data storage system. If they match according to a reasonably chosen proximity criterion, the values are transferred to the resource manager for subsequent proactive workload management. Otherwise, a switching signal is sent to the preprocessing module, and the values of the time series data $w_1(t)$, $w_2(t)$ are retransmitted to the DVM block, which forms a second set of OMF functions for the training block [9].

Thus, if the forecast results turn out to be unsatisfactory when implementing a less computationally expensive hybrid method, sets of OMF functions for $w_1(t)$, $w_2(t)$ re-obtained using the DVM method.

6. CONCLUSION

The article is devoted to the study of the problem of predicting the workload of virtualized data centres in conditions of noisy values of its metrics due to objective factors related to the resource management process of physical machines. The problems associated with the noise of workload data, in particular the problem of "noisy neighbours" and the "aging" of software, are summarized.

To solve these problems, in order to improve the quality of forecasting, an approach is proposed to pre-process retrospective (pre-stored) workload data represented by time series of performance metrics, for example, using processor cores of physical machines. An approach to reducing the noise component of the time series values is presented based on a class of methods for the mode decomposition of signals. The features, advantages and disadvantages of methods of decomposition into empirical and variational modes are considered.

A hybrid approach to preprocessing a time series of retrospective data on the level of processor core utilization based on the use of variants of the method of modal signal decomposition is proposed.

REFERENCES

- [1] Sepich N. et al. The impact of task workload on cybersickness. *Frontiers in Virtual Reality*, vol. 3, art. no. 943409, 2022. DOI: 10.3389/frvir.2022.943409.
- [2] Naveen S. et al. *Workload Consumption Metric Forecasting*, 2023. DOI: 10.52458/978-81-955020-5-9-104.
- [3] Dubey P, Tiwari A.K., Raja R. *Compute Services in AWS*, 2023. 132 p. DOI: 10.2174/9789815165821123010009.
- [4] Tugnait J. A Data-cleaning approach to robust multisensor detection of improper signals. *IEEE Access*, vol. 7, pp. 124956-124966, 2019. DOI: 10.1109/ACCESS.2019.2938856.
- [5] Jia X. et al. Analysis of noise and its characteristics in avalanche photodiode. *AIP Advances*, vol. 9, no. 14, 2024. DOI: 10.1063/5.0229293.

- [6] Chen C. et al. Exploring the relationship between pavement surface texture and friction based on Hilbert-Huang Transformation. *Tribology International*, vol. 187, art no. 108737, 2023. DOI: 10.1016/j.triboint.2023.108737.
- [7] Jinsong Y. et al. A response reconstruction method based on empirical mode decomposition and modal synthesis method. *Mechanical Systems and Signal Processing*, vol. 184, art no. 109716, 2023. DOI: 10.1016/j.ymssp.2022.109716.
- [8] Xie M., Guo H. Self-Adaptive Alternating Direction Method of Multipliers for Image Denoising. *Applied Sciences*, vol. 14, no. 22, art no 104272024. DOI: 10.3390/app142210427.
- [9] Romansky R. Empirical evaluation of the transfer of information resources in active learning. *International Journal on Information Technologies and Security*, vol. 15, no. 1, 2023, pp. 39-48. DOI: 10.59035/WZFU1905.

Information about the authors:

Oleg Jakovlevich Kravets – Dr. Sci (IT), professor of Voronezh State Technical University, areas of scientific research - system analysis, optimization, simulation of complex objects.

Yuliia Vladimirovna Moiseeva – postgraduate student of of Voronezh State Technical University, areas of scientific research – system analysis, optimization, simulation of complex objects.

Elena Igorevna Mutina – PhD in Technical Sciences, associate professor of Moscow State University of Technology “STANKIN”, Moscow, areas of scientific research – system analysis, optimization, simulation of complex objects.

Olga Yurievna Zaslavskaya - Dr. Sci (Pedagogy), professor of Moscow City University, areas of scientific research – information technology, system analysis, modelling.

Alexey Vladislavovich Gorshkov – PhD in Technical Sciences, General manager of MAKS Research and Production Center, Moscow, areas of scientific research – system analysis, optimization, simulation of complex objects.

Yuriy Viktorovich Redkin – PhD, associate professor of Admiral Ushakov Maritime State University, Novorossiysk, areas of scientific research - wireless networks, digital signal processing, control and data processing systems.

Igor Victorovich Atlasov - Dr. Sci (Phys.-Math.), professor of Moscow University of Ministry of Internal Affairs of Russian Federation named by V. Ja. Kikot, areas of scientific research – system analysis, optimization, simulation of complex objects.

Manuscript received on 04 December 2025