Cluster computing performance in the context of nonextensive statistics

Paweł Dymora¹, Mirosław Mazurek¹, Rafał Gawron¹

¹(Department of Distributed Systems, Rzeszow University of Technology, Poland)

Abstract: This work presents the problem of clusters computing performance in the context of nonextensive statistics. Simulation studies have demonstrated that the cluster computing process is self-similar or long-range dependent and the Hurst coefficient estimation can be used to specify computing performance. The paper presents a comparison of several methods for estimating Hurst coefficient, characterized by simplicity of implementation and convergence results.

Keywords: long-memory effect, complex system, cluster, stationarity, Hurst coefficient, nonextensive statistics

I. INTRODUCTION

Nowadays, computers are an integral part of our lives. We are witnessing a significant technological progress and the creation of more and more new services. We have more and more efficient devices, and today mobile phones have the computing power comparable to that recently used servers. The demand for computing power continues to increase, and when individual servers are not able to provide us with the required level of service, are combined into groups called clusters. One group of clusters, providing high computing power are HPC clusters (High Performance Computing) [8, 10]. They represent an important part of business, economy, industry and science. The offered range of its possibilities is very wide but the demand for computing power continues to grow. Initially used for the calculation of individual servers, later mainframes and when they also were not sufficient it began to connect servers in a clusters. An important feature associated with the performance of the entire computing system is adequate assessment of its effectiveness and efficiency. It should be noted that not only the architecture and interrelationships of the various elements on the physical level is important, but the operating environment and the software is also important. Moreover, the input data used in the computation must be specially prepared to obtain maximal computation power of the cluster. Therefore, in addition to attempts of building and configuration of the efficient computing cluster a very important part is carrying out of various performance tests. However, we come to the point where adequate analysis of the obtained simulations results is a key element [7, 9, 10].

The computer or cluster systems consists of many interdependent subsystems. There are two possible approaches for their analysis. The first one is based on the idea of reductionism proposed by Descartes. Error! Reference source not found. Reference source not found. and can be considered as a still ruling paradigm in the case of computer science and engineering. The second one can be related to the still new idea of the complex systems approach, where in order to understand the behaviour of such systems one needs to have the knowledge about behaviour of system components and also, that is more important, how they act together [4, 5]. This specific paradigm change can be even shown in the case of the idea of Turing machines and new approach for considerations in the case of interactive processing – even the opinion that “the computer engineering is not a mathematical science” was presented. It should be noted that Turing machine is a mathematical idea while its implementations are the physical ones, but if the physical nature of computer systems was indicated, there is a need to have an appropriate physical (thermodynamical) basis for deliberations in such a case [2, 5]. In [11] it has been shown that the analysis of processes in computer systems can be based on non-extensive thermodynamics. However, this analysis considers only spatial correlations, meanwhile in this paper we will focus cluster computing performance in the context nonextensive statistics [6].

II. METHODOLOGY OF THE SYSTEM RESEARCH

The article presents results of the research on the analysis of the computational load on the cluster system in the context of nonextensive statistics. The obtained results allow to specify the characteristics of the system with regard to the onlytoyshardware architecture and software, but also in relation to the processed input. In this context, we can say that the computing system presents the characteristics of the complex or simple system which has a direct impact on the understanding of the processes occurring in it. The result of knowledge can be used to optimize the entire computing system and subsystems. The complex system is a theorem difficult to unambiguously explain. It is a system which consists of many diverse and autonomous, but mutually related and independent components joined together. It can be characterized by properties not directly arising from its basic components, making it difficult to even not possible to describe the method of classical physics or mathematics. Research on the analytical description of methods of systems with nonextensive characteristics of data and properties led to the creation on statistical mechanicals based on the self-similarity. Using these methods we are able to adequately assess the characteristics of the system and to calculate the specific parameters of the system, which makes the system more than the sum of the parts. It is difficult to predict the effects of complex system behaviour because these results are not very sensitive to initial data and small noise [1, 3]. The system is constantly evolving and changing, and the number of interactions between the system elements is essential for the characteristics of the system. In order to determine whether we are dealing with complex or simple system must consider the statistical properties of the system: stationarity, autocorrelation and self-similarity index [3].
The first examined property of the system is stationarity. In the simplest terms, the process can be considered stationary if examining its properties at any time will have the same value. We can distinguish strict stationarity and stationarity in a wide range. Time series are strictly stationary if any allowable data \( \{t_1, \ldots, t_k\} \) and any \( h \in \mathbb{Z} \) cumulative set \( \{x_{t_1}, \ldots, x_{t_k}\} \) distribution is identical to the set \( \{x_{t_1+h}, \ldots, x_{t_k+h}\} \) distribution. For the strict stationarity, the mean value and variance (also called variation measure) are constant in time. For the time series with finite variances, weakly stationary or otherwise called stationary over a wide range, the average of its elements, is constant in time and autocovariance \( \{s, t\} \) depends only on the difference \( h = |t - s| \). Stationarity condition in a wider sense of random process is that the first and second moments would not change in time. In other words, if successive values change in time, the process is non-stationary.

Another analysed parameter is the autocorrelation. Autocorrelation is a statistical tool for analysing the function describing the degree of a given time series elements depends on previous values in the same time series. In statistics, the autocorrelation of a random process describes the process correlation at different time points. When \( x_t \) the value of the process in time, \( t \) is equal to the average value \( \mu \), then the obtained result must match the trend line.

Self-similarity index also known as Hurst coefficients is a real number taking values from 0 to 1. Describes the behavior of the process in time and is associated with the autocorrelation and fractal dimension. However, it is problematic, to precisely calculate it so that we know that the process is weakly stationary or otherwise called stationary in the same time series. In other words, if successive values change in time, the process is non-stationary.

Another analysis of parameter studied is the autocorrelation. Autocorrelation is a statistical tool for analysing the function describing the degree of a given time series elements depends on previous values in the same time series. In statistics, the autocorrelation of a random process describes the process correlation at different time points. When \( x_t \) the value of the process in time, \( t \) is equal to the average value \( \mu \), then the obtained result must match the trend line.

For Hurst exponent, the following formula can be performed with four methods: absolute moments, aggregated variance, periodogram and method of residuals of regression. The method of absolute moments is the aggregated time series of the original time series \( X = \{X_t, i \geq 1\} \) by dividing into blocks of size \( m \) and averaging operation:

\[
X^m(k) = \frac{1}{m} \sum_{i=|k-1|m+1}^{km} X(i)
\]

where \( X^m(k) \) denotes the aggregated time series and \( k \) is the index of the block. Then, for the prepared data, the sum of absolute moments of the aggregated time series is calculated with the following formula:

\[
\frac{1}{N/m} \sum_{k=1}^{N/m} |X^m(k)|
\]

where \( N \) is the number of data in the test series.

Method of aggregated variance unlike the previous simplifies that instead of calculating the sum of absolute moments of aggregated series, calculates the variance of sampling rate according to the formula:

\[
\text{Var} X^m = \frac{1}{m} \sum_{k=1}^{m} (X^m(k))^2 - \left( \frac{1}{m} \sum_{k=1}^{m} X^m(k) \right)^2
\]

Calculations using the residuals of regression begin with data division of the blocks with size \( m \), and within each block, we calculate the sum of partial data which will be marked as \( Y(i), i = 1, 2, \ldots, m \). Then, the obtained result must match the trend line calculated by least square method. In this case, the residual sum is calculated as the point of deviation from the given reference point in a created trend line. Residuals sampling variance is calculated with the following formula:

\[
\text{Var} = \frac{1}{m} \sum_{i=1}^{m} (Y(i) - i * a - b)^2
\]

where \( a \) and \( b \) are coefficients of the fit line. The obtained results should be averaged for large \( m \), the calculation result is proportional to \( m^2 \).

The first task in order to examine the cluster computing performance with the use of nonextensive statistics was to
build the HPC cluster and perform a series of tests to gain data for further analysis. For the base cluster nodes we selected following machines: the Sun Fire X4270 and Sun Blade X6240. The cluster is based on the Beowulf architecture [9], combining servers with different processor architecture. This carries the risk associated with the uneven work of both servers and thus also lowers results in some performance tests. However, it permits building of universal cluster configuration which does not require the configuration of specialized and identical devices. Both servers are connected to a Gigabit Ethernet network with the use of the following device: Alcatel OmniSwitch 6850. The required network bandwidth for configuration and testing of designed cluster is 1 Gb/s. The less network bandwidth prevents form using the full CPU computing power, creating a bottleneck of the system. Figure 1 shows a logical diagram of the designed HPC cluster.

The designed HPC cluster configuration schema.

**IV. RESULTS OF INVESTIGATIONS**

In this section we present the calculation of the statistical properties of the cluster computing system on the base of data collected during performance tests. We want to obtain the information whether the cluster computing system has the characteristic of the complex or the simple system. In order to do this we must determine few statistical features of the system on the base of its computation process characteristics. The first examined feature was to determine whether the process is a stationary or not. The study uses data in the form of system load time series. Analysing the stationarity chart presented in Fig. 2 the first visible thing is that characteristics of the graph is similar to the original load chart based on the performed computing system load analysis.

Graph presented in Fig. 2 shows that our tested process is not stationary. In addition to stationary graph also shows the variation characteristic of the tested process. The highest value increase can be observed during a drastic decrease in load due to the completion of the performed tests.

The next studied parameter is the autocorrelation. Results are definitely different from zero which indicates the correlation of the tested process. The obtained data are shown in Fig. 3.

Hurst parameter study was performed with the use of four methods: absolute moments, the aggregate variance, periodogram and residuals of regression methods. All analytical methods were implemented in a computer program. When the studies
were performing on the original test data we did not get the expected results, because the obtained values did not fit into the expected ranges. Then, according to the fractional Brownian motion studies were carried out on modified data presented as increments calculated from the original test time series. Diagram in Fig. 4 shows the increases which were calculated and used for further analyses of the Hurst coefficient.

The first method – the Hurst coefficient estimation with the absolute moments method was performed with the use of increases data (other methods also). The graph in the Fig. 5 shows the characteristic of the Hurst coefficient on the selected range of samples because in the whole samples range the graph would be illegible. Hurst coefficient estimation with the absolute moments method result: H = 0.657836.

The second method – the Hurst coefficient estimation with the aggregate variance method was performed. The obtained results were presented in the Fig. 6, the graph shows only a selected samples range. The estimated Hurst coefficient for the whole data range is: \( H = 0.997984 \).

The third method – the Hurst coefficient estimation with the residuals of regression method was performed and presented in the Fig. 7. The estimated Hurst coefficient for the whole data range is: \( H = 0.932676 \).
The last method – the Hurst coefficient estimation with the periodogram method was performed and presented in the Fig. 8. The estimated Hurst coefficient for the whole data range is: $H = 0.969878$. During program execution we must additionally define the parameter which denotes the cutoff point.

V. CONCLUSION

In this paper we have presented a methods of nonextensive statistics used for performance analysis of a cluster computing system. The main purpose of these methods is classification of the systems as a complex or simple system category. The complex system is characterized by a lack of stationarity, autocorrelation and Hurst exponent belonging to $0 < H < 1$ and $H \neq 0.5$. The analysis of the stationarity in Fig. 2, shows a graph resembling the tested time series. Successive values don’t still remain at the same level so that we can conclude that the test process is nonstationary. Another studied system feature was autocorrelation. The obtained values are different from zero, showing that the tested data are characterized by a correlation of data, such as in complex systems. The last studied property is self-similarity index, which was obtained using four methods. In most cases, the results were in the range of $0.5 < H < 1$ qualifying tested process as a process with a long memory. All presented properties and their estimation proves that created and tested HPC cluster system belongs to the group of complex systems. It can be concluded that the tested system is sensitive on the input data and the disruption caused by the other software operating in system or hardware limitations. Based on the obtained results and the knowledge that our system is classified as complex, the process of optimizing a computing cluster as also standalone servers should take into account the characteristics of long-term processes, in particular the effect of long memory. Therefore, in addition to the selection of appropriate physical elements as increased memory, cache or choosing the right processor also are important parameters of specific programs code or configuration parameters. The selection of the appropriate parameters were made by a number of test runs, configuration trials and appropriate modifications of testing algorithms. With this actions we have achieved a partial improvements of the next tests results, and thus the overall performance improvements of the of the cluster system.

The research presented in this work was done on the hardware that was supported by Grant POPW.01.03.00-18-012/09 The Operational Program Development of Eastern Poland.

REFERENCES


