International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol. 3, Issue. 6, Nov - Dec. 2013 pp-3872-3877 ISSN: 2249-6645

Cluster computing performance in the context of non extensive statistics

Paweł Dymora[1], Mirosław Mazurek[1], Rafał Gawron[1]
[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 configurationof 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.** 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 resultsof the research on theanalysis of the computationalload on the cluster systemin the context ofnonextensivestatistics. The obtained resultsallows to specifythe characteristics of the systemwith regardnot onlytoitshardware architectureorsoftware, butalso in relation totheprocessed input. In thiscontext, we cansaythatthecomputing systempresents the characteristics ofa complexorsimplesystemwhich has a directimpact on theunderstanding of the processesoccurring in it. The resultin knowledge can be used tooptimize the entirecomputing systemandits subsystems. Thecomplex system is a theorem difficultto unambiguouslyexplains.It is a systemwhich consists ofmany diverseand autonomous, butmutuallyrelatedandinterdependentcomponentsjoined together.It canbe characterized bypropertiesnotdirectly arising fromits basic components, making it difficultor evennotpossibleto describeusing themethodsof classicalphysics ormathematics.Research on the analyticaldescriptionmethodsof systems withverylarge amounts of dataand propertiesled to the creationof statistical mechanicsbasedon the self-similarity. Using thesemethodswe areindeedable to accuratelyanswerandcalculate thespecificproperties, butwith a certain probabilityand precisionwe can answera series of questionsrelated to oursystem. We have to agreewith the statementof the Aristotle: "*The whole ismorethan thesum of the parts*"which, in relationto the subjectof complex systemsmakes us awarethat the systemsarenon-additiveandnon-linear. It is difficult topredict theeffectsof complex systemsbehaviourbecause theoverall resultsare very sensitive toinitial dataand small noise [1, 3]. The systemis constantly evolvingandchanging,and the number ofinteractions between the system elements causes that for the characteristics calculationsin this type of systemsweuseprobabilistic methods.In order to determinewhetherwe are dealing witha complex or simple systemwe must considerthreestatistical propertiesof the system:stationarity, autocorrelationandself-similarityindex [3].

The first examined property of the system is stationarity. In the simplestterms, the process can be calledstationaryifexamining itspropertiesat any timewill havethe same value. We can distinguishstrictstationarityandstationarityina wide range. Time series are strictlystationaryif for anyallowable data $\{t_1, \dots, t_k\}$andany $h \in Z$ cumulative set $\{x_{t_1}, \dots, x_{t_k}\}$distributionis identical tothe set$\{x_{t_1+h}, \dots, x_{t_k+h}\}$distribution.For the strictstationarityappears thatthe mean value and variance(also called variation measure)are constant in time. The time serieswith finitevarianceis weaklystationary orotherwise calledstationaryover a wide rangeif the averageof its elements$\mu_t$is constantin time and autocovariance$\gamma(s, t)$depends only on the difference $h = |t - s|$.Stationarityconditionin a wider sense of random process is that the first and secondmoment would not changein time.In otherwords,ifsuccessive valueschange in time [5, 11].

Anotheranalysed parameteristhe autocorrelation. Autocorrelationis astatistical toolfor analysing the function describing the degree of a given time series elementsdepends onpreviousvaluesin the sametime series. In statistics, the autocorrelationof a random process describes theprocesscorrelationat different time points. When$x_t$the value ofthe process in time *t*is equal toμandvariation σ^2than formulato determinethe autocorrelationcanbe represented as:

$$R(t, s) = \frac{E[(X_t - \mu)(X_s - \mu)]}{\sigma^2}$$

Self-similarityindexalso known asHurstcoefficientis a real numbertaking valuesfrom 0 to 1.Describes the behaviour of the processin time and isassociated with theautocorrelation andfractal dimension. However, it is problematic, to precisely calculate it so that we are talking about the coefficient estimationvalue.There are manycomputation methods which withdifferent efficienciesallows toestimate the value ofthe index.Hurstexponentcanbe seenin three cases[3]:

1. ***H =0.5***means thatbetween the valuesof the testedtime series, there is no correlation, the valuesare random.
2. ***H >0.5***means that process has long-termrelationships. Such a caseis calledpersistentwith positiveautocorrelation andrepresents situations wherethere has beenan increase inthe value of the time seriesand there is alikelihood of itsrecurrenceequal to*H*.
3. ***H<0.5***analogouscaseofanti-persistent behaviour and thevalue increase is likely to bepreceded bya decline.

Hurstexponent studies can be performed with four methods: absolute moments, aggregatedvariance,periodogramand methodsof residuals of regression.The methodof absolute momentscreatesthe aggregated time series of the originaltime series$X = \{X_1, i \geq 1\}$by dividinginto blocksof size*m* andaveraging operation:

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

where$X^m(k)$denotes theaggregated time series and *k*is the index ofthe block.Then,for the prepareddata the sum of absolute momentsof the aggregatedtime series is calculated with the following formula:

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

where*N*is the number ofdatain the testedseries.

Method of aggregated variance unlike the previousimplies thatinstead of calculatingthe sum of theabsolute momentsof aggregated series, calculatesthe variance ofsampling rateaccording to the formula:

$$VarX^{(m)} = \frac{1}{\frac{N}{m}} \sum_{k=1}^{\frac{N}{m}} \left(X^{(m)}(k)\right)^2 - \left(\frac{1}{\frac{N}{m}} \sum_{k=1}^{\frac{N}{m}} \left(X^{(m)}(k)\right)\right)^2$$

Calculationsusingthe residuals of regressionbegin with datadivisionon the blocks with size*m*,and within eachblockwe calculate sum of partial data which will be markedas: $Y(i), i = 1,2, \dots, m$. Then,the obtained resultsmustmatch thetrend line calculated by least square method. In this case the residuumis calculated as the point ofdeviationfroma givenreference point in a created trend line. Residuals sampling variance is calculated with the following formula:

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

where*a* and *b*are coefficients of the fit line. The obtained results should be averagedfor large*m*,the calculation resultisproportional to m^{2H}[5, 11].

Periodogrammethodhas a different approachthan presented inthe previous examples. In this case,the Fouriertransformis usedfor the obtained frequencydependence data. Statisticscalledperiodogramgeneratesthe following formula:

$$I(\lambda) = \frac{1}{2\pi N} \left| \sum_{j=1}^{N} X_j e^{ij\lambda} \right|^2$$

III. HPC CLUSTER CONFIGURATION

The first task in order to examine the cluster computing performance with the use of nonextensive statistics was to build the HPC cluster an 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.

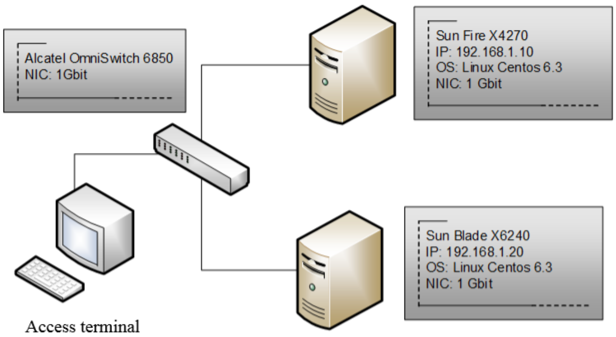

Fig. 1.The designed HPC cluster configuration schema.

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 stationaritychart 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.

Graphpresented in Fig.2.showsthat ourtestedprocess is notstationary. In addition tostationarygraphalso shows thevariation characteristic of the tested process.The highest value increase can be observedduringa drastic decreasein load due tothe completion ofthe performed tests.

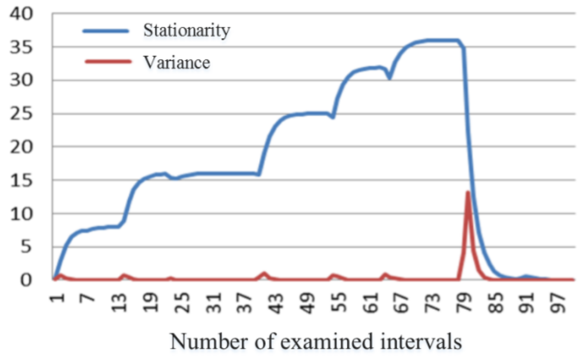

Fig. 2.Stationarity and the variance of the examined load time series.

The next studied parameter is the autocorrelation.Results aredefinitelydifferent from zerowhich indicatesthe correlationof the testedprocess. The obtaineddata are showninFig. 3.

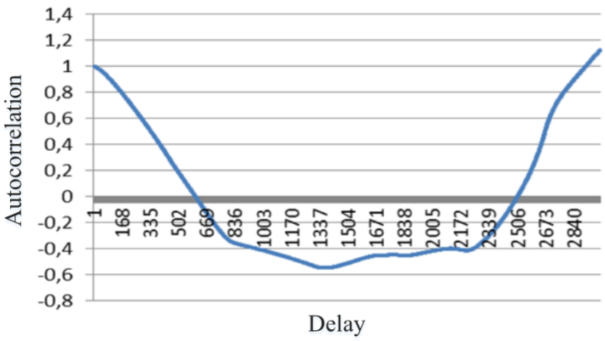

Fig. 3.Autocorrelation of the examined load time series.

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 programs. 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.

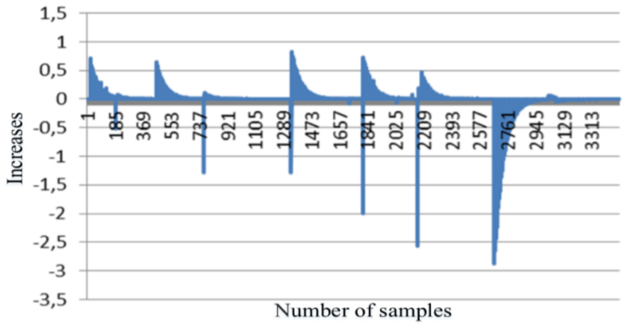

Fig. 4.Increases of the examined load time series.

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.

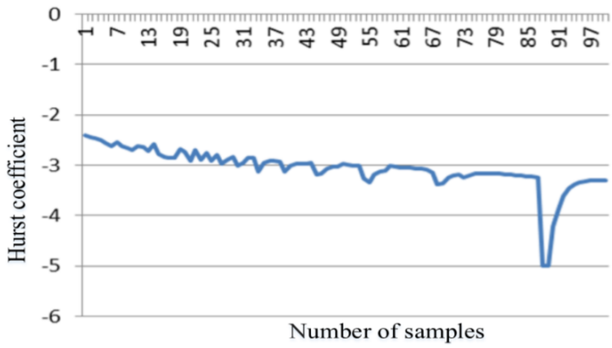

Fig. 5.Hurst coefficient estimation with the absolute moments method.

The second method – the Hurst coefficient estimation with the aggregate variance methodwas 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.

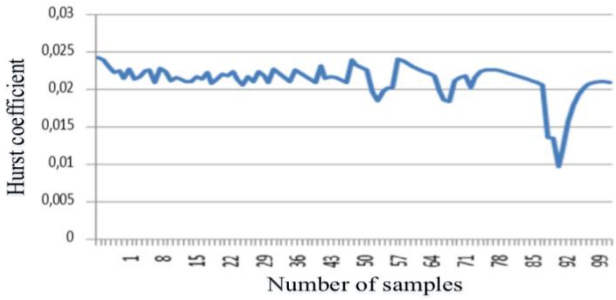

Fig. 6.Hurst coefficient estimation with the aggregate variancemethod.

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.

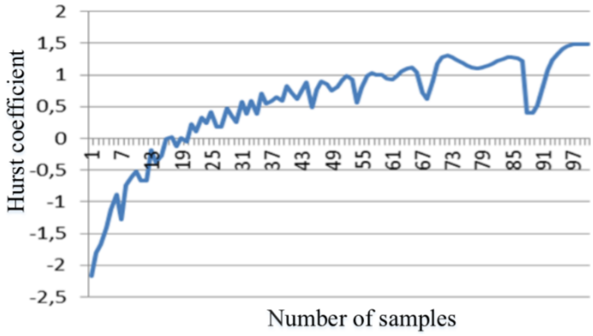

Fig. 7.Hurst coefficient estimation with the residuals of regression method.

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.

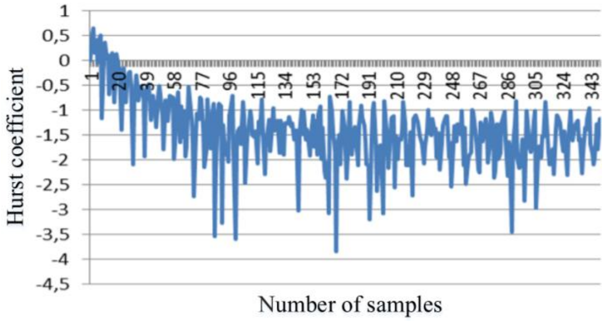

Fig. 8. Hurst coefficient estimation with the periodogram method.

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 testeddata 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 testedprocess 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

[1] P. Dymora, M. Mazurek, *Delay analysis in wireless sensor network protocols*, PAK 2013 nr 10, s. 1054-1056, 2013.

[2] Strzałka B., Mazurek M., Strzałka D., *Queue Performance in Presence of Long-Range Dependencies – an Empirical Study*, International Journal of Information Science, 2012, 2(4), pp. 47-53.

[3] Yan R., Wang Y., *Hurst parameter for security evaluation of LAN traffic,* Information technology Journal 11(20, 2012, pp. 269-275.

[4] Dymora P., Mazurek M., Strzałka D., *Long-range dependencies in memory pages reads during man-compute system interaction*, Annales UMCS Informatica XII (2) 2012, pp. 49-58.

[5] Strzałka D., *Non-extensive statistical mechanics – a possible basis for modelling processes in computer memory system*, ActaPhysicaPolonica A 117(4), 2010, pp. 652–657

[6] Dymora P., Mazurek M., Strzałka D., *Statistical mechanics of memory pages reads during man–computer system interaction*, MetodyInformatykiStosowanej, vol. 1/2011 (26), 2011, pp. 15-21.

[7] Chuan-Lin Lai, Chao-Tung Yang, *Construct a Grid Computing Environment on Multiple Linux PC Clusters*, International

Conference on Open Source, 2003.
[8] C. T. Yang, S. S. Tseng, M. C. Hsiao, *A Portable paralleling compiler with loop partitioning*, 1999.
[9] Beowulf Cluster Computing with Linux Thomas Sterling, Massachusetts Institute of Technology, 2002.
[10] Dymora P., Mazurek M., Strzałka D., Piękoś M., Influence of batch structure on cluster computing performance – complex systems approach, Annales UMCS, Informatica. Volume 12, Issue 1, Pages 57–66, 2013.
[11] Strzałka D., Grabowski F., Non-Extensive Thermodynamics of Algorithmic Processing - the Case of Insertion Sort Algorithm, in Thermodynamics, ed. Tadashi Mizutani, InTech (2011)