ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 2011. No. 8(114)
ELEKTRONIKA IR ELEKTROTECHNIKA

SYSTEM ENGINEERING, COMPUTER TECHNOLOGY
T 120
SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS

Forecasting of GRID Job Waiting Time from Imputed Time Series

K. Sutiene
Department of Mathematical Research in Systems, Kaunas University of Technology, Studentų g. 50, Kaunas, Lithuania, e-mail: kristina.sutiene@ktu.lt
G. Vilutis, D. Sandonavicius
Department of Computer Network, Kaunas University of Technology, Studentų g. 50, Kaunas, Lithuania, e-mail: gytis.vilutis@ktu.lt

crossref http://dx.doi.org/10.5755/j01.eee.114.8.706

Introduction

It is increasingly confronted with the problem of choosing suitable clusters for execution of jobs in GRID networks for commercial and scientific purposes. Although many studies have been carried out and published while looking for the best solution, this problem is still relevant. WMS (Workload Management System) services, which operate according to particular algorithms of a cluster search, are responsible for choice of suitable clusters for jobs sent to GRID network. It is difficult for WMS algorithms to evaluate precisely the following: how long a job will be executed in the particular cluster; whether accessible cluster satisfies requirements for a job execution (the amount of working nodes, the minimal processor speed and other); how long a job is queued waiting for a cluster; reliability of operation of services that constitute accessible clusters.

In continuation of research in this area, it has been chosen to improve the concept of QoGS algorithm [1], which operation is based on the search of the best cluster, using qualitative parameters. While improving this algorithm, it is necessary to evaluate the fact that a part of values of parameters, describing operation of GRID network, varies high in time. This is the reason why monitoring systems, determining parameters of GRID network, and which would provide information about values of parameters at the current time, are necessary. However, one of the most varying parameters and the one which is the most difficult to determine is time T_Q^C when a job is waiting in a queue Q of a cluster C. The value of parameter T_Q^C is calculated when its terms are known: a number of jobs queued J_{NR} and the average time of execution of one job queued $\overline{J_C}$. Determination of parameter $\overline{J_C}$ causes most problems since it is impossible to work out the value of this parameter with usual monitoring systems. Meanwhile, specialized systems provide value of this parameter with certain latency: from 0.01 hour (when a cluster is free) up to 70 hours (when a cluster is loaded fully). One of the possible solutions is to forecast values of parameters that vary high in time. GRID cluster for a job execution will be chosen more precisely (more suitably), using forecasted values of parameters that describe clusters in QoGS algorithm, since the average of time series of parameters has been used so far.

Having received data from specialized and usual monitoring systems, it is possible to calculate parameter value $\overline{J_C}$ and in this way to fill in time series necessary for forecasting. However, due to unstable work of usual GRID monitoring systems, moments in time appear when values of parameter J_{NR} are not received and a lack of data appears in time series of parameter $\overline{J_C}$. Another reason, which determines a lack of data in time series of parameter $\overline{J_C}$, is moments in time when there are no jobs in GRID cluster queues. Due to this reason, in order to make forecasting of parameter $\overline{J_C}$, it is necessary to impute not only missing values but also to evaluate nature of variation of this parameter.

Forecasting is widely and successfully applied in various areas, such as forecasting of traffic flows [2], forecasting of workload of information systems [3] or forecasting of processes that take place in GRID networks [4–8]. The main task of research carried out in GRID networks is to improve operation of WMS or other scheduling service, ensuring equal distribution of jobs [4, 5, 7]. Methods presented in papers are used for forecasting of such dynamic parameters as: the level of network workload [9], frequency of a cluster failure or a cluster workload [4, 7]. Despite the fact that forecasting in paper [5] is made using widely applied methods (such as Exponential Smoothing, Running Mean, Sliding Median, Last or Last2), it is practically impossible to make

forecasting of averaged time of one job queued. Forecasting in paper [4] is made using Markov chains. However, due to a lack of application of Markov chains that future data depends only on the present and does not depend on the past data, forecasting may be made only for a very short period (up to 2 hours)

A problem of forecasting time of a job queued and time of a job execution is examined in papers [6] and [8]. However, algorithms of finding similar jobs used for forecasting time of a job execution are not implemented in most GRID networks. Methods described in this literature collect historic data of jobs solved, such as: group name, user name, executable name of job, a number of CPUs, requested run time and time of the day when a job arrives to a cluster. Having done self-training according to this data, it is possible to forecast that execution time of the job, which arrived, will be the same as that of the most similar executed job. This method serves perfectly for computing, having GRID network with a very limited amount of clusters. Since QoGS method is intended for GRID networks with an unlimited amount of clusters, therefore, when improving a chosen method, it is necessary to take into consideration that the necessary data, according to the described methods [6] and [8], is not accessible in many cases or its samples are too small.

Pattern analysis of statistical data

The data used in this paper have been taken from BalticGrid network. Fig. 1 shows the statistical observations of parameter $\overline{J_C}$ obtained from the particular cluster using the monitoring system. These data were observed over 1-hour intervals during a period from 2/28/2011 00:00 AM (Monday) until 4/1/2011 00:00 AM (Friday). The percent of missing values is 28 percent. These statistical observations are employed throughout this paper to demonstrate the considered problem and the proposed solution.

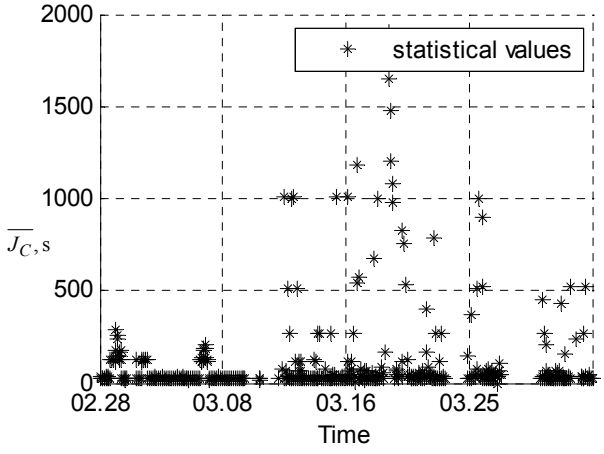

Fig. 1. Historical observations of job's waiting time in a cluster

Having performed the analysis of time series of parameter $\overline{J_C}$, it has been observed that the data have the nature of seasonality: time series exhibits repeating intra-day and inter-day periodicities. Weekday pattern (Fig. 2) is computed by averaging values of $\overline{J_C}$ every 7-day periods, whereas hourly pattern (Fig. 3) is obtained by averaging values every 24-hour periods.

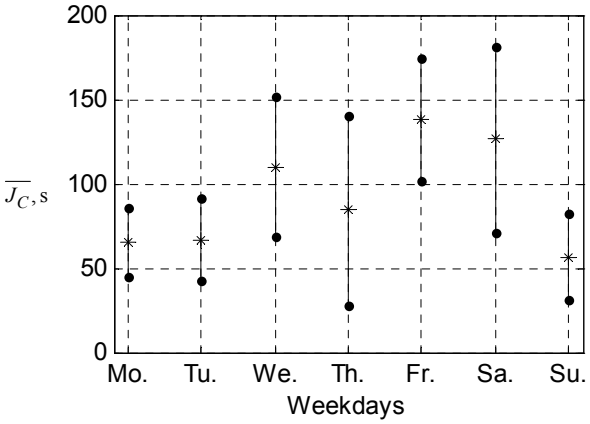

Fig. 2. Weekday pattern of observed values

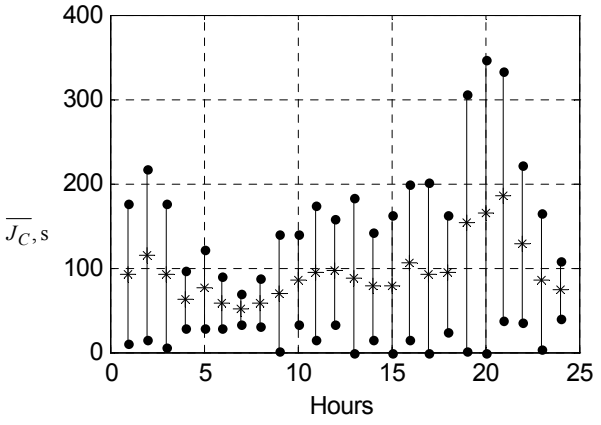

Fig. 3. Hourly pattern of observed values

Time series used in forecasting must contain observations from a short period. In comparison, Fig. 4 shows how distribution of values of parameter $\overline{J_C}$ differs subject to in which season the given data were received.

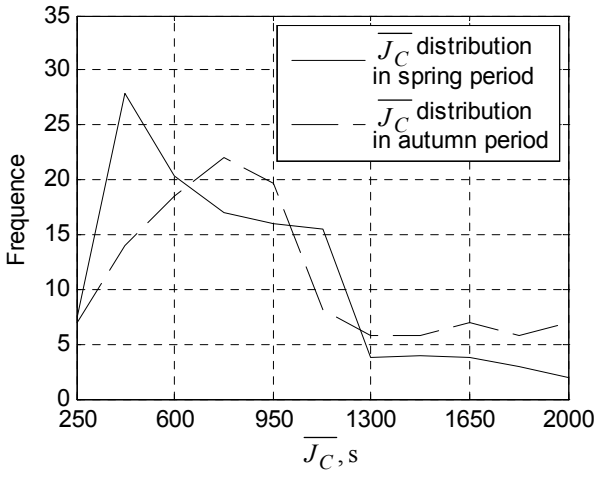

Fig. 4. Distribution of values of parameter $\overline{J_C}$

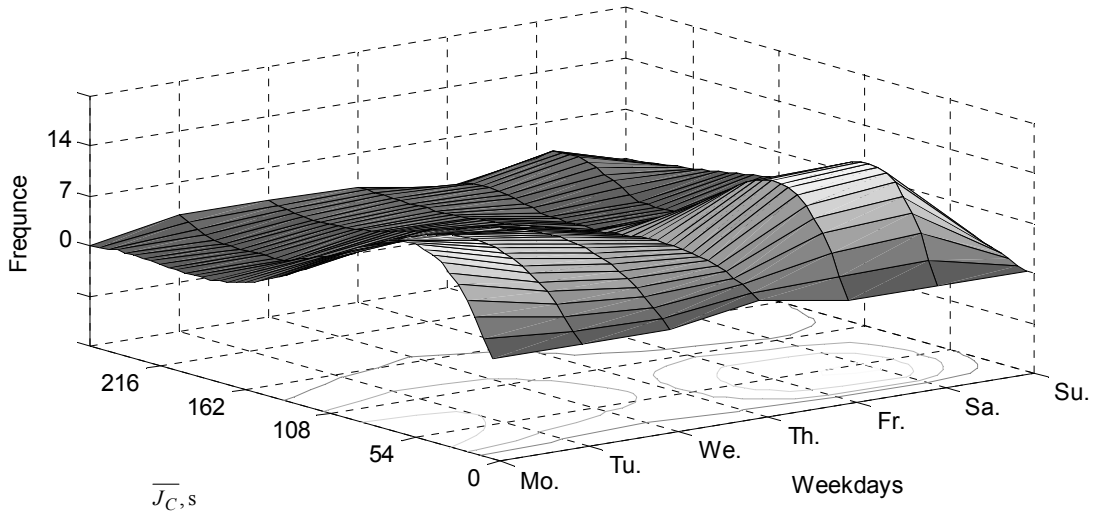

Fig. 5. Spatial polygon of a week's variation of parameter $\overline{J_C}$

Therefore, it is necessary to take into consideration the fact that the length of time series, according to which forecasting is made, would be taken from a short term. Both curves in Fig. 4 were obtained using the same size of samples. It is apparent in this figure that jobs in BalticGrid in the period of spring required less GRID cluster capacity, whereas in the period of autumn the demand of this capacity was more distributed. This could be determined by the fact that most jobs in the period of spring have shorter execution time. Forecasting, which is made using time series of a half-year, will provide more inaccuracies than using time series of data of 2-3 months.

When analysing the distribution of observations received in BalticGrid and drawing spatial polygon for each weekday (Fig. 5), the following tendencies are noticed: having made forecasting, more precise parameter values may be expected on Friday and Monday because on other days high deviation of data is apparent (especially on Thursday). It should also be emphasized that the highest probability to receive value of parameter $\overline{J_C}$ on different days of the week varies insignificantly. Spatial polygon of such nature is formed when a large amount of jobs, which do not require much cluster capacity, prevails in GRID network.

Imputation of missing values in a time series of job's waiting time

It has been already considered that the forecasting of future workload serves as useful information in a cluster selection decisions and provides guidelines in a task scheduling. The forecasting methods used in this paper are based on time series approach and combine the current state, as well as historical records of $\overline{J_C}$ values. Since the statistical data of parameter $\overline{J_C}$ contains missing values (Fig. 1), usual forecasting methods fail to produce forecasts. Thus, the primary task is to impute missing values in historical data.

The proposed way for the imputation of missing values is based on a seasonal pattern observed in the accumulated statistical data as presented in Section "Pattern Analysis of Statistical Data**".**

Let $Z(t)$ denote the observed parameter $\overline{J_C}$ in the particular cluster at time t. A time series of $Z(t)$ is modelled as the sum of two periodic patterns $W(t), D(t)$ plus a white noise $\varepsilon(t)$

$$Z(t) = W(t) + D(t) + \varepsilon(t). \quad (1)$$

Based on curve fitting procedure, the best fitted periodical function for a weekly seasonal pattern is the sum of two sinusoids

$$W(t) = \sum_{k=1}^{2} a_k \sin\left(b_k \frac{2\pi}{7} t + c_k\right), \quad (2)$$

where parameters a_k, b_k, c_k, $k = \overline{1,2}$ are chosen best based on the scatter of weekly averaged values for parameter $\overline{J_C}$. The modelled situation is depicted in Fig. 6.

As the next step, the weekly seasonal pattern is filtered out from the accumulated statistical data. Then, the daily periodical behaviour is described as Fourier function

$$D(t) = h + \sum_{k=1}^{2}\left[m_k \cos\left(k\omega \frac{2\pi}{7\cdot 24} t\right) + n_k \sin\left(k\omega \frac{2\pi}{7\cdot 24} t\right)\right], \quad (3)$$

where parameters h, ϖ, m_k, n_k, $k = \overline{1,2}$ are estimated based on curve fitting procedure applied to the filtered data.

The hourly values of parameter $\overline{J_C}$ and the fitted function are displayed in Fig. 7.

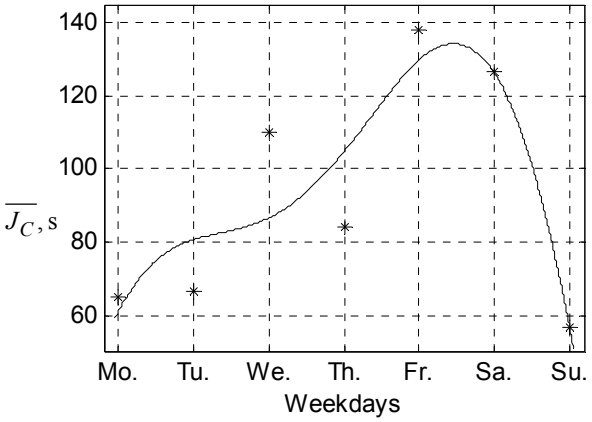

Fig. 6. Fitted curve for a weekday pattern

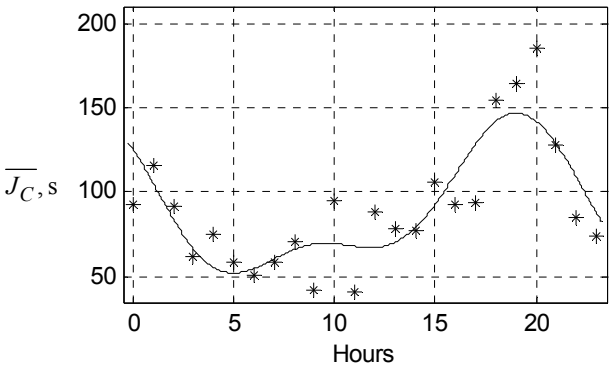

Fig. 7. Fitted curve for a hourly pattern

The combination of functions $W(t)$ and $D(t)$ describes the deterministic part of a seasonal model, while the stochastic part is described as a white noise $\varepsilon(t)$. The obtained seasonal model is used to impute the missing values in observations of parameter $\overline{J_C}$. Fig. 8 displays the time series of $\overline{J_C}$, where missing values are already generated from the imputation model.

The imputed values plus the historical observations of parameter $\overline{J_C}$ become as raw data for a forecasting procedure.

Short-term forecasting of values for job's waiting time

The analysis of historical data showed that there is significant dependency in time between values of parameter $\overline{J_C}$. Thus, forecasting model should incorporate an autoregressive structure to explain the inter-day correlations, as well as intra-day correlations. This gives a task to forecast the system workload over the periods of the day for several days in future.

In this paper, time series forecasting approaches [10], such as autoregressive (AR) model, nonlinear autoregressive (nAR) model, autoregressive moving average (ARMA) model, seasonal autoregressive integrated moving average (SARIMA) model are employed. Since historical data are characterized by a seasonal pattern, the prediction based on curve-fitting (CF) technique is also considered. The historical records with imputed values (Fig. 8) over a given period are employed to train the certain model.

The orders of AR, MA and ARMA components are determined by examining the autocorrelation function ACF and the partial autocorrelation function PACF. The usage of nonlinear function as a summed series of nonlinear units, such as wavelet networks, allows to overcome drawbacks in AR model. In the deseasonalizing process of SARIMA model, two seasonal indexes are computed, such as weekly (for each weekday) and daily (for each hour) indexes. The curve-fitting technique is employed to choose the best fit among periodical functions because of seasonal pattern observed in historical data. The estimated function is the sum of sinusoids

$$F(x) = \sum_{k=1}^{5} a_k \sin(b_k x + c_k), \tag{4}$$

where optimal parameters a_k, b_k, c_k are estimated for the best fit of function $F(x)$ over the model's trained period.

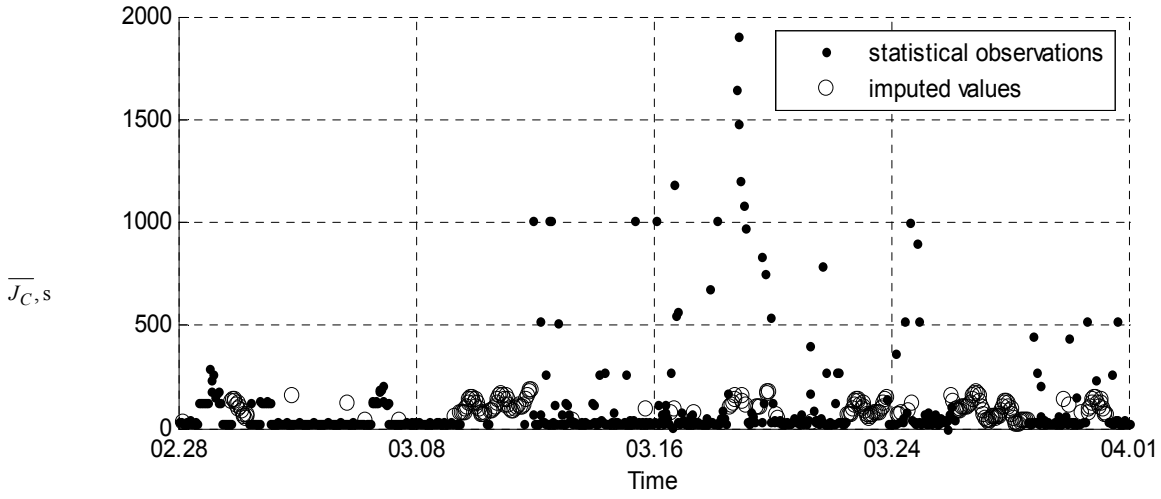

Fig. 8. Historical observations with imputed missing values

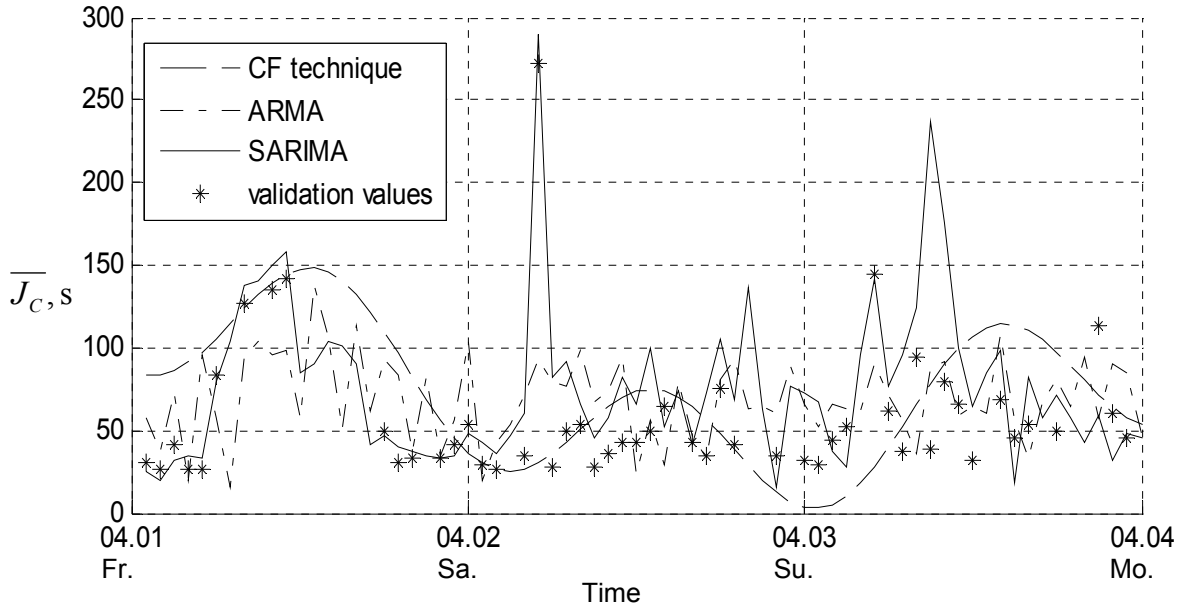

Fig. 9. Forecasts of parameter $\overline{J_C}$ over the validation period

The validation of forecasts is performed for three days from 4/1/2011 00:00 AM (Friday) until 4/4/2011 00:00 AM (Monday). The forecasting methods are compared based on the validation measure, such as root mean square error (RMSE). The heuristic weighted RMSE (wRMSE) is introduced in order to enhance the estimation of error at the beginning of validation period since the forecasts are most important for the next few hours. wRMSE is computed based on formula

$$wRMSE(\hat{x}) = \sqrt{\mathbf{E}\left(w(\hat{x}-x)^2\right)}, \qquad (5)$$

where w – array of weights, x – validation data, $\hat{x}$ – forecasts of parameter $\overline{J_C}$.

Table 1 shows the computed values of RMSE and wRMSE over the validation period for each forecasting model. RMSE and wRMSE are estimated only for existing validation values, since the validation period also contains missing values.

Table 1. Validation measures of $\overline{J_C}$ forecasts obtained from the certain model

	AR	nAR	ARMA	SARIMA	CF	Best/ Worst
RMSE	48.40	44.87	41.81	39.54	51.17	SARIMA/ CF
wRMSE	26.51	24.58	22.90	21.66	16.09	CF/ AR

The last column records the best / worst model based on minimum / maximum value of RMSE or wRMSE respectively in rows. SARIMA and ARMA models demonstrate good validation results comparing to other forecasting models. CF technique is only advanced if the beginning of validation period is enhanced. Fig. 9 displays forecasts of parameter $\overline{J_C}$ obtained from well-suited forecasts models. These forecasts are demonstrated over the validation period.

Conclusions

Having performed literature analysis, it was noticed that forecasting possibilities of parameter of averaged time of a job queued in GRID cluster, when there is no full time series, are not examined enough.

The in-depth research of parameter of averaged time of a job queued in GRID cluster has been carried out. It has been determined that this parameter has three types of seasonalities. Due to seasonalities set, time series suitable for forecasting cannot be longer than 2-3 months.

Time series forecasting approaches, such as AR, nAR, ARMA and SARIMA, as well as CF technique are compared over the validation period. Forecasting validation results indicate that inclusion of seasonal pattern in forecasting model achieve superior performance on accuracy of future projections.

It was noticed that job's waiting time in a queue was shortened if forecasting method SARIMA was used while employing QoGS algorithm. This allows to infer that the concept of QoGS algorithm is reasonable and requires its more rigorous research using simulation methods.

References

1. **Plėštys R., Kavaliūnas R., Vilutis G., Sandonavičius D., Vaškevičiūtė R.** The measurement of grid QoS parameters // Proceedings of the 29th international conference on Information Technology Interfaces. – Dubrovnik, 2007. – P. 703–707.

2. **Balsys K., Valinevičius A., Eidukas D.** Traffic Flow Detection and Forecasting // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – No. 5 (101). – P. 91–94.
3. **Wolski R.** Forecasting network performance to support dynamic scheduling using the network weather service // High Performance Distributed Computing. – Portland, 1997. – P. 316–325.
4. **Shi L., Guo L., Yang S., Wu B.** A Markov Chain Based Resource Prediction in Computational Grid // Proceedings of Fourth International Conference on Frontier of Computer Science and Technology. – Shanghai, 2009. – P. 119–124.
5. **Sonmez O., Yigitbasi N., Iosup A., Epema D.** Trace–Based Evaluation of Job Runtime and Queue Wait Time Predictions in Grids // Proceedings of HPDC '09 Proceedings of the 18th ACM international symposium on High performance distributed computing ACM. – New York, 2009. – P. 111–120.
6. **Smith W.** Prediction Services for Distributed Computing // IEEE International Parallel and Distributed Processing Symposium (IPDPS'2007). – Long Beach, 2007. – P. 1–10.
7. **Piro R. M., Guarise A., Patania G., Werbrouck A.** Using historical accounting information to predict the resource usage of grid jobs. Future Generation Computer Systems, 2009. – Vol. 25. – No. 6. – P. 499–510.
8. **Li H., Groep D., Wolters L.** Efficient Response Time Predictions by Exploiting Application and Resource State Similarities // GRID '05 Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing. – IEEE Computer Society. – Washington, 2005. – P. 234–241.
9. **Eidukas D., Valinevičius A., Vilutis G., Kilius Š., Vasylius T.** Duomenų perdavimo tinkle apkrautumo skaičiavimas // Elektronika ir Elektrotechnika. – Kaunas: Technologija, 2005. – Nr. 8(64). – P. 22–26.
10. **Tsay R. S.** Analysis of Financial time series // Wiley interscience. – Chicago: Wiley, 2005. – 251 p.

Received 2011 08 29

K. Sutiene, G. Vilutis, D. Sandonavicius. Forecasting of GRID Job Waiting Time from Imputed Time Series // Electronics and Electrical Engineering. – Kaunas: Technologija, 2011. – No. 8(114). – P. 101–106.

Most attention in the paper is paid to GRID quality parameter, which specifies averaged time of execution of one job waiting in a cluster queue. Detailed analysis of time series of this parameter, which is received from monitoring systems of academic BalticGrid network, is presented. Seasonalities of variation of parameter value are shown graphically. Problematic forecasting cases of this parameter are reviewed. Forecasting methodology, which involves a model of data imputation, is proposed. The experiment is carried out, using forecasting methods that are prevailing in practice (AR, nAR, ARMA, SARIMA, CF technique). Ill. 10, bibl. 10, tabl. 1 (in English; abstracts in English and Lithuanian).

K. Šutienė, G. Vilutis, D. Sandonavičius. Užduoties laukimo Grid ištekliuje trukmės prognozavimas naudojant atstatytos laiko eilutės reikšmes // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2011. – Nr. 8(114). – P. 101–106.

Daugiausia dėmesio skiriama Grid kokybės parametrui, kuris nusako vidutinę vienos užduoties, laukiančios ištekliaus eilėje, vykdymo trukmę. Pateikiama detali šio parametro laiko eilutės analizė, kuri yra gauta iš akademinio BalticGrid tinklo stebėjimo sistemų. Grafiškai parodomas parametro vertės kitimo sezoniškumas. Apžvelgiami probleminiai šio parametro prognozavimo atvejai. Siūloma prognozavimo metodika, kuri apima ir duomenų atkūrimo modelį. Eksperimentas atliktas su labiausiai praktikoje paplitusiais prognozavimo metodais (AR, nAR, ARMA, SARIMA, CF technique). Il. 10, bibl. 10, lent. 1 (anglų kalba; santraukos anglų ir lietuvių k.).