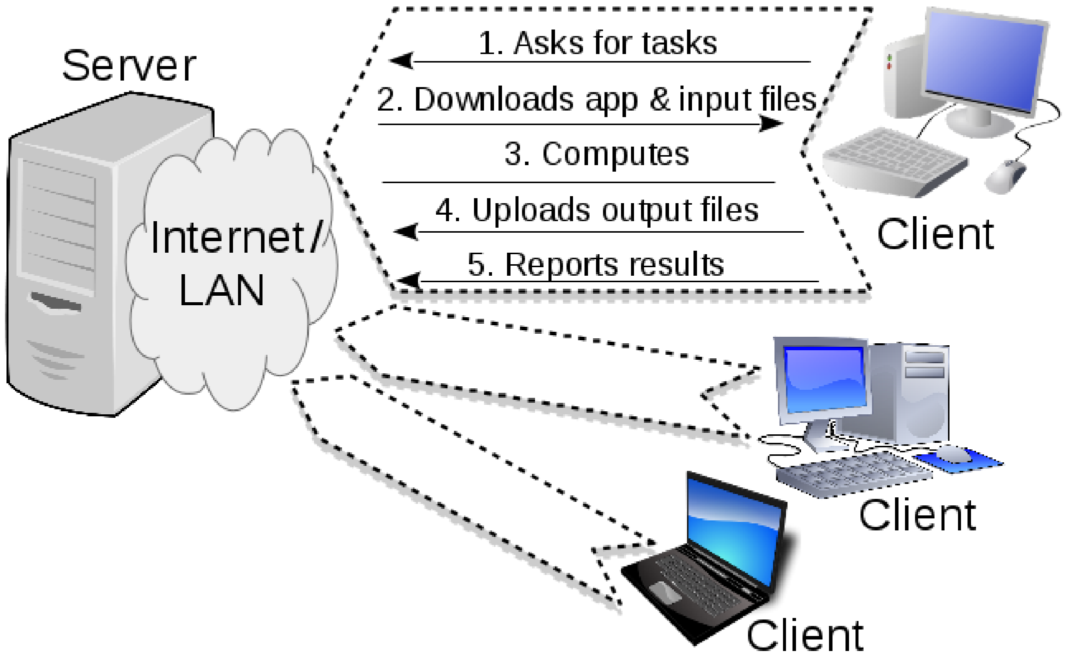
Dynamic forecasting of the completion time of a computational experiment in a Desktop Grid

V.S. Litovchenko (PetrSU)
E.E. Ivashko (KRC of RAS)

Introduction

- Desktop Grid is a form of distributed high-throughput computing system, which uses idle time of non-dedicated geographically distributed computing nodes connected over low-speed network.
- A number of advantages:
  - high scalability
  - fault-tolerance
  - low cost for deploying and maintenance
- Desktop Grid systems are intended to solve computationally intensive problems.
- The BOINC software platform - the most popular Desktop Grid software.

BOINC



- The BOINC platform has a client-server architecture.
- The client part can work on an arbitrary number of computers with various hardware and software characteristics.
- The server supports the simultaneous operation of a large number of independent projects.

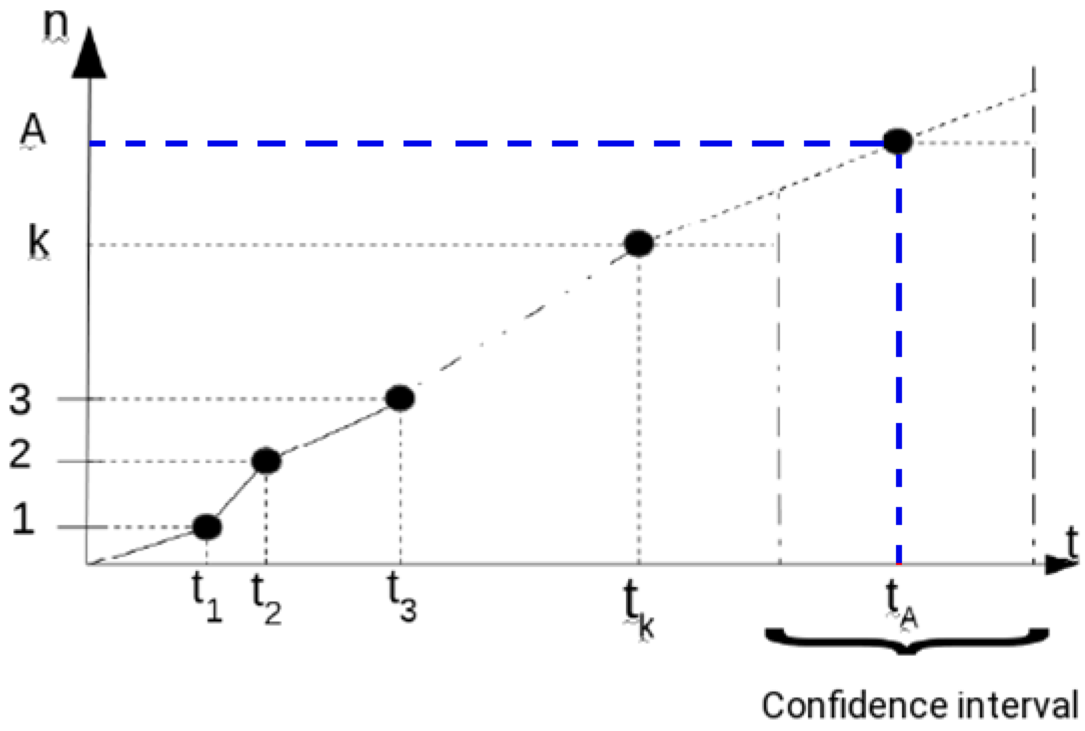
Problem description

- Desktop Grid projects are based on computational experiments.
- Computational experiment is a set of tasks for which the results analysis could be started only when finished all the tasks of the computational experiment.
- This raises a problem of a tasksbag (set of tasks) runtime estimation: for a scientist, it is important to know when he could be able to start the results processing.
- A number of peculiarities which complicate a tasksbag runtime estimation:
    - High hardware and software heterogeneity;
    - Low reliability of computing nodes;
    - Uncertainty of processing time.
- A tasksbag runtime estimation is an important problem for computational projects based on Desktop Grids.

Problem description

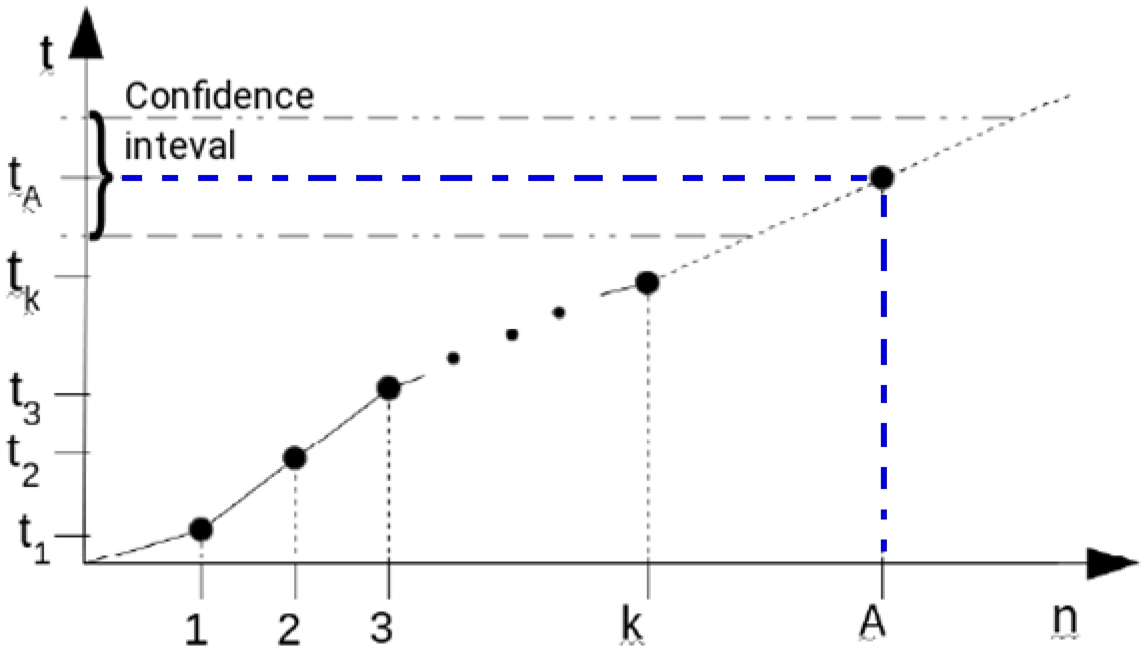
- We consider a BOINC-based Desktop Grid, consisting of a number of computing nodes.
- A computational experiment of N tasks takes place; we need to construct a forecast on completion time of the computational experiment.
- To make a forecast one should determine a functional dependence reflecting to time series. This functional dependence is called a forecast model.
- We assume that change in performance is linear.

Problem description



- Consider a cumulative process of results retrieving. This process is described by a time series: $Z(t) = Z(t_1), Z(t_2), \dots, Z(t_k)$, discrete time points: $t_1 < t_2 < \cdots < t_k$
- At the point t_k (forecast point) one should estimate a time point t_k, at which observed value $Z(t_p)$ will exceed a specified value A.

Problem description



- Assume that there is a functional dependence between previous and future values of the process: $Y(t) = F(Y_{t-1}, Y_{t-2}, Y_{t-3}, \dots) + \varepsilon_i$, here ε_i – is a random error with a normal law of distribution. This dependence is piecewise linear with up trend.
- For convenience, turn to considering a process: $Y_i = (Y(t))^{-1}, i = 1, \dots, k$ which describes time points of i-th result receiving.

Statistical model

This is a linear regression model, which is described by the following formula:

$$y_i = a \cdot x_i + b + \varepsilon_i$$

here, *a, b* - coefficients of the regression, ε_i - white noise, $i > 1$

The least-squares deviation method minimizes sum of errors squared magnitudes of observed and estimated values using the following formula:

$$\sum_{i=1}^{n} (y_i - (a \cdot x_i + b))^2 \rightarrow min$$

The optimal coefficients *a* and *b* are defined by the following formulae:

$$a = \frac{\sum_{i=1}^{k} y_i \cdot \sum_{i=1}^{k} i^2 - \sum_{i=1}^{k} i \cdot \sum_{i=1}^{k} i \cdot y_i}{k \cdot \sum_{i=1}^{k} i^2 - (\sum_{i=1}^{k} i)^2},$$

$$b = \frac{k \cdot \sum_{i=1}^{k} i \cdot y_i - \sum_{i=1}^{k} y_i \cdot \sum_{i=1}^{k} i}{k \sum_{i=1}^{k} i^2 - (\sum_{i=1}^{k} i)^2}.$$

Confidence interval

- Having right statistical model and keeping the trend, observed values and extrapolated point forecast are mismatching due to:
  - inexact parameters of the model;
  - random error;

- These errors can be shown as a forecast confidence interval.
- The confidence interval is an interval, in which the true value falls with a certain degree of probability.

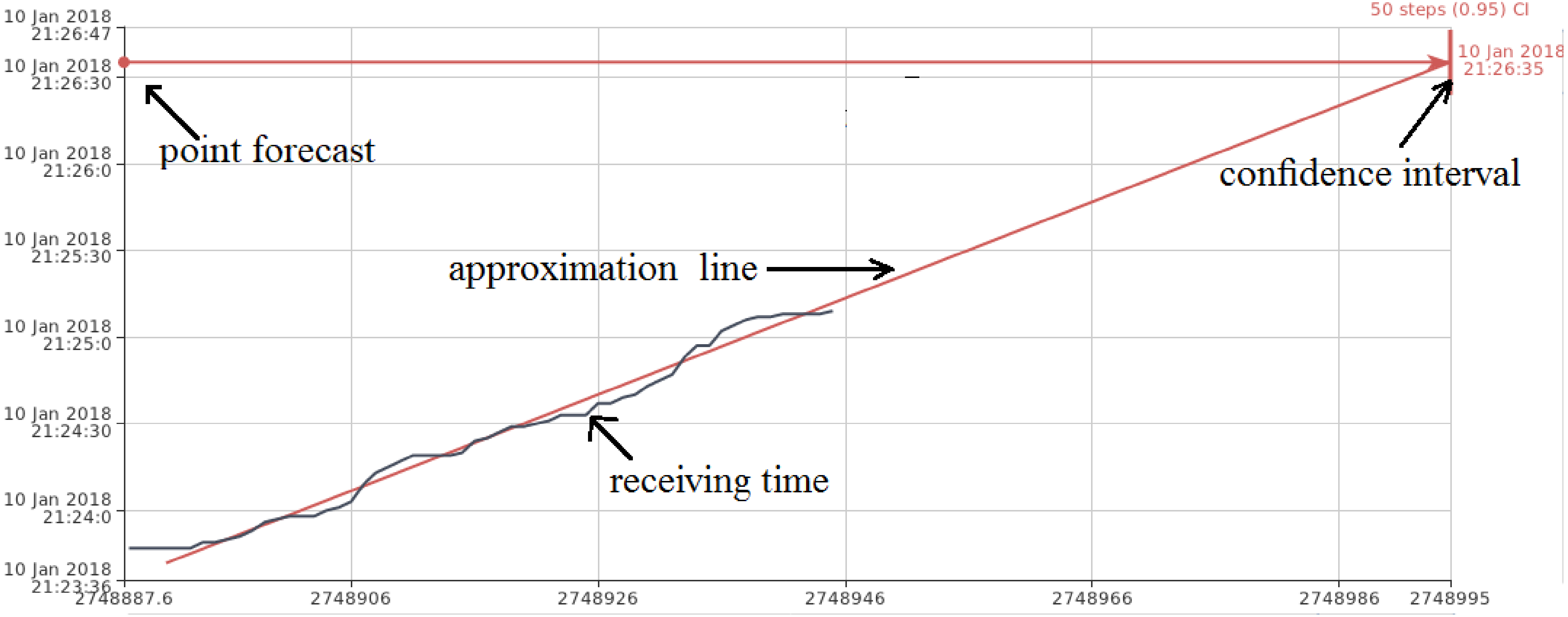
Confidence interval

$$\hat{y}_{k+p} \pm t_\gamma \cdot \sqrt{\underbrace{\frac{\sum_{t=1}^{k}(y_t - \hat{y}_t)^2}{k-1}}_{S_y^2} \cdot \left(\frac{k+1}{2} + \frac{(k+p-\bar{t})^2}{\sum_{t=1}^{k}(t-\bar{t})^2}\right)}$$

Here

- y_{k+p} - point forecast at the moment $k+p$, where k – number of observed values and p – look-ahead period;
- t_γ - value of Student's t-statistics;
- S_y^2 - mean-square distance between observed and forecasted values;
- $t = 1,2,\ldots,k$ - process step;
- $\bar{t} = \frac{k+1}{2}$ - mean step.
- y_t - observed values; $\hat{y}_t$ - forecasted values;

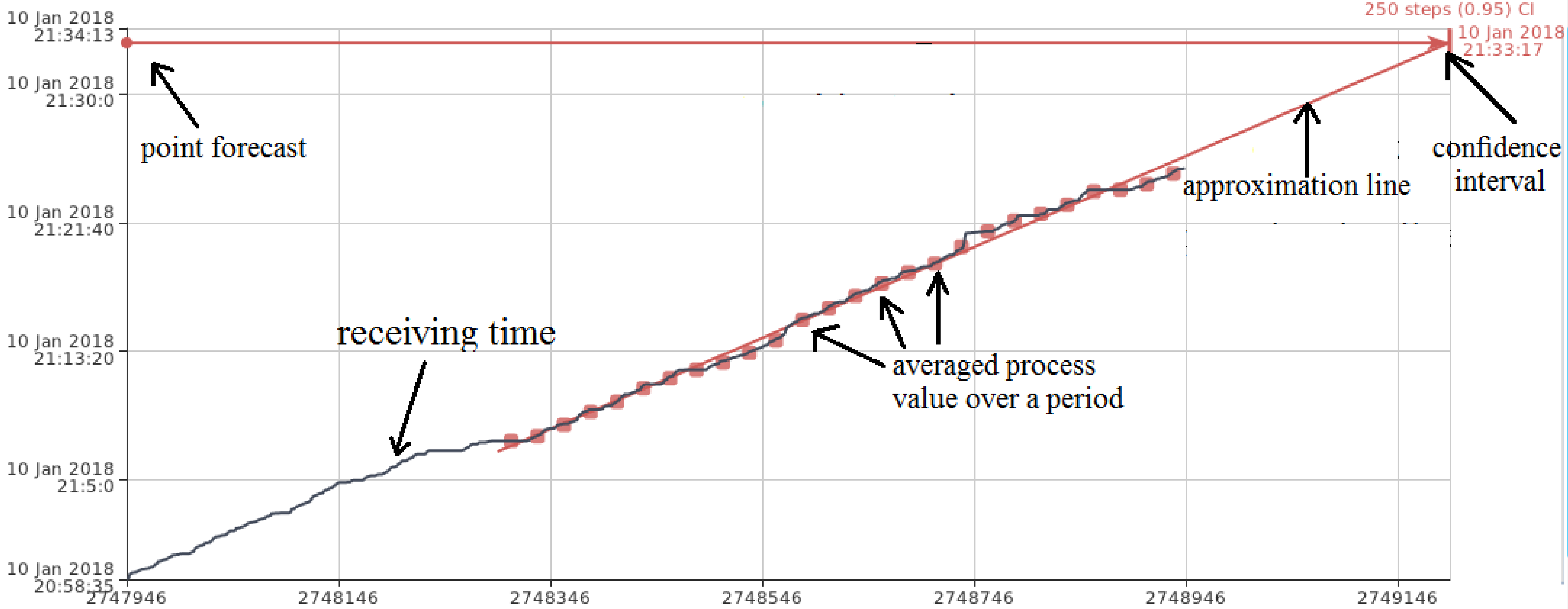
Experiments Analysis



Forecasting with p = 50 and confidence interval 0.95.

RakeSearch project, two months period, 117 thousand of records.

Experiments Analysis

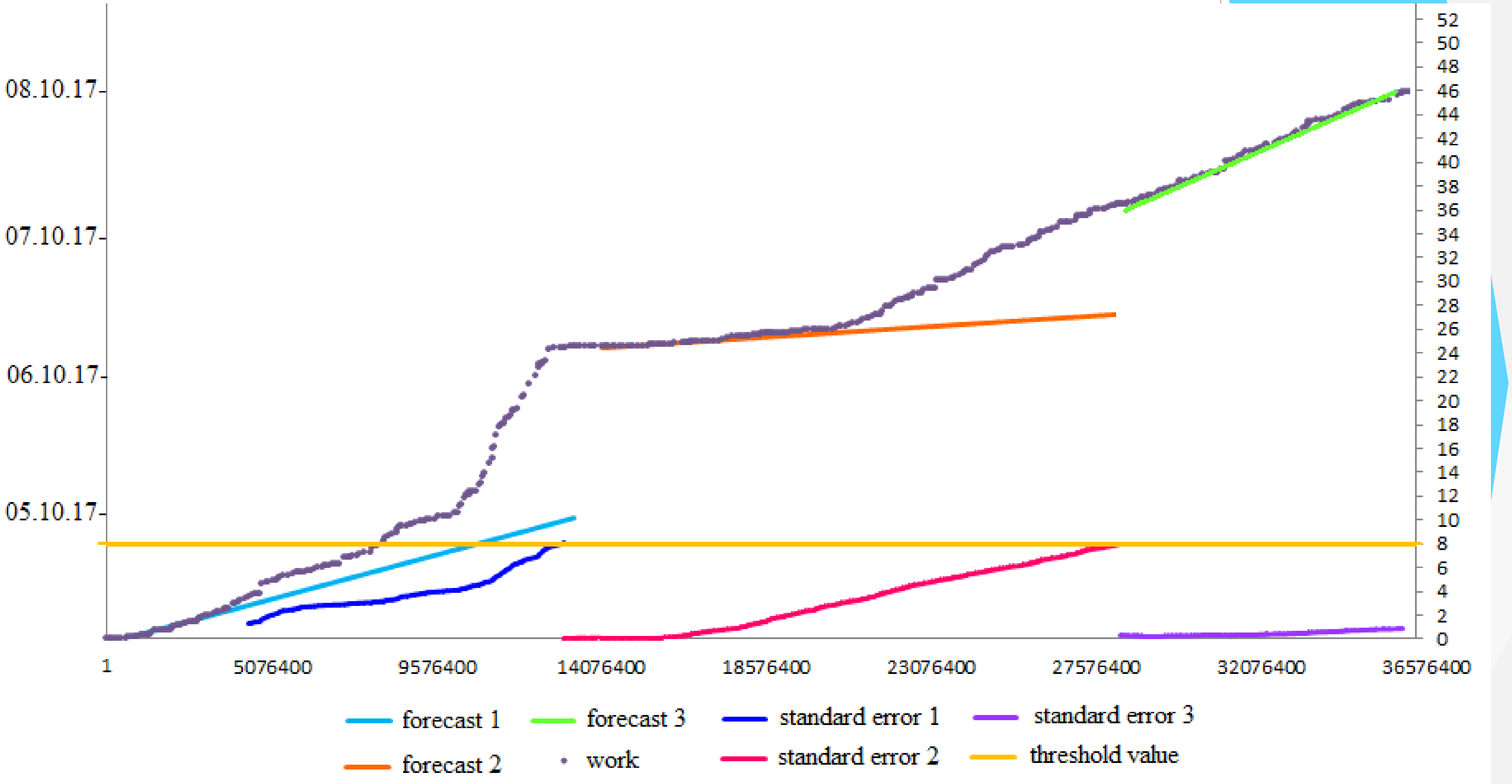


Forecasting with p = 250, confidence interval 0.95.

Statistical error

- Statistical error is the standard deviation of the observed values from the predicted and is calculated by the formula: $\varepsilon_i = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{|y_i - y_i^*|}{y_i} \cdot 100\%$, where y_i^* - calculated values.
- If the accumulated error is exceeded at a certain level, it is considered that the forecast does not correspond to the real process anymore.
- Construction of a tasksbag runtime estimation consists of three stages:

1. a point estimation of the completion time and the corresponding confidence interval are constructed
2. then the accumulation of the statistical error is tracked during the calculations
3. when the statistical error exceeds a certain threshold, the forecast is recalculated with an update of the point estimation of the completion time and the confidence interval.


08.10.17
07.10.17
06.10.17
05.10.17
52
50
48
46
44
42
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1
5076400
9576400
14076400
18576400
23076400
27576400
32076400
36576400
forecast 1
forecast 3
standard error 1
standard error 3
forecast 2
work
standard error 2
threshold value

Conclusion

- We presented the statistical approach to a batch of tasks runtime estimation in a Desktop Grid, which consists of choosing a statistical model, forecasting and construction confidence interval, tracking statistical error accumulation and recalculation of the forecast.
- Basing on the described mathematical and statistical models, we develop an algorithm and a BOINC-module. The module integrated into a volunteer computing project to provide a real-time tasksbag runtime estimation.

Thank you for your attention!

Dynamic forecasting of the completion time of a computational experiment in a Desktop Grid

V.S. Litovchenko

va.lentina97@yandex.ru

E.E. Ivashko

ivashko@krc.karelia.ru