A Gaussian Approximation of Runtime Estimation in a Desktop Grid Project

Oleg Lukashenko
(joint work with Evsey Morozov, Alexander Rumyantsev, Evgeny Ivashko)

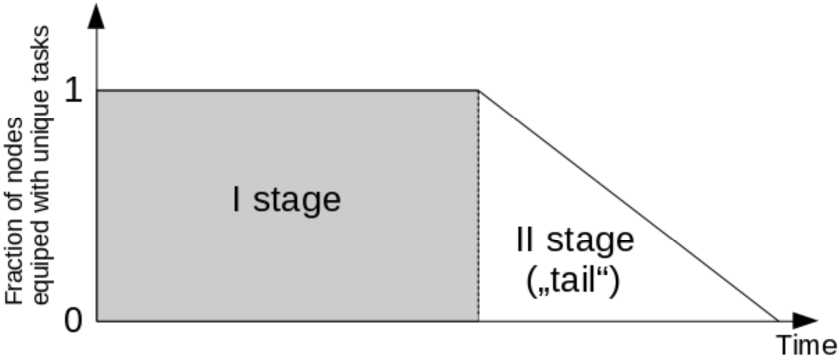
Institute of Applied Mathematical Research
of the Karelian Research Centre of RAS

August 29, 2017

Outline

- We propose a novel stochastic model describing the dynamics of a Desktop Grid project with many hosts and many workunits to be preformed.
- The model is based on the approximation of the basic process by means of the asymptotics of the superposed on-off sources both with light- and heavy-tailed distributions of the hosts' working sessions.
- It is well-known that, after an appropriate scaling, the limiting process describing the summary workload in the system turns out to be *Brownian motion* (BM), when the sources are light-tailed, while it becomes *Fractional Brownian motion* (FBM) when sources are heavy-tailed.
- Then the problem runtime estimation reduces to calculation of the hitting time of the given threshold the summary work to be done.

Two stages of project completion in a Desktop Grid:



- At the first stage the number of workunits is greater than the number of hosts, and thus each host can obtain at least one workunit.
- At the second stage, all available workunits are dispatched and there are available (idle) hosts.

On-Off model:

- Each host is characterized by the service rate R (the amount of processed work per unit time) and the following *on-off* process $\{I(t),\ t \geq 0\}$,

$$I(t) = \begin{cases} R, & t \in \text{ on-period}, \\ 0, & t \in \text{ off-period}. \end{cases} \tag{1}$$

- During an *on-period* each host is processing the work, while it keeps silence (inactive) during the following *off-period*. It is assumed that, for each host, on periods and off periods separately are iid.

- Assume that there are n types of hosts, and N_i is the number of the ith type of hosts.
- Each type $i = 1, \ldots, n$ is characterized by the service rate R_i and corresponding *on-off* process $\{I^{(i)}(t),\ t \geq 0\}$,

$$I^{(i)}(t) = \begin{cases} R_i,\ t \in \text{ on-period}, \\ 0,\ t \in \text{ off-period}. \end{cases} \tag{2}$$

- The on-off processes generated by different hosts are assumed to be independent.

The cumulative processed work:

- The aggregated amount of the work (*cumulative processed work*) served by all hosts during time interval $[0, t]$ is defined as

$$A(t) = \int_0^t \left(\sum_{i=1}^{n} \sum_{k=1}^{N_i} I_k^{(i)}(u) \right) du,$$

where $I_k^{(i)}$ are independent copies of $I^{(i)}$, $i = 1, \ldots, n$.

Distribution of on-off periods:

- The statistical behavior of the cumulative processed work crucially depends on the distribution of on-off periods. Let, for the ith type of hosts, F^i_{on}, F^i_{off} be the distribution functions of on- and off-period respectively. Denote by

$$\mu^i_{on},\ \sigma^i_{on},\ \mu^i_{off},\ \sigma^i_{off}$$

  the mean length and standard deviation of the ith type of on-, off-period, respectively, $i = 1, \ldots, n$.
- We assume that the following conditions hold as $x \to \infty$:

$$\begin{aligned} 1 - F^i_{on}(x) &\sim x^{-\alpha^i_{on}} L^i_{on}(x), \quad \alpha^i_{on} \in (1,2) \\ 1 - F^i_{off}(x) &\sim x^{-\alpha^i_{off}} L^i_{off}(x), \quad \alpha^i_{off} \in (1,2) \qquad (3) \end{aligned}$$

Gaussian approximation: theoretical motivation

Functional limit theorem argumentation leads to the following approximation for sufficiently large values N_i and T:

$$A(tT) \approx T \left(\sum_{i=1}^{n} R_i N_i \frac{\mu_{on}^i}{\mu_{on}^i + \mu_{off}^i} \right) t + \sum_{i=1}^{n} T^{H_i} R_i \sqrt{L_i(T) N_i} c_i B_{H_i}(t), \quad (4)$$

where c_i are positive constants, L_i are slowly varying at infinity functions (expressed in the terms of given parameters) and B_{H_i} are independent fractional Brownian motions with the Hurst parameters H_i defined as

$$H_i = \frac{3 - \min(\alpha_{on}^i, \alpha_{off}^i)}{2} \in \left(\frac{1}{2}, 1 \right), \ i = 1, \ldots, n.$$

- If σ^i_{on}, $\sigma^i_{off} < \infty$ for all $i = 1, \ldots, n$, then the limiting process of the cumulative processed work becomes

$$T\left(\sum_{i=1}^{n} \frac{R_i N_i \mu^i_{on}}{\mu^i_{on} + \mu^i_{off}}\right) t + \left(\sqrt{T} \sum_{i=1}^{n} R_i \sqrt{N_i} c_i\right) W(t), \qquad (5)$$

where $W(t)$ is a Wiener process, and constants c_i are expressed by the original predefined parameters

Gaussian approximation

- Given above asymptotic results provide theoretical motivation to consider the following model:

$$A(t) = mt + \sigma X(t), \tag{6}$$

  where m and σ are positive constants (can be estimated from observations); X is a centered Gaussian process, which describes random fluctuations of the cumulative processed work around the linearly increasing mean.
- The second term in expression (6) can be negative, so the trajectory of the process $A(t)$ in general is not monotonically increasing, what does not agree with the physical sense of the *processed work*.

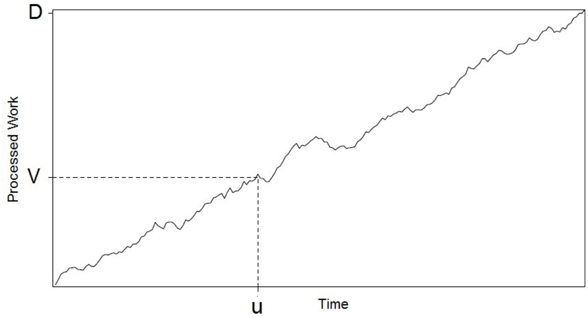
Runtime as a hitting time of Gaussian process

- Let D be the required amount of work to be done in the Desktop Grid project, and we denote by τ_D the corresponding runtime of the project.
- The runtime can be defined as the corresponding *hitting time* of the process $A(t)$, that is,

$$\tau_D = \min\{t : \ A(t) \geq D\}. \tag{7}$$

- In case of Wiener process the probability density function of r. v. τ_D is available in explicit form.
- In case of fractional Brownian motion we are forced to use Monte-Carlo simulation to estimate expected value or other quantitative characteristics.

Runtime as a hitting time of Gaussian process



- From a practical point of view, we are interested in estimation the following conditional expectation

$$\gamma_{D,V}(u) = \mathbb{E}[\tau_D | \tau_V = u], \tag{8}$$

i. e., the expected runtime given that the fraction $V \leq D$ had been completed up to time instant u.

Concluding remarks:

- In this work a novel Gaussian model to describe the activity of the DG with a large number of hosts is suggested.
- Possible future research directions:
  - Validation of this model using statistical data describing real DG projects.
  - Comparison with some well-known direct statistical prediction techniques (for example based on autoregressive models).
  - Asymptotic analysis of the hitting time of fractional Brownian motion.

Thank you for attention!