# Classical Regeneration Based Approaches for Output Analysis of Multiserver Systems Simulation

Irina Peshkova

Petrozavodsk State University,
Lenina Pr. 33, Petrozavodsk, Russia
iaminova@petrsu.ru
http://www.petrsu.ru

**Abstract.** We present and discuss three types of regeneration that are based on classical regeneration for steady-state simulation and output analysis of multiserver systems. First method is artificial regeneration based on exponential splitting property of some distributions. Regenerative envelopes allow to construct upper and lower bound for QoS characteristic of the system. Resampling of New Better than Used inter-arrivals or New Worse than Used service times accelerates simulation. All this techniques are reduced to classical regeneration and allow to receive shorter regeneration cycles.

**Keywords:** Multiserver Systems · Regenerative Simulation· Artificial Regeneration · Splitting · Regenerative Envelopes · NBU distributions.

## 1 Introduction

Modern computer systems, as a rule, have a multiserver or multicore architecture, which is caused both by the physical features of the modern technological process of microchip production and by the need to create high-performance computers suitable for solving computationally complex problems. Data transfer to the users in modern networks also requires coordinated performance of several devices, as well as work with several servers, for solving problems of reducing delays and optimizing throughput. By this reason, reliable estimation of Quality of Service (QoS) parameters of multiserver systems is an important problem to study.

Regeneration of stochastic process is one of the powerful and efficient methods of steady-state simulation and output analysis of complex queueing systems [3,5,8,7,6,11].

Standard classical regeneration method uses visits of the process (describing the systems behavior) to a fixed return state as regeneration points. The trajectory between such points can be divided into independent (for classical regeneration) and identically distributed groups (in general, they can be weakly dependent for wide-sense regeneration). It allows to apply well-developed classical statistical methods based on the special form of the Central Limit Theorem (CLT) for confidence estimation of QoS parameters.

However, in multiserver systems standard regeneration is not guaranteed by stability condition of the system and regeneration points may be too rare to be useful in practice or even they might not exist. We study an alternative constructive method of obtaining the regeneration points.

The so-called artificial regeneration uses the splitting property [2] and can be applied for estimation of multiserver systems performance in two ways. First method is based on constructing a new, equivalent to the original, system by using the splitting property, which allows to change the internal structure of continuous valued random variables in such a way to obtain a sequence of times at which the distribution has some specific properties, and these times occur with positive probability (possibly by enriching the probability space). Second technique uses coupling method and the epochs when the process has a given splitting distribution as regenerative times [1]. If the splitting is based on exponential distribution, then the system at this particular points has memoryless property, and thus, regenerates in classical sense in both cases.

The regenerative envelopes method is reduced to replacing the original system based on the coupling method [9] and the monotonicity properties of the studied processes with a pair of minorant and majorant systems that regenerate in the standard classical sense [10]. Then we can use upper (lower) bounds of confidence intervals for QoS characteristic of majorant (minorant) system for the confidence interval construction of the original system.

We discuss the possibility of accelerating the simulation for multiserver systems in which service times and iterarrival times have New Worse than Used distribution or New Better than Used distribution. If service times have New Worse than Used distribution we resample service time each time when the remaining service time exceeds a fixed constant $a$. If interarrival times has New Better than Used distribution we resample input each time when the remaining interarrival time is less than a fixed constant $b$. In both cases we obtain more frequent classical regeneration that allows to accelerate simulation.

## 2 Artificial regeneration based on splitting property in multiserver systems

The density $f$ of continuous random variable (r.v.) $T$ can be split if there exists some $0 < p < 1$ and density $f_0$, such that:

$$f \geqslant p f_0. \tag{1}$$

We define a new Bernoulli random variable $I$ called *splitting indicator* such that $\mathrm{P}(I = 1) = p$. We construct a new splitted r.v. $T'$ (stochastically equivalent to the r.v. $T$) as follows:

$$T' = IT_0 + (1 - I)T_1, \tag{2}$$

where $T_0$ has density $f_0$, and $T_1$ has density

$$f_1 = \frac{f - pf_0}{1-p}. \tag{3}$$

Now we can replace the original r.v. $T$ with a triple $[T_0, T_1, I]$.

The particular case of splitting, the exponential splitting, could be efficiently applied in simulation due to the memoryless property of exponential distribution.

We say that a positive valued r.v. $T$ with density $f$ is *exponentially splitted* if there exist constants $\lambda > 0$, $\tau_0 \geqslant 0$ and $0 < p < 1$ such that

$$f(x) \geqslant p\lambda e^{-\lambda(x-\tau_0)}, \quad x \geqslant \tau_0. \tag{4}$$

That is, splitting representation $T'$ of exponentially splitted r.v. $T$ is a triple $(T_0, T_1, I)$ where $T_0$ is left truncated exponentially distributed r.v. with truncation point $\tau_0$, rate $\lambda$, density

$$f_0(x) = \begin{cases} 0, & x \leqslant \tau_0, \\ \lambda e^{-\lambda(x-\tau_0)}, & x > \tau_0, \end{cases}$$

and $T_1$ has density

$$f_1(x) = \begin{cases} \frac{f(x)}{1-p}, & x \leqslant \tau_0, \\ \frac{f(x)-p\lambda e^{-\lambda(x-\tau_0)}}{1-p}, & x > \tau_0. \end{cases}$$

Now we describe the application of exponential splitting technique to discrete event simulation of multiserver systems.

Consider a stochastic process $\Theta = \{\mathbf{X}(t), \mathbf{T}(t)\}_{t\geqslant 0}$, with discrete (vector) component $\mathbf{X} = (X_1, \ldots, X_n)$ and continuous component $\mathbf{T} = (T_1, \ldots, T_m) \geqslant \mathbf{0}$ (hereafter we denote vector-valued variables in bold and assume the dimension is clear from context, if not sub-indexed explicitly). Moreover, we assume that $\mathbf{T}$ decreases *linearly* with time, that is, $\mathbf{T}(t+\delta) = \mathbf{T}(t) - \delta\mathbf{1}$, for $\delta > 0$, and define an *event* as a time epoch $t_i$ such that $T_{j^*}(t_i) = 0$ for some $j^* \in \{1, \ldots, m\}$ (in this case we say that the $i$th event occurring at $t_i$ is of $j^*$th type). We assume that the discrete component $\mathbf{X}$ is changed only at event epochs, e.g. by some recurrent relation

$$\mathbf{X}(t_i+) = G_{j^*}\left(\mathbf{X}(t_i-)\right), \tag{5}$$

where $G_{j^*}$ is the recursion related to type $j^*$ event, or, in general, assume the change is governed by some Markovian probability

$$\mathrm{P}_{j^*}(x, x') = \mathrm{P}_{j^*}\{\mathbf{X}(t_i+) = x' | \mathbf{X}(t_i-) = x\}.$$

The component $\mathbf{X}(t)$ does not change if $\mathbf{T}(t) > \mathbf{0}$, that formally means

$$\mathbf{X}(t) = \mathbf{X}(t_i+), i = \max\{k : t_k < t\}.$$

We also assume that in the vector $\mathbf{T}$ the zeroed at $t_i-$ component $T_{j^*}$ is initialized at $t_i+$ from some distribution with density $f_{j^*}$, that is,

$$f_{j^*}(u, x, x') = \mathrm{P}(T_{j^*}(t_i+) \in du | \mathbf{X}(t_i-) = x, \mathbf{X}(t_i+) = x'). \tag{6}$$

Now we assume, that if $\mathbf{X}(t) = x^*$ for some specific $x^*$, then the process $\Theta$ allows to perform exponential splitting of the continuous component $\mathbf{T}$, that is for each continuous component $T_j$ there exist constants $\tau_0(j), \lambda(j), p(j)$ such that (4) holds for the density $f_j(u, x, x^*)$, with corresponding density $f_{0,j}(u) = \lambda_j e^{-\lambda_j(u-\tau_0(j))}$ and $f_{1,j}$ defined appropriately. At that, we introduce a split process $\Theta' = \{\mathbf{X}(t), \mathbf{T}'(t), \mathbf{Z}(t)\}_{t\geqslant 0}$ enhanced with auxiliary phase variable $\mathbf{Z} \in \{0, 1, 2\}^m$ denoting the phases of components of $\mathbf{T}$ defined componentwise in a recursive manner as follows

$$Z_{j^*}(t_i+) = \begin{cases} 1, & \text{if } \mathbf{X}(t_i+) = x^* \text{ and } I_{j^*}(t_i+) = 1, \\ 2, & \text{otherwise}, \end{cases} \tag{7}$$

$$Z_k(t_i+) = Z_k(t_i-), k \neq j^*, \tag{8}$$

where $t_i$ is the epoch of type $j^*$ event, and $I_{j^*}$ is the splitting indicator corresponding to the component $T'_{j^*}$ with success probability $p(j^*)$. However, starting from $Z_{j^*}(t_i+) = 1$ at time $t_i$, the component $Z_{j^*}$ makes a step between the events:

$$Z_{j^*}(t) = \begin{cases} 1, \, t_i < t < t_i + \tau_0(j^*) \\ 0, \, t \geqslant t_i + \tau_0(j^*), \end{cases} \tag{9}$$

that is, the phase $Z_{j^*}$ is changed from one to zero when not less than $\tau_0(j^*)$ time passes since the epoch $t_i$ corresponding to the last event of $j^*$th type (we call this time epoch a pseudo-event). In case $Z_j \neq 1$, the component $Z_j$ does not change during inter-event time.

We define the artificial regeneration epochs as the (increasing) sequence of *pseudo-event* epochs $\{\beta_k\}_{k\geqslant 1}$

$$\beta_{k+1} = \min\{t > \beta_k : \mathbf{X}(t+) = x^*, \mathbf{Z}(t+) = \mathbf{0}\}. \tag{10}$$

This method of artificial regeneration could be applied to performance estimation of $m$-server system with energy efficiency control and single waiting room. The discrete and continuous components of the simulated process $\Theta$ we define in the following way. The discrete components $\tau_0(t)$ is the queue size at time $t$, and $X_i(t)$ is the mode of $i$th server, $i = 1, \dots, m$, $X_1(t) = 1$ if the server is active, and $X_1(t) = 0$ if the server is in the sleep mode. Continuous components: $T_0(t)$ is the time before the next arrival epoch, and $T_i(t)$ is the remaining times of current activity (service/sleep) of $i$th server, $i = 1, \dots, m$. For the regeneration epoch, we set $x^* = \{k, 1, \dots, 1\}$ for some $k \geqslant 0$ and recall that artificial regeneration occurs at such pseudo-event epoch $t$ that $\mathbf{X}(t+) = x^*$ and all the exponentially split components $T_0, \dots, T_m$ are in exponential phases, that is, $\mathbf{Z}(t+) = \mathbf{0}$.

The second approach of defining the regeneration times is to construct stochastically equivalent system with the same input, but sampling from either $f_0$ or $f_1$ at every random variable initialization time epoch. Then the original r.v. $T$ is replaced by a stochastically equivalent splitting representation $T' = IT_0 + (1-I)T_1$. Coupling method allows to construct new artificial system, which is equivalent to the original system, and by this reason has equivalent QoS parameters.

To construct the artificial regeneration points we suggest the following method. Sample the sequence of splitting indicators $I_i, i \geqslant 1$ as a Bernoulli 0-1 r.v. with success probability $p$. Sample service times $\tilde{T}_j$ either from density $f_0$, or from $f_1$, using relation (2). Fix integer $\nu_0$ (fixed number of customers in system) and define

$$\beta_{n+1} = \inf\Big\{k > \beta_n : \nu_k = \nu_0 \leqslant m, \\ I_i = 1, i \in \mathcal{M}_k,\ \mathcal{M}_{\beta_n} \cap \mathcal{M}_k = \emptyset\Big\}, \quad n \geqslant 0, \tag{11}$$

where $\mathcal{M}_n = \{i : t_i \leqslant t_n < t_i^d\}$ is the set of the customers presented in new system at arrival instant $t_n$ and $t_i^d$ is the departure instant of the $i$th customer. Note that at the constructed regeneration epoch, the remaining work possesses memoryless property, since all the tasks being served have exponentially distributed service times. It is easy to see that $\Theta_n := \{\nu_n, I_j, j \in \mathcal{M}_{\beta_n}\},\ n \geqslant 1$ is a Markov process, and that distribution of $\Theta_{\beta_n}$ is independent of $n \geqslant 1$ and pre-history $\{\Theta_k,\ k < \beta_n\}$.

## 3 Regenerative envelopes for multiserver systems simulation

Now we consider an infinite buffer FCFS $m$-server $GI/G/m$ queueing system $\Sigma$, with the renewal input with instants $t_n$, the iid interarrival times $L_n = t_{n+1} - t_n$ and the iid service times $S_n,\ n \geqslant 1$. Denote $\nu_n$ the number of customers at instant $t_n^-$, $Q_n$ the number of customers waiting in the queue at instant $t_n^-$, so $Q_n = \max(0, \nu_n - m)$.

The main idea of the regenerative envelopes method is the construction of two new systems: a *majorant* system $\overline{\Sigma}$ and a *minorant* system $\underline{\Sigma}$ with the same input as in the original system $\Sigma$, but (stochastically) smaller (in minorant system) or larger (in majorant system) service times.

We describe the construction of $m$-server majorant system $\tilde{\Sigma}$. The corresponding variables in the system $\tilde{\Sigma}$ we endow with tildes. Assume that the service time distributions in both systems are ordered as $F_S(x) \leqslant F_{\tilde{S}}(x),\ x \geqslant 0$, that is $S \geqslant_{st} \tilde{S}$ (*stochastically*). Moreover, by construction, $L =_{st} \tilde{L}$. Then the *coupling method* allows to take $\tilde{t}_n = t_n$ and $S_n \geqslant \tilde{S}_n$ w.p.1.

Define the set $\overline{\mathcal{M}}_n = \{i : t_i \leqslant t_n < \overline{z}_i\}$ of the customers, which are being served in the system $\overline{\Sigma}$ at instant $t_n$, $\overline{S}_i(n)$ – the remaining service time of customer $i$ at instant $t_n^-$ in $\overline{\Sigma}$.

Fix arbitrary integer $\nu_0 \geqslant 0$ (number of customers in the system), constants $0 \leqslant a \leqslant b < \infty$ and define

$$\overline{\beta}_{n+1} = \inf\Big\{k > \overline{\beta}_n : \overline{\nu}_k = \nu_0,\ \overline{S}_i(k) \in (a,\ b),\ \overline{\mathcal{M}}_{\overline{\beta}_n} \bigcap \overline{\mathcal{M}}_k = \varnothing,\ i \in \overline{\mathcal{M}}_k\Big\}. \tag{12}$$

Then, at the (discrete) instant $k = \overline{\beta}_{n+1}$, we replace all the non-zero remaining times $\overline{S}_i(k), i \in \overline{\mathcal{M}}_k$, by the *upper bound* $b$.

It is easy to see that $\overline{\mathbf{\Theta}}_n := \{\overline{\nu}_n,\, \overline{S}_i(n), i \in \overline{\mathcal{M}}_n\},\, n \geqslant 1$, is a Markov process and the distribution of $\overline{\mathbf{\Theta}}_{\overline{\beta}_n}$ is independent of $n$ and pre-history $\{\overline{\Theta}_k,\, k < \overline{\beta}_n\},\, n \geqslant 1$.

The process $\{\overline{\mathbf{\Theta}}_n,\, n \geqslant 1\}$ classically regenerates at the instants $\{\overline{\beta}_k\}$ and has the i.i.d *regeneration cycles* $\{\overline{\mathbf{\Theta}}_k,\, \overline{\beta}_n \leqslant k < \overline{\beta}_{n+1}\}$ with i.i.d cycle lengths $\alpha_n = \overline{\beta}_{n+1} - \overline{\beta}_n,\, n \geqslant 1$.

In a minorant system $\underline{\Sigma}$ we define the regenerative moments $\underline{\beta}_n, n \geqslant 0$ in a similar way (with possibly other values $a, b, \nu_0$). At each such instant $\underline{\beta}_n$ we replace the remaining times $\underline{S}_i(k), i \in \underline{\mathcal{M}}_k$, by the lower bound $a$.

The systems $\Sigma$, $\overline{\Sigma}, \underline{\Sigma}$ have zero initial state, and moreover (in an evident notation) $\overline{L}_n = \underline{L}_n = L_n,\, \overline{S}_n \geqslant S_n \geqslant \underline{S_n},\, n \geqslant 1$. Then it follows from coupling property that

$$\overline{\nu}_n \geqslant \nu_n \geqslant \underline{\nu}_n,\; \overline{Q}_n \geqslant Q_n \geqslant \underline{Q}_n,\; n \geqslant 1, \tag{13}$$

and it allows us to use regenerative simulation of $\overline{\Sigma}$ and $\underline{\Sigma}$ to obtain an upper and lower bounds of the mean stationary queue size (number of customers) in original system.

$$\mathsf{E}\underline{Q} \leqslant \mathsf{E}Q \leqslant \mathsf{E}\overline{Q}.$$

## 4 Accelerated simulation of multiserver networks with New Better than Used distributions

In this section we will discuss how the properties of New Better (Worse) than Used distributions of input or service times could be applied to accelerate regeneration simulation.

We say that the distribution $F$ of r.v. $T$ has *increasing (decreasing) failure rate IFR (DFR)* if

$$F_t(x) = P(T \leqslant t + x | T > t) = P(T \leqslant x) = \frac{F(t+x) - F(t)}{\overline{F}(t)} \tag{14}$$

is increasing (decreasing) in $t$. Remark, that

$$\lim_{x \to 0} \frac{F_t(x)}{x} = \frac{f(t)}{\overline{F}(t)} = r(t). \tag{15}$$

Function $r(t)$ is called *failure rate* and it is easy to see that $\overline{F}(t) = e^{-\int_o^t r(x)dx}$. Weibull and Gamma distributions with shape parameter $\alpha > 1$ are IFR, with $\alpha < 1$ are DFR.

It is known [4], that if $ET = \mu_1$ and $F$ is IFR (DFR), then

$$\overline{F}(t) \geqslant (\leqslant)\; e^{\frac{-t}{\mu_1}},\; t < \mu_1. \tag{16}$$

We say that the distribution $F$ of r.v. $T$ has *increasing (decreasing) failure rate in average IFRA (DFRA)* if

$$\frac{1}{t}\int_0^t r(u)du \ \ increases\ in\ t. \tag{17}$$

Distribution $F$ of r.v. $T$ is called *New Better (Worse) than Used NBU (NWU)* if $\overline{F}_t \leqslant \overline{F}$ or

$$\overline{F}(t+x) \leqslant \overline{F}(t) \cdot \overline{F}(x). \tag{18}$$

The following relation between IFR, IFRA and NBU holds.

$$IFR \Rightarrow IFRA \Rightarrow NBU.$$

Let $S$ be the (generic) service time (with distribution B), $L$ – (generic) interarrival time (with distribution A) in $m$-server system $\Sigma$. Define r.v. $S_t$ - remaining service time at $t^-$ and $\overline{B}_t(x) = P(S \geqslant t+x|S \geqslant t) = P(S_t \geqslant x)$, r.v. $L_t$ - remaining interarival time at $t^-$ and $\overline{A}_t = P(L \geqslant t+x|L \geqslant t) = P(L_t \geqslant x)$.

If service times $S$ have NBU distribution or interarrival times $L$ have NWU distribution in the original system $\Sigma$ then we can apply coupling method to construct a new system $\tilde{\Sigma}$ using property (14) of such distributions.

If $S$ has NWU (NBU) distribution, then $\overline{B}_t \geqslant (\leqslant)\ \overline{B}$ or

$$P(S_t \geqslant x) \geqslant (\leqslant)\ P(S \geqslant x).$$

Then we construct new system $\tilde{\Sigma}$ with the same input, but each time when $S_t \geqslant a = const$ we resample $S$ with d.f. $B$. Using classical regeneration approach for new system $\tilde{\Sigma}$ we obtain lower (upper) bound for QoS parameter of the original system $\Sigma$ (for example, queue size).

If $L$ has NBU (NWU) distribution, then $\overline{A}_t \leqslant (\geqslant)\ \overline{A}$ or

$$P(L_t \geqslant x) \leqslant (\geqslant)\ P(L \geqslant x).$$

Then we construct new system $\tilde{\Sigma}$ with the same service times, but each time when $L_t \leqslant a = const$ we resample $L$ with d.f. $A$. In this case we can use classical regeneration approach for new system $\tilde{\Sigma}$ to obtain lower (upper) bound for QoS parameter of the original system $\Sigma$.

In the case of NWU service times or NBU interarrival times we obtain more frequent classical regenerations that allows to accelerate simulation.

## 5 Conclusion

We present three techniques which allow to construct and speed-up classic regenerations in multiserver systems: artificial regeneration based on splitting property, regenerative envelopes and accelerated simulation for systems with NWU service times of NBU input Our future work will focus on the comparison of the simulation results efficiency.

**ACKNOWLEDGEMENTS**

The publication was carried out under state order to the Karelian Research Centre of the Russian Academy of Sciences (Institute of Applied Mathematical Research KRC RAS). This research is partially supported by RFBR project 18-07-00147 and by the Ministry of Education and Science of Russian Federation (project no. 14.580.21.0009, unique identifier RFMEFI58017X0009).

## References

1. Andradóttir, S., Calvin, J.M., Glynn, P.W.: Accelerated Regeneration for Markov Chain Simulations. Probability in the Engineering and Informational Sciences **9**(04), 497–523 (1995). https://doi.org/10.1017/S0269964800004022
2. Andronov, A.: Artificial regeneration points for stochastic simulation of complex systems. In: Simulation Technology: Science and Art. 10th European Simulation Symposium ESS'98. pp. 34–40. SCS, Delft, The Netherlands (1998)
3. Asmussen, S.: Applied probability and queues. Springer, New York (2003)
4. Barlow, R.E., Proschan, F.: Mathematical theory of reliability. With contributions by L. C. Hunter. The SIAM Series in Applied Mathematics, John Wiley and Sons (1965)
5. Crane, M.A., Lemoine, A.J.: An Introduction to the Regenerative Method for Simulation Analys, Lecture Notes in Control and Information Sciences, vol. 4. Springer, Berlin, Heidelberg (1977). https://doi.org/10.1007/BFb0007339
6. Foss, S.G., Kalashnikov, V.V.: Regeneration and renovation in queues. Queueing Systems **8**(1), 211–223 (1991), `http://link.springer.com/article/10.1007/BF02412251`
7. Glynn, P.: Wide-sense regeneration for Harris recurrent Markov processes: an open problem. Queueing Systems **68**(3-4), 305–311 (2011). https://doi.org/10.1007/s11134-011-9238-x
8. Glynn, P.W., Iglehart, D.L.: Simulation methods for queues: An overview. Queueing Systems **3**(3), 221–255 (1988), `http://link.springer.com/article/10.1007/BF01161216`
9. Morozov, E.: Coupling and stochastic monotonicity of queueing processes. PetrSU, Petrozavodsk (2013)
10. Morozov, E., Rumyantsev, A., Peshkova, I.: Monotonicity and stochastic bounds for simultaneous service multiserver systems. In: Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2016 8th International Congress on. pp. 294–297. IEEE (2016), `http://ieeexplore.ieee.org/abstract/document/7765374/`
11. Sigman, K., Wolff, R.W.: A review of regenerative processes. SIAM Review **35**(2), 269–288 (1993)