Asymptotic Analysis of Queueing Models based on Synchronization Method

Afanaseva L.G.

Department of Mathematics and Mechanics,
Lomonosov Moscow State University

Moscow, 23-27 October 2017
MSU, RUDN

Introduction

This talk is focused on the stability conditions of the multichannel queueing systems with heterogeneous servers and a regenerative input flow $X(t)$. The main idea is constructing an auxiliary service process $Y(t)$ which is also a regenerative flow and defining the common points of regeneration for the both processes $X(t)$ and $Y(t)$. Then the traffic rate of the system is defined in terms of the mean of the increments of these processes on the common regeneration period. It allows to use well-known results from the renewal theory to find the instability and stability conditions. The possibilities of the proposed approach are demonstrated by examples.

Plan

1. Model description
2. Synchronization
3. Stability Analysis
4. Queueing system with unreliable servers
5. Stability Analysis of a Multiserver model with Simultaneous Service and a Regenerative Input Flow
6. Conclusion

Model description

We consider a queueing system with a regenerative input flow $X(t)$ (see Thorisson, 2000, Afanasyeva and Bashtova, 2014).
For $X(t)$ let
$\theta_i^{(1)}$ – the ith regeneration point ($\theta_0^{(1)} = 0; i = 0, 1, \ldots$);
$\tau_i = \theta_i^{(1)} - \theta_{i-1}^{(1)}$ – the ith regeneration period;
$\xi_i^{(1)} = X(\theta_i^{(1)}) - X(\theta_{i-1}^{(1)})$ the number of customers arrived during the ith regeneration period.
Assume that $\mathrm{E}\tau_1^{(1)} < \infty, \mathrm{E}\xi_1^{(1)} < \infty$.
The limit

$$\lambda_X = \lim_{t\to\infty} \frac{X(t)}{t} = \frac{\mathsf{E}\xi_1^{(1)}}{\mathsf{E}\tau_1^{(1)}}$$

is the rate of $X(t)$.
We consider the discrete time queueing system as well as continuous time.

Define an auxiliary process $Y(t)$ as the number of customers that can be served during the time-interval $[o, t)$ under assumption that there are always customers for service during this interval.

Assumption 1

$Y(t)$ is a regenerative flow not depending on $X(t)$. For $Y(t)$ we put:
$\{\theta_n^{(2)}\}_{n=1}^{\infty}(\theta_0^{(2)} = 0)$ is the sequence of regeneration points;
$\tau_n^{(2)} = \theta_n^{(2)} - \theta_{n-1}^{(2)}$ - the nth regeneration period

Assumption 2

For the continuous-time case $Y(t)$ is a strongly regenerative flow, i.e.

$$\tau_n^{(2)} = v_n^{(1)} + v_n^{(2)} \tag{1}$$

where $P(v_n^{(1)} > x) = e^{-\delta x}(\delta \in (0, \infty))$, $v_n^{(1)}$ and $v_n^{(2)}$ are independent random variables and $Y(\theta_{n-1}^{(2)} + v_n^{(1)}) - Y(\theta_{n-1}^{(2)}) = 0$.
For the discrete-time case $X(t)$ and $Y(t)$ are aperiodic regenerative flows, i.e.

$$GCD\{k : P(\theta_1^{(i)} = k) > 0\} = 1, \quad i = 1, 2.$$

Synchronization of $X(t)$ and $Y(t)$

We determine common points of regeneration $\{T_n\}_{n=1}^{\infty}$ for the both processes $X(t)$ and $Y(t)$ putting in the discrete-time case

$$T_n = \min\left\{\theta_j^{(1)} > T_{n-1} : \bigcup_{l=1}^{\infty}\{\theta_j^{(1)} = \theta_l^{(2)}\}\right\}, T_0 = 0 \qquad (2)$$

and in the continuous-time case

$$T_n = \min\left\{\theta_j^{(1)} > T_{n-1} : \bigcup_{l=1}^{\infty}\{\theta_{l-1}^{(2)} < \theta_j^{(1)} \leqslant \theta_{l-1}^{(2)} + \nu_l^{(1)}\}\right\}, T_0 = 0. \quad (3)$$

Lemma 1

The sequence $\{T_n\}_{n=1}^{\infty}$ consists of common regeneration points for $X(t)$ and $Y(t)$ and

1. for the continuous-time case

$$\mathrm{E}(T_n - T_{n-1}) = \delta \mathrm{E}\tau_1^{(1)} \mathrm{E}\tau_1^{(2)} < \infty$$

2. for the discrete-time case

$$\mathrm{E}\, T_1 = \mathrm{E}\theta_1^{(1)} \mathrm{E}\theta_1^{(2)} < \infty.$$

Let

$$\Delta_Y(n) = Y(T_n) - Y(T_{n-1}),$$
$$\Delta_X(n) = X(T_n) - X(T_{n-1}).$$

Then

$$\lambda_X = \frac{E\Delta_X(n)}{E(T_n - T_{n-1})}, \lambda_Y = \frac{E\Delta_Y(n)}{E(T_n - T_{n-1})}.$$

We define the traffic rate as follows.

$$\rho = \frac{\lambda_X}{\lambda_Y} = \frac{E\Delta_X(n)}{E\Delta_Y(n)}.$$

Stability Analysis

We define the stochastic flow $\tilde{Y}(t)$ as the number of customers really leaving the system during time interval $[0, t)$.

Condition 1

The following stochastic inequalities take place

$$\tilde{\Delta}_Y(n) = \tilde{Y}(T_n) - \tilde{Y}(T_{n-1}) \leqslant \Delta_Y(n), \ (n = 1, 2, ...).$$

Let $Q(t)$ be the number of customers at the system at time t so that

$$Q(t) = Q(0) + X(t) - \tilde{Y}(t).$$

Condition 2

There are two possible cases:

(i) $Q(t)$ is a stochastically bounded process, i.e.

$$\lim_{x \to \infty} \liminf_{t \to \infty} P(Q(t) \leqslant x) = 1;$$

(ii) $Q(t) \xrightarrow[t \to \infty]{P} \infty$.

Condition 3

If Condition 2(ii) takes place then for any $\epsilon > 0$ there is n_ϵ such that for $n > n_\epsilon$

$$E\tilde{\Delta}_Y(n)\mathbb{I}(Q(t) \geqslant m \text{ for all } t \in [T_n, T_{n+1})]) \geqslant E\Delta_Y(n) - \epsilon. \qquad (4)$$

Theorem 1

1. Let Condition 1 be fulfilled. If $\rho \geqslant 1$ then

$$Q(t) \xrightarrow[t\to\infty]{P} \infty,$$

   i.e. the system is instable

2. Let Conditions 2 and 3 be fulfilled. If $\rho < 1$ then $Q(t)$ is a stochastically bounded process.

Queueing systems with unreliable servers

Here we consider a continuous-time queueing system with a regenerative input flow $X(t)$ and m heterogeneous servers which may be not available for operation from time to time. Assume that $n_i(t) = 0$ if at time t the ith server is in unavailable state and $n_i(t) = 1$ otherwise ($i = \overline{1, m}$).

Condition 4

The stochastic process $\overrightarrow{n}(t) = (n_1(t), \ldots, n_m(t))$ is a strongly regenerative one with regeneration points $\{\theta_n^{(2)}\}_{n=1}^{\infty}$ ($\theta_0^{(2)} = 0$), $\tau_n^{(2)} = \theta_n^{(2)} - \theta_{n-1}^{(2)}$, $E\tau_n^{(2)} < \infty$ with an exponential phase $\nu_n^{(1)}$ so that $\tau_n^{(2)} = \nu_n^{(1)} + \nu_n^{(2)}$. We also assume that $n_i(\theta_{n-1}^{(2)} + t) = 0$ for $t \in [0, \nu_n^{(1)}]$, $i = \overline{1, m}$.

Service times of customers served by the ith server constitute a sequence $\{\eta_{in}\}_{n=1}^{\infty}$ of iid random variables, $b_i = \mathsf{E}\eta_{in} < \infty$, $B_i(x) = \mathsf{P}(\eta_{in} \leqslant x), (i - \overline{1, m})$.
It is possible that an unavailable period starts while a customer is receiving service. Then the service of the customer is immediately interrupted. There are various disciplines for continuation of the service after restoration (see Gaver, 1962). We consider the preemptive repeat different service discipline (D_1) and preemptive resume service discipline (D_2). In the former case service is repeated from the start and the service time after restoration is independent of the original service time. In the latter case service continuous after restoration. Under Condition 4 the regenerative process $\overrightarrow{n}(t)$ has the limit distribution. Denote

$$\pi_i = \lim_{t \to \infty} \mathsf{P}(n_i(t) = 1).$$

Let $Y^{(d)}(t)$ be an auxiliary service process for discipline $D^{(d)}(d = 1, 2)$. Then $Y^{(1)}(t)$ is a regenerative flow with points of regeneration $\{\theta_n^{(2)}\}_{n=1}^{\infty}$. For discipline D_2 we need the additional

Condition 5

Service times have the first exponential phase, i.e.

$$\eta_{in} = \eta_{in}^{(1)} + \eta_{in}^{(2)}$$

where $\eta_{in}^{(1)}$ and $\eta_{in}^{(2)}$ are independent random variables and $P(\eta_{in}^{(1)} > x) = e^{-\alpha_i x}\,(\alpha_i \in (0, \infty))$.

Then as regeneration points for $Y^{(2)}(t)$ we take subsequence $\{\theta_{n_k}^{(2)}\}_{k=1}^{\infty}$ of the sequence $\{\theta_n^{(2)}\}_{n=1}^{\infty}$ such that at time $\theta_{n_k}^{(2)}$ interrupted services were in the first exponential phase. Now basing on Lemma 1 we see that traffic rate $\rho^{(2)}$ for discipline D_2 has a form

$$\rho^{(2)} = \frac{\lambda x}{\sum_{i=1}^{m} \frac{\pi_i}{b_i}}. \tag{5}$$

Consider discipline D_1. Let $\{s_{in}^{(2)}\}_{n=1}^{\infty}$ be moments of the ith server becoming unavailable and $\{s_{in}^{(1)}\}_{n=1}^{\infty}$ be moments of the ith server coming back to service. We assume that

$$0 = s_{i0}^{(2)} < s_{i1}^{(1)} < s_{i1}^{(2)} < \dots .$$

Then $u_{in}^{(2)} = s_{in}^{(1)} - s_{in-1}^{(2)}$ – the length of the nth blocked and $u_{in}^{(1)} = s_{in}^{(2)} - s_{in}^{(1)}$ – the length of the nth available period. Let $u_{in} = u_{in}^{(1)} + u_{in}^{(2)}$ be the length of the nth cycle for server i and ν_{ij} the number of cycles for the ith server during the jth regeneration period. For the continuous-time case we assume $\mathsf{E}\nu_{ij} < \infty$. Let $(u_{i,j}^{(1)}(k), u_{i,j}^{(2)}(k))(k = \overline{1, \nu_{ij}})$ be the kth cycle, so that the jth regeneration period for the ith server is defined by the vector

$$\{(u_{i,j}^{(1)}(k), u_{i,j}^{(2)}(k)), k = \overline{1, \nu_{ij}}, \tau_j^{(2)}\}.$$

We introduce the counting process

$$\mathscr{K}_i(t) = \max\{n : \sum_{s=1}^{n} \eta_{is} \leqslant t\},$$

the renewal function

$$H_i(t) = \mathsf{E}\mathscr{K}_i(t).$$

Then the traffic rate $\rho^{(1)}$ for discipline D_1 has a form

$$\rho^{(1)} = \lambda_X \mathsf{E}\tau_1^{(2)} \left(\sum_{i=1}^{m} \mathsf{E} \sum_{l=1}^{\nu_{i1}} H_i(u_{i,1}^{(2)}(l)) \right)^{-1}.$$

Denote $Q^{(d)}(t)(d = 1, 2)$ the number of customers at the system a time t.

Condition 6

i. $\mathsf{P}\{\max\limits_{1 \leqslant l \leqslant \nu_{i1}} u_{i,1}^{(2)}(l) > x\} > 0$ for any x;

ii. $\mathsf{P}\{\xi_1 = 0, \tau_1^{(1)} > 0\} + \mathsf{P}\{\xi_1 = 1, t_1 + \nu_{11} < \tau_1^{(1)}\} > 0$

where t_1 is the arrival time of the customer on the first regeneration period $\tau_1^{(1)}$ of the input flow $X(t)$ and ν_{11} is the service time by the first server.

Corollary 1

If $\rho^{(d)} > 1$ then $Q^{(d)}(t) \xrightarrow[t\to\infty]{P} \infty$.

Under some additional assumptions the statement of the Corollary is true for the case $\rho^{(d)} = 1$.

Corollary 2

Let Conditions 4,6 for the system with discipline D_1 and Conditions 4, 5, 6 for the system with discipline D_2 be fulfilled and $\rho^{(d)} < 1$. Then process $Q^{(d)}(t)$ is a stable one $(d = 1, 2)$.

For a queueing system with independent renewal processes of interruptions (see Morozov et al, 2011) we have

$$\rho^{(2)} = \lambda_X \left(\sum_{i=1}^{m} \frac{a_i^{(2)}}{a_i b_i} \right)^{-1}, \quad \rho^{(1)} = \lambda_X \left(\sum_{i=1}^{m} a_i^{-1} \mathsf{E} H_i(u_{i,1}^{(2)}) \right)^{-1},$$

where $a_i^{(2)}(a_i^{(1)})$ - the mean of the working (unavailable) period for the ith server; $a_i = a_i^{(1)} + a_i^{(2)}$

We have from Theorem 1 for this system

Corollary 3

If $\rho^{(d)} > 1$ then $Q^{(d)}(t) \xrightarrow[t\to\infty]{P} \infty$. If $\rho^{(d)} < 1$ and for a system with discipline D_2 Condition 6 holds then $Q^{(d)}(t)$ is a stochastically bounded process as $t \to \infty$ $(d = 1, 2)$.

As before, under additional assumptions, $Q^{(d)}(t) \xrightarrow[t\to\infty]{P} \infty$ if $\rho^{(d)} = 1$.

Let us note that for discipline D_2 our stability condition for the case $b_i = b$ coincides condition (23) obtained in (Morozov et all, 2011) for the system with recurrent input flow that is $GI|G|m$.

Stability Analysis of a Multiserver model with Simultaneous Service and a Regenerative Input Flow

This part is devoted to the stability analysis of a multi-server queueing system in which each new customer requires a random number of servers simultaneously and a random service time is identical at all occupied servers. For this model with a regenerative input flow we deduce stability criterion using synchronization method. The most crucial attribute of the systems which provide a random number of servers per customer is that a customer cannot begin service until all required servers are available. Therefore servers may be idle even when there are customers waiting to enter service. Queueing systems belonging to this class are found in a variety of contexts. In computer systems, buffers and other temporary storage devices are used for programs and date of varying dimensions. The increasing interest to multi-server systems with simultaneous service is motivated by the modeling of high performance clusters (HPC) and cloud/distributed computing containing a huge number of servers working in parallel.

Queueing systems with concurrent service have been also considered in a number of works. Here we refer to pioneering papers of Brill and Green, 1984, Whitt, 1985 and more late works, Rumyantsev and Morozov, 2015, Morozov and Rumyantsev, 2016. Detail analysis of available results in this domain of the queueing theory and extensive list of references are given in the paper Rumyantsev and Morozov, 2015.

Let us note that the mentioned papers (except Morozov and Rumyantsev, 2016, where $MAP|M|s$ cluster model is considered) deal with exponential distributions of inter-arrival and service times and authors use the matrix analysis of the system.

The main contribution of the represented research is an extension of the stability criterion to the cluster model with a regenerative input flow. The class of these flows is very broad and includes MMP, MAP, DSPP with intensity that is a regenerative process. The detail description of the regenerative flows and processes one may find in (Thorisson, 2000 and Afanaseva and Bashtova, 2014).

Model Description

We consider two models S_1 and S_2 with a regenerative input flow $X(t)$ with the rate $\lambda = \lim\limits_{t\to\infty} \frac{X(t)}{t}$ w.p.1. Customer i occupies ζ_i servers simultaneously for an exponentially distributed service time η_i with rate μ, $(i \geq 1)$ for the model S_1 and for the model S_2

$$\eta_i = \eta_i^{(1)} + \ldots + \eta_i^{(r)} (r > 1),$$

where $\{\eta_i^{(k)}\}_{k=1}^{r}$ are independent identically distributed(iid) random variables with rate μ_k for $\eta_i^{(k)} (k = \overline{1, r})$. We call $\eta_i^{(l)}$ the lth phase of the service time. All ζ_i servers occupied by customer i are simultaneously released upon completion of their service. The sequence $\{\zeta_i\}_{i=1}^{\infty}$ consists of iid random variables with given distribution

$$p_j = \mathsf{P}(\zeta = j), \quad j = 1, \ldots m, \quad \sum_{j=1}^{m} p_j = 1.$$

Auxiliary processes

In this section we define an auxiliary process $Y(t)$ that will be used in our analysis. We think $Y(t)$ as the number of customers that can be served in the system if there are always customers for service. It means that $Y(t)$ is defined by the sequences $\{\eta_n\}_{n=1}^{\infty}$ and $\{\zeta_n\}_{n=1}^{\infty}$ and does not depend on the input flow $X(t)$. Customer n occupies ζ_n servers simultaneously. We call customer n class-i one if $\zeta_n = i$. For the model S_1 we introduce the stochastic process $U_1(t)$ with the state space

$$\mathcal{K}_1 = \{\overrightarrow{k} = (k_1, \ldots, k_j, k_{j+1}); \quad \sum_{i=1}^{j} k_i \leq m, \quad \sum_{i=1}^{j+1} k_i > m\}$$

and for the model S_2 the stochastic process $U_2(t)$ with the state space

$$\mathcal{K}_2 = \{\overrightarrow{k} = (k_1, e_1, \ldots, k_j, e_j, k_{j+1}); \quad \sum_{i=1}^{j} k_i \leq m, \quad \sum_{i=1}^{j+1} k_i > m, e_i = \overline{1, r}\}$$

Under condition that there are always customers for service we put $U_1(t) = \overrightarrow{k} = (k_1, \ldots, k_j, k_{j+1})$ if there are j customers on the servers at instant t, the ith serving customer has a class $k_i (i = \overline{1, j})$ and the first customer in the queue has a class k_{j+1}. For the process $U_2(t)$ coordinates $k_i (i = \overline{1, j+1})$ have the same meaning and e_i is the number of the phase of the service time for the ith customer at instant t. We note that $U_i(t)(i = 1, 2)$ is a Markov chain with a finite set of states and there are limits

$$\lim_{t \to \infty} \mathrm{P}(U_i(t) = \overrightarrow{k}) = \mathrm{P}_i(\overrightarrow{k}), \quad (i = 1, 2).$$

The auxiliary process $Y_i(t)(i = 1, 2)$ is a regenerative flow (Afanaseva and Bashtova, 2014) and there exists the rate

$$\lambda_{Y_i} = \lim_{t \to \infty} \frac{Y_i(t)}{t} \quad w.p.1.$$

For the state $\overrightarrow{k} = (k_1, \ldots, k_j, k_{j+1})$ of the Markov chain $U_1(t)$ define $j(\overrightarrow{k}) = j$ as the number of customers on the servers and for $U_2(t)$ define $g(\overrightarrow{k})(\overrightarrow{k} \in \mathcal{K}_2)$ as the number of customers on the servers which are in the last (the rth) phase of the service time, i.e.

$$g(\overrightarrow{k}) = g(k_1, e_1, \ldots, k_j, e_j, k_{j+1}) = \sum_{i=1}^{j} \mathbb{I}(e_i = r).$$

Here $\mathbb{I}(A)$ is an indicator function of the event A. Then the rates of the auxiliary processes $Y_1(t)$ and $Y_2(t)$ are given by the formulas

$$\lambda_{Y_1} = \mu \sum_{\overrightarrow{k} \in \mathcal{K}_1} j(\overrightarrow{k}) \mathsf{P}_1(\overrightarrow{k}), \quad \lambda_{Y_2} = \mu_r \sum_{\overrightarrow{k} \in \mathcal{K}_2} g(\overrightarrow{k}) \mathsf{P}_2(\overrightarrow{k}). \tag{6}$$

Now we define the traffic rate for the system S_i as follows

$$\rho^{(i)} = \frac{\lambda_X}{\lambda_{Y_i}}, \quad (i = 1, 2). \tag{7}$$

Intuitively, it is clear that the system is stable when $\rho^{(i)} < 1$ and it is unstable otherwise. The main stability result of the paper consists of the formal proof of this fact.

Synchronization of the input flow and auxiliary service processes

First we construct the common regeneration points for the input flow $X(t)$ and auxiliary process $Y_i(t)$. We fix an arbitrary state $\overrightarrow{k}$ of the Markov chain $U_i(t)$ assuming that $\mathsf{P}_i(\overrightarrow{k}) > 0$. Let $\{t_n^{(\overrightarrow{k})}\}_{n=1}^{\infty}$ be the moments of hits $U_i(t)$ into the state $\overrightarrow{k}$ so that

$$t_n^{(\overrightarrow{k})} = \min\{t_j > t_{n-1}^{(\overrightarrow{k})} : U_i(t_j + 0) = \overrightarrow{k}\}, \quad (t_0^{(\overrightarrow{k})} = 0), \quad n = 1, 2, \ldots$$

Here $\{t_j\}_{j=1}^{\infty}$ is the sequence of the timing of each jump $U_i(t)$. Then $\{t_{n+1}^{(\overrightarrow{k})} - t_n^{(\overrightarrow{k})}\}_{n=1}^{\infty}$ is a sequence of iid random variables and $t_n^{(\overrightarrow{k})}$ is the nth regeneration point of $Y_i(t)$.

Moreover $\mathsf{E}(t_{n+1}^{(\overrightarrow{k})} - t_n^{(\overrightarrow{k})}) < \infty$ and $Y_i(t)$ is the strongly regenerative flow. It means that its regeneration period is of the form

$$t_{n+1}^{(\overrightarrow{k})} - t_n^{(\overrightarrow{k})} = \eta_n^{(\overrightarrow{k})} + \nu_n^{(\overrightarrow{k})}.$$

Here $\eta_n^{(\overrightarrow{k})}$ and $\nu_n^{(\overrightarrow{k})}$ are independent random variables and $\eta_n^{(\overrightarrow{k})}$ is the time which $U_i(t)$ is in the state $\overrightarrow{k}$. Since $U_i(t)$ is a Markov chain, $\eta_n^{(\overrightarrow{k})}$ has an exponential distribution with rate $j(\overrightarrow{k})\mu$ for $U_1(t)$ and $\sum_{i=1}^{j} \mu_{e_i}$ for $U_2(t)$.

Let us define the common regeneration points $\{T_n^{(\overrightarrow{k})}\}_{n=1}^{\infty}$ for $Y_i(t)$ and $X(t)$ with regeneration points $\{\theta_n\}_{n=1}^{\infty}$ as

$$T_n^{(\overrightarrow{k})} = \min\{\theta_l > T_{n-1}^{(\overrightarrow{k})} : \bigcup_{i=1}^{\infty}(t_j^{(\overrightarrow{k})} \leqslant \theta_l < t_j^{(\overrightarrow{k})} + \eta_j^{(\overrightarrow{k})})\}, n = 1, 2, \ldots, \tag{8}$$

$$T_0^{(\overrightarrow{k})} = 0.$$

For fix $\overrightarrow{k} \in \mathcal{K}_i$ with $P_i(\overrightarrow{k}) > 0 (i = 1, 2)$ we define

$$\Delta_X^{(\overrightarrow{k})}(n) = X(T_{n+1}^{(\overrightarrow{k})}) - X(T_n^{(\overrightarrow{k})}),$$

and

$$\Delta_{Y_i}^{(\overrightarrow{k})}(n) = Y_i(T_{n+1}^{(\overrightarrow{k})}) - Y_i(T_n^{(\overrightarrow{k})})$$

$n = 1, 2, \ldots$. Then $\{\Delta_X^{(\overrightarrow{k})}(n)\}_{n=1}^{\infty}$ and $\{\Delta_{Y_i}^{(\overrightarrow{k})}(n)\}_{n=1}^{\infty}$ are sequences of iid random variables and w.p.1

$$\lambda_X = \lim_{t\to\infty} \frac{X(t)}{t} = \frac{\mathsf{E}\Delta_X^{(\overrightarrow{k})}(1)}{\mathsf{E}(T_2^{(\overrightarrow{k})} - T_1^{(\overrightarrow{k})})}, \quad \lambda_{Y_i} = \lim_{t\to\infty} \frac{Y_i(t)}{t} = \frac{\mathsf{E}\Delta_{Y_i}^{(\overrightarrow{k})}(1)}{\mathsf{E}(T_2^{(\overrightarrow{k})} - T_1^{(\overrightarrow{k})})}.$$

Therefore the traffic rate defined by (7) can be rewritten as follows

$$\rho^{(i)} = \frac{\mathsf{E}\Delta_X^{(\overrightarrow{k})}(1)}{\mathsf{E}\Delta_{Y_i}^{(\overrightarrow{k})}(1)}. \tag{9}$$

Note that it is true for any state $\overrightarrow{k} \in \mathcal{K}_i$ such that $\mathsf{P}_i(\overrightarrow{k}) = \lim_{t\to\infty} \mathsf{P}(U_i(t) = \overrightarrow{k}) > 0 (i = 1, 2)$.

Stability criterion

Theorem 2

Let $Q(t)$ be the number of customers at the system $S_i (i = 1, 2)$ at instant t. Then $Q(t)$ is a stable process if and only if

$$\rho^{(i)} < 1.$$

The proof is based on results obtained in the first part of the talk (Theorem 1).
In order to calculate the traffic rates we need to find the limit distributions of the control processes $U_i(t)(i = 1, 2)$. Unfortunately, for $U_2(t)$ we could not do it up to now but late we give an example demonstrating how it can be done. Let t_n be the moment of departure of the nth customer. Then

$$\lim_{n\to\infty} \mathsf{P}(U_1(t_n) = \overrightarrow{k}) = \pi_{\overrightarrow{k}} = p_{k_1} \cdot \cdots \cdot p_{k_{j+1}}. \tag{10}$$

Consider a renewal process $\{\zeta_j\}_{j=1}^{\infty}$ and a counting process

$$N(n) = \max\{j : Z_j \leqslant n\}, \quad \text{where} \quad Z_j = \zeta_1 + \cdots + \zeta_j, \quad Z_0 = 0.$$

Then we have

$$\lambda_Y = \mu \left(\sum_{j=1}^{m} \frac{1}{j} \sum_{\overrightarrow{k} \in \mathcal{K}_{|}} \prod_{i=1}^{j+1} p_{k_i} \right)^{-1} = \mu \left(\sum_{j=1}^{m} \frac{1}{j} \mathsf{P}(Z_j \leq m, Z_{j+1} > m) \right)^{-1} =$$

$$= \mu \left(\sum_{j=1}^{m} \frac{1}{j} \mathsf{P}(N(m) = j) \right)^{-1} = \mu \left(\mathsf{E} \frac{1}{N(m)} \right)^{-1}.$$

Example. Queueing systems with two servers

Here we consider the systems S_1 and S_2 with two servers and two-phase distribution of the service time in S_2. Thus, the service time η in S_2 has a form $\eta = \eta^{(1)} + \eta^{(2)}$ where $\eta^{(1)}$ and $\eta^{(1)}$ are independent exponentially distributed random variables with rates μ_1 and μ_2 respectively. The service time in S_1 has an exponential distribution with the same mean $\mu^{-1} = \frac{1}{\mu_1} + \frac{1}{\mu_2} = \frac{1+\delta}{\mu_2}$ where $\delta = \frac{\mu_2}{\mu_1}$. We shall compare traffic rates for these systems and show that $\rho^{(2)} > \rho^{(1)}$ if $0 < \delta < \infty$. For S_1 we have

$$\rho^{(1)} = \lambda_X \frac{(1+\delta)(2-p_1^2)}{2\mu_2} \tag{11}$$

where $p_1 = \mathsf{P}(\zeta = 1)$.

To obtain $\rho^{(2)}$ we have to calculate the limit distribution $\mathsf{P}_2(\overrightarrow{k}) = \lim\limits_{t\to\infty} \mathsf{P}(U_2(t) = \overrightarrow{k})$. As the state space for $U_2(t)$ we take the set $\mathcal{K}_2 = \{(2, e_1), (1, e_1, 1, e_2), (1, e_1), e_i = 1, 2\}$. The state $(2, e_1)$ means that there is one serving customer which occupies two servers and e_i is the phase of the service time. The interpretation of the residual states is evident. Let us number the states by such a way

$$\{1\} = (2,1); \quad \{2\} = (2,2); \quad \{3\} = (1,1,1,1); \quad \{4\} = (1,2,1,2);$$

$$\{5\} = (1,2,1,1); \quad \{6\} = (1,1); \quad \{7\} = (1,2).$$

One may easy verify that

$$x_1 = \frac{2\delta(1+\delta)(1-p_1)}{(2-p_1^2)(1+2\delta)+2\delta^2}, \quad x_2 = \frac{x_1}{\delta}, \quad x_3 = \frac{(1+\delta-p_1)p_1^2}{2(1+\delta)(1-p_1)}x_1,$$

$$x_4 = \frac{p_1^2}{(1+\delta)(1-p_1)}x_1, \quad x_5 = \frac{p_1^2}{2\delta(1+\delta)(1-p_1)}x_1,$$

$$x_6 = \frac{p_1(1+\delta-p_1)}{1+\delta}x_1, \quad x_7 = \frac{p_1}{\delta}x_1.$$

$$\lambda_{Y_2} = \mu_2(x_2 + 2x_4 + x_5 + x_7) = 2\mu_2 \frac{1+\delta}{(2-p_1^2)(1+\delta)^2 - p_1^2\delta(1-p_1)}.$$

Therefore, the traffic rate

$$\rho^{(2)} = \rho^{(1)} \left(1 - \frac{\delta p_1^2(1-p_1)}{(2-p_1^2)(1+\delta)^2}\right).$$

We see that $\rho^{(2)}{=}\rho^{(1)}$ in the four cases: 1) $p_1 = 0$; 2) $p_1 = 1$; 3) $\delta = 0$; 4) $\delta = \infty$.
In the first case we have a classical model $Reg|G|2$, in the second case the system $Reg|G|1$ and in the rest cases service time in the system S_2 has an exponential distribution. In the rest cases the traffic rate $\rho^{(2)}$ is less then $\rho^{(1)}$ and $\rho^{(2)}(\delta)$ takes the maximum value when $\delta = 1$, that is $\mu_2 = \mu_1$.

Conclusion

In this talk we considered stability problem for multiserver queues with a regenerative input flow. Let us note that stability analysis is an essential and challenging stage of the investigation of stochastic models. However stability conditions may be of independent interest. In particular, the stability criterion of the multi-server model can be used for the capacity planning of a model and estimation of the upper bound of potential energy saving. The main contribution of this paper is an extension of the stability criterion to the model with a regenerative input flow. The method we use has the following steps. Firstly, we define an auxiliary process $Y(t)$ that is the number of customers which are served at the system if always there are customers for service. Secondly, assuming that this process is a regenerative flow not depending on the input flow $X(t)$ under some additional conditions we construct the common points of regeneration of $Y(t)$ and $X(t)$. This step we call synchronization of the processes.

This approach allows us to use results from the renewal theory for the stability analysis of the systems satisfying additional conditions. One may think that these conditions are too restrictive to be useful for the analysis of the real models. Therefore we apply our approach to stability analysis of two classical systems: the system with interruptions and the system with simultaneous service. It follows from our results that the structure of the input flow does not effect on the stability condition. One has to know only the intensity of this flow to estimate the traffic rate. But the distribution of the service time plays an essential role.

- L.G.Afanasyeva, E.E. Bashtova (2014) Coupling method for asymptotic analysis of queues with regenerative input and unreliable server. Queueing Systems, 76(2):125–147.
- L.Afanasyeva, A.Tkachenko (2014) Multichannel queueing systems with regenerative input flow. Theory of Probability and Its Applications, 58(2):174–192.
- S.Asmussen (2003) Applied Probability and Queues, Springer-Verlag.
- P.Brill, L.Green (1984) Queues in which customers receive simultaneous service from a random number of servers: A system point approach. Management Science, 30(1): 51–68.
- N.M. Van Dyk(1989) Blocking of finite inputs which require simultaneous servers with general think and holding times Operation Research Letters, 8(1): 45–52.
- D.Gaver Jr (1962) A waiting line with interrupted service, including priorities. Journal of the Royal Statistical Society. Series B (Methodological), 24:73–90.

- A.Krishnamoorthy, P.Pramod, S. Chakravarthy (2012) Queues with interruptions: a survey, TOP 1-31doi:10.1007/s11750-012-0256-6.
- E.Morozov, D.Fiems, H.Bruneel (2011) Stability analysis of multiserver discrete-time queueing systems with renewal-type server interruptions. Performance Evaluation, 68(12):1261–1275. doi:10.2016/j.peva.2011.07.002.
- E.Morozov, A.Rumyantsev (2016) Stability Analysis of a $MAP|M|s$ Cluster model by Matrix-Analytic Method. European Workshop on Performance Engineering, 63-76.
- A.Rumyantsev, E.Morozov (2015) Stability criterion of a multi-server model with simultaneous service. Annals of Operations Research, 1-11.
- H.Thorisson (2000) Coupling, Stationary and Regeneration, Springer, New York.
- W.Whitt (1985) Blocking when service is required from several facilities simultaneously. AT&T Technical Journal, 64(8): 1807-1856.