

A Heterogeneous Queuing System with Mixed Operating Modes as a Model of Serverless Computing in Cloud Environments

OLEKSANDR O. KYRYCHENKO, IHOR MALYK, OKSANA L. KYRYCHENKO

Department of Mathematical Problems of Control and Cybernetics,

Yuriy Fedkovych Chernivtsi National University,

2 Kotsiubynskoho Street, Chernivtsi,

UKRAINE

Abstract: - The paper considers the model of a heterogeneous queuing system $H_K(t)/M/\infty$ and limit evolutions in the averaging and diffusion approximation scheme. The input process of the system under study is modelled as a mixture of independent heterogeneous Poisson processes, which reflects the behaviour of modern cloud serverless platforms, where tasks arrive simultaneously from different event sources (HTTP gateways, object stores, message queues, event buses). Based on the apparatus of semi-Markovian random evolutions, the weak convergence of the queue length process $L(t)$ in two asymptotic schemes: averaging (law of large numbers) and diffusion approximation (central limit theorem), with the Ornstein–Uhlenbeck limit process, is proved. As a result, simple analytical estimates of the queue length $L(t)$ and the probability of exceeding the maximum length $P\left(\sup_{t \in [0, T]} L(t) < L_{MAX}\right)$ are constructed. Also, a necessary and sufficient condition for the queue's finiteness is expressed in terms of the parameters of the input and task-processing processes. The results obtained enable us to develop analytical methods for parameter estimation and predictive scaling for complex serverless systems in cloud environments, without the need to train models or neural networks on large amounts of data.

Key-Words: - queue; heterogeneous queuing system; Poisson process; semi-Markov process; queue length estimation; model; modelling; cloud computing; serverless.

Received: April 23, 2026. Revised: June 29, 2026. Accepted: August 3, 2026. Published: October 26, 2026.

1 Introduction

The motivation for this work is to use stochastic process models to analyze the characteristics of systems with random mixed modes of operation. Such systems naturally arise in modern cloud computing, where the same computing platform simultaneously serves heterogeneous task flows: synchronous API calls, asynchronous events from object stores, messages from queues, and other types of triggers. Each event source has its own statistical characteristics, so aggregated traffic cannot be accurately described by a single Poisson flow, whether of constant or variable intensity. In turn, the standard reactive automatic scaling mechanisms of such platforms lead to problems of cold start, inefficient use of resources and increased delays in queue systems, which makes the task of building analytical models of complex structure capable of providing quality of service assessments relevant. Queueing-theoretic models of this kind are widely used for capacity planning and service performance evaluation across a broad range of applied domains, from the utilization of airport

runways under overload conditions [1] to modern distributed computing infrastructures.

To more clearly illustrate the system, let us consider two examples that simulate real-world situations, described by almost the same assumptions, but with very different content. In the first example, we consider a queuing system with a non-uniform arrival process, $M(t)/M/c$, where the arrival intensity is $\lambda(t)$. In this case, the number of tasks in the interval $[t_1, t_2]$ will be proportional to the average intensity of requests in this interval, i.e.

$$E(N(t_2) - N(t_1)) = \int_{t_1}^{t_2} \lambda(t) dt \geq 0. \quad (1)$$

The physical and applied meaning of relation (1) is clear, taking into account the properties of the Poisson process $N(t)$ for the receipt of applications with variable intensity $\lambda(t)$. Let us consider the second example of task receipt, which is based on the mixing of several streams of application receipt, each treated as a stream of tasks from different sources. As a mathematical model, it will be assumed that the mathematical model of the receipt

of applications will be a mixture of the independent processes of receipt of applications $N_1(t), \dots, N_K(t)$ with the corresponding weights $w_1(t), \dots, w_k(t)$, i.e.

$$N(t) = \sum_{i=1}^K w_i(t) N_i(t) \quad (2)$$

where $w_i(s) \in (0,1]$ – normalized weights, i.e.

$$\sum_{i=1}^K w_i(t) = 1, t \geq 0.$$

The two examples considered above are very similar but reflect conceptually opposite applied problems with different task-arrival natures (applications). The difference in the considered approach lies precisely in distinguishing between task arrivals of different natures, varying in intensity or complexity. This concept is very similar to the Markov-modulated Poisson Process (MMPP) [2], since in the MMPP process, the main averaging factor is the average intensity of task arrivals, which is determined by the following relation

$$\tilde{\lambda}(t) = \sum_{i=1}^K w_i(t) \lambda_i(t), \quad (3)$$

where $\lambda_i(t)$ is the intensity of the arrival from the i -th channel with the probability of arrival $w_i(t)$ at time t . It is clear that the random processes $N(t)$ defined by the relation (2) and the random process $\tilde{N}(t)$ (3) have different probabilistic characteristics; the corresponding QS (queuing system) described by these two processes under the conditions of all other characteristics are different. This fact, for example, can be shown by calculating the variances of the random processes $N(t)$ and $\tilde{N}(t)$ under the condition of pairwise independence of the random processes $N_1(t), \dots, N_K(t)$:

$$\begin{aligned} \text{Var}(N(t)) &= \sum_{i=1}^K w_i^2(t) \lambda_i(t); \\ \text{Var}(\tilde{N}(t)) &= \sum_{i=1}^K w_i(t) \lambda_i(t). \end{aligned}$$

Since $w_i(t) \in [0,1]$, we always have $w_i^2(t) \leq w_i(t)$, so:

$$\text{Var}(N(t)) \leq \text{Var}(\tilde{N}(t)).$$

This difference is significant in applied problems, as the variance of the input stream directly affects the accuracy of queue-length estimates and the probability of exceeding concurrent execution limits. The MMPP model assumes that all event sources form a single stream with an averaged intensity, whereas model (2) accounts for each source generating an independent stream with its own statistical characteristics. That is

why model (2) is better suited to systems with heterogeneous triggers than the classical MMPP. As shown below, variance is an important factor in analyzing limit evolutions within the diffusion approximation scheme [3].

Therefore, the main mathematical object of the study of this work is a heterogeneous queuing system, which is described by the following characteristics:

- The arrival process characterizes the input process (2) with known parameters $(w_i(t), \lambda_i(t), i = 1, \dots, K), t \geq 0$; this approach allows us to explicitly take into account the heterogeneity of event sources that form the load on the system;

- The processing time is characterized by an exponential distribution with an infinite number of servers (i.e., sufficiently many servers) and a processing intensity $\mu_i, i \geq 0$. We will assume that task processing on the servers is independent; that is, task processing times are independent. We will also assume that each server is characterized by the cost c_i per unit of time and α_i – the average number of rejected tasks per unit of time under the condition of the operation of the corresponding server;

- For simplicity, we will denote by $A(t)$ the set of active servers at time t , i.e. those servers that are either processing a task or are ready (paid) to process the task, by $A_W(t)$ the set of active servers at time t , which are processing a task at this time. Since in the infinite-server mode, the power $A_W(t)$ coincides with the number of tasks in processing, the total intensity of completions at time t is equal to:

$$\mu(t) = \sum_{i \in A_W(t)} \mu_i = \bar{\mu} L(t),$$

$\bar{\mu} = \frac{1}{E(S)}$ (S the averaged processing time of a task);

- The number of tasks in the system queue is unlimited (Mode 1) and limited L_{MAX} (Mode 2). This assumption allows us to estimate the number of rejected tasks, and therefore the number of unprocessed tasks;

- The number of processes of task arrival K based on formula (2) is a constant, i.e. this parameter does not depend on time;

- The intensities of arrival $\lambda_i(t)$ are bounded functions, i.e. $0 \leq \lambda_i(t) \leq \lambda_{MAX}$;

- $\Lambda(T)$ та $M(T)$ – average arrival and completion intensities over $[0, T]$, $\lambda(t)$ and $\mu(t)$ – the corresponding instantaneous intensities.

Such assumptions directly correspond to the typical architecture of modern serverless platforms, in which processing is characterized by a practically unlimited number of isolated containers

(corresponding servers), and the limitation of Mode 2 arises from synchronous calls with hard limits on the number of simultaneous executions. That is why the analytical theory developed below has direct applicability to optimization problems, predictive autoscaling, and SLA estimation for such systems.

The novelty of this work lies in the development of a theoretical basis for obtaining the coefficients of the limit process for a heterogeneous queuing system $H_K(t)/M/\infty$ with mixed operating modes in a closed form through the initial parameters of the model, while in the previous work of the authors [4] only the formulation of the necessary and sufficient condition for the boundedness of the queue and the theorem on weak convergence in the averaging and diffusion approximation schemes are given, without proofs and without calculating these coefficients. Also, estimates of the probability of exceeding the limit on simultaneous executions are obtained in explicit form, and a system with a bounded queue $H_K(t)/M/L_{MAX}$ is modelled as an Ornstein–Uhlenbeck process. The obtained results are analytical, do not require training data and provide a guarantee of the quality of service in a closed form.

2 Literature Review

A significant number of modern scientific works are devoted to applying the central limit theorem to the analysis of QS with a large or random number of servers, a characteristic feature of platforms with automatic scaling. In [5], the central limit theorem for a QS consisting of a random number of servers N with intensities $\mu_k, k \geq 1$, is considered, and the main assumption is made regarding the independence of these servers, i.e., the work is based on one of the main assumptions of the central limit theorem. The authors used the following assumption regarding the characteristics of the intensities:

$$\mu_k^N = \tilde{\mu}_k + \frac{1}{\sqrt{n}} \hat{\mu}_k, k \geq 1,$$

where the pair $(\tilde{\mu}_k, \hat{\mu}_k), \tilde{\mu}_k > 0$ completely determines the asymptotic behavior of the intensity of the k -th server at $N = n$. For the input process, the intensity of arrivals is determined by the number of available servers, with the following asymptotes:

$$\lim_{n \rightarrow \infty} \frac{\lambda^n}{n} = \lambda > 0$$

and

$$\lim_{n \rightarrow \infty} \frac{\lambda^n - n\lambda}{\sqrt{n}} = \hat{\lambda}.$$

The number of arrivals up to time t is determined by the following relation $\sup\{l > 0 | \sum_{i=1}^l \frac{U_i}{\lambda^n} \leq t\}$, where U_i are strictly positive independent identically distributed random variables with $EU_i = 1$. Then, according to this work, the number of applications in the system $X^n(t)$ as $n \rightarrow \infty$ has the following limit

$$X(t) = X_0 + \sigma W(t) + \beta t + \gamma \int_0^t \xi(s) ds,$$

where $W(t), t \geq 0$, is a standard Wiener process, and the parameters σ, β and γ depend on the initial parameters of the process for receiving and processing tasks.

In [6], the central limit theorem for queuing systems is also considered, and it is assumed that the QS is described by a Markov chain $Q^n(k)$, defined for the arrival rate $\lambda_n = n$ and the order processing rate $\mu_n = a(n + \sqrt{n})$. In this case, the normalized and centered random process

$$Y^n(t) = X^n([(\lambda_n + \mu_n)t]),$$

where $X^n(k) = \frac{Q^n(k) - \lambda_n}{a\sqrt{n}}$ converges weakly to the Wiener process $W(t)$, or more precisely, the convergence

$$\begin{aligned} \lim_{n \rightarrow \infty} P(Y^n(t) \leq x | Y^n(0) = y) = \\ = \frac{1}{\sqrt{2\pi\sigma^2 t}} \int_{-\infty}^x e^{-\frac{1}{2\sigma^2 t}(z-y)^2} dz. \end{aligned}$$

Another work that studies the asymptotic properties of QS for a large number of servers and a rapid increase in the arrival intensity λ_n is [7], which considers G/G/m QS and their asymptotic characteristics. One of the most useful results of this work is the approximation of the general QS G/G/m using the simpler and better-studied M/M/c QS, for example, the relation for estimating the waiting time in both systems. One of the approximations in this work is the approximation of the waiting time for the GI/M/m model:

$$\begin{aligned} P(W(GI/M/m; c_a^2) > 0) \approx \\ \approx \min \left\{ 1, \frac{1 - \Phi(2(1 - \rho)\sqrt{m}(1 + c_a^2)^{-1})}{1 - \Phi((1 - \rho)\sqrt{m})} P(W(M/M/m)) \right\}, \end{aligned}$$

where $c_a^2 = \frac{DX}{(EX)}$ is the square of the coefficient of variation for the random variable X , which determines the receipt of applications for a certain fixed time interval (for the exponential distribution $c_a^2 = 1$). Also, important results of this work include a comparative analysis of different QS, and the authors consider the $H_2/M/m$ QS, which is the simplest process built on mixtures.

A simpler approach to determining the optimal number of servers for a QS with a random task-processing intensity μ is described in [8]. In this case, the authors propose a simple approximation of the number of servers $m \approx E\left(\frac{\lambda}{\mu}\right) + z_\alpha \sigma_{\frac{\lambda}{\mu}}$, where $\sigma_{\frac{\lambda}{\mu}}$ is the standard deviation for the random variable $\frac{\lambda}{\mu}$, z_α is the quantile of level α for the standard normal distribution. In the case of some restrictions on the distribution of the random variable $\frac{\lambda}{\mu}$ (for example, a normal distribution), an adjustment to the optimal number of servers is considered:

$$m \approx E\left(\frac{\lambda}{\mu}\right) + z_\alpha \sqrt{E\left(\frac{\lambda}{\mu}\right) + D\left(\frac{\lambda}{\mu}\right)}.$$

In [9], the problem of determining the optimal number of servers for the so-called fuzzy queueing system is considered, namely, the M/M/s:K model, where K is the maximum permissible number of servers and s is the actual number of servers in operation.

In this work, the following assumption is made regarding the intensity of application receipt and processing:

$$\lambda_n = \begin{cases} (K-n)\lambda, n = 0, \dots, K, \\ 0, n > K, \end{cases}$$

$$\mu_n = \begin{cases} n\mu, n = 1, \dots, s; \\ s\mu, n > s. \end{cases}$$

Unlike previous works, this work is based on an optimization problem that minimizes the cost of processing requests, i.e.

$$E(CT) = E(CS) + E(CE) \rightarrow \min,$$

where CS is the average total cost per unit of time, CE is the average cost of processing a request per unit of time. The work defines the transition probabilities of the corresponding discrete Markov chain with continuous time based on the following Kolmogorov-Chapman equation

$$p'(t) = A(\lambda, \mu)p(t),$$

where $A(\lambda, \mu)$ is the generator of the Markov process defined based on the intensities λ_n and μ_n , $p_i(t)$ is the probability that at time t the system contains exactly i requests and

$$\pi = \lim_{t \rightarrow \infty} p(t)$$

is the system's stationary distribution. The work presents an algorithm for estimating the optimal number of servers based on the optimization problem defined above and the stationary distribution π .

In [10], the $M^X/M/m$ queueing system with an infinite number of servers $m = \infty$ is considered. Moreover, the paper assumes that this QS is described by a Markov process $J(t), t \geq 0$, with generator Q ; independent non-negative random variables determine the number of simultaneous requests X_i , satisfying $E \log(X_i + \varepsilon) < \infty$ for some $\varepsilon > 0$. The main result of the paper is the fact that the random variable $N^{\frac{\beta}{2}} \left(\frac{M^{(N)}}{N} - \rho \right)$, where $\beta = \min(\alpha, 1)$, α is the scaling factor for the arrival time, $M^{(N)}$ is the normalized number of tasks in the queue, converges to a normal distribution with zero mean and a variance depending on the QS parameters and the scaling factor α . Using this statement, the authors proposed an estimate of the number of servers in the QS based on the queue length:

$$P(M^{(N)} \leq x) \approx \Phi_{(0, \sigma^2)} \left(N^{\frac{\beta}{2}} \left(\frac{x}{N} - \rho \right) \right),$$

where $\Phi_{(0, \sigma^2)}(x)$ is the distribution function of the normal distribution with parameters $(0, \sigma^2)$. Thus, based on this statement, it is possible to estimate the optimal number of servers for a bounded process $M^{(N)}$ on the finite interval $[0, T]$.

In [11], a scheme for estimating the optimal number of clusters based on a normal distribution is also considered for a closed-loop queueing system with M steps, where r_i parallel servers with intensity μ_i . Describe the i -th step. It is shown that the joint probability that at each step $i = 1, \dots, M$ will approximately obey a normal distribution, i.e.

$$P(n_1, n_2, \dots, n_K) \approx$$

$$\approx \left(\prod_{i=1}^K \gamma_i(n_i, \alpha_i) \right) \left(\sum_{i=1}^M \sigma_i^2 \right) *$$

$$* \left(\sum_{i=K+1}^M \sigma_i^2 \right)^{-1} e^{-(N_K - \bar{N}_K)^2 (\sum_{i=K+1}^M \sigma_i^2)^{-1}},$$

where n_i is the number of unprocessed tasks at the i -th step, K is the number of initial steps for which the evaluation was performed, N_K is the number of unprocessed requests at these steps, σ_i^2 is the variance of the number of unprocessed tasks at the i -th step. Thus, this result allows us to estimate the optimal number of servers for a multi-step QS based on the preliminary evaluation and determination of the total number in the QS

$$P(L \leq N) = \sum_{n_1 + n_2 + \dots + n_M \leq N} P(n_1, n_2, \dots, n_K).$$

In [12], a QS with M servers is considered, and the key assumption is that servers can fail and be

sent for repair. An additional assumption is also made about the presence of a certain number of servers $M_a = L - M$, which are in reserve and can immediately replace servers that have failed. Similar to previous works, the article considers a non-uniform continuous-time Markov chain that describes a QS with the following arrival and processing intensities

$$\lambda_n = \begin{cases} M\lambda + (S - n)\alpha, n = 0, \dots, S, \\ (L - n)\lambda, n = S + 1, \dots, M + S = L, \\ 0, n > L, \end{cases}$$

$$\mu_n = \begin{cases} i\mu, n = i, \dots, M + S; \\ 0, n > M + S. \end{cases}$$

In [13], the problem of minimizing the cost of task execution under queue constraints in the QS is considered, with the cost measured by the number of servers (backup servers S and active servers R). The general optimization problem has the following form

$$T_{cost}(S, R) \rightarrow \min_{S, R}$$

provided that

$$A.V. = \sum_{i=0}^R \sum_{j=0}^S \pi_{i,j} > A,$$

where $T_{cost}(S, R)$ is the total cost of servicing (reservation) servers per unit of time, π is the stationary probabilities of the corresponding random process describing the QS. The preliminary analysis of the problem showed that the objective function $T_{cost}(S, R)$ is not convex; therefore, the problem cannot be solved using convex programming methods. Also, in this work, an analysis of three server usage policies was conducted, and the optimal solutions for each policy were examined. It was determined that the optimal policy is the no-delay server-repair policy when repair is available. The authors in [13] also use the stationary distributions of the corresponding Markov process to analyze the optimal number of servers for the $M/M/(S, s)$ queuing system. These systems are characterized by the fact that starting with S servers, the server that performed this task leaves the system simultaneously with the completed task, until the number of servers is equal to s , after which the number of servers is again restored to the number S . The paper describes the generating operator of the Markov process for the processes of receipt and processing of tasks with intensities λ and μ respectively; it is shown that the system has a non-degenerate stationary distribution at:

$$\lambda < \mu \frac{S - s}{\sum_{j=s+1}^S \frac{1}{j}}$$

and the stationary distribution of the corresponding Markov process is defined by the relation $\pi_{S+n} =$

$\pi_S R^n$, where R is the solution of the second-order auxiliary matrix equation $R^2 A_2 + R A_1 + A_0 = 0$.

In [14], an analysis of the QS was carried out, for which the main parameter is the process of incoming applications, which is characterized by a piecewise-stable random process $A(t)$ with probability 1, and in this case, unlike previous works, $A(t)$ is a semi-Markov process defined based on the Markov recovery process (A_n, τ_n) . One of the main assumptions in this work is the assumption of the averaged dynamics of $A(t)$

$$\left(\lim_{t \rightarrow \infty} \frac{A(t)}{t} = \frac{\lambda}{a} \right),$$

where $\lambda^{-1} = E(\tau_n - \tau_{n-1})$, $a^{-1} = E(A_n - A_{n-1})$. The main result of the work establishes the weak convergence of the process for the number of unprocessed tasks in the system to the Wiener process $W(t)$ with appropriate centering and normalization as $T \rightarrow \infty$. Thus, based on this limit theorem, it is once again possible to estimate the optimal number of servers by imposing a limit on the number of unprocessed requests over a sufficiently large interval $[0, T]$.

In [15], the main attention is paid to a two-stage tandem queuing system, in which tasks arrive according to a Poisson process with rate λ , and the processing intensities are μ_1 and μ_2 respectively, in the first and second steps. The costs of staying in the queue are h_1 and h_2 in the first and second steps, respectively. The main idea of the work is to use Markov decision processes over the set of admissible policies Π , where each policy $\pi \in \Pi$ determines both the server and the queue position for each task. The work focuses on minimizing the average cost of staying in the queue, which is defined by the following relation:

$$g(\pi, x) = \lim_{n \rightarrow \infty} \sup \frac{V_n(\pi, x)}{n},$$

where

$$V_n(\pi, x) = E_x \left[\sum_{k=0}^{n-1} c(X_k, d_k(X_k)) \right]$$

is the average cost of being in the queue for policy π at the initial value x , where X_k is the placement of tasks in queues at step k , d_k is the placement of tasks in queues for the k -th step, $c(x, d(x))$ is the cost of being in the queue. The paper shows that the described QS has a non-degenerate stationary distribution if $\lambda(\mu_1^{-1} + \mu_2^{-1}) < 2$; it is also shown that the optimization problem $g(\pi, x) \rightarrow \min_{\pi \in \Pi}$ has at least one solution, which in particular is optimal for each step.

Another result on parallel QS is considered in [16], which assumes that the intensity determines the input process λ , and that n servers are used in parallel, each with m_i cores and processing speed s_i . The input intensity is distributed among the servers in proportion to λ_i , where $\lambda = \lambda_1 + \dots + \lambda_n$. The paper determines that the optimal characteristics of the parallel QS are QS with the same processing speeds $s_1 = \dots = s_n$ for the idle-speed model, and $s_i = \left(\frac{\lambda_i}{\beta m_i}\right)^\alpha$, where α is an additional parameter for the power of the processor at speed s , namely s^α .

Thus, analysis of modern scientific research shows that asymptotic analysis, parameter optimization for queuing systems, and Markov models provide the theoretical basis for developing adequate mathematical models of complex queuing systems, including modern cloud serverless platforms. However, the vast majority of the cited works treat the input flow as a single process with an averaged intensity, which, as shown above, leads to a systematic underestimation of dispersion and can result in inadequate estimates of queue characteristics in systems with several independent sources of events. That is why in this work, the attention is focused on the mixture model (2) and the asymptotic analysis of its characteristics in two schemes – averaging and diffusion approximation.

3 Main Results

In [17], the authors propose the following notation for this QS – $(M(t)/M/\infty)^K|M$ for the case with an infinite queue and $(M(t)/M/L_{MAX})^K|M$ for the QS with a finite queue L_{MAX} , with the last letter denoting the independence of the task-processing processes. In our case, taking into account the fact that we are considering the normalized Poisson process (2), we will use the following notation for the QS – $H_K(t)/M/\infty$ or $H_K(t)/M/L_{MAX}$ for the cases of a finite and infinite queue. The block diagram of the QS under consideration is shown in Figure 1.

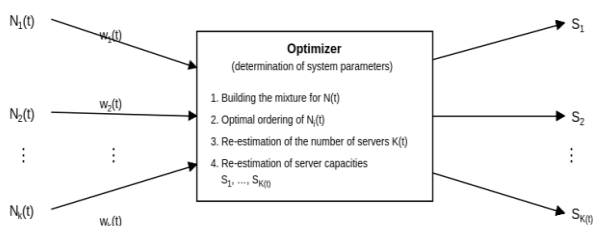


Fig. 1: Structural diagram of the QS for the cases of finite and infinite queues

Source: created by the authors

First, let us consider the result that determines the conditions under which the queue $L(t)$ grows to infinity over time in Mode 1. As was noted in classical works, the existence of a stationary regime of operation of the QS in the case $(w_i(t), \lambda_i(t))_{i=1}^K = (w_i, \lambda_i)_{i=1}^K$ or the boundedness of the queue $P(\forall t \geq 0: L(t) < \infty) = 1$ are determined exclusively by the intensities of the arrival and the task-processing processes. For further consideration, we will define the averaged intensities on a finite interval

$$0 \leq \Lambda(T) = \frac{1}{T} \int_0^T \sum_{i=1}^K w_i(t) \lambda_i(t) dt; \quad (4)$$

$$0 \leq M(T) = \frac{1}{T} \int_0^T \sum_{i \in A(t)} \mu_i dt. \quad (5)$$

Using these definitions, we define the conditions for the boundedness of the queue for $H_K(t)/M/\infty$ QS.

Lemma 1 [4]. For the queue $H_K(t)/M/\infty$ to be bounded, it is necessary and sufficient that

$$\lim_{T \rightarrow \infty} \frac{\Lambda(T)}{M(T)} < 1.$$

Proof. First, we note that the result of Lemma 1 is intuitively clear, since the average arrival intensity will be strictly less than the average processing intensity. Consider the random process

$$X(t) = N(t) - N_S(t),$$

where $N_S(t)$ – defines the process of processing tasks, i.e.

$$N_S(t) = \sum_{i \in A(t)} N_i(t),$$

where $N_i(t), i \geq 1$ – independent Poisson random processes with intensities μ_i . The considered process will go to $-\infty$, since

$$EX(T) = E(N(T)) - E(N_S(T)) = T(\Lambda(T) - M(T))$$

and under the conditions of independence of Poisson processes

$$\begin{aligned} D(X(T)) &= D(N(T)) + D(N_S(T)) = \\ &= \sum_{i=1}^K \int_0^T w_i^2(s) \lambda_i(s) ds + \int_0^T \sum_{i \in A(t)} \mu_i ds. \end{aligned}$$

Considering the fact that the weights $w_i(t) \in [0,1]$, we obtain

$$\begin{aligned} D(X(T)) &\leq \sum_{i=1}^K \int_0^T w_i(s) \lambda_i(s) ds + \int_0^T \sum_{i \in A(t)} \mu_i ds = \\ &= T(\Lambda(T) + M(T)). \end{aligned}$$

Using Chebyshev's inequality, we obtain

$$\begin{aligned} P\left(X(T) > T(\Lambda(T) - M(T)) + a\sqrt{T(\Lambda(T) + M(T))}\right) &\leq \\ &\leq P\left(|X(T) - T(\Lambda(T) - M(T))| \geq a\sqrt{T(\Lambda(T) + M(T))}\right) \leq \frac{1}{a^2}, a > 0. \end{aligned}$$

Taking into account the fact that $\Lambda(T) < M(T)$ for sufficiently large T and substituting $a = \sqrt[4]{T}$, we obtain that:

$$T(\Lambda(T) - M(T)) + \sqrt[4]{T}\sqrt{T(\Lambda(T) + M(T))} \rightarrow -\infty.$$

On the other hand

$$\begin{aligned} P\left(X(T) > T(\Lambda(T) - M(T)) + \sqrt[4]{T}\sqrt{T(\Lambda(T) + M(T))}\right) &\leq \\ &\leq \frac{1}{\sqrt[4]{T}} \rightarrow 0. \end{aligned}$$

Let us now consider the sequence of points $t_i = T \cdot h \cdot i$, $h = \frac{1}{N}$, $i = 1, \dots, N$ and $\Delta X(t_i) = X(t_i) - X(t_{i-1}) - E(X(t_i) - X(t_{i-1}))$. Using Etemadi's inequality, we obtain:

$$\begin{aligned} P\left(\max_{i=1, \dots, N} |X(t_i) - E(X(t_i))| \geq 3\sqrt[4]{T}\sqrt{T(\Lambda(T) + M(T))}\right) &\leq \\ &\leq 3 \max_{i=1, \dots, N} P\left(|X(t_i) - E(X(t_i))| \geq \sqrt[4]{T}\sqrt{T(\Lambda(T) + M(T))}\right). \end{aligned}$$

However,

$$\begin{aligned} \max_{i=1, \dots, N} P(|X(t_i) - E(X(t_i))| \geq \sqrt[4]{T}\sqrt{T(\Lambda(T) + M(T))}) &\geq \\ &\geq \sqrt[4]{T}\sqrt{T(\Lambda(T) + M(T))} \rightarrow 0 \end{aligned}$$

for arbitrary N , which, taking into account the piecewise stability of $X(t)$, proves $P(X(t) \rightarrow -\infty) = 1$.

It should also be noted that

$$P(\forall t \geq 0: L(t) < \infty) = P(X(t) \rightarrow -\infty),$$

which proves Lemma 1.

In the future, we will focus on the random processes describing the number of tasks in the queue, denoted by $L(t)$ and the number of failures, denoted $\alpha(t)$. Note that the average number of rejections, according to our assumptions, will be equal

$$E\alpha(t) = \sum_{i \in A_W(t)} \alpha_i.$$

Let us move on to a more detailed consideration of the dynamics of the random process $L(t)$, $t \geq 0$. The growth of this process on the interval $[t, t + \Delta]$ for $\Delta \ll 1$ will be determined by the following relation

$$P(L(t + \Delta) - L(t) = m | L(t) = k) = \begin{cases} \sum_{i \geq \max(m-k, 0)} \frac{(\Delta\lambda(t))^i e^{-\Delta\lambda(t)}}{i!} \frac{(\Delta\mu(t))^{i+k-m} e^{-\Delta\mu(t)}}{(i+k-m)!} \\ \text{where } m \geq -k + 1 \\ \sum_{i=0}^{\infty} \sum_{j=i+k}^{\infty} \frac{(\Delta\lambda(t))^i e^{-\Delta\lambda(t)}}{i!} \frac{(\Delta\mu(t))^j e^{-\Delta\mu(t)}}{j!}, \\ \text{where } m = -k. \end{cases}$$

where

$$\mu(t) = \sum_{i \in A(t)} \mu_i.$$

Let us consider in more detail the conditional mathematical expectation of growth, provided that $L(t) = k$:

$$\begin{aligned} E(L(t + \Delta) - L(t) | L(t) = k) &= \\ &= \sum_{m=-k}^{\infty} m P(L(t + \Delta) - L(t) = m | L(t) = k) = \\ &= -k * e^{-\Delta(\lambda(t) + \mu(t))} \sum_{i=0}^{\infty} \sum_{j=i+k}^{\infty} \frac{(\Delta\lambda(t))^i}{i!} \frac{(\Delta\mu(t))^j}{j!} + \\ &+ e^{-\Delta(\lambda(t) + \mu(t))} \sum_{m=-k+1}^{\infty} m \sum_{i \geq \max(m-k, 0)} \frac{(\Delta\lambda(t))^i}{i!} \frac{(\Delta\mu(t))^{i+k-m}}{(i+k-m)!}. \end{aligned}$$

For $k = 0$ we obtain

$$\begin{aligned} E(L(t + \Delta) - L(t) | L(t) = 0) &= \\ &= \sum_{m=1}^{\infty} m \sum_{i=m}^{\infty} \frac{(\Delta\lambda(t))^i e^{-\Delta\lambda(t)}}{i!} \frac{(\Delta\mu(t))^{i-m} e^{-\Delta\mu(t)}}{(i-m)!} > 0. \end{aligned}$$

Using the properties of the Skellam distribution, we conclude that $E(L(t + \Delta) - L(t) | L(t) = 0) =$

$$\Delta\lambda(t)P(N_{in} - N_{out} \geq 0) - \Delta\mu(t)P(N_{in} - N_{out} \geq 2),$$

where $N_{in} \sim \text{Poiss}(\Delta\lambda(t))$, $N_{out} \sim \text{Poiss}(\Delta\mu(t))$.

For sufficiently large $k > 0$ based on the definition of probabilities for the Poisson distribution for $\lambda(t) < \mu(t)$, we obtain that

$$\begin{aligned} \phi(k, \Delta, t) &= \\ &= E(L(t + \Delta) - L(t) | L(t) = k) = e^{-\Delta(\lambda(t) + \mu(t))} * \end{aligned}$$

$$* \left(\begin{aligned} & -k \sum_{i=0}^{\infty} \sum_{j=i+k}^{\infty} \frac{(\Delta\lambda(t))^i (\Delta\mu(t))^j}{i! j!} + \\ & \sum_{m=-k+1}^{-1} m \sum_{i \geq \max(m-k, 0)} \frac{(\Delta\lambda(t))^i (\Delta\mu(t))^{i+k-m}}{i! (i+k-m)!} + \\ & \sum_{m=1}^{k-1} m \sum_{i \geq \max(m-k, 0)} \frac{(\Delta\lambda(t))^i (\Delta\mu(t))^{i+k-m}}{i! (i+k-m)!} + \\ & \sum_{m=k}^{\infty} m \sum_{i \geq \max(m-k, 0)} \frac{(\Delta\lambda(t))^i (\Delta\mu(t))^{i+k-m}}{i! (i+k-m)!} \end{aligned} \right).$$

It is clear that the second term exceeds the third in absolute value, and the first exceeds the fourth, which implies $E(L(t+\Delta) - L(t) | L(t) = k) < 0$ for sufficiently large k .

Lemma 2 [4]. Let $0 < \lambda(t) < \mu(t)$. Then the conditional mathematical expectation $\phi(k, \Delta, t)$ has the following properties:

- $\phi(0, \Delta, t) > 0$;
- $\phi(k, \Delta, t)$ is a strictly decreasing function in k ;
- $\lim_{k \rightarrow \infty} \phi(k, \Delta, t) < 0$ for arbitrary $\Delta > 0, t \geq 0$.

Proof. Before proving Lemma 2, we note that $\phi(k, \Delta, t)$ can be defined in terms of independent random variables $N_{in} \sim \text{Poiss}(\Delta\lambda(t)), N_{out} \sim \text{Poiss}(\Delta\mu(t))$, namely

$$\phi(k, \Delta, t) = E(\max(-k, N_{in} - N_{out})).$$

The first fact follows from the definition of the function $\phi(k, \Delta, t)$, i.e.

$$\phi(0, \Delta, t) = E(\max(0, N_{in} - N_{out})) \geq 0$$

because $\max(0, N_{in} - N_{out}) \geq 0$. For the second one, the difference between

$$\begin{aligned} \phi(k, \Delta, t) - \phi(k+1, \Delta, t) &= \\ & E(\max(-k, N_{in} - N_{out})) - \\ & E(\max(-k-1, N_{in} - N_{out})) = \\ & E \left(\max(-k, N_{in} - N_{out}) - \right. \\ & \left. \max(-k-1, N_{in} - N_{out}) \right) = \\ & P(N_{in} - N_{out} \leq -k-1) \geq 0, \end{aligned}$$

because $\max(-k, N_{in} - N_{out}) - \max(-k-1, N_{in} - N_{out}) = 0$ for $N_{in} - N_{out} \geq -k$. Hence,

$$\phi(k, \Delta, t) - \phi(k+1, \Delta, t) \geq 0$$

or

$$\phi(k, \Delta, t) \geq \phi(k+1, \Delta, t).$$

To prove the last statement, we note that for large k

$$\begin{aligned} \lim_{k \rightarrow \infty} \phi(k, \Delta, t) &= \lim_{k \rightarrow \infty} E(\max(-k, N_{in} - N_{out})) = \\ & \lim_{k \rightarrow \infty} E(N_{in} - N_{out}) < 0 \end{aligned}$$

by assumption $0 < \lambda(t) < \mu(t)$.

Lemma 2 is proved.

Let us now consider the piecewise-constant semi-Markov process $\tilde{L}(t) = L(t_n), t \in [t_n, t_{n+1})$,

generated by the Markov renewal process $(\tilde{L}_n = L(t_n), \tau_n = t_n - t_{n-1})$ [3]. Note that, based on the construction, this semi-Markov process is defined by the semi-Markov kernel

$$Q(k, A, t) =$$

$$= P(\tilde{L}_n \in A | \tilde{L}_{n-1} = k, \tau_{n-1} = t) I_{\{t \leq \tau_n < t+\Delta\}}. \quad (6)$$

Using this process, we can find a limit representation for the semi-Markov process $\tilde{L}(t)$, which, based on the separability property of $L(t)$, will also be a limit representation for $L(t)$. For this purpose, we will use the averaging scheme [18], [19] and the diffusion approximation scheme [3], [20]. For the averaging scheme, we will define a family of semi-Markov processes $\tilde{L}^\varepsilon(t), t \geq 0$, for $\varepsilon > 0$, generated by the Markov renewal process $(\tilde{L}_n^\varepsilon = \varepsilon L_n, \tau_n^\varepsilon = \varepsilon \tau_n)$. For the diffusion approximation scheme, we will define a family of semi-Markov processes $\tilde{L}^\varepsilon(t), t \geq 0$, for $\varepsilon > 0$, which is defined by the following equality

$$\tilde{L}^\varepsilon(t) = \varepsilon L\left(\frac{t}{\varepsilon^2}\right) - \frac{\rho(t)}{\varepsilon} = \frac{1}{\varepsilon} \left(\varepsilon^2 \tilde{L}\left(\frac{t}{\varepsilon^2}\right) - \rho(t) \right),$$

where the function $\rho(t)$ will be defined below as a solution to the averaging equation, to formulate the main statement, similarly to [3], [21], we introduce the concept of a compensating operator for a semi-Markov process, defined as a Markov renewal process.

Definition 1. The compensating operator $\Gamma\phi(x), x \in R$ for a semi-Markov random process $\xi(t), t \geq 0$, which is defined by a semi-Markov kernel $Q(x, A, t), x \in R, A \in \beta_R, t \geq 0$, is called the operator

$$\Gamma\phi(x) = \frac{E\{\phi(\xi_n) - \phi(\xi_{n-1}) | \xi_{n-1} = x, \tau_{n-1} = t\}}{E\{\theta_n | \xi_{n-1} = x, \tau_{n-1} = t\}}. \quad (7)$$

Then the following statement holds. We also denote the intensity of the time spent in the state by:

$$b(x, t) = \frac{1}{E\{\theta_n | \xi_{n-1} = x, \tau_{n-1} = t\}}.$$

Theorem 1. The following convergences occur:

- In the averaging scheme, the random process $\tilde{L}^\varepsilon(t), t \in [0, T]$ weakly converges as $\varepsilon \rightarrow +0$ to the solution of the differential equation

$$\frac{d\rho(t)}{dt} = \frac{\phi(\rho(t), \Delta, t)}{\Delta} \quad (8)$$

on the finite interval $[0, T]$ with the initial condition $\rho(0) = L(0)$;

- In the diffusion approximation scheme, the random process $\tilde{L}^\varepsilon(t), t \in [0, T]$ weakly converges to the diffusion process $\tilde{L}^0(t), t \in$

$[0, T]$ as $\varepsilon \rightarrow +0$, which is defined by the Ito stochastic differential equation

$$d\tilde{L}^0(t) = C_1(\tilde{L}^0(t), \rho(t), t)dt + \sqrt{B(\rho(t), t)}dW(t) \quad (9)$$

on the finite interval $[0, T]$ with the initial condition $\tilde{L}^\varepsilon(0) = \tilde{L}^0(0)$, where $C_1(\tilde{L}^0(t), \rho(t), t)$, $B(\rho(t), t)$ are defined in the proof of the theorem.

Proof. To verify the convergence conditions for the averaging scheme, let us check the necessary conditions from [18]. First, the jumps $\tilde{L}_n - \tilde{L}_{n-1}$ have finite second-order moments, since, based on the construction of the random process $L(t)$,

$$\begin{aligned} D(\tilde{L}_n - \tilde{L}_{n-1}) &\leq D(X(t_n) - X(t_{n-1})) = \\ &= \int_{t_{n-1}}^{t_n} (\lambda(s) + \mu(s))ds < \infty. \end{aligned}$$

Also, the jump $\theta_n = \tau_n - \tau_{n-1} = \Delta$ satisfies the condition of uniform integrability and boundedness:

$$\begin{aligned} \sup_{k \geq 0} \int_T^{\infty} \bar{F}_k(t)dt &= 0; \\ Ee^{-\varepsilon\theta_n} &= e^{-\varepsilon\Delta} \leq 1 - \frac{\Delta}{2}\varepsilon, \varepsilon \ll 1. \end{aligned}$$

Thus, the statement about convergence in the averaging scheme is true.

Let us now consider the convergence study in the diffusion approximation scheme. To do this, we note that, according to the construction of the semi-Markov process $\tilde{L}^\varepsilon(t)$, $t \geq 0$, we obtain the following relations for the moments:

$$\begin{aligned} E\{\tilde{L}_n^\varepsilon | \tilde{L}_{n-1}^\varepsilon = x + \varepsilon x_1, \tau_{n-1} = t\} &= \\ &= \tilde{\alpha}(x, t) + \varepsilon \tilde{\alpha}_1(x, x_1, t) + o(\varepsilon); \\ E\{(\tilde{L}_n^\varepsilon - \tilde{L}_{n-1}^\varepsilon)^2 | \tilde{L}_{n-1}^\varepsilon = \\ x + \varepsilon x_1, \tau_{n-1} = t\} &= \tilde{B}(x, t) + o(1); \\ E\{\tilde{L}_n^\varepsilon | \tilde{L}_{n-1}^\varepsilon = \rho(t) + \varepsilon x_1, \tau_{n-1} = t\} &= \\ = E\{\tilde{L}_n^\varepsilon | \tilde{L}_{n-1}^\varepsilon = \rho(t), \tau_{n-1} = t\} + o(\varepsilon). \end{aligned}$$

Taking into account the definition of $L(t_n)$, we define the above parameters:

$$\begin{aligned} \tilde{\alpha}(x, t) &= x + (\lambda(t) - \mu(t)), \\ \tilde{\alpha}_1(x, x_1, t) &= x_1, \\ \tilde{B}(x, t) &= \lambda(t) + \mu(t) + (\lambda(t) - \mu(t))^2. \end{aligned}$$

Using this representation of moments and the definition of the compensating operator on the test functions $\phi \in C^2(R^1)$, we obtain:

$$\begin{aligned} E\{\phi(\tilde{L}_n^\varepsilon) - \phi(\tilde{L}_{n-1}^\varepsilon) | \tilde{L}_{n-1}^\varepsilon = x + \varepsilon x_1, \tau_{n-1} = t\} &= \\ \frac{E\{\theta_n | \tilde{L}_{n-1}^\varepsilon = x + \varepsilon x_1, \tau_{n-1} = t\}}{\varepsilon^{-1}b(x + \varepsilon x_1, t)} * \\ E\left\{\phi'(\tilde{L}_{n-1}^\varepsilon)(\tilde{L}_n^\varepsilon - \tilde{L}_{n-1}^\varepsilon) + \frac{1}{2}\phi''(\tilde{L}_{n-1}^\varepsilon)(\tilde{L}_n^\varepsilon - \tilde{L}_{n-1}^\varepsilon)^2 \middle| \tilde{L}_{n-1}^\varepsilon = x + \varepsilon x_1, \tau_{n-1} = t\right\}. \end{aligned}$$

Taking into account the fact that the averaged trajectory $\rho(t)$ is given by the differential equation (8) at $x_1 = \rho(t)$, we obtain

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0+} \Gamma^\varepsilon \phi(x) &= b(\rho(t), t) \tilde{\alpha}_1(\rho(t), x, t) \phi'(x) + \\ &+ \frac{1}{2} b(\rho, t) (\tilde{B}(\rho(t), t) - \tilde{\alpha}^2(\rho(t), t)) \phi''(x). \end{aligned}$$

The same result we obtain that all the convergence conditions in the diffusion approximation scheme [19] are satisfied, and therefore, there is a weak convergence of $\tilde{L}^\varepsilon(t)$, $t \geq 0$, to the solution of the Ito stochastic differential equation

$$d\tilde{L}^0(t) = C_1(\tilde{L}^0(t), \rho(t), t)dt + \sqrt{B(\rho(t), t)}dW(t),$$

where

$$\begin{aligned} C_1(x, \rho, t) &= b(\rho, t) \tilde{\alpha}_1(\rho, x, t), \\ B(\rho, t) &= b(\rho, t) (\tilde{B}(\rho, t) - \tilde{\alpha}^2(\rho, t)). \end{aligned}$$

Theorem 1 is proved.

Remark 1. Due to the time scaling $\frac{t}{\varepsilon^2}$ for the process $\tilde{L}^\varepsilon(t)$, $t \geq 0$, in the diffusion approximation scheme, the estimate of the deviation of the prelimit and limit evolution based on the Berry-Essen inequalities has the following form

$$\sup_{x \in R} |P(\tilde{L}^\varepsilon(t) \leq x) - P(\tilde{L}^0(t) \leq x)| \leq \frac{C\varepsilon}{\sqrt{t}},$$

where the constant C depends on the third moment of the nested chain $L(t_n)$, i.e.

$$C = C_0 \frac{E|L(t_n) - L(t_{n-1})|^3}{(\tilde{B}(\rho, t) - \tilde{\alpha}^2(\rho, t))^{\frac{3}{2}} |\lambda(t) - \mu(t)|},$$

provided $\lambda(t) \neq \mu(t)$, where the constant C_0 does not depend on the parameters of the model distribution.

Remark 2. As noted above, Mode 2 describes the system's operation, taking into account the queue constraint $L(t) \leq L_{MAX}$. In this case, Lemma 2 remains true, and the limiting diffusion process $\tilde{L}_{ref}^0(t)$, $t \geq 0$ will be described by the stochastic differential Ito equation taking into account reflection, i.e.

$$\begin{aligned} d\tilde{L}_{ref}^0(t) &= \\ C_1(\tilde{L}_{ref}^0(t), \rho(t), t)dt &+ \sqrt{B(\rho(t), t)}dW(t) - dR_t, \end{aligned}$$

where the process R_t describes the 'return' of the limiting process $\tilde{L}_{ref}^0(t)$ to the zone $[0, L_{MAX}]$, i.e.

$$R_t = \max_{0 \leq s \leq t} (\tilde{L}^0(s)),$$

where $\tilde{L}^0(t)$, $t \geq 0$, is the limiting process in the diffusion approximation scheme without restriction, specified in Theorem 1.

Consider the following example of the functioning of $H_K(t)/M/L_{MAX}$ system with the following parameters

$$\lambda(t) = t, \mu(t) = 0.2t^{1.8}, L_{MAX} = 11, \\ \tilde{L}^0(0) = \tilde{L}_{ref}^0(0) = 1.$$

The results of the simulation of the limit trajectories based on the Euler–Maruyama method [22] for $\tilde{L}^0(t), t \geq 0, \tilde{L}_{ref}^0(t), t \geq 0$ are shown in Figure 2.

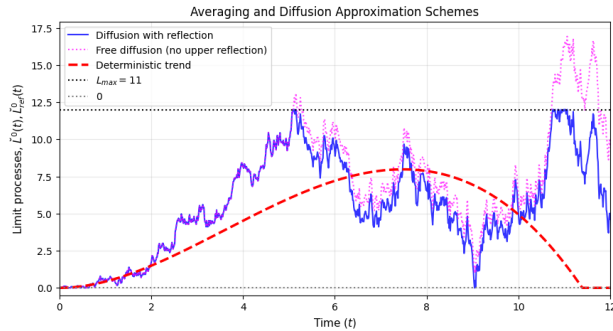


Fig. 2: Dynamics of Limiting Evolutions $\tilde{L}^0(t), t \geq 0, \tilde{L}_{ref}^0(t), t \geq 0$

Source: created by the authors

Based on the proved theorem, the inhomogeneous process $L(t), t \geq 0$, obeys the law of large numbers in the averaging scheme and the central limit theorem in the diffusion approximation scheme. In addition, the limiting process in the diffusion approximation scheme must be an Ornstein–Uhlenbeck process. It therefore has some asymptotic properties that will be decisive for our problem of approximating the queue length in the QS $H_K(t)/M/\infty$. One such property concerns the presence of an equilibrium point for the limiting processes (8) and (9). First, let us analyze the averaged equation (8), which is determined by the properties of the function $\phi(k, \Delta, t)$. According to Lemma 2, the averaged equation (8) will increase for small $\rho(t)$ and decrease for large $\rho(t)$, which means the presence of an equilibrium point for each fixed t :

$$\rho_0(\Delta, t) := \{x \in R_+ : \phi(x, \Delta, t) = 0\}.$$

The function $\rho_0(\Delta, t)$ is an equilibrium function for the solution $\rho(t)$ and in addition, the closeness of $\rho_0(\Delta, t)$ will be dictated by the ratio $\frac{\Lambda(t)}{M(t)}$, in the case $\frac{\Lambda(t)}{M(t)} \ll 1$, $\rho_0(\Delta, t)$ will be close to 0, which will correspond to the averaged trajectory of the random process $L(t), t \geq 0$. In addition, if the condition of Lemma 1 is met, the function $\rho_0(\Delta, t)$ is bounded. From an applied point of view, the value of $\rho_0(\Delta, t)$ reflects a typical stationary level of servers simultaneously occupied. It can serve as a guideline for planning the number of pre-reserved resources.

Note that $\rho_0(\Delta, t)$ is defined based on the averaged equation (8), the initial value $\rho_0(\Delta, t) \approx \rho(t)$ for large t and based on Lemma 1, $\rho_0(\Delta, t)$ can be easily found by the method of dividing the segment in half.

On the other hand, the Ornstein–Uhlenbeck process, given by the stochastic Ito differential equation (9), has the same equilibrium function $\rho_0(\Delta, t)$, i.e., the averaged solution (9) will be close to the deterministic function $\rho_0(\Delta, t)$ on average. Using the given reasoning and the specification of the random process $\tilde{L}^\varepsilon(t)$, we can state the following approximation for the random process $L(t), t \geq 0$

$$L(t) = \frac{1}{\varepsilon} \tilde{L}^\varepsilon(\varepsilon^2 t) + \frac{\rho(\varepsilon^2 t)}{\varepsilon^2} \approx \frac{1}{\varepsilon} \tilde{L}^0(\varepsilon^2 t) + \frac{\rho(\varepsilon^2 t)}{\varepsilon^2},$$

where $\tilde{L}^0(\varepsilon^2 t)$ is defined as a solution of the Ito stochastic differential equation

$$\tilde{L}^0(\varepsilon^2 t) = L(0) + \int_0^{\varepsilon^2 t} C_1(\tilde{L}^0(s), \rho(s), s) ds + \\ + \int_0^{\varepsilon^2 t} \sqrt{B(\rho(s), s)} dW(s)$$

or using considerations regarding $\rho_0(\Delta, t)$, we obtain

$$\tilde{L}^0(\varepsilon^2 t) \approx L(0) + \int_0^{\varepsilon^2 t} C_1(\tilde{L}^0(s), \rho_0(\Delta, s), s) ds + \\ + \int_0^{\varepsilon^2 t} \sqrt{B(\rho_0(\Delta, s), s)} dW(s)$$

and

$$L(t) \approx \frac{1}{\varepsilon} \left(L(0) + \int_0^{\varepsilon^2 t} C_1(\tilde{L}^0(s), \rho_0(\Delta, s), s) ds + \right. \\ \left. + \int_0^{\varepsilon^2 t} \sqrt{B(\rho_0(\Delta, s), s)} dW(s) \right) + \frac{\rho(\varepsilon^2 t)}{\varepsilon^2}.$$

According to this approximation

$$L(t) \sim N \left(\frac{1}{\varepsilon} \left(L(0) + E \left(\int_0^{\varepsilon^2 t} C_1(\tilde{L}^0(s), \rho_0(\Delta, s), s) ds \right) \right) \right. \\ \left. + \frac{\rho(\varepsilon^2 t)}{\varepsilon^2}, \frac{1}{\varepsilon^2} \int_0^{\varepsilon^2 t} B(\rho_0(\Delta, s), s) ds \right),$$

or in a more simplified form

$$L(t) \sim N \left(\rho_0(\Delta, t), \frac{1}{\varepsilon^2} \int_0^{\varepsilon^2 t} B(\rho_0(\Delta, s), s) ds \right). \quad (10)$$

Note that, by the definition of ρ_0 we have $\phi(\rho_0(\Delta, t), \Delta, t) = 0$, hence the drift vanishes at the equilibrium point and $\frac{\rho(\varepsilon^2 t)}{\varepsilon^2} \rightarrow \rho_0(\Delta, t)$ as $\varepsilon \rightarrow +0$.

Thus, a simple estimate was obtained for the analysis of the reliability intervals of a random process $L(t), t \geq 0$, which characterizes the queue length in a heterogeneous QS $H_K(t)/M/\infty$ [4]. Formula (10) has direct practical significance: it states that the distribution of the number of tasks simultaneously in the system approximately obeys a normal distribution with mean $\rho_0(\Delta, t)$ and variance determined by the integral of the diffusion coefficient B . On this basis, it is possible to construct confidence intervals for $L(t)$ and use them to optimize the system configuration without resorting to resource-intensive Monte Carlo simulations or neural network training.

The Ito equation (9) for the limit process also allows us to estimate the probability of exceeding the maximum queue length. Using Doob's inequality, we obtain an estimate of the form

$$P\left(\sup_{t \in [0, T]} L(t) < L_{MAX}\right), \quad (11)$$

where L_{MAX} is the maximum queue length in a heterogeneous QS $H_K(t)/M/L_{MAX}$, which in applied problems corresponds to a hard limit on simultaneous executions (such as the reserved concurrency limit in modern cloud platforms), the estimate (11) in combination with (10) allows us to formulate the optimization problem of the form “find the minimum number of pre-reserved servers that provides a given probability of failure-free operation on the interval $[0, T]$ ” as an explicit analytical problem with a unique solution. Similar optimization statements, in which the level of redundancy is chosen from a cost-benefit criterion built on reliability indicators of a semi-Markov system, are considered in [23]; unlike those approaches, the estimate (11) is obtained in closed analytical form directly from the limit diffusion process and does not require the numerical solution of a system of regeneration equations.

4 Conclusion

The paper presents an asymptotic analysis of the heterogeneous QS $H_K(t)/M/\infty$ based on the application of semi-Markovian random evolutions and properties of averaged processes in the averaging and diffusion approximation scheme, which are analogues of the law of large numbers and the central limit theorem for random processes. The studied model is based on the assumption that the input flow of tasks is a mixture of independent heterogeneous Poisson processes from different sources of events with their own statistical characteristics, which significantly distinguishes it

from classical models of the MMPP type, in which only the intensity is averaged, but not the variance of the input flow.

As a result of the analysis, it was possible to find necessary and sufficient conditions for the fact that the queue will be finite, taking into account the initial characteristics of the input process, which is defined as a mixture of heterogeneous Poisson processes with intensity proportional to the number of tasks currently being processed (Lemma 1). The properties of the conditional mathematical expectation of the queue length increment (Lemma 2) were investigated, and the existence of an equilibrium point $\rho_0(\Delta, t)$, was established, which determines the typical system load level and serves as a guideline for configuring automatic scaling mechanisms.

The second important result of this work is the analysis of the queue length process $L(t), t \geq 0$ based on the limit evolutions in the averaging and diffusion approximation scheme (Theorem 1). The averaging scheme defines a deterministic trajectory for the system's average load (ordinary differential equation (8)), and the diffusion approximation scheme describes random deviations from this trajectory via the stochastic Ito differential equation (9), whose solution is the Ornstein–Uhlenbeck process. Using the necessary and sufficient condition of the boundedness of the queue (Lemma 1), it was possible to construct a fairly simple method for estimating the confidence interval for the queue length in a heterogeneous QS $H_K(t)/M/\infty$ based on the approximation (10). It should also be noted that setting $L(t), t \geq 0$ as a solution to the stochastic differential equation of Ito allows us to estimate the probability

$P\left(\sup_{t \in [0, T]} L(t) < L_{MAX}\right)$, where L_{MAX} is the maximum queue length in the inhomogeneous QS $H_K(t)/M/L_{MAX}$. An estimate of the difference between the limiting and prelimit evolutions was also obtained using the Barry-Ensen inequality.

The practical significance of the results lies primarily in providing a closed analytical basis for parameter estimation and predictive resource planning for systems with complex structures, in particular, modern cloud serverless platforms. Unlike approaches based on machine learning and neural networks, which require significant volumes of historical data sets and do not provide analytical guarantees of service quality, the proposed analytical approach allows obtaining explicit formulas for optimal system parameters and confidence intervals for key performance

characteristics directly based on estimates of the input process parameters $\{w_i(t), \lambda_i(t)\}$ and the server configuration. The obtained estimates can be used, in particular, to determine the minimum (i.e., optimal) number of pre-reserved servers that provides a given probability of failure-free operation over a finite observation interval.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the author(s) utilized artificial intelligence (AI) technologies solely for the purpose of linguistic refinement and grammatical correction. Following the application of these tools, the author(s) comprehensively reviewed, verified, and edited the resulting text. The author(s) maintain full accountability and responsibility for the accuracy, integrity, and final content of the publication.

References:

- [1] Ricardianto P., Putra A. P., Majid S. A., Fachrial P., Samosir J., Adi E. N., Wardana A., Rafi S., Ozali I., Endri E. Evaluation of the Two Runway Queuing System: Evidence from Soekarno-Hatta International Airport in Indonesia. *WSEAS Transactions on Systems and Control*. Vol. 17, 2022, pp. 142–152. DOI: 10.37394/23203.2022.17.16.
- [2] Fischer W., Meier-Hellstern K. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*. Vol. 18, Issue 2, 1993, pp. 149–171. DOI: 10.1016/0166-5316(93)90035-S.
- [3] Koroliuk V. S., Koroliuk V. V., Limnios N., Queuing Systems with Semi-Markov Flow in Average and Diffusion Approximation Schemes. *Methodology and Computing in Applied Probability*. Vol. 11, 2009, pp. 201–209. DOI: 10.1007/s11009-008-9081-7.
- [4] (text in Ukrainian) Kyrychenko O., Malyk I., Kyrychenko O., Queue length estimates in heterogeneous queueing systems with variable operating modes. (Оцінки довжини черги в неоднорідних системах масового обслуговування зі змінними режимами роботи). *Scientific Progress: Theories, Applications and Global Impact* : Proceedings of the 3rd International Scientific and Practical Conference, Braga, Portugal, March 2–4, 2026 / European Open Science Space. EOSS, 2026, pp. 258–260. doi: 10.70286/EOSS-02.03.2026.009.258-260.
- [5] Atar R. Central limit theorem for a many-server queue with random service rates. *Annals of Applied Probability*. Vol. 18, № 4, 2008, pp. 1548–1568. DOI: 10.1214/07-AAP497.
- [6] Lalchandani A. P. Some Limit Theorems in Queuing Theory. *Technical Report*. No. 29, 1967. 82 p.
- [7] Whitt W. Approximations for the GI/G/m Queue. *Production and Operations Management*. Vol. 2, Issue 2, 1993, pp. 114–161. DOI: 10.1111/j.1937-5956.1993.tb00094.x.
- [8] Grassmann W. K. Finding the Right Number of Servers in Real-World Queuing Systems. *INFORMS Journal on Applied Analytics*. Vol. 18, Issue 2, 1988, pp. 94–104. DOI: 10.1287/inte.18.2.94.
- [9] Pardo M. J., de la Fuente D. Optimal selection of the service rate for a finite input source fuzzy queueing system. *Fuzzy Sets and Systems*. 2008. Vol. 159, Issue 3. P. 325–342. URL: 10.1016/j.fss.2007.05.014.
- [10] Yajima M., Phung-Duc T. A Central limit theorem for a Markov-modulated infinite-server queue with batch Poisson arrivals and binomial catastrophes. *Performance Evaluation*. Vol. 129, 2019, pp. 2–14. DOI: 10.1016/j.peva.2018.10.002.
- [11] Gordon W. J., Newell G. F. Closed Queuing Systems with Exponential Servers. *Operations Research*. Vol. 15, № 2, 1967, pp. 254–265. DOI: 10.1287/opre.15.2.254.
- [12] Ke J.-C., Lee S.-L., Liou C.-H. Machine Repair Problem in Production Systems with Spares and Server Vacations. *RAIRO – Operations Research*. Vol. 43, № 1, 2009, pp. 35–54. DOI: 10.1051/ro/2009004.
- [13] Nair A. N., Jacob M. J., Krishnamoorthy A. The Multi-Server M/M/(s, S) Queueing Inventory System. *Annals of Operations Research*. Vol. 233, 2015, pp. 321–333. DOI: 10.1007/s10479-013-1405-5.
- [14] Lebedev E., Livinska H. Multi-channel Queueing Networks with Input Flow Controlled by Semi-Markov Process. *Mathematics in Computer Science*. Vol. 13, 2019, pp. 333–340. DOI: 10.1007/s11786-019-00398-4.
- [15] Ahn H., Duenyas I., Lewis M. E. Optimal control of a two-stage tandem queueing system with flexible servers. *Probability in the Engineering and Informational Sciences*. Vol. 16, Issue 4, 2002, pp. 453–469. DOI: 10.1017/S0269964802164047.
- [16] Cao J., Li K., Stojmenovic I. Optimal Power Allocation and Load Distribution for Multiple Heterogeneous Multicore Server Processors

across Clouds and Data Centers. *IEEE Transactions on Computers*. Vol. 63, Issue 1, 2014, pp. 45–58. DOI: 10.1109/TC.2013.122.

- [17] Massey W. A., Whitt W. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems*. Vol. 13, 1993, pp. 183–250. DOI: 10.1007/BF01158933.
- [18] Malyk I. V., Compensating Operator and Weak Convergence of Semi-Markov Process to the Diffusion Process without Balance Condition. *Journal of Applied Mathematics*. Art. 563060, 2015, 7 p. DOI: 10.1155/2015/563060.
- [19] Limnios N., Swishchuk A. Weak Convergence of DTSMRE in Series Scheme. In: *Discrete-Time Semi-Markov Random Evolutions and Their Applications*. Birkhäuser, Cham, 2023. P. 55–81. DOI: 10.1007/978-3-031-33429-0_4.
- [20] Limnios N., Wu B. Diffusion Approximation of Loss Queueing Systems. *Methodology and Computing in Applied Probability*. Vol. 27, Art. 20, 2025. DOI: 10.1007/s11009-025-10141-1.
- [21] Koroliuk V. S., Limnios N. *Stochastic Systems in Merging Phase Space*. Singapore: World Scientific, 2005, 348 p. DOI: 10.1142/5979.
- [22] Kloeden P. E., Platen E. *Numerical Solution of Stochastic Differential Equations*, Springer Berlin, Heidelberg, 1992, 636 p. DOI: 10.1007/978-3-662-12616-5.
- [23] Prawar Naithani A., Arora H. D., Ekata. Optimizing System Efficiency and Reliability: Integrating Semi-Markov Processes and Regenerative Point Techniques for Maintenance Strategies in Plate Manufacturing. *WSEAS Transactions on Mathematics*. Vol. 23, 2024, pp. 633–642. DOI: 10.37394/23206.2024.23.67.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Oleksandr O. Kyrychenko – Conceptualization and Methodology: the model of a heterogeneous queueing system with mixed operating modes as an analytical description of serverless platforms, and the averaging and diffusion approximation framework. Formal analysis: formulation of Lemma 1, Lemma 2 and Theorem 1 and the proofs, jointly with the co-authors. Writing – original draft.

Ihor Malyk – Methodology: semi-Markov random evolutions and the compensating-operator framework. Formal analysis: the technical conditions of weak convergence, the derivation of the generator of the limiting diffusion and of the coefficients C1 and B. Writing – review & editing.

Oksana L. Kyrychenko – Formal analysis: the conditional expectation of the queue-length increment and the monotonicity in Lemma 2. Validation: the variance estimates and the comparison with the MMPP approximation. Writing – review & editing.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0

https://creativecommons.org/licenses/by/4.0/deed.en_US