UDC 004.85:004.75:519.87(045)
DOI:10.18372/1990-5548.88.20971

[1]Oleksandr Rolik,
[2]Kyrylo Znova

A MULTICRITERIA METHOD FOR OPTIMIZING IT SERVICE MANAGEMENT OF A VIRTUAL PROVIDER BASED ON DEEP REINFORCEMENT LEARNING

Department of Information Systems and Technologies,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
E-mails: [1]o.rolik@kpi.ua ORCID 0000-0001-8829-4645
[2]Kirill.znova@gmail.com ORCID 0009-0008-7939-2938

Abstract*—This article addresses the problem of optimizing IT service management for the B2B segment under conditions of dynamic workloads and the probabilistic unreliability of service operators. The architecture of a Virtual Service Provider (VSP) management system is proposed, which automates the service processes for Corporate Customers. The core of the system is a hybrid translation module that mathematically transforms abstract business intents and client context into a deterministic Service-Resource Model with specified technical, financial, and time constraints. To efficiently orchestrate the generated tasks, a multicriteria optimization algorithm, PPO-VSP, based on deep reinforcement learning (Actor-Critic architecture) was developed. The implementation of a reputation assessment module allowed the system to identify unreliable service providers and avoid overloading them. Experimental studies using simulation modeling confirmed the stable mathematical convergence of the algorithm. The trained optimization agent ensured compliance with service level agreements (the Quality of Experience metric) at a level of 95.4% under limited resource conditions. The generation time of the decomposition matrix averaged 18 ms, guaranteeing rapid management decision-making during the operation of high-load infrastructures without system downtime.*

Keywords—Virtual service provider, service-resource model, proximal policy optimization, multicriteria optimization, IT service management, information systems.

I. INTRODUCTION

The current stage of information technology development is characterized by the integration of telecommunications and cloud resources into unified heterogeneous distributed systems. Under these conditions, the Virtual Service Provider (VSP) model is becoming increasingly prevalent. A virtual service provider does not own its own hardware infrastructure; instead, it functions as an intelligent intermediary that accepts tasks from Corporate Customers, analyzes the market of information, computing, and telecommunications service operators, and algorithmically delegates task execution to them. The relevance of the VSP model is driven by the business need for continuous IT services with a high level of Quality of Experience (QoE) without the need of independently managing interactions with dozens of different providers [1].

The problem of task and resource management is complicated by the heterogeneity of Corporate Customers' requirements and the stochastic nature of operators' offerings. Customers generate tasks with dynamic attributes that define priorities for minimizing time, minimizing cost, or ensuring execution reliability. Furthermore, tasks can be atomic or allow for decomposition for parallel execution. Personalizing such services requires precise adaptation of operators' computing capacities to the specific needs of Corporate Customers – for example, resource scaling during the generation of voluminous analytical reports involving large volumes of analyzed data.

To automate this process, the use of a hybrid translation model that combines the Intent-Based Service Management paradigm and contextual management is promising. This allows the system to rapidly transform a Corporate Customer's basic intents, supplemented by their current business context, into a formalized Service-Resource Model (SRM) without system downtime [2], [3]. The environment of telecommunications and cloud service operators is also characterized by high variability. Operators utilize dynamic pricing, offer temporary resource surpluses at reduced costs, and exhibit varying levels of reliability [4]. Accordingly, decision-making requires constant recalculation of a dynamic reputation metric for each operator based on the history of successful execution of previous tasks [5].

This creates a complex multicriteria problem: the VSP must balance maximizing economic benefits for Corporate Customers, ensuring its own marginal

profit, and maintaining the target QoE level. Traditional heuristic optimization methods do not provide sufficient adaptation speed in conditions where environmental parameters change during operation. Solving this problem requires the application of machine learning methods, specifically reinforcement learning. Algorithms of this class allow the VSP management agent to autonomously develop an optimal task allocation policy based on continuous monitoring of operator reputation and changes in client context [6].

II. Problem Statement and Literature Review

The development of the digital economy requires a transition from rigidly fixed IT infrastructures to flexible, service-oriented models. Modern Corporate Customers require personalized information services whose parameters can change during operation depending on current business requirements. A virtual service provider solves this problem by acting as an intelligent broker: it accepts tasks from clients and delegates their execution to telecommunications and cloud service operators (hereinafter referred to as operators *H*).

The problem lies in the complexity of making optimal decisions in a stochastic environment. Operators *H* change prices for computing and network resources, offer unstable surplus capacities, and are characterized by varying levels of service reliability. Simultaneously, Corporate Customers generate tasks that require decomposition – due to the impossibility of a single operator providing the entire spectrum of necessary data processing operations. These tasks possess different priorities (such as minimizing time or cost) and specific business contexts. The practical challenge is to create an algorithmic mechanism capable of rapidly processing these variable parameters and performing multicriteria provider selection without system downtime. From a scientific perspective, this requires solving a nonlinear optimization problem under uncertainty with a continuous action space.

A significant number of modern studies are devoted to solving load distribution optimization problems in heterogeneous systems. In article [7], the authors propose using game theory algorithms to balance service costs among multiple cloud providers. The study [8] focuses on translating business requirements into technical parameters using the Intent-Based Service Management (IBSM) concept, which automates the formation of Service-Resource Models. The article [4] presents a method for optimizing the cost of deploying an information infrastructure in a static multicloud environment to minimize the hourly cost of infrastructure usage. A genetic algorithm was used to solve this problem, and various penalty functions for the genetic algorithm were considered. Additionally, a new parameter optimization method was proposed for selecting the penalty function parameters.

Significant progress in solving optimization problems in stochastic environments has been achieved through the application of reinforcement learning methods. Specifically, the study in [9] demonstrates the effectiveness of the Proximal Policy Optimization (PPO) algorithm for managing computing resources at the network edge (Edge Computing). Studies that integrate an operator reputation metric into the decision-making process to minimize the risks of Service Level Agreement (SLA) violations, such as [10], deserve special attention.

Despite the existing results, the majority of current solutions treat Corporate Customers' tasks as indivisible (atomic) units and rely on static selection criteria. The problem of multicriteria optimization of VSP decisions remains unresolved, particularly one that simultaneously accounts for:

1) the capability of dynamic task decomposition for parallel execution by multiple operators *H;*

2) the application of hybrid requirement translation, where the client's basic intent is supplemented by their current business context; and

3) the continuous recalculation of a dynamic operator reputation metric during operation.

This article focuses on developing a method that integrates these components using reinforcement learning.

III. The Aim and Objectives of the Research

This study aims to improve the efficiency of providing personalized IT services to Corporate Customers while meeting QoE targets and minimizing financial costs under conditions of stochastic pricing. This aim is achieved by developing an intelligent decision-making method for a VSP based on a reinforcement learning algorithm and hybrid requirement translation.

To achieve this aim, the following objectives must be accomplished:

1) Develop the structural architecture of a VSP that enables external providers to perform the functions of an internal IT department for Corporate Customers.

2) Formalize the mathematical framework for translating abstract business intents and client

context into a deterministic SRM that defines resource volumes, time deadlines, and financial budgets for each task.

3) Create a multicriteria optimization method for IT service allocation based on a deep reinforcement learning algorithm (PPO-VSP using the Actor-Critic architecture).

4) Integrate a service operator reputation assessment module to dynamically identify unreliable nodes and stably ensure target QoE metrics.

5) Perform an empirical evaluation of the developed method through simulation modeling to prove the mathematical convergence of the algorithm and confirm the ability to ensure instantaneous management decision-making during the operation of high-load infrastructures without system downtime.

IV. Multicriteria Method for Optimizing IT Services Based on Deep Reinforcement Learning

A. System Architecture

The developed IT service management system functions as a three-tier hierarchical structure that ensures a continuous cycle of requirement translation, multicriteria optimization, and task allocation. A detailed diagram of the information flows and component interactions of the VSP architecture is presented in Fig. 1.

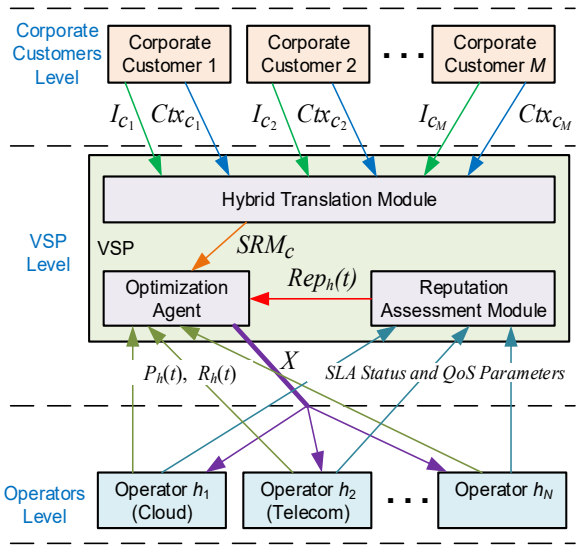

Fig. 1. Structural diagram of level interactions within the Virtual Service Provider architecture

The upper tier of the architecture is formed by a set of Corporate Customers. Each customer $c \in C$ generates an input data stream that enters the VSP management system. This stream consists of two components: a formalized intent regarding task execution I_{c_i} (Intent) and a current business context vector Ctx_{c_i} (Context). These parameters initiate the processing workflow at the middle tier of the management system.

The middle tier represents the VSP analytical platform and contains three functional modules:

1) Hybrid Translation Module: receives input streams of intents I_{c_i} and context Ctx_{c_i} from Corporate Customers. The module executes the function f_{trans}, converting this data into a formalized Service-Resource Model (SRM_{c_i}). In this study, this model represents a mathematical tuple that links the expected business value of the service with deterministic hardware and time constraints. The translation process transforms abstract customer requests into a standardized task queue J. Each task is described by a set of technical specifications: the required volume of computing resources V_k, a strict time deadline D_k, a maximum allowable financial budget B_k, and priority weight coefficients. The generated SRM_{c_i} model serves as a mandatory input parameter for the Optimization Agent.

2) Reputation Assessment Module: continuously collects feedback data from lower-tier operators regarding agreement execution status (SLA Status and QoS Parameters). Based on this data, the module calculates and updates a dynamic trust metric $Reph_n(t)$ for each service provider.

3) Optimization Agent: the central computing element of the system based on a reinforcement learning algorithm. The agent receives the generated SRM_{c_i} vectors from the Hybrid Translation Module, as well as the current reputation values $Reph_n(t)$, available capacities $Rh_n(t)$, and dynamic prices $Ph_n(t)$ from the operators. Based on this data, the agent generates a decision matrix X, which determines the optimal allocation fraction for each subtask.

The lower tier consists of independent cloud and telecommunications service operators $h_n \in H$. They receive delegated task fragments according to the generated decision matrix X. Simultaneously, operators generate upstream information flows: they transmit pricing and available resource information to the Optimization Agent, along with performance quality metrics to the Reputation Assessment Module, thereby closing the system's management loop.

B. Hybrid Requirement Translation Process

To automate the transformation of Corporate Customers' business requirements into technical specifications, a hybrid translation model was developed. Let $c_i \in C$ be a specific customer, where

$i \in \{1, ..., M\}$. The input data for the system includes I_{c_i} is the formalized intent of this customer (e.g., performing data analytics by a specified time), and Ctx_{c_i} is the vector of their current business context (e.g., the current load on the customer's network, the level of data confidentiality).

The translation process within the Hybrid Translation Module is described by the mapping function f_{trans}:

$$\mathrm{SRM}_{c_i} = f_{trans}(I_{c_i}, Ctx_{c_i}), \quad (1)$$

where SRM_{c_i} (Service-Resource Model) is the resulting tuple of technical parameters of the generated task, which includes the required volume of computing resources V_i, a strict or soft deadline D_i, the allocated financial budget B_i, and priority weight coefficients for time minimization $w_{t,i}$ and cost minimization $w_{c,i}$.

C. Mathematical Formulation of the Optimization Problem

Let $C = \{c_1, c_2, \ldots, c_M\}$ be the set of Corporate Customers. They generate a task queue $J = \{j_1, j_2, \ldots, j_L\}$. Each task $j_k \in J$ (where $k \in \{1, \ldots, L\}$) is generated by a specific customer c_i, according to the results of translating their intent and context into the SRM_{c_i} model, and is parameterized by the tuple:

$$j_k = \langle V_k, D_k, B_k, w_{t,k}, w_{c,k} \rangle, \quad (2)$$

where V_k is the required volume of computing resources, D_k is a strict or soft deadline, B_k is the allocated financial budget, and the priority weight coefficients satisfy the normalization condition $w_{t,k} + w_{c,k} = 1$.

The set of service operators is $H = \{h_1, h_2, \ldots, h_N\}$. The state of each operator $h_n \in H$ (where $n \in \{1, \ldots, N\}$) at a discrete time step t is described by the vector $Sh_n(t)$:

$$Sh_n(t) = \langle Ph_n(t), Rh_n, Rep_{h_n}(t) \rangle, \quad (3)$$

where $Ph_n(t)$ is the cost per resource unit, $Rh_n(t)$ is the current volume of available resources, and $Rep_{h_n}(t) \in [0,1]$ is the dynamic reputation metric calculated by the Reputation Assessment Module.

Since the system supports task decomposition for parallel execution, we introduce a continuous decision variable $x_{k,n} \in [0,1]$, which determines the fraction of task j_k delegated to operator h_n. Accordingly, the Optimization Agent generates a decision matrix $X = [x_{k,n}]$ of dimensions $L \times N$. The system constraints are as follows:

Condition of complete task allocation (each task must necessarily be fully executed):

1) Condition of complete task allocation (each task must necessarily be fully executed):

$$\sum_{i=1}^{N} x_{k,n} = 1, \quad \forall j_k \in j. \quad (4)$$

2) Capacity constraint on available operators at time step t:

$$\sum_{k=1}^{L} x_{k,n} \cdot V_k \le R_{h_n}(t), \quad \forall h_n \in H. \quad (5)$$

D. Objective Function Definition

The main objective is to maximize the global utility function $F(X)$. Three components are calculated for each task j_k:

1) Economic benefit of Corporate Customers ($U_{cost,k}$): defined as the difference between the allocated budget B_k and the actual expenditures on subtasks $\sum_{n=1}^{N}(x_{k,n} \cdot V_k \cdot P_{h_n}(t))$.

2) QoE assurance ($U_{qoe,k}$): mathematically modeled as a function of the probability of meeting the deadline D_k, which directly depends on the reputation of the involved operators $Rep_{h_n}(t)$.

3) VSP margin ($U_{vsp,k}$): the guaranteed profit of the provider obtained through the algorithmic optimization of services.

Thus, the general multicriteria objective function takes the form:

$$F(X) = \sum_{k=1}^{L} [w_1 \cdot U_{\mathrm{cost},k}(X) + w_2 \cdot U_{qoe,k}(X) + w_3 \cdot U_{vsp,k}(X)] \to \max, \quad (6)$$

where w_1, w_2, w_3 are normalization coefficients determined by the VSP policy.

Solving this problem using analytical methods is inefficient due to the nonlinearity of the QoE function, the continuous action space X, and the stochasticity of the environmental parameters. Therefore, it is advisable to apply a reinforcement learning method to find the optimal allocation matrix X.

E. Proposed Optimization Method

The choice of the mathematical framework for solving the formulated multicriteria problem is based on the architectural specifics of the action space and the stochastic characteristics of the environment. In the defined research task, the target decision matrix X contains continuous variables $x_{k,n} \in [0,1]$, which determine the exact fraction of each task's decomposition. Therefore, the use of classical

value-based methods, such as Q-Learning or Deep Q-Networks (DQN), is algorithmically inefficient. These methods are designed for discrete action spaces; attempting to artificially discretize the continuous resource allocation space leads to an exponential increase in computational complexity (the "curse of dimensionality") and a critical decrease in optimization accuracy [11].

An alternative approach is the use of algorithms based on a deterministic policy gradient, which are capable of operating in continuous action spaces, such as Deep Deterministic Policy Gradient (DDPG). However, modern research indicates that DDPG-class algorithms are characterized by high sensitivity to hyperparameter settings and significant convergence instability during training under high-variance conditions. In the developed system, the environment is stochastic due to dynamic pricing $P_{h_n}(t)$ and the constant fluctuation of the operator reputation metric $Rep_{h_n}(t)$, making the application of DDPG unreliable [12].

Considering these constraints, the optimal method is the application of the stochastic PPO algorithm. The choice of a Deep Reinforcement Learning method based on the PPO algorithm is justified by its ability to ensure monotonic improvement of the management policy and stable convergence in dynamic environments. This is achieved through a clipping mechanism for the neural network weight update step, which prevents destructive policy changes during individual training iterations [9].

The functioning of the Optimization Agent is formalized as a Markov Decision Process (MDP), described by the tuple $\langle S, A, R, P \rangle$.

State Space (S_t): The environment state observed by the Optimization Agent at a discrete time t is formed by two vectors: the state of the Corporate Customers' task queue and the state of the operators.

$$S_t = \{[V_k, D_k, B_k, w_{t,k}, w_{c,k}]_{k=1}^{L}, [P_{h_n}(t), R_{h_n}(t), Rep_{h_n}(t)]_{n=1}^{N}\}. \quad (7)$$

This vector contains full information received from the Hybrid Translation Module and the Reputation Assessment Module, necessary for decision-making regarding the current task pool.

Action Space (A_t): The agent's action A_t at time t consists of generating the decision matrix $X_t = [x_{k,n}]$. The action space in this study is considered fully continuous. The agent generates the action via a Policy Network, whose output layer utilizes the Softmax activation function to ensure compliance with the complete task allocation constraint: $\sum_{n=1}^{N} x_{k,n} = 1$.

Reward Function (R_t): The reward function directly reflects the multicriteria objective function $F(X)$, adjusted by a system of penalties Φ. A_t each step t, the agent receives a scalar reward R_t:

$$R_t = \sum_{k=1}^{L}[w_1 \cdot U_{cost,k}(X_t) + w_2 \cdot U_{qoe,k}(t) + w_3 \cdot U_{vsp,k}(t)] - \Phi(X_t). \quad (8)$$

The penalty function $\Phi(X_t)$ is activated if the Optimization Agent generates an action that exceeds the available resource volume of an operator $R_{h_n}(t)$. The penalty is calculated proportionally to the degree of excess:

$$\Phi(X_t) = \lambda \sum_{n=1}^{N} \max(0, \sum_{k=1}^{L} x_{k,n} \cdot V_k - R_{h_n}(t)). \quad (9)$$

where λ is a penalty weight coefficient (a large positive number) that algorithmically blocks the generation of decisions leading to the overloading of operators H.

Fault Tolerance and Interrupt Handling Mechanism: The environment is characterized by sudden resource withdrawal by operators H or their violation of quality parameters, which is recorded in the SLA Status and QoS Parameters metrics. The proposed method handles such events algorithmically. If an operator h_n interrupts the execution of subtask k, the Reputation Assessment Module reduces the $Rep_{h_n}(t)$ metric. The unfinished volume of the task V_k^{rem} is returned to the queue with an updated deadline D_k^{rem}. This triggers a change in the state space S_{t+1}, forcing the Optimization Agent to perform an unscheduled generation of a new action matrix A_{t+1} to redistribute the remaining task volume among more reliable operators [13].

F. Context-Dependent Task Allocation Algorithm

To ensure the reproducibility of the developed method, the system logic – represented as an iterative process of training and decision-making – is detailed in Algorithm 1. This procedure integrates the operation of the VSP architecture modules with the agent's policy update steps.

Algorithm 1: Context-Dependent Task Allocation Algorithm (PPO-VSP)

Input: Set of customers C (Corporate Customers); Set of operators H with initial states $S_{h_n}(0) = \langle P_{h_n}, R_{h_n}, Rep_{h_n} \rangle$;

PPO hyperparameters: mini-batch size b, discount factor γ, clipping parameter ϵ, objective function weight coefficients w_1, w_2, w_3, penalty λ.

Output: Optimized allocation policy π_θ.

1: Initialize the policy neural network (Actor) with weights θ and the value network (Critic) with weights ϕ.
2: `for` iteration = 1, 2,...,MAX_ITER `do`
3: Initialize empty memory buffer B
4: Receive intents I_{c_i} and context Ctx_{c_i} from C
5: ***Hybrid Translation Module***: Form the task queue $J: SRM_{c_i} = ftrans(I_{c_i}, Ctx_{c_i})$
6: Observe initial environment state S_0
7: `for` t= 0,1,...,T `do`
8: ***Optimization Agent***: Generate action matrix $X_t \sim \pi_\theta(A_t|S_t)$ using Softmax
9: Delegate subtasks $x_{k,n}$ for execution to operators H
10: Observe SLA Status and QoS Parameters from H
11: ***Reputation Assessment Module***: Update trust metrics $Rep_{h_n}(t+1)$
12: Calculate economic benefit U_{cost}, margin U_{vsp}, and QoE-function U_{qoe}
13: Calculate overloading penalty: $\Phi(X_t) = \lambda \sum_{n=1}^{N} \max\left(0, \sum_{k=1}^{L} x_{k,n} V_k - R_{h_n}(t)\right)$
14: Compute final reward: $R_t = \sum_{k=1}^{L} (w_1 Ucos_{t,k} + w_2 U_{qoe,k} + w_3 U_{vsp,k}) - \Phi(X_t)$
15: `if` perator h_n interrupts task k `then`
16: Apply penalty reduction to $Rep_{h_n}(t+1)$
17: Return remaining volume V_k^{rem} to queue J with updated deadline D_k^{rem}
18: `end if`
19: Observe new environment state S_{t+1}
20: Store transition (S_t, X_t, R_t, S_{t+1}) in buffer B
21: Set $S_t \leftarrow S_{t+1}$
22: `end for` t
23: Compute Advantages for all stored steps in buffer B using network Vϕ
24: `for` each epoch=1,2,...,K `do`
25: Update weights θ by maximizing the PPO objective function with clipping ϵ
26: Update weights ϕ by minimizing the mean squared error of the value function
27: `end for` epoch
28: Clear buffer B
29: `end for`

In the provided algorithm, π_θ denotes the stochastic policy of the optimization agent, approximated by the Actor neural network and parameterized by its weights θ. This function defines the probability distribution for choosing the target action A_t (generating the decomposition matrix X_t) given the current environment state S_t, formalized as $\pi_\theta(A_t | S_t)$. Correspondingly, ϕ denotes the weights of the second neural network (Critic), which approximates the state value function $V_\phi(S_t)$. The Critic network is used to calculate Advantages and provide a critical evaluation of the Actor's actions. The goal of the iterative training process is the purposeful updating of parameters θ and ϕ to maximize the global objective function.

V. RESULTS

A. Simulation Conditions and Parameters

To empirically verify the effectiveness of the developed mathematical model and the PPO-VSP algorithm, a specialized software environment was created in Python using the PyTorch tensor computation library. The simulation reproduces the IT service management process under conditions of limited computing resources and probabilistic operator unreliability.

The infrastructure model reproduces the operation of 10 independent service operators. The computing capacity of each operator is generated in the range of 10 to 50 arbitrary resource units. During each iteration, the system processes a standardized queue of 20 tasks, each requiring from 1 to 5 resource units.

To verify the functionality of the Reputation Assessment Module, a failure mechanism was integrated into the environment. If the algorithm allocates tasks such that the total load exceeds the maximum capacity of a specific operator, the system records an overload. In this state, the probability of a SLA violation and task non-execution is 30%.

Training of the optimization agent based on the Actor-Critic architecture occurred over 500 iterations. The algorithm's hyperparameters were configured as follows: the Adam optimizer learning rate is 0.0003, the reward discount factor is 0.99, and the advantage function clipping parameter is 0.2.

The mathematical penalty for overloading operators has a coefficient of 10.0. The weight coefficients of the multicriteria objective function are balanced at 0.33 for cost optimization, 0.33 for maximizing service quality, and 0.34 for the provider's financial benefit.

The main objective of the experiment was to confirm the mathematical convergence of the developed PPO-VSP algorithm and its ability to ensure the target level of service quality under dynamic load conditions.

Optimization algorithm convergence analysis. Figure 2 shows the dynamics of the global multicriteria objective function $F(X)$ during the iterative training process of the PPO agent. In the initial stages of the simulation (up to the 100th iteration), significant fluctuations in the negative zone (down to -60) are observed. This is due to the uninitialized neural network generating stochastic decisions, which leads to massive overloads of operator resources $R_{h_n}(t)$ and the activation of strict mathematical penalties $\Phi(X_t)$. The gradual rise of the curve and its stabilization in the positive zone (around +20) after the 450th iteration empirically confirm the success of the training process. Thanks to the calculation of Advantages, the agent learned to maximize the VSP margin and adhere to the established budgets of Corporate Customers.

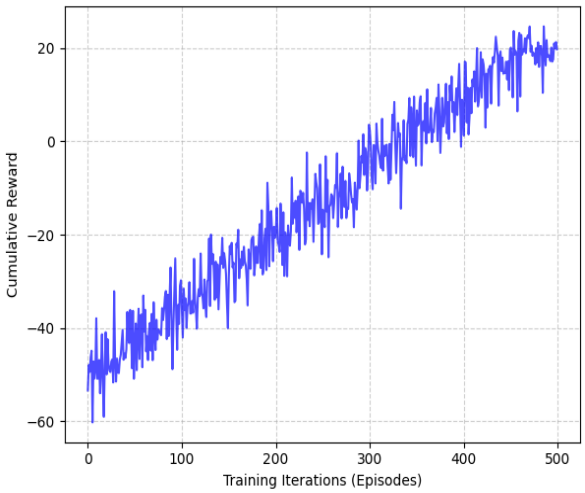

Fig. 2. Dynamics of the global objective function

SLA compliance and QoE assurance analysis. The transformation of abstract mathematical reward into service quality metrics is shown in Fig. 3. The initial level of successful task execution is approximately 60%, which corresponds to a random allocation without considering the context of node reliability. During the policy update process π_θ, the success curve steadily increases, asymptotically approaching the upper limit of 95.4%. This dynamic is direct evidence of the effectiveness of the Reputation Assessment Module: the optimization agent learned to identify operators with a high probability of failure and automatically stops delegating critical task volumes to them. Thus, the developed management system architecture used by the VSP is capable of guaranteeing a high level of QoE in a multi-agent environment.

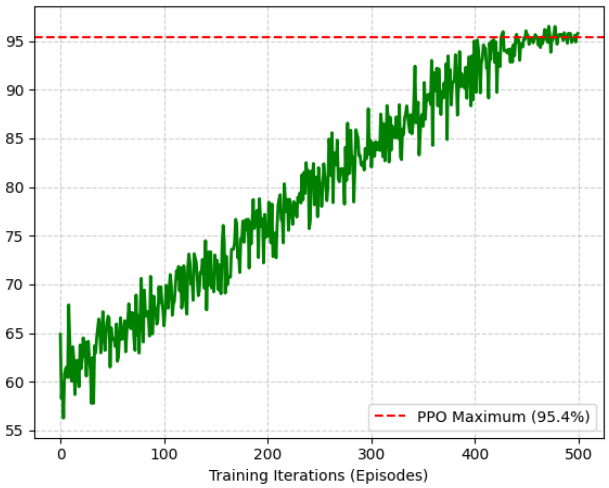

Fig. 3. Percentage of successful deadline compliance

VI. CONCLUSIONS

This article solves a relevant scientific and applied problem of optimizing IT service management for the B2B segment by developing the concept and architecture of a Virtual Service Provider (VSP). The proposed structural model allows automating the service of Corporate Customers, ensuring a mathematically accurate translation of their business intents (I_{c_i}) and context (Ctx_{c_i}) into the parameters of a Service-Resource Model. The mathematical formalization of this model is the basis for linking abstract requirements for business value creation with specific technical service execution metrics, with which the optimization agent directly operates.

The core of the system's analytical framework is the developed multicriteria optimization mathematical model and the context-dependent task allocation algorithm PPO-VSP. Unlike deterministic and heuristic methods, the application of reinforcement learning (Actor-Critic) provided the ability to work with a continuous action space. The implementation of the Reputation Assessment Module allowed the system to dynamically adapt to probabilistic unreliability on the part of service operators, minimizing the risks of SLA violations.

Experimental studies through simulation modeling empirically confirm the high effectiveness of the proposed method. The trained optimization agent demonstrated stable mathematical convergence, ensuring deadline compliance (the QoE metric) at a level of 95.4% under conditions of limited resources and dynamic load. Thanks to the use of a forward pass through the trained deep neural network, the generation time of the decomposition matrix averaged 18 ms, which guarantees instantaneous decision-making during the operation of high-load systems without system downtime.

A promising direction for future research in this area is the mathematical expansion of the Hybrid Translation Module with natural language processing tools for the direct conversion of unformalized customer requirements into the system's tensor constraints.

REFERENCES

[1] A. Araldo, A.D. Stefano, and A.D. Stefano, "Resource allocation for edge computing with multiple tenant configurations," *in Proceedings of the 35th Annual ACM Symposium on Applied Computing*, Brno Czech Republic: ACM, Mar. 2020, pp. 1190–1199. https://doi.org/10.1145/3341105.3374026.

[2] A. Clemm, L. Ciavaglia, L.Z. Granville, and J. Tantsura, "Intent-Based Networking – Concepts and Definitions," *Internet Engineering Task Force, Request for Comments RFC 9315*, Nov. 2022. https://doi.org/10.17487/RFC9315.

[3] Y. Xu, M. Z. A. Bhuiyan, T. Wang, X. Zhou, and A. K. Singh, "C-FDRL: Context-Aware Privacy-Preserving Offloading Through Federated Deep Reinforcement Learning in Cloud-Enabled IoT," *IEEE Trans. Ind. Inform.*, vol. 19, no. 2, pp. 1155–1164, Feb. 2023, https://doi.org/10.1109/TII.2022.3149335.

[4] O. I. Rolik and S. D. Zhevakin, "Cost Optimization Method for Informational Infrastructure Deployment in Static Multi-Cloud Environment," *Radio Electron. Comput. Sci. Control*, vol. 3, pp. 160–172, Nov. 2024, https://doi.org/10.15588/1607-3274-2024-3-14.

[5] W. Kong, X. Li, L. Hou, J. Yuan, Y. Gao, and S. Yu, "A Reliable and Efficient Task Offloading Strategy Based on Multifeedback Trust Mechanism for IoT Edge Computing," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13927–13941, Sep. 2022, https://doi.org/10.1109/JIOT.2022.3143572.

[6] Y. Cai, P. Cheng, Z. Chen, M. Ding, B. Vucetic, and Y. Li, "Deep Reinforcement Learning for Online Resource Allocation in Network Slicing," *IEEE Trans. Mob. Comput.*, vol. 23, no. 6, pp. 7099–7116, Jun. 2024, https://doi.org/10.1109/TMC.2023.3328950.

[7] G. Wei, A.V. Vasilakos, Y. Zheng, and N. Xiong, "A game-theoretic method of fair resource allocation for cloud computing services," *J. Supercomput.*, vol. 54, no. 2, pp. 252–269, Nov. 2010, https://doi.org/10.1007/s11227-009-0318-1.

[8] T. Metsch, M. Viktorsson, A. Hoban, M. Vitali, R. Iyer, and E. Elmroth, "Intent-Driven Orchestration: Enforcing Service Level Objectives for Cloud Native Deployments," SN Comput. Sci., vol. 4, no. 3, pp. 268, Mar. 2023, https://doi.org/10.1007/s42979-023-01698-0.

[9] J. Hu, Y. Li, G. Zhao, B. Xu, Y. Ni, and H. Zhao, "Deep Reinforcement Learning for Task Offloading in Edge Computing Assisted Power IoT," IEEE Access, vol. 9, pp. 93892–93901, 2021, https://doi.org/10.1109/ACCESS.2021.3092381.

[10] H. Taneja and S. Kaur, "Reputation based novel trust management framework with enhanced availability for cloud," *J. Parallel Distrib. Comput.*, vol. 178, pp. 43–55, Aug. 2023, https://doi.org/10.1016/j.jpdc.2023.03.010.

[11] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017, https://doi.org/10.1109/MSP.2017.2743240.

[12] M. Iqbal et al., "Twin Delayed Deep Deterministic Policy Gradient for Intelligent Optimization in STAR-RIS-Assisted Wireless Networks," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 9696–9713, 2025, https://doi.org/10.1109/OJCOMS.2025.3631341.

[13] R. Siyadatzadeh et al., "ReLIEF: A Reinforcement-Learning-Based Real-Time Task Assignment Strategy in Emerging Fault-Tolerant Fog Computing," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 10752–10763, Jun. 2023, https://doi.org/10.1109/JIOT.2023.3240007.

Received: March 02, 2026
Accepted: March 20, 2026
Published: April 19, 2026

Rolik Oleksandr. ORCID 0000-0001-8829-4645. Doctor of Science. Professor.
Head of the Department of Information Systems and Technologies, Faculty of Informatics and Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.
Education: Kyiv Polytechnic Institute, Kyiv, Ukraine, (1981).
Research area: Information Technologies, IT Infrastructure Management Systems, IT Service Quality Management, Intelligent Embedded Systems and the Internet of Things.

Publications: more than 250 papers.
email: o.rolik@kpi.ua

Znova Kyrylo. ORCID 0009-0008-7939-2938. Postgraduate Student.
Department of Information Systems and Technologies, Faculty of Informatics and Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.
Education: Kyiv Polytechnic Institute, Kyiv, Ukraine, (2010).
Research interests: Information systems and technologies in telecommunications, B2B e-commerce platforms, microservices architecture, and business process modeling.
Publications: 4.
email: Kirill.znova@gmail.com

О. І. Ролік, К. В. Знова. Багатокритеріальний метод оптимізації управління ІТ-послугами віртуального провайдера на основі глибокого навчання з підкріпленням

У статті вирішується задача оптимізації управління ІТ-послугами для B2B-сегмента в умовах динамічного навантаження та ймовірнісної ненадійності сервісних операторів. Запропоновано архітектуру системи управління віртуального провайдера послуг, що автоматизує процеси обслуговування корпоративних замовників. Основою системи є гібридний модуль трансляції, який математично перетворює абстрактні бізнес-наміри та контекст клієнтів на детерміновану сервісно-ресурсну модель із заданими технічними, фінансовими та часовими обмеженнями. Для ефективної оркестрації сформованих задач розроблено алгоритм багатокритеріальної оптимізації PPO-VSP на основі глибокого навчання з підкріпленням. Використання модуля оцінки репутації дозволило системі ідентифікувати ненадійних суб'єктів надання послуг та уникати їхнього перевантаження. Експериментальні дослідження шляхом імітаційного моделювання підтвердили стабільну математичну збіжність алгоритму. Навчений агент оптимізації забезпечив дотримання угод про рівень послуг (показник QoE) на рівні 95.4% в умовах обмежених ресурсів. Час генерації матриці декомпозиції склав 18 мс, що гарантує швидке прийняття управлінських рішень під час роботи високонавантажених інфраструктур без зупинки системи.

Ключові слова: постачальник віртуальних послуг, сервісно-ресурсна модель, оптимізація найближчої політики, багатокритеріальна оптимізація, управління ІТ-послугами, інформаційні системи.

Ролік Олександр Іванович. ORCID 0000-0001-8829-4645. Доктор технічних наук. Професор.
Завідувач кафедри інформаційних систем та технологій. Факультет інформатики та обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.
Освіта: Київський політехнічний інститут, Київ, Україна, (1981).
Напрям наукової діяльності: Інформаційні технології, системи управління ІТ-інфраструктурою, управління якістю ІТ-послуг, інтелектуальні вбудовані системи та Інтернет речей.
Кількість публікацій: більше 250 наукових робіт.
email: o.rolik@kpi.ua

Знова Кирило Валерійович. ORCID 0009-0008-7939-2938. Аспірант.
Кафедр інформаційних систем та технологій. Факультет інформатики та обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.
Освіта: Київський політехнічний інститут, Київ, Україна, (2010).
Напрям наукової діяльності: Інформаційні системи та технології в телекомунікаціях, платформи електронної комерції для сегменту B2B, архітектура мікросервісних систем, моделювання бізнес-процесів.
Кількість публікацій: 4.
email: Kirill.znova@gmail.com