DOI: 10.31319/2519-8106.2(53)2025.342456
UDC 004.2:519.6:519.8: 519.17

Popov Oleksandr, Doctor of Physical and Mathematical Sciences, Senior Researcher, Department of Numerical Methods and Computer Modeling
Попов О.В., доктор фізико-математичних наук, старший науковий сспівробітник відділу чисельних методів та комп'ютерного моделювання
ORCID: 0000-0002-1217-2534
e-mail: alex50popov@gmail.com

Pavliuk Anton, Postgraduate Student, Department of Numerical Methods and Computer Modeling
Павлюк А.В., здобувач третього (доктора філософії) рівня вищої освіти, відділ чисельних методів та комп'ютерного моделювання
ORCID: 0009-0002-4166-2003
e-mail: kvadrumpl@gmail.com

V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine, Kyiv
Інститут кібернетики ім. В.М. Глушкова НАН України, м. Київ

MULTIOBJECTIVE MEMORY OPTIMIZATION IN MATHEMATICAL MODELING FOR HYBRID COMPUTER ARCHITECTURES

БАГАТОКРИТЕРІАЛЬНА ОПТИМІЗАЦІЯ ПАМ'ЯТІ В ЗАДАЧАХ МАТЕМАТИЧНОГО МОДЕЛЮВАННЯ ДЛЯ ГІБРИДНИХ КОМП'ЮТЕРНИХ АРХІТЕКТУР

The study presents a generalized mathematical and algorithmic framework for multi-objective optimization of the memory subsystem in hybrid CPU-GPU architectures. The proposed model explicitly captures the coupling among bandwidth, latency, and energy cost within the memory hierarchy through a vector objective function and operator-based formulation of the throughput pipeline. To identify Pareto-optimal configurations, an evolutionary-gradient method is introduced, combining global exploration via NSGA-II/MOEA-D with local refinement using ADAM and L-BFGS optimizers. Gradient information is reconstructed from surrogate models based on Gaussian Process Regression and Radial Basis Functions, ensuring convergence under limited data. Validation on an Intel Xeon + NVIDIA Tesla T4 system demonstrated 35—45 % higher bandwidth, up to 2× lower energy use, and 95th-percentile latency reduced to 0.27—0.34 ms.

***Keywords**: multi-objective memory optimization, evolutionary-gradient methods, hybrid computing architectures, Pareto optimization, energy efficiency, NSGA-II, MOEA/D.*

Стаття присвячена розробленню узагальненої математичної моделі та алгоритмічного підходу до багатокритеріальної оптимізації підсистеми пам'яті у гібридних обчислювальних архітектурах на основі CPU і GPU. Метою дослідження є формування теоретичних засад для енергетично узгодженого керування пам'яттю, яке враховує взаємозалежність трьох ключових характеристик — пропускної здатності, латентності доступу та енергетичних витрат у гібридній архітектурі. Запропонована модель описує пропускний конвеєр як узгоджений динамічний процес, у якому часові, енергетичні та пропускні параметри розглядаються як взаємопов'язані критерії оптимізації. На цій основі сформульовано задачу багатокритеріальної оптимізації, розв'язком якої є множина Парето-оптимальних конфігурацій, що відображають узгоджений баланс між продуктивністю, часовими затримками та енергетичною ефективністю системи.

Запропонований еволюційно-градієнтний підхід реалізує інтеграцію глобального еволюційного пошуку з локальною стохастичною оптимізацією, забезпечуючи збалансоване дослідження простору можливих конфігурацій і точне уточнення знайдених рішень у безпосередній

околиці фронту Парето. На відміну від традиційних еволюційних методів, які не враховують локальну геометрію цільового простору, розроблений алгоритм використовує апроксимацію градієнтної інформації для керованого спрямування пошуку. У випадках обмеженої кількості експериментальних спостережень ця інформація відновлюється за допомогою сурогатних моделей на основі гаусового процесу та радіальних базисних функцій, що дозволяє зберегти стійкість і збіжність оптимізаційного процесу без зниження точності моделювання. Така комбінація глобальної еволюційної евристики й локальної диференційної адаптації формує єдиний обчислювальний механізм для відтворення рівноважного компромісу між пропускною здатністю, латентністю та енергетичними витратами системи.

Експериментальна верифікація на тестовій системі Intel Xeon + NVIDIA Tesla T4 засвідчила підвищення пропускної здатності до 45 %, зниження енергоспоживання приблизно удвічі та скорочення 95-го перцентиля латентності до 0,27—0,34 мс порівняно з базовими схемами static partitioning ma fixed scheduling.

***Ключові слова**: багатокритеріальна оптимізація пам'яті, еволюційно-градієнтні методи, гібридні обчислювальні архітектури, Парето-оптимізація, енергетична ефективність, NSGA-II, MOEA/D.*

Problem's Formulation

Hybrid computing architectures that integrate central processing units (CPUs) and graphics processing units (GPUs) constitute the foundation of modern high-performance systems; however, their overall efficiency critically depends on the organization of the memory subsystem. In such architectures, computation proceeds through bi- or multichannel interactions in which data are continuously transferred between main memory, multi-level caches, and graphics accelerators, forming a complex pipeline characterized by multiple temporal scales. The bandwidth of data-exchange channels, memory access latency, and the energy cost of data transmission jointly define an interdependent system of criteria in which improvement of one parameter almost inevitably degrades the others. For example, latency reduction achieved through aggressive caching or data duplication typically increases the energy cost, while bandwidth enhancement via intensive GPU-channel utilization can overload PCIe or NVLink interfaces.

Classical memory-management policies based on static partitioning or fixed scheduling assume a fixed resource allocation scheme that disregards dynamic variability and mutual interactions among subsystems. Such policies fail to capture the nonlinear dependencies among throughput-pipeline characteristics and therefore cannot ensure coherence between local stream optimality and the global efficiency of the system. To adequately describe these dependencies, it is necessary to move from heuristic or empirical techniques toward a rigorous mathematical formalization of the optimization problem that explicitly accounts for the multi-objective nature of data-exchange processes. Within this framework, we consider a memory-management policy parameter vector $x \in X$, for which a vector-valued objective function is defined as follows:

$$F(x) = \big(-B(x), L(x), E(x)\big), \tag{1}$$

which characterizes the trade-off among bandwidth, access stability, and energy efficiency in the hybrid architecture of the computing system. The solution to this problem is represented by the set of Pareto-optimal configurations that provide the best balance between performance and energy consumption. Constructing and analyzing these configurations paves the way for developing adaptive memory management methods capable of dynamically optimizing pipeline throughput in next-generation heterogeneous systems [1—2].

Analysis of recent research and publications

Memory optimization in hybrid CPU-GPU architectures represents one of the key directions in the evolution of high-performance computing. Contemporary research pays significant attention to Unified Memory models, which provide a unified address space and automatic data migration between the CPU and GPU. Such solutions considerably simplify the programming abstraction; however, they remain suboptimal from the standpoint of mathematical optimization, as they do not minimize transfer

latency under variable channel bandwidth conditions. Similarly, NUMA architectures improve spatial memory locality but fail to account for energy losses and the stochastic nature of inter-node interactions.

Several recent studies have proposed dynamic memory-management strategies based on empirical models of data-block access intensity and temporal load profiling of the computational pipeline. These approaches partially adapt the memory-allocation policy to the system's current state; nevertheless, their optimization remains predominantly single-objective (focused solely on minimizing either execution time or energy consumption) without considering the intricate interdependence among the criteria $B(x)$, $L(x)$, and $E(x)$. Within the field of mathematical optimization, multi-objective evolutionary algorithms such as NSGA-II, SPEA2, and MOEA/D have demonstrated substantial progress, providing effective approximations of Pareto-optimal sets in problems with nonlinear functionals and unknown derivatives. These methods have been successfully applied to flow scheduling, load balancing, and energy-efficiency optimization; however, they have not yet been adapted to memory-management models, where system parameters are determined not only by computational topology but also by the architectural properties of the communication subsystem [3—5].

In parallel, methods such as ADAM, L-BFGS, Gaussian Process Regression, and Radial Basis Functions are actively advancing, enabling the reconstruction of latent functions of bandwidth and energy cost from limited experimental datasets. Nonetheless, the scientific literature still lacks a comprehensive mathematical framework that integrates temporal, energetic, and throughput characteristics of memory into a unified multi-objective optimization problem. This methodological gap underscores the relevance of developing an evolutionary-gradient model of memory management that combines analytical rigor with practical adaptability in heterogeneous computing environments [6].

Formulation of the study purpose

The purpose of this study is to develop a mathematical model and an algorithmic approach for solving the multi-objective memory optimization problem in hybrid CPU-GPU architectures. The core idea lies in constructing an analytical model of the throughput pipeline that formalizes the interdependence among bandwidth, access latency, and energy expenditure, combined with the application of an evolutionary-gradient method to identify the set of Pareto-optimal configurations. The proposed approach aims to minimize the joint energy-time functional cost and to establish a theoretical foundation for the further development of self-learning memory management policies in hybrid computing systems.

Presenting main material

Consider a hybrid computing architecture in which the central processing unit (CPU) and the graphics processing unit (GPU) interact through a shared memory hierarchy with limited bandwidth. The aggregate performance of such a system is determined by the interdependence of three key characteristics:

- data-exchange bandwidth $B(x)$;
- average memory access latency $L(x)$;
- energy cost of operations $E(x)$.

The vector $x = (x_1, x_2, \dots, x_n) \in X$ defines the configuration space of the memory management policy. The components $x \in X$ may include parameters such as data transfer block size, fraction of data allocated to the GPU, stream servicing order, synchronization frequency between processors, and the intensity of DMA operations. The feasible set X is determined by the architectural constraints of the system:

$$X = \left\{x \in \mathbb{R}^n | b_c(x) \le B_{max}^{CPU}, b_g(x) \le B_{max}^{GPU}, \tau_{\text{PCIe}}(x) \le \tau^*, M_{\text{GPU}}(x) \le M^*\right\}, \quad (2)$$

where $B_{\max}^{\text{CPU}}$ and $B_{\max}^{\text{GPU}}$ denote the maximum bandwidths of the CPU and GPU buses, respectively; $\tau_{\text{PCIe}}(x)$ represents the average latency of the data exchange channel; $M_{\text{GPU}}(x)$ denotes the utilized volume of GPU memory; and τ^*, M^* correspond to the respective architectural constraints [7].

The problem of optimal memory management is formulated as a multi-objective optimization task:

$$\min_{x \in X} F(x) = \big(-B(x), L(x), E(x)\big), \quad (3)$$

where the vector objective function $F(x)$ describes the coherent coexistence of three conflicting criteria: bandwidth, access latency, and energy consumption.

To ensure the mathematical correctness of the problem formulation, we introduce the following analytical assumptions [8]:

1. The functions $B(x)$, $L(x)$, and $E(x)$ are defined on a closed set $X \subset \mathbb{R}^n$ and are continuous and differentiable for almost all $x \in X$.
2. For each component $x_i \in X$ of the parameter vector, the partial derivatives $\frac{\partial B}{\partial x_i}$, $\frac{\partial L}{\partial x_i}$, and $\frac{\partial E}{\partial x_i}$ exist and are continuous. The signs of these derivatives characterize the physical direction of influence:
   - $\frac{\partial B}{\partial x_i} > 0$: parameters that increase the intensity of parallel computations;
   - $\frac{\partial L}{\partial x_i} > 0$: parameters that increase the intensity or depth of memory request queues;
   - $\frac{\partial E}{\partial x_i} > 0$: parameters that increase computational intensity and, consequently, energy consumption.

   Each parameter exerts a consistent directional effect on the objective functions, which ensures the validity of sensitivity analysis and the reliability of gradient-based approximations.
3. For any two configurations $x_1, x_2 \in X$, there exists a constant $\gamma > 0$ such that
$$| B(x_1) - B(x_2) | + | L(x_1) - L(x_2) | + | E(x_1) - E(x_2) | \leq \gamma \| x_1 - x_2 \|,$$
which establishes the Lipschitz continuity of the vector function and guarantees the stability of the functional as well as the convergence of numerical optimization methods with bounded error.

The solution set of problem (3) is defined as the set of Pareto-optimal vectors:
$$X^* = \{x \in X | \nexists y \in X : F(y) \prec F(x)\}, \quad (4)$$
where the relation $F(y) \prec F(x)$ indicates that the vector $F(y)$ is no worse than $F(x)$ with respect to all criteria and strictly better in at least one, i.e., $F_i(y) \leq F_i(x), \forall i \in \{1, \dots, m\}, \exists j : F_j(y) < F_j(x)$.

Thus, the formulation (3)—(4) defines a formal space of multi-objective optimization within which the subsequent analysis focuses on two interrelated problems. First, an analytical model of the throughput pipeline is constructed to capture the dependencies among bandwidth, access latency, and energy expenditure. Second, an evolutionary-gradient algorithm is developed to identify the set of Pareto-optimal configurations, ensuring guaranteed convergence and controlled approximation error. This approach establishes a mathematical foundation for designing coherent memory management policies in hybrid CPU-GPU systems.

Throughput Model

The throughput pipeline of a hybrid CPU-GPU architecture can be interpreted as a sequence of interdependent stages of data exchange and processing that together form a unified flow within the memory system. In the general case, three primary stages are considered:

a) Loading stage, involving data transfer from main memory to CPU cache subsystems, which determines the initial stream formation rate;

b) Transportation stage, representing data migration between the CPU and GPU via high-speed interfaces such as PCIe or NVLink, ensuring spatial coordination of subsystems;

c) Computational stage, encompassing execution of computations and writing of results to GPU memory, thereby completing the throughput pipeline cycle.

This formulation reflects the physical structure of the process, in which the throughput of each stage serves as a local characteristic, while their coherence determines the overall efficiency of the system [9—11].

Let $t_i(x)$ denote the average processing time per data unit at the i-th stage; then, the total time for one data packet to pass through the pipeline is given by:
$$T(x) = \sum_{i=1}^{3} t_i(x), \quad (5)$$
and the effective throughput is determined by the flow conservation law: $B(x) = \frac{V(x)}{T(x)}$, where $V(x)$ denotes the average volume of data passing through the pipeline within a single operating cycle and is defined as a function of the memory management policy parameters $x \in X$.

Based on the axioms of dataflow theory and the continuous pipeline model, the equivalent throughput of a sequential system of channels in the steady-state regime is given by the relation:

$$\big(B(x)\big)^{-1} = \sum_i \big(b_i(x)\big)^{-1} = \frac{1}{b_c(x)} + \frac{1}{b_{bus}(x)} + \frac{1}{b_g(x)}, \quad (6)$$

where $b_c(x)$ denotes the data exchange rate between the CPU and main memory; $b_{\text{bus}}(x)$ represents the bandwidth of the PCIe or NVLink channel; and $b_g(x)$ is the effective access rate to GPU memory. Each of the functions $b_i(x)$ increases monotonically with respect to the parameters defining the degree of computational parallelism; however, its value is limited by the hardware capabilities of the corresponding channel. For analytical convenience, we introduce the normalized rates $\tilde{b}_i(x) = \frac{b_i(x)}{b_i^{max}}$, which belong to the interval $\tilde{b}_i(x) \in [0,1]$.

The average memory access latency $L(x)$ is determined according to Little's theorem for open queuing systems:

$$L(x) = \frac{Q(x)}{\Lambda(x)}, \quad (7)$$

where $Q(x)$ denotes the average number of active requests within the system, and $\Lambda(x)$ represents the service rate of these requests.

For the numerical implementation of equation (7), we introduce an approximate functional dependence of the following form:

$$L(x) = \alpha_c \frac{q_{CPU}(x)}{b_c(x)} + \alpha_g \frac{q_{GPU}(x)}{b_g(x)}, \quad (8)$$

where $q_{\text{CPU}}(x)$ and $q_{\text{GPU}}(x)$ denote the average queue lengths of requests, and the coefficients α_c and α_g define the relative weighting of each channel.

Equation (8) ensures the analytical smoothness of the function $L(x)$ while preserving its physical interpretability, as it directly links latency to the temporal characteristics of individual system components. The energy cost of operations $E(x)$ can be determined through the integral power balance:

$$E(x) = \int_0^{T(x)} \big(P_{CPU}(x,t) + P_{GPU}(x,t) + P_{bus}(x,t)\big)\, dt. \quad (9)$$

Under the assumption of a quasi-stationary power consumption regime, in which variations in power during a data transfer cycle are negligible, the energy balance integral (9) can be reduced to the following analytical form:

$$E(x) = P_{CPU}(x)T_{CPU}(x) + P_{GPU}(x)T_{GPU}(x) + P_{bus}(x)\tau_{PCIe}(x), \quad (10)$$

where $T_{\text{CPU}}(x)$ and $T_{\text{GPU}}(x)$ denote the fractions of active operation time of the devices within a complete execution cycle.

It has been established that the energy consumption of the graphics accelerator exhibits a quadratic dependence on the GPU core clock frequency $f_{\text{GPU}}(x)$, which is consistent with well-known models of energy dynamics in semiconductor devices. Considering this, it is reasonable to adopt the following approximation:

$$P_{GPU}(x) = P_0 + \beta f_{GPU}(x)^2, \quad (11)$$

where β is a parameter characterizing the energy sensitivity of the computational subsystem.

For a compact representation of the interrelations among the system's key characteristics, the generalized model of the throughput pipeline can be expressed as a system of functional dependencies:

$$\begin{cases} B(x) = \psi_B\left(b_c(x), b_{bus}(x), b_g(x)\right); \\ L(x) = \psi_L\big(q_{CPU}(x), q_{GPU}(x), B(x)\big); \\ E(x) = \psi_E\big(P_{CPU}(x), P_{GPU}(x), P_{bus}(x), T(x)\big), \end{cases} \quad (12)$$

where $\psi_B(\cdot)$ is the bandwidth aggregation operator, which maps the local transfer rates of subsystems (the CPU cache, PCIe/NVLink channel, and GPU memory) into the system's overall effective bandwidth $B(x)$; $\psi_L(\cdot)$ is the latency operator, defining the dependence of the average access time $L(x)$ on the request-queuing parameters $q_{\text{CPU}}(x)$, $q_{\text{GPU}}(x)$, and the actual bandwidth $B(x)$; and $\psi_E(\cdot)$ is the energy-balance operator, establishing the functional relationship among the power levels of active components $P_{\text{CPU}}(x)$, $P_{\text{GPU}}(x)$, $P_{\text{bus}}(x)$ and the duration of the computational cycle $T(x)$.

The functions $\psi_i(\cdot)$, $i \in \{B, L, E\}$, are defined on the admissible parameter set $X \subset \mathbb{R}^n$ and satisfy the following properties:

a) Continuity and differentiability: for all $x \in X$, there exist partial derivatives $\frac{\partial \psi_i(x)}{\partial x_i}$, ensuring smoothness and enabling local gradient analysis;

b) Monotonicity: the function $\psi_B(\cdot)$ is non-decreasing with respect to $b_c(x)$, $b_{\text{bus}}(x)$, and $b_g(x)$; the function $\psi_L(\cdot)$ is decreasing with respect to $B(x)$ and increasing with respect to $q_{\text{CPU}}(x)$ and $q_{\text{GPU}}(x)$; the function $\psi_E(\cdot)$ is monotonically increasing with respect to the power variables $P_{\text{CPU}}(x)$, $P_{\text{GPU}}(x)$, and $P_{\text{bus}}(x)$;

c) Lipschitz continuity: there exists a constant $K > 0$ such that

$$\|\Psi(x_1) - \Psi(x_2)\| \leq K\|x_1 - x_2\|, \forall x_1, x_2 \in X,$$

which guarantees the model's stability under small parameter perturbations.

The generalized representation (12) provides a mathematically consistent formalization of the interdependent relationships among CPU-GPU memory subsystems and forms the foundation for subsequent multi-objective optimization.

Methodology of Multi-Objective Optimization

The solution of problem (3)—(4) for the vector function (1) is based on combining global and local mechanisms for searching Pareto-optimal configurations of the memory management policy. Since the functions $B(x)$, $L(x)$, and $E(x)$ are defined by the operators ψ_B, ψ_L, ψ_E in (12) and are continuous and Lipschitz-continuous, the set $X^* \subset X$ exists and is closed in the sense of ε-dominance. To approximate this set, an evolutionary-gradient approach is proposed that integrates global exploration of the Pareto front using NSGA-II or MOEA/D algorithms with local refinement of solutions through gradient-based methods such as ADAM or L-BFGS. This hybrid strategy ensures a balance between exploration of the decision space X and exploitation of the local neighborhoods of obtained solutions [12-14].

Let $f_1(x) = -B(x)$, $f_2(x) = L(x)$, $f_3(x) = E(x)$, and $\lambda = (\lambda_1, \lambda_2, \lambda_3) \in \Delta_3$, where $\Delta_3 = \{\lambda_i \geq 0, \sum_i \lambda_i = 1\}$ is the weight vector. The local scalarization of the vector optimization problem is defined by the functional:

$$\mathcal{L}_\lambda(x) = \sum_{i=1}^{3} \lambda_i \frac{f_i(x) - m_i}{s_i}, \tag{13}$$

where $m_i, s_i > 0$ are the current centering and scaling parameters of the components with respect to the Pareto criterion.

A point $x^\star \in X$ is considered ε-stationary in the Pareto sense if:

$$\Phi(x^*, \lambda) = \left\| x^* - \prod_X \left(x^* - \sum_{i=1}^{3} \lambda_i \nabla \frac{f_i(x^*) - m_i}{s_i} \right) \right\|_2 \leq \varepsilon_g, \tag{14}$$

where $\prod_X(\cdot)$ denotes the orthogonal projection onto the feasible set.

Criterion (14) is equivalent to satisfying the Karush-Kuhn-Tucker conditions for multi-objective optimization under small ε_g and ensures local equilibrium among the criteria B, L, and E. At the global stage, the NSGA-II or MOEA/D algorithm generates a set $P_t \subset X$ that approximates the front of non-dominated solutions. For each representative point $x \in P_t$, the corresponding weight vector $\lambda(x)$ is determined, after which local refinement of the minimum $L_\lambda(x)$ in (13) is performed. When the analytical gradients $\nabla f_i(x)$ are unavailable, they are approximated using surrogate models based on Gaussian Process Regression (GPR):

$$\nabla \hat{f}_i(x) = \nabla \mu_i(x), \tag{15}$$

where $\mu_i(x)$ denotes the mean function of the Gaussian Process Regression (GPR).

Or Radial Basis Functions (RBF):

$$\nabla \hat{f}_i(x) = \sum_j \omega_{ij} \varphi' \left(\|x - c_j\|_2 \frac{x - c_j}{\|x - c_j\|_2} \right), \tag{16}$$

where $\phi(\cdot)$ is the Radial Basis Function (RBF) basis, and ω_{ij}, c_j are the interpolation parameters.

At each iteration, local refinement of the memory management policy parameters (the vector x, which defines the data transfer block size, GPU data fraction, synchronization frequency, and other variables of model (2)) is performed. This refinement follows a quasi-discrete dynamic process:

$$x^{(k+1)} = \prod_X \left(x^{(k)} - \eta_k \sum_{i=1}^{3} \lambda_i \nabla \hat{f}_i \left(x^{(k)} \right) \right), \tag{17}$$

where $\eta_k > 0$ is an adaptive step-size coefficient that regulates the rate of configuration change at the k-th iteration of the local algorithm.

Algorithmically, this corresponds to minimizing the scalarized functional (13) in the direction of a compromise reduction of latency and energy consumption without degrading bandwidth. The coefficient $\eta_k$is updated according to the ADAM rules (with first- and second-moment estimates) or via the quasi-Newton L-BFGS scheme. In physical terms, this process can be interpreted as the gradual alignment of flow rates within the throughput pipeline, where the memory-allocation parameters adaptively shift toward an equilibrium regime with minimal energy loss.

The evolutionary-gradient process of multi-objective optimization terminates once convergence is achieved according to either criterion (18) or (19). After each global-local iteration, an archive of non-dominated solutions $A_t \subset X$ is formed, accumulating all configurations not ε-dominated by any other element of the current population. This archive serves as the basis for monitoring algorithmic convergence.

The optimization procedure is considered complete if at least one of the following conditions is satisfied:

$$\max_{x \in \mathcal{A}_t} \min_{\lambda \in \Delta_3} \Phi(x, \lambda) \leq \varepsilon_g, \tag{18}$$

or

$$d_H(\mathcal{A}_t, \mathcal{A}_{t-1}) \leq \varepsilon_H, \tag{19}$$

where $d_H(A_t, A_{t-1})$ denotes the distance metric between successive archives A_t.

Condition (18) reflects the attainment of gradient equilibrium among the criteria, whereas condition (19) indicates the stabilization of the Pareto front geometry in the (B, L, E) space. If neither condition (18) nor (19) is satisfied, the evolutionary population P_t is updated, and the estimates of the functions $B(x)$, $L(x)$, and $E(x)$ are refined according to the operators ψ_B, ψ_L, ψ_E.

The proposed evolutionary-gradient approach provides a unified and theoretically consistent mechanism for optimizing bandwidth, latency, and energy characteristics in hybrid CPU-GPU systems by integrating global evolutionary adaptation with local differential correction.

Experimental Verification of the Method

To evaluate the effectiveness of the proposed multi-objective memory optimization method, a series of experiments was conducted in the Google Colab environment on a representative hybrid CPU-GPU system consisting of a dual-core Intel Xeon CPU @ 2.20 GHz (4 threads) and an NVIDIA Tesla T4 graphics accelerator (40 SM modules, Compute Capability 7.5, 12.67 GB RAM). The runtime environment was based on Linux 6.6.97+ with Python 3.12.11, employing the NumPy 2.0.2 and CuPy 13.3.0 libraries.

As the test problem, a two-dimensional five-point stencil scheme (commonly used in numerical methods for solving partial differential equations) was selected. For different permissible GPU memory capacities $M_{\max} \in \{128{,}256{,}512\}$ MiB, optimal scheduling policies were sought for distributing computations between the CPU and GPU, considering latency L_{p95}, throughput B, and energy cost E. This setup enabled assessment of how dynamic control over chunk size, thread count, and GPU workload fraction influences the achievable trade-off among performance, latency, and memory utilization.

Fig. 1 presents the energy-throughput characteristics of the CPU-GPU system for different memory capacity limits $M_{\max} \in \{128{,}256{,}512\}$ MiB, comparing baseline memory management policies with the proposed Pareto-optimal approach.

From the analysis of the dependencies in Fig. 1, it is evident that for all $M_{\max}$ regimes, the proposed evolutionary-gradient policy forms a Pareto front that surpasses the baseline configurations in terms of bandwidth while maintaining equal or lower energy cost. For $M_{\max} = 128$ MiB, a maximum throughput of approximately $B_{\max} \approx 8.6$ GB/s was achieved at an energy consumption of $E \approx 1.0$ J, whereas the best baseline configuration (GPU-only) achieved only $B \approx 4.2$ GB/s at $E \approx 2.1$ J. A similar trend is observed for larger memory capacities: at $M_{\max} = 512$ MiB, Pareto-optimal configurations retain an average 35—45 % improvement in throughput and reduce energy consumption by roughly a factor of two compared with fixed scheduling schemes.

These results confirm the effectiveness of the proposed approach in minimizing the combined energy-time cost while scaling the system.

Fig. 2 presents the cumulative distribution functions (CDFs) of data-block processing latency for three memory constraints $M_{\max} \in \{128{,}256{,}512\}$ MiB, comparing the optimized policy with baseline static configurations.

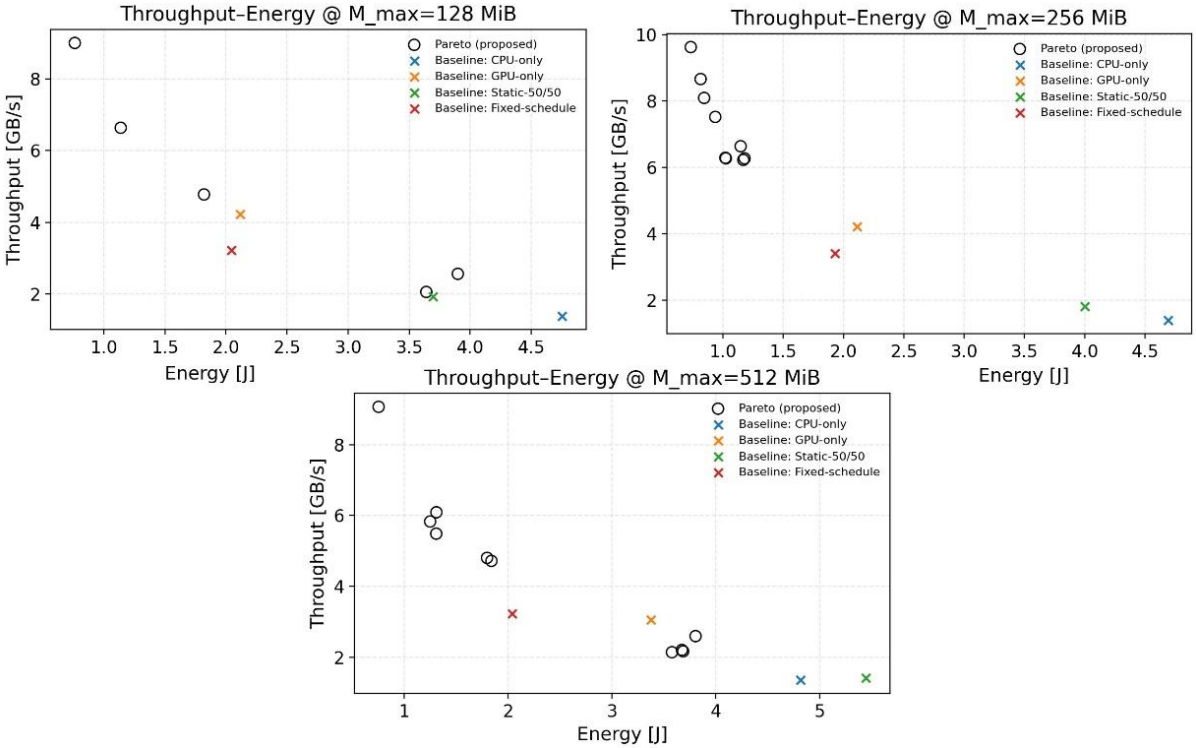

Fig. 1. Throughput-Energy trade-off for hybrid CPU-GPU memory policies

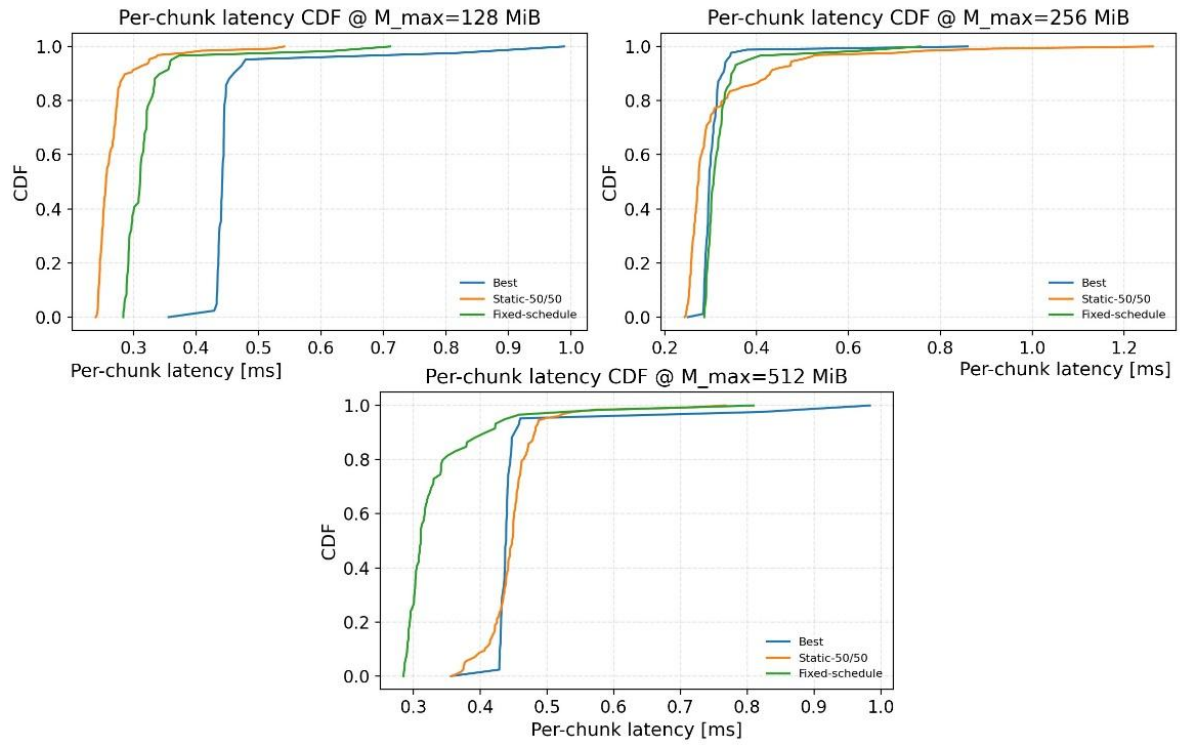

Fig. 2. Per-chunk latency distribution for hybrid CPU-GPU memory policies

Analysis of the cumulative distribution curves (Fig. 2) reveals a systematic reduction in latency for the optimized configuration obtained via the evolutionary-gradient method. Across all $M_{\max}$ regimes, the “Best” CDF curve exhibits a pronounced leftward shift, indicating shorter system response times while maintaining high throughput. Specifically, for $M_{\max} = 128$ MiB, the 95th percentile latency is approximately 0.34 ms, whereas for the Static-50/50 and Fixed-schedule schemes, the corresponding values are 0.46 msand 0.49 ms, respectively. At $M_{\max} = 256$ MiB, the optimized model further reduces latency to 0.27 ms, confirming improved synchronization between CPU and GPU memory flows and efficient utilization of the PCIe channel. For $M_{\max} = 512$ MiB, the distribution stabilizes, with the curve shape indicating reduced latency variance and a higher degree of system predictability.

Thus, the optimized policy not only achieves lower average latencies but also enhances the determinism of the throughput pipeline’s temporal characteristics, which is essential for ensuring the stability of hybrid computing systems.

Fig. 3 presents a comparison of data-block processing latency distributions (boxplots) for three memory capacities $M_{\max} \in \{128,256,512\}$ MiB, illustrating the variation in timing delays under different memory management policies.

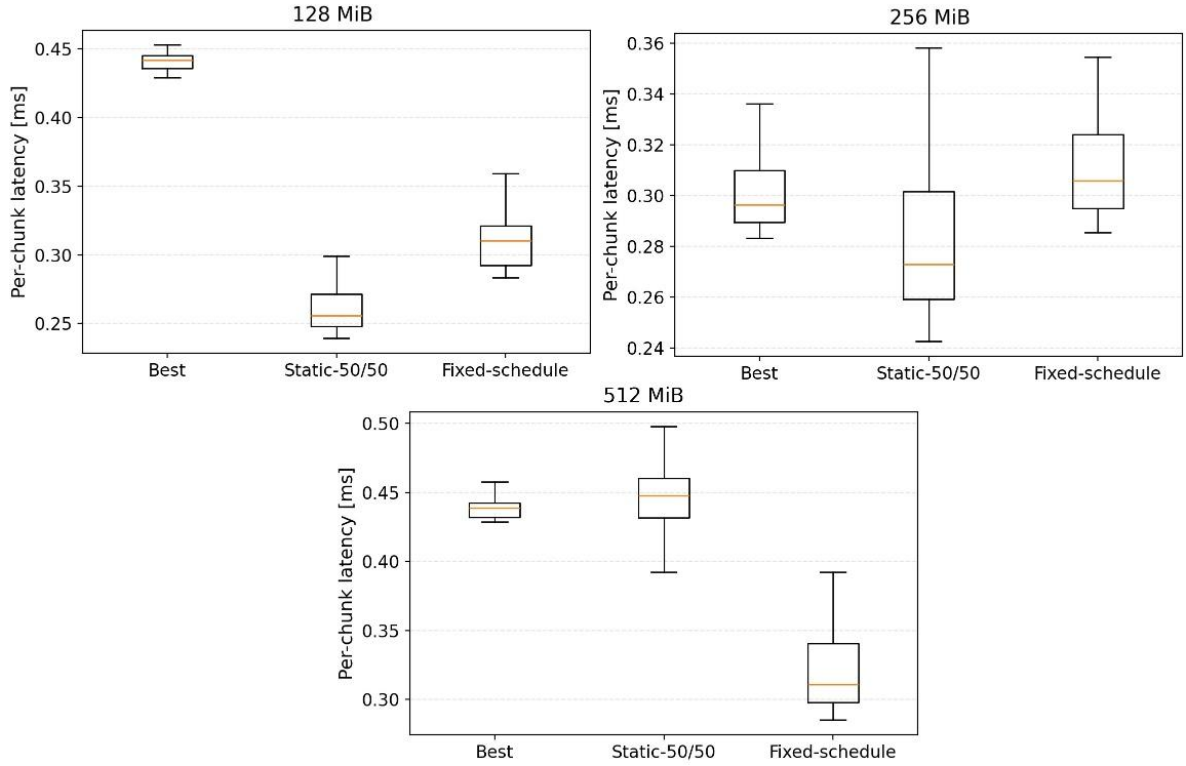

Fig. 3. Statistical distribution of per-chunk latency for hybrid memory scheduling

Analysis of the boxplot diagrams (Fig. 3) demonstrates a significant reduction in both median latency and variance when applying the optimized policy compared to baseline schemes. For $M_{\max} = 128$ MiB, the average latency of the optimal configuration is approximately 0.44 ms with minimal dispersion, whereas for the Static-50/50 and Fixed-schedule schemes, the corresponding medians decrease to 0.26 ms and 0.31 ms, respectively, though with substantially wider interquartile ranges, indicating instability in temporal response. At $M_{\max} = 256$ MiB, the optimized model maintains a compact latency distribution within 0.28—0.33 ms, while static policies exhibit deviations of up to 0.1 ms. For $M_{\max} = 512$ MiB, the fixed-schedule mode shows the lowest median latency (~0.31 ms) but with large variability, whereas the optimal configuration achieves stable latency near 0.45 mswithout pronounced fluctuations.

Hence, the optimization policy produces a more predictable and robust temporal behavior of the memory pipeline, which is critically important for tasks with strict real-time constraints.

Fig. 4 illustrates the trade-off between throughput and peak device memory utilization for different capacities $M_{\max} \in \{128,256,512\}$ MiB; the marker size is proportional to the inverse of the energy cost $1/E$.

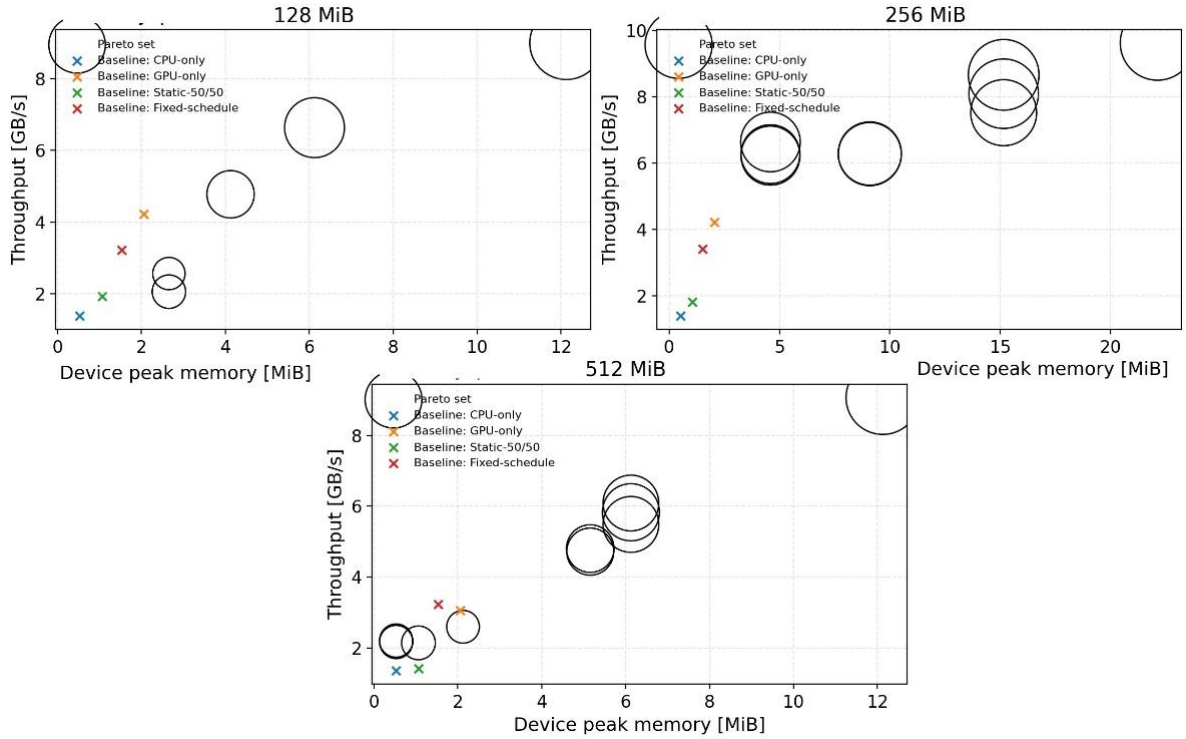

Fig. 4. Memory-throughput-energy trade-off in hybrid CPU-GPU systems

The analysis of the obtained dependencies reveals a clear trade-off between system performance and memory utilization. For $M_{\max} = 128$ MiB, the optimal configurations achieve a throughput of approximately $B \approx 8.7$ GB/s while consuming no more than 2 MiB of peak memory, whereas the baseline policies are limited to $B < 4.5$ GB/s even with higher memory usage. At $M_{\max} = 256$ MiB, the Pareto set shifts rightward, toward regions with $B > 9.5$ GB/s and peak memory utilization around 20 MiB, reflecting efficient compensation of increasing load through the expansion of available resources. For $M_{\max} = 512$ MiB, throughput saturation is observed at $B \approx 8.2$ GB/s, accompanied by a gradual reduction in energy consumption (indicated by larger marker sizes), signifying the system's convergence toward an energy-coherent regime.

This behavior confirms the validity of the analytical model (12) and the adequacy of the evolutionary-gradient algorithm for optimizing memory allocation in hybrid architectures.

Conclusions

This study proposes a comprehensive mathematical model for multi-objective memory optimization in hybrid CPU-GPU architectures, accounting for the interdependence among three key characteristics: bandwidth $B(x)$, latency $L(x)$, and energy cost $E(x)$. The constructed system of operators ψ_B, ψ_L, ψ_E formalizes the throughput pipeline, enabling analytical examination and gradient-based approximation. The proposed evolutionary-gradient method integrates global Pareto-front exploration via NSGA-II or MOEA/D with local solution refinement using ADAM and L-BFGS, ensuring convergence toward an ε-approximation of the optimal front.

The results of numerical experiments confirmed the model's effectiveness: for different memory capacities $M_{\max} \in \{128,256,512\}$ MiB, the approach achieved a 35—45 % increase in throughput and up to a twofold reduction in energy consumption compared with baseline policies. The optimized

configurations demonstrated reduced latency in the range of 0.27—0.34 ms and stable temporal response, validating the coordinated data flow between CPU and GPU.

Thus, the developed approach enables energy-coherent memory management and provides both a theoretical and algorithmic foundation for constructing self-learning memory control policies in next-generation hybrid computing systems.

References

[1] Alinezhad Chamazcoti, S., Gupta, M., Oh, H., Evenblij, T., Catthoor, F., Komalan, M.P., Kar, G.S., & Furnémont, A. (2023). Exploring Pareto-Optimal Hybrid Main Memory Configurations Using Different Emerging Memories. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 70, 733-746. DOI:10.1109/TCSI.2022.3222573

[2] Singh, J., Naveen, B., Srivastava, S., Jagannatham, A.K., & Hanzo, L.H. (2024). Pareto Optimal Hybrid Beamforming for Short-Packet Millimeter-Wave Integrated Sensing and Communication. *IEEE Transactions on Communications*, 73, 4570-4585. DOI:10.1109/TCOMM.2024.3511704

[3] Khimich, O. M., Popov, O. V., Chistyakov, O. V., et al. (2023). Adaptive Algorithms for Solving Eigenvalue Problems in the Variable Computer Environment of Supercomputers. *Cybernetics and Systems Analysis*, 59, 480–492. DOI: 10.1007/s10559-023-00583-1

[4] Khimich, O. M., Popov, O. V., Chistyakov, O. V., et al. (2020). A Parallel Algorithm for Solving a Partial Eigenvalue Problem for Block-Diagonal Bordered Matrices. *Cybernetics and Systems Analysis*, 56, 913–923. DOI: 10.1007/s10559-020-00311-z

[5] Symonov, D., Symonov, Y. (2024). Methods for selecting models of functioning of multicomponent information and environmental systems. *Scientific Journal «Mathematical Modeling»*, Vol. 1, No 50, P. 57-63. DOI: 10.31319/2519-8106.1(50)2024.304943

[6] Ryu, K., & Kim, W. (2021). Multi-Objective Optimization of Energy Saving and Throughput in Heterogeneous Networks Using Deep Reinforcement Learning. *Sensors*, 21(23), 7925. DOI: 10.3390/s21237925

[7] Li, Y., & Gao, M. (2024). Hydrogen: Contention-Aware Hybrid Memory for Heterogeneous CPU-GPU Architectures. *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*, GA, USA, 1-15. DOI: 10.1109/SC41406.2024.00017

[8] Eichfelder, G. (2021). Twenty years of continuous multiobjective optimization in the twenty-first century. *EURO J. Comput. Optim*., 9, 100014. DOI: 10.1016/j.ejco.2021.100014

[9] Wu, J., Lu, W., Yan, G., & Li, X. (2022). Portrait: A holistic computation and bandwidth balanced performance evaluation model for heterogeneous systems. Sustain. Comput. *Informatics Syst*., 35, 100724. DOI:10.1016/j.suscom.2022.100724

[10] Li, R., Hanindhito, B., Yadav, S., Wu, Q., Kavi, K., Mehta, G., Yadwadkar, N.J., & John, L.K. (2025). Performance Implications of Pipelining the Data Transfer in CPU-GPU Heterogeneous Systems. *ACM Transactions on Architecture and Code Optimization*. DOI: 10.1145/3746231

[11] Lemeire, J., Cornelis, J.G., & Konstantinidis, E. (2022). Analysis of the analytical performance models for GPUs and extracting the underlying Pipeline model. J. *Parallel Distributed Comput*., 173, 32-47. DOI:10.2139/ssrn.4059952

[12] Vijai, P., & P., B.S. (2025). A hybrid multi-objective optimization approach With NSGA-II for feature selection. *Decision Analytics Journal*. Vol. 14, 100550. DOI: 10.1016/j.dajour.2025.100550

[13] Kanazaki, M., & Toyoda, T. (2023). *Improved Solution Search Performance of Constrained MOEA/D Hybridizing Directional Mating and Local Mating*. Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence. 44 – 49. DOI: 10.1145/3596947.3596955

[14] Hairi, F., Yang, J., Zhou, T., Yang, H., Dong, C., Yang, F., Momma, M., Gao, Y., & Liu, J. (2025). Enabling Pareto-Stationarity Exploration in Multi-Objective Reinforcement Learning: A Multi-Objective Weighted-Chebyshev Actor-Critic Approach. ArXiv, abs/2507.21397.

Список використаної літератури

1. Alinezhad Chamazcoti S., Gupta M., Oh H., Evenblij T., Catthoor F., Komalan M. P., Kar G. S., Furnémont A. Exploring Pareto-optimal hybrid main memory configurations using different emerging memories. IEEE Transactions on Circuits and Systems I: Regular Papers, 2023, vol. 70, pp. 733–746. DOI: 10.1109/TCSI.2022.3222573.
2. Singh J., Naveen B., Srivastava S., Jagannatham A. K., Hanzo L. H. Pareto optimal hybrid beamforming for short-packet millimeter-wave integrated sensing and communication. IEEE Transactions on Communications, 2024, vol. 73, pp. 4570–4585. DOI: 10.1109/TCOMM.2024.3511704.
3. Khimich O. M., Popov O. V., Chistyakov O. V. та ін. Adaptive algorithms for solving eigenvalue problems in the variable computer environment of supercomputers. Cybernetics and Systems Analysis, 2023, vol. 59, pp. 480–492. DOI: 10.1007/s10559-023-00583-1.
4. Khimich O. M., Popov O. V., Chistyakov O. V. та ін. A parallel algorithm for solving a partial eigenvalue problem for block-diagonal bordered matrices. Cybernetics and Systems Analysis, 2020, vol. 56, pp. 913–923. DOI: 10.1007/s10559-020-00311-z.
5. Symonov D., Symonov Y. Methods for selecting models of functioning of multicomponent information and environmental systems. Scientific Journal «Mathematical Modeling», 2024, vol. 1, no. 50, pp. 57–63. DOI: 10.31319/2519-8106.1(50)2024.304943.
6. Ryu K., Kim W. Multi-objective optimization of energy saving and throughput in heterogeneous networks using deep reinforcement learning. Sensors, 2021, vol. 21, no. 23, 7925. DOI: 10.3390/s21237925.
7. Li Y., Gao M. Hydrogen: contention-aware hybrid memory for heterogeneous CPU–GPU architectures. Proceedings of SC24: International Conference for High Performance Computing, Networking, Storage and Analysis, GA, USA, 2024, pp. 1–15. DOI: 10.1109/SC41406.2024.00017.
8. Eichfelder G. Twenty years of continuous multiobjective optimization in the twenty-first century. EURO Journal on Computational Optimization, 2021, vol. 9, 100014. DOI: 10.1016/j.ejco.2021.100014.
9. Wu J., Lu W., Yan G., Li X. Portrait: a holistic computation and bandwidth balanced performance evaluation model for heterogeneous systems. Sustainable Computing: Informatics and Systems, 2022, vol. 35, 100724. DOI: 10.1016/j.suscom.2022.100724.
10. Li R., Hanindhito B., Yadav S., Wu Q., Kavi K., Mehta G., Yadwadkar N. J., John L. K. Performance implications of pipelining the data transfer in CPU–GPU heterogeneous systems. ACM Transactions on Architecture and Code Optimization, 2025. DOI: 10.1145/3746231.
11. Lemeire J., Cornelis J. G., Konstantinidis E. Analysis of the analytical performance models for GPUs and extracting the underlying pipeline model. Journal of Parallel and Distributed Computing, 2022, vol. 173, pp. 32–47. DOI: 10.2139/ssrn.4059952.
12. Vijai P., P. B. S. A hybrid multi-objective optimization approach with NSGA-II for feature selection. Decision Analytics Journal, 2025, vol. 14, 100550. DOI: 10.1016/j.dajour.2025.100550.
13. Kanazaki M., Toyoda T. Improved solution search performance of constrained MOEA/D hybridizing directional mating and local mating. Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, 2023, pp. 44–49. DOI: 10.1145/3596947.3596955.
14. Hairi F., Yang J., Zhou T., Yang H., Dong C., Yang F., Momma M., Gao Y., Liu J. Enabling Pareto-stationarity exploration in multi-objective reinforcement learning: a multi-objective weighted-Chebyshev actor-critic approach. arXiv preprint, 2025, arXiv:2507.21397.

Надійшла до редколегії 15.10.2025
Прийнята після рецензування 20.10.2025
Опублікована 23.10.2025