

AI-PASS: An AI-Based Predictive, Adaptive and Energy Aware Task Scheduling Framework for Heterogeneous Parallel Computing

Nidhi Chahal^{1,*}, Purshotam Jha², Sonia³

Abstract

Heterogeneous parallel computing systems are made up of a combination of multicore CPUs, GPUs, and specialized devices to boost computational power and utilize resources. But changes in workload intensity and processor capability, energy consumption and system conditions are dynamic, making efficient task scheduling difficult. Traditional scheduling approaches are based primarily on task properties or the state of the resources employed and can be sub-optimal when workloads are evolving. This paper introduces a predictive and feedback-based scheduling framework, named AI-PASS (AI-based Predictive, Adaptive and Sustainable Scheduler), which integrates workload prediction with adaptive taskresource selection. A novel Adaptive Task Suitability Score (ATSS) is used to predict execution time, processor utilization, energy usage, thermal condition, communication overhead and compatibility of tasks and resources. The weights associated are dynamically changed based on the condition of the system. AI-PASS can be described as being closed loop with monitoring, prediction, characterization of task, selection of resources, execution, measurement, and correction based feedback. The goal of the framework is to minimize makespan, energy usage and resource imbalance, and increase utilization. Further, simulation experiments are conducted to compare the performance of AI-PASS with Round Robin, Min-Min, and HEFT on heterogeneous CPU-GPU workloads which are measured in terms of makespan, energy consumption, throughput, CPU utilization, load imbalance and scheduling overhead.

Keywords: Heterogeneous computing, parallel computing, task scheduling, artificial intelligence, workload prediction, energy-aware scheduling, CPU-GPU scheduling, adaptive scheduling, resource allocation, ATSS

*Author for Correspondence

Nidhi Chahal
E-mail: Nidhi.ece@cgce.edu.in

¹Assistant Professor, Department of Electronics and Communication Engineering, Chandigarh Engineering College-CGC, Landran, Mohali, Punjab, India

^{2,3}Student, Department of Electronics and Communication Engineering, Chandigarh Engineering College-CGC, Landran, Mohali, Punjab, India

Received Date: September 10, 2026
Accepted Date: September 15, 2026
Published Date: September 20, 2026

Citation: Nidhi Chahal, Purshotam Jha, Sonia. AI-PASS: An AI-Based Predictive, Adaptive and Energy Aware Task Scheduling Framework for Heterogeneous Parallel Computing. Journal of Operating Systems Development & Trends. 2026; 13(2): 28–34

INTRODUCTION

With the growing need for high-performance computing, heterogeneous parallel computing has become an important method for running computationally intensive applications. Modern systems usually feature multiple CPU cores, GPUs, and dedicated accelerators for increased performance and resource utilization. However, different processing resources make task scheduling more complex, as the execution efficiency of a task may vary from one processing resource to another.

Typical scheduling approaches are based on the characteristics of a task (i.e., priority), availability

of a processor, or estimated execution time. All these methods operate well when the workload is stable, but their efficiency may decline if the workload intensity, processor utilization, communication needs, or energy consumption fluctuate over time. This could lead to longer execution times, an unbalanced use of resources, or unnecessary energy consumption in a heterogeneous environment if a task is assigned only based on the currently available processor.

High-performance processors easily consume a lot of energy; therefore, energy efficiency is an important factor in parallel computing. Therefore, recent research has focused on energy-shaping scheduling algorithms that consider both execution performance and energy consumption. Meanwhile, the advent of artificial intelligence and machine learning methods has opened up possibilities for forecasting workload behavior and enabling intelligent resource management.

However, it is insufficient to predict the optimum processing resource for a single task from the workload. To determine the schedule, the scheduler must consider several factors: expected run time, resource utilization, energy requirement, thermal condition, communication overhead, compatibility of the task and available processor. The challenge is then to tie predictive information with an adaptive task/resource selection mechanism.

To solve these issues, this study introduces a task scheduling approach for heterogeneous parallel computing systems called AI-PASS (AI-based Predictive, Adaptive, and Sustainable Scheduler). AI-PASS integrates workload prediction and a multi-factor resource-selection process referred to as the Adaptive Task Suitability Score (ATSS). Performance, energy, thermal, communication, and utilization parameters are used to determine the suitability of each available resource for a specific task in ATSS.

AI-PASS is not a fixed scheduling policy; instead, it dynamically varies the importance of these parameters based on the current system condition. The framework also allows for the incorporation of execution feedback to allow for differences between the predicted and actual behavior of the system to affect future scheduling decisions. Hence, the proposed approach is a continuous cycle of monitoring, predicting, scheduling, executing, measuring, and adapting.

The primary aims of this research are as follows: to design a predictive scheduling mechanism; enhance task selection from heterogeneous resources; minimize unnecessary energy consumption; achieve efficient resource utilization; and assess the performance of adaptive scheduling when the workload changes.

The proposed work has the following significant features:

- Predictive and adaptive heterogeneous task scheduling framework.
- Multi-Factor Resource Selection Adaptive Task Suitability Score (ATSS).
- Dynamic weighting of performance, energy, thermal and communication factors.
- An adaptive scheduling decision feedback mechanism based on execution behavior.
- A simulation-based approach was used to compare AI-PASS with traditional scheduling methods.

The proposed framework is designed to offer a feasible solution for combining AI with heterogeneous task scheduling, ensuring a trade-off between the computational performance and energy efficiency.

RELATED WORK

In recent years, research on heterogeneous task scheduling has focused on enhancing computational performance, minimizing energy consumption, and optimizing resource utilization. Traditional HEFT methods primarily rely on task priorities and estimated execution times to select resources but may not perform well when a workload changes dynamically [1].

Heterogeneous multicore systems have emerged with energy awareness as an important scheduling issue. Liu et al. integrated task prioritization, core-aware mapping, predictive dynamic voltage and

frequency scaling (DVFS), and optimized scheduling and energy consumption [2]. Likewise, Huda suggested the hybrid PSO-LSTM approach, which integrates workload prediction and energy-efficient task scheduling [3]. The use of AI, especially deep reinforcement learning, has been further enhanced to make scheduling decisions in heterogeneous computing systems [4, 5]. The energy-aware selection of task mapping and resources between performance and energy has been studied in other studies [6, 7–9]. Recent reinforcement learning-based methods have also shown better adaptability and resource allocation for variable workloads [10, 11–13].

However, most current methods are targeted at specific parts, such as prediction, energy optimization, resource selection, or reinforcement learning. The complementary approach of workload prediction, resource suitability and execution feedback is less consistently treated. To fill this void, the proposed AI-PASS framework incorporates the concepts of workload forecasting, adaptive resource assessment, energy awareness, and feedback-based scheduling using an Adaptive Task Suitability Score (ATSS).

RESEARCH GAP AND PROBLEM FORMULATION

Research Gap

Several heterogeneous scheduling approaches have made significant progress in improving execution performance and energy efficiency. However, some limitations of this study remain. Traditional heuristic approaches generally depend on predefined scheduling rules and may become less effective when workload conditions change rapidly. Recent learning-based approaches have demonstrated the potential of intelligent scheduling and load-aware resource management; however, prediction and scheduling decisions are not always integrated into a unified mechanism [14, 15].

Energy-efficient scheduling has also received considerable attention in heterogeneous computing and edge environments. However, energy optimization may focus primarily on reducing power consumption without sufficiently considering future workload conditions and the changing states of available resources [16].

A further limitation is that scheduling decisions may consider only a limited number of parameters. In a heterogeneous environment, processor utilization, execution time, energy consumption, thermal conditions, communication overhead, and task-resource compatibility can simultaneously influence the scheduling quality. Therefore, a scheduling mechanism that dynamically adapts to the importance of these factors can provide better flexibility under changing system conditions.

Therefore, the identified research gap is the lack of a lightweight and unified scheduling mechanism that integrates workload prediction, multi-factor resource suitability, adaptive decision weighting, and execution feedback into a continuous scheduling loop. AI-PASS is designed to address this gap through its predictive and feedback-based scheduling architecture.

PROPOSED AI-PASS FRAMEWORK

AI-PASS is designed as a predictive and adaptive scheduling framework for heterogeneous parallel systems. It selects a suitable processing resource for each incoming task based on the current system conditions and predicted workload. The framework consists of workload monitoring, prediction, task characterization, adaptive resource selection, and feedback control.

Workload Monitoring

The monitoring module continuously collects the processor utilization, queue length, execution time, energy consumption, temperature, and communication activity to represent the current system state.

Workload Prediction

Recent workload observations are used to estimate the near-future system state. The state is represented as follows:

$$S_t = [U_t, Q_t, E_t, T_t]$$

Where, U_t , Q_t , E_t , and T_t denote utilization, queue length, energy consumption, and thermal state. The predicted state is:

$$S_{t+1} = P(S_t, S_{t-1}, \dots, S_{t-k})$$

Where, P is the prediction model and k is the observation history length.

Task Characterization

Each task is analyzed using computational intensity, execution time, memory, parallelism, and communication requirements.

Adaptive Resource Selection

ATSS evaluates execution time, utilization, energy, thermal condition, communication overhead, and task-resource compatibility. The resource with the highest score was selected.

Feedback Adaptation

Actual performance is compared with predictions, and the resulting feedback updates future scheduling decisions (Figure 1). Thus, AI-PASS follows the following cycle:

Monitor → Predict → Characterize → Score → Schedule → Execute → Measure → Adapt (Figure 1).

Overall Objective

The framework aims to reduce the makespan, energy consumption, and load imbalance while improving resource utilization with a reasonable scheduling overhead.

ADAPTIVE SUITABILITY SCORE AND SCHEDULING ALGORITHM

The Adaptive Task Suitability Score (ATSS) is the main decision mechanism of AI-PASS. It evaluates the suitability of each resource for a given task using performance, utilization, energy, thermal, communication, and compatibility factors.

For task t_{ij} and resource r_{ij} :

$$ATSS_{ij} = w_1 P_{ij} + w_2 U_{ij} + w_3 E_{ij} + w_4 T_{ij} + w_5 C_{ij} + w_6 K_{ij}$$

where P, U, E, T, C, K represent normalized performance, resource availability, energy efficiency, thermal suitability, communication suitability, and task-resource compatibility, respectively. All values were normalized between 0 and 1, with

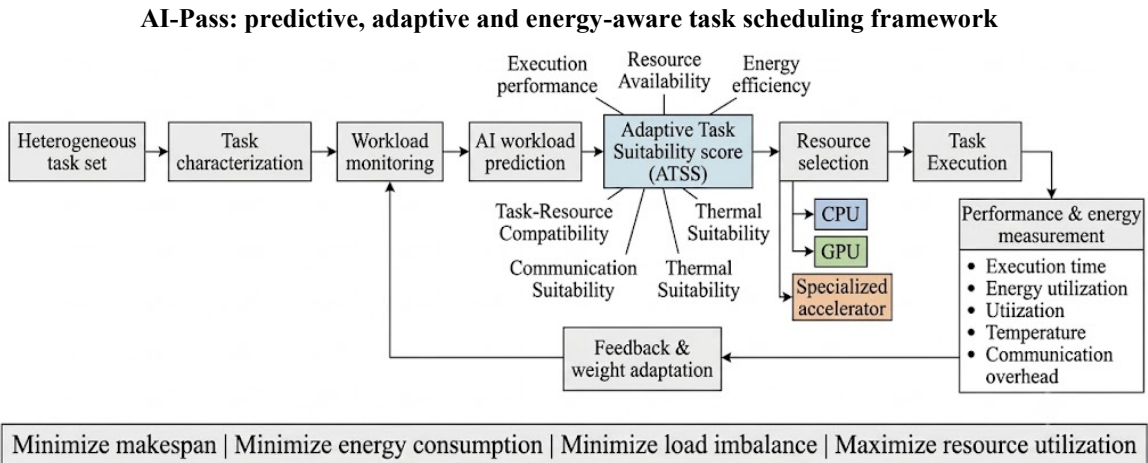


Figure 1. Closed-loop architecture of the proposed AI-PASS framework.

$$\sum_{k=1}^6 w_k = 1$$

Unlike fixed-weight methods, AI-PASS dynamically adjusts the weights according to current system conditions. The resource with the highest score was selected.

$$r^* = \arg \max_{r_j \in R} ATSS_{ij}$$

Scheduling Algorithm

The AI-PASS scheduling process consists of the following steps:

1. Initialize resources and monitoring parameters.
2. Collect current workload and resource-state information.
3. Predict the near-future workload.
4. Characterize the next task.
5. Calculate ATSS for all eligible resources.
6. Select the resource with the highest ATSS.
7. Execute the task and measure actual performance.
8. Compare predicted and actual results and update scheduling weights.
9. Repeat until all tasks are completed.

This feedback-based process allows the selected resource to change according to the workload and system conditions, providing adaptive scheduling instead of a fixed resource assignment policy.

EXPERIMENTAL METHODOLOGY

The AI-PASS framework was evaluated using a simulated heterogeneous computing environment containing CPU cores and GPU resources with different processing capabilities. A heterogeneous task set consisting of different computational requirements, execution times, and resource preferences was considered. Low, medium, and high workload intensities were used to represent changing system conditions. The workload prediction module used historical resource utilization and task-arrival information, and the predicted state was provided to the ATSS-based resource-selection mechanism.

AI-PASS was compared with three scheduling methods: Round Robin (RR), which distributes tasks sequentially; Min-Min, which selects tasks according to the minimum expected completion time; and HEFT, which uses task priorities and heterogeneous execution costs.

Performance was evaluated using makespan, energy consumption, resource utilization, throughput, load imbalance, and scheduling overhead. All methods used the same workload and resource configuration to ensure consistent comparison under different system-load conditions.

RESULTS AND DISCUSSION

AI-PASS was tested with 200 tasks with various computational demands in a simulated heterogeneous CPU-GPU environment. Round Robin (RR), Min-Min, HEFT, and AI-PASS were performed with the same workload for easy comparison (Table 1).

The best execution time obtained by AI-PASS was the lowest makespan, which was approximately 36.8% faster than RR and showed a slight improvement in throughput over Min-Min. In this simulation, its energy consumption was 17.3% less than that of RR, but HEFT had the lowest energy consumption.

Overall, the results showed that AI-PASS is able to make a more balanced scheduling decision by considering workload prediction, suitability of resources, energy, and feedback. Because the evaluation is simulation-based, real CPU-GPU validation is necessary for practical confirmation.

Table 1. Performance comparison of scheduling methods.

Scheduling method	Makespan (ms)	Energy	Resource utilization	Throughput
Round Robin	2145.26	3432.34	79.17%	0.093
Min-Min	1365.05	2865.71	99.97%	0.147
HEFT	1571.00	2512.65	79.14%	0.127
AI-PASS	1354.86	2837.13	99.67%	0.148

LIMITATIONS AND FUTURE RESEARCH

The present study had some limitations. First, the evaluation is simulation-based and may not fully represent real CPU-GPU systems, where factors such as memory contention, GPU kernel overhead, cache effects, power state transitions, and thermal throttling can influence performance. Second, workload prediction depends on the quality of historical data, and the prediction accuracy may decrease under sudden or unfamiliar workload patterns. The use of normalized parameters in the ATSS may also influence scheduling decisions depending on the selected normalization ranges and adaptive weights.

Future work will focus on validating AI-PASS on real heterogeneous CPU-GPU platforms and task-based programming environments. Reinforcement learning can be explored for the automatic adjustment of ATSS weights, whereas lightweight prediction models can reduce scheduling overhead. Further extensions may consider distributed CPU-GPU-FPGA systems and edge accelerators. Additional factors, such as reliability, memory bandwidth, data locality, and carbon-aware energy consumption, can also be incorporated to improve the applicability of cloud and edge environments [17].

CONCLUSION

In this study, we introduce a framework for task scheduling in heterogeneous parallel computing, named AI-PASS, which schedules tasks with the assistance of Artificial Intelligence by considering adaptivity and energy efficiency. The framework integrates workload anticipation with multi-factor task-resource selection, accounting for performance, resource availability, energy, thermal condition, communication, and task-resource compatibility, using the Adaptive Task Suitability Score (ATSS). It has an adaptive weighting and feedback loop that enables scheduling decisions to factor in changing conditions in the system.

The simulation results revealed that the proposed method, AI-PASS, delivered a minimum makespan of 1354.86 ms with a maximum throughput of 0.148 for the evaluated methods. However, HEFT was able to achieve a lower energy consumption in the tested scenario, indicating a trade-off between individual scheduling goals. In conclusion, AI-PASS offers a well-rounded scheduling strategy and serves as a steppingstone for further testing on real CPU-GPU hardware in the future.

REFERENCES

1. Topcuoglu H, Hariri S, Wu MY. Performance-effective and low-complexity task scheduling for heterogeneous computing. *IEEE Trans Parallel Distrib Syst.* 2002;13(3):260-74.
2. Liu Y, Qu H, Chen S, Feng X. Energy efficient task scheduling for heterogeneous multicore processors in edge computing. *Sci Rep.* 2025;15(1):11819.
3. ul Huda Z. An enhanced hybrid PSO-LSTM framework for energy-aware task scheduling in heterogeneous cloud computing environments. *J Emerg Paradigms Comput Syst.* 2025;1(1).
4. Sun C, Yang T, Lei Y. DDDQN-TS: a task scheduling and load balancing method based on optimized deep reinforcement learning in heterogeneous computing environment. *Int J Intell Syst.* 2022;37(11):9138-72.
5. Shen W, Lin W, Wu W, Wu H, Li K. Reinforcement learning-based task scheduling for heterogeneous computing in end-edge-cloud environment. *Cluster Comput.* 2025;28(3):179.
6. Nesmachnow S, Dorronsoro B, Pecero JE, Bouvry P. Energy-aware scheduling on multicore heterogeneous grid computing systems. *J Grid Comput.* 2013;11(4):653-80.

7. Moulik S, Chaudhary R, Das Z. HEARS: a heterogeneous energy-aware real-time scheduler. *Microprocess Microsyst.* 2020;72:102939.
8. Moulik S, Das Z, Devaraj R, Chakraborty S. SEAMERS: a semi-partitioned energy-aware scheduler for heterogeneous multicore real-time systems. *J Syst Archit.* 2021;114:101953.
9. Lin Z, Li C, Tian L, Zhang B. A scheduling algorithm based on reinforcement learning for heterogeneous environments. *Appl Soft Comput.* 2022;130:109707.
10. Chen J, He Y, Zhang Y, Han P, Du C. Energy-aware scheduling for dependent tasks in heterogeneous multiprocessor systems. *J Syst Archit.* 2022;129:102598.
11. Chen G, Qi J, Sun Y, Hu X, Dong Z, Sun Y. A collaborative scheduling method for cloud computing heterogeneous workflows based on deep reinforcement learning. *Future Gener Comput Syst.* 2023;141:284-97.
12. Song Y, Li C, Tian L, Song H. A reinforcement learning based job scheduling algorithm for heterogeneous computing environment. *Comput Electr Eng.* 2023;107:108653.
13. Deng Z, Yan Z, Huang H, Shen H. Energy-aware task scheduling on heterogeneous computing systems with time constraint. *IEEE Access.* 2020;8:23936-50.
14. Yao Z, Ding L, Zhang H, Li H, Chen L, Li Z. Deep reinforcement learning for job scheduling on load-aware heterogeneous cluster. *J Big Data.* 2025;12(1):276.
15. Idouar Y, Cassagne A, Pilla LL, Sopena J, Bouyer M, Orhan D, et al. Energy-aware scheduling strategies for partially-replicable task chains on heterogeneous processors. *Parallel Comput.* 2026:103206.
16. Tan Q, Chen W, Liu D. Task scheduling for heterogeneous multi-core processors based on deep reinforcement learning. *Int J Intell Syst.* 2025;2025(1):7562400.
17. Idouar Y, Cassagne A, Pilla LL, Sopena J, Bouyer M, Orhan D, Lacassagne L, Galayko D, Barthou D, Jegou C. Energy-aware scheduling strategies for partially-replicable task chains on heterogeneous processors. *Parallel Computing.* 2026 Jul 2:103206.