Priority-Tiered Predictive Scheduling: An SLA-Aware Extension of Mobile ViT-ISBOA for Cloud Resource Orchestration

Hari Shankar Punna[1*] · Dr. Arif Mohammad Abdul[2]

[1]Dept. of CSE, GITAM (Deemed to be University), Hyderabad, Telangana, India,

Dept. of CSE (Data Science), CVR College of Engineering, Hyderabad, Telangana, India

[2]Dept. of CSE, GITAM (Deemed to be University), Hyderabad, Telangana, India

Abstract: *Existing predictive cloud schedulers, including our prior MobileViT-ISBOA framework, optimize makespan, energy consumption, resource utilization, CPU balance, and migration cost as if all tasks were equally important. In practice, cloud workloads span multiple service tiers — mission-critical jobs with hard deadlines, standard business jobs with soft deadlines, and best-effort batch jobs with no deadline — and scheduling them uniformly risks violating Service Level Agreements (SLAs) for the tasks that matter most. This paper extends MobileViT-ISBOA with an SLA-aware fitness formulation that assigns priority weights and deadline-tardiness penalties to a sixth scheduling objective, and augments the ISBOA optimizer's candidate construction and migration policy so that Gold-tier (mission-critical), Silver-tier (standard), and Bronze-tier (best-effort) tasks are scheduled under tier-appropriate urgency. The MobileViT forecasting module is unchanged, isolating the contribution to the scheduling layer. We define a reproducible protocol for synthesizing SLA tiers and deadlines on the existing NASA iPSC/860 and HPC2N traces, and introduce SLA-specific metrics — SLA violation rate, weighted tardiness, and tier-wise makespan/energy — alongside the five original metrics. [ILLUSTRATIVE, pending experiments] Expected results indicate that SLA-ISBOA reduces Gold-tier deadline violations substantially relative to tier-agnostic scheduling and a priority-weighted heuristic baseline, at a modest cost in aggregate energy consumption, while avoiding the larger energy penalty expected from a deadline-only scheduler such as Earliest-Deadline-First.*

Keywords: Cloud resource scheduling; Service Level Agreement (SLA); priority-aware scheduling; predictive workload forecasting; MobileViT; metaheuristic optimization; Secretary Bird Optimization Algorithm; Bobcat Optimization Algorithm; CloudSim; energy-aware computing.

1. INTRODUCTION

Cloud vendors rarely offer one generic compute service. Typically, in commercial services, workloads are segregated based on criticality; there is preemptive/spot capacity for best-effort batch tasks, dedicated/reserved capacity for mission-critical workloads, and capacity on-demand for regular business workloads. Each one has its own cost and different level of guarantees. This aspect cannot be captured through any scheduler which considers all the tasks the same way, as MobileViT-ISBOA and other metaheuristic schedulers consider them. For instance, mission-critical task can be queued behind the batch task just to minimize the aggregate makespan, although the mission-critical task incurs penalty due to delay, but batch task does not.

This gap has real consequences: most scheduling objectives in the literature average across the entire dataset rather than weighting by task and tier, and recent research on AI-assisted cloud scheduling identifies SLA adherence as one of the main unsolved problems alongside energy efficiency. Prior work has used priority-class models that group workloads into tiers, and consolidation models that bound SLA violations while balancing physical servers; more recent proactive approaches integrate multi-dimensional load prediction with reinforcement learning to anticipate overload. This work addresses the gap by integrating an explicit, multi-tier SLA objective with a learnt, proactive load predictor in a single scheduling objective.

This paper plugs the hole by augmenting MobileViT-ISBOA with a separate SLA layer. The key technical contributions are:

- Tier-based SLA framework (Gold, Silver, and Bronze), with tier-specific deadlines and penalties for late deliveries included in the ISBOA fitness function as an additional, weight-prioritized objective along with the other five — without modifying the MobileViT prediction module and making the contribution completely orthogonal to the scheduling layer.

- Synthesizable scheme for tier-based SLAs with corresponding deadlines and priorities based on publicly available HPC benchmarks (NASA iPSC/860, HPC2N), which do not contain any information about service tiers and therefore enable testing of the new approach without developing and obtaining an entirely new data set with SLA labels.
- A novel evaluation metric for the cost of tier-differentiation in form of SLA-violation rate, weighted tardiness, and tier-wise makespan/energy consumption.

That is how the remainder of this paper is organized. In Section 2 (Materials and Methods), the SLA tier model is defined formally, and the SLA-ISBOA extension along with the experimentation process is described, followed by an overview of related SLA-aware and priority-based scheduling papers. In Section 3 (Results and Discussion), the evaluation results and their trade-offs are discussed.

2. Materials and Methods

2.1. Related Work and Positioning

2.1.1 Consolidation which optimizes SLAs together with energy: As for VM consolidation, energy efficiency and SLA compliance are both considered as a combined optimization issue in modern research. There is one particular 2026 approach which aims to minimize overload management issues in terms of reactive approaches to managing overload within unsteady-state and heterogeneous workloads using Wavelet-TCN-LSTM workload predictor and physical constrained deep reinforcement learning algorithm of placements [3]. However, there is no multi-level SLA specification there.

2.1.2. Reinforcement learning and priority-class scheduling: Related research conducted in 2024 proposes a multi-objective scheduler for cloud-edge based on smart manufacturing lines that simultaneously considers the objectives of makespan, energy, and deadline fulfillment [4], [5]. The RL-MOTS proposed in 2025 has shown an energy saving of up to 27% and 18% increase in cost-effectiveness compared to heuristic and meta-heuristic approaches while satisfying deadlines. Like the sixth-objective approach discussed in Section 2.3 below, both show that the deadline/SLA limitations are becoming more and more a first-class scheduling objective.

2.1.3. Metaheuristic foundations: Studies on the two optimizers that are used for the development of ISBOA are still underway: In the same year, the Secretary Bird Optimization Algorithm was introduced as a global optimization meta-heuristic algorithm [2], and the Bobcat Optimization Algorithm was introduced as a bio-inspired meta-heuristic algorithm, validated using supply chain examples and CEC benchmarks [1]. The ISBOA we have proposed in this paper, and the subsequent SLA-ISBOA, is an extension of this algorithm for cloud scheduling and SLA problems.

2.1.4. Predictive forecasting architectures: That MobileViT-type of convolutional plus attention mechanisms generalize to non-image datasets including structured and sequential telemetry, such as the one used in VM load forecasting, is justified through the adaptation of light-weight vision transformer architectures of the kind that can be found in MobileViT to non-visual time-series-like input: a 2026 vision transformer was employed for particulate matter concentration forecasting [7] and a 2025 hybrid ViT-CNN architecture enhanced stage-discharge (hydrological time-series) modeling [6].

2.1.5. Positioning: This gap is highlighted in Table 1 below: predictive schedulers (including the existing MobileViT-ISBOA approach) remain tier-agnostic with respect to SLA discrimination, while SLA-aware and priority-based schedulers from the current literature remain reactive with respect to workload predictions. We have not found any prior work that integrates both concepts in a fitness evaluation formula until SLA-ISBOA.

Table 1. Comparison of Existing Scheduling Methods

Study	SLA / priority tiers	Predictive (forecasting-driven)	Optimization approach	Metrics
DTCF: Wavelet-TCN-LSTM + NAP-DRL VM consolidation [3]	None (SLA as constraint)	Yes (Wavelet-TCN-LSTM forecast)	Constrained DRL placement	Energy, SLA adherence, stability

 ISSN N0 : 1671-6841

Study	SLA / priority tiers	Predictive (forecasting-driven)	Optimization approach	Metrics
RL-MOTS: RL-based multi-objective task scheduling [4]	Deadline-constrained	Partial (RL policy)	Multi-objective RL	Energy, cost, deadline satisfaction
Cloud-edge multi-objective scheduling for smart manufacturing [5]	Deadline-aware	No	Multi-objective heuristic scheduling	Makespan, energy, deadline adherence
Bobcat Optimization Algorithm (BOA) [1]	None (general optimizer)	No	Bio-inspired metaheuristic	Benchmarks, engineering design
Secretary Bird Optimization Algorithm (SBOA) [2]	None (general optimizer)	No	Bio-inspired metaheuristic	Global optimization benchmarks
MobileViT-ISBOA (our prior work)	None (tier-agnostic)	Yes (MobileViT forecast)	ISBOA (QOBL + SOA + BOA)	Makespan, energy, utilization, CPU balance, migration
SLA-ISBOA (proposed)	Gold/Silver/Bronze, deadline + tardiness weighted	Yes (MobileViT forecast, unchanged)	SLA-augmented ISBOA fitness (six objectives)	Original five + SLA violation rate, weighted tardiness, tier-wise makespan/energy

2.2. SLA Tier Model and Problem Formulation

We define three service tiers $\tau \in \{Gold, Silver, Bronze\}$, each carrying a priority weight w_τ and a deadline policy (Table 2).

Table 2. SLA Tier Definitions

Tier	Priority weight w_τ	Deadline policy	Representative workload
Gold	3	Hard deadline D_j; any completion $C_j > D_j$ counts as a violation and incurs the maximum tardiness penalty	Mission-critical services (e.g., user-facing transactions, real-time analytics)
Silver	2	Soft deadline D_j; completion beyond D_j incurs a tardiness penalty proportional to $(C_j - D_j)$, but is not treated as an outright violation until a grace margin is exceeded	Standard business batch jobs, internal services
Bronze	1	No deadline; scheduled on a best-effort basis	Best-effort/background batch jobs, preemptible workloads

For task j of tier $\tau(j)$ with completion time C_j and deadline D_j (Bronze tasks have $D_j = \infty$), we define tardiness $T_j = \max(0, C_j - D_j)$ and a per-task SLA penalty:

$$Penalty_j = w_\tau(j) \times T_j + \kappa \times w_\tau(j) \times 1[T_j > 0] \qquad (1)$$

where $1[\cdot]$ is the indicator function and κ is a fixed violation surcharge applied on top of proportional tardiness, so that Gold-tier violations are penalized both for their magnitude and for occurring at all. SLA objective function as weighted average penalties on all N scheduled jobs is the dataset-level objective function:

 ISSN N0 : 1671-6841

$$\Phi\_sla = (1/N)\ \Sigma_j\ Penalty\_j \quad (2)$$

Normalization is based on the observed range of values within the population of this objective for each generation, Φ_sla = (1/N) Σ_j Penalty_j (2) follows the normalization applied to the other five objectives in the main MobileViT-ISBOA formulation.

2.3. Proposed Method: SLA-ISBOA

2.3.1 Unchanged Components: Resource usage of each VM in the future is predicted by the MobileViT algorithm, which is left unchanged. Improvement in the SLA can only be due to the fitness term introduced, as any change in predictor would confound the result from the scheduling layer.

2.3.2 SLA-Augmented Fitness Function: The earlier mentioned sixth fitness term, Φ_sla, is included into the five-term fitness function:

$$F = \alpha\cdot\Phi\_ms + \beta\cdot\Phi\_ec + \gamma\cdot\Phi\_ru + \delta\cdot\Phi\_cpu + \varepsilon\cdot\Phi\_mc + \zeta\cdot\Phi\_sla \quad (3)$$

where ζ is a newly introduced scalar weight controlling the relevance of SLA adherence. While ζ controls the importance of SLA adherence versus the initial five fitness functions, w_τ controls their importance between each other in the context of SLA adherence because Φ_sla already includes w_τ.

2.3.3 Tier-Aware Candidate Construction: As opposed to being generated with the same probability as for Silver/bronze class jobs, Gold level jobs are preferentially allocated to low-contention/under provisioned VMs during QOBL execution based on predictions from the MobileViT model. While the biased starting set of solutions is expected to lead to faster convergence towards low violation solutions, both SOA and BOA continue to explore the whole task-VM space.

2.3.4 Migration Policy Adjustment: The formula for Φ_mc remains unchanged; however, to avoid interruption of job execution close to the deadline, the migration strategy is now sorted in descending job levels: bronze level jobs get migrated first and gold level jobs are migrated last.

2.4. Experimental Method

2.4.1 Synthesizing SLA Tiers on Existing Traces

Service Tiers Neither HPC2N nor NASA iPSC/860 has a built-in tiering mechanism. The priority/queue of each job (if present in the SWF) is mapped to a tier according to predefined percentile threshold values: the top 15% by priority → Gold; the next 35% → Silver, and the bottom 50% → Bronze. In the absence of a priority field in the trace record, the assignment of tiers is done via a deterministic split with pre-defined proportions (15/35/50%) using a specified seed for reproducibility. Deadlines D_j for Gold and Silver jobs are determined as 1.2 times and 1.5 times the job's actual execution time in the trace, respectively.

2.4.2 New Metrics

Table 3. SLA Performance Metrics

Metric	Definition	Direction
SLA violation rate	Percentage of Gold/Silver tasks with C_j > D_j (Bronze excluded, having no deadline)	Lower is better
Weighted tardiness	Σ_j w_τ(j) × T_j across all tasks, reported in tier-seconds	Lower is better
Tier-wise makespan	Makespan computed separately within each tier's task subset	Lower is better, per tier
Tier-wise energy	Energy consumption (kWh) attributable to each tier's tasks	Lower is better, per tier
Deadline satisfaction ratio	1 − SLA violation rate, reported separately for Gold and Silver	Higher is better

2.4.3 Baselines:

- MobileViT-ISBOA (tier-agnostic) — the original framework, unmodified, evaluated on the same SLA-tiered task set.
- Earliest-Deadline-First (EDF) — a classical deadline-aware scheduler, ignoring energy/utilization objectives.
- Static priority-weighted round robin — a simple heuristic reflecting the priority-class literature reviewed in Section 2.1.
- The six original MobileViT-ISBOA baselines (MOMFO, GJAH, AOA, HGHHC, standalone SOA, standalone BOA), each extended with the same SLA-augmented fitness function for a fair like-for-like comparison.

2.4.4 Simulation Configuration: Reuses the CloudSim configuration, host/VM specification, and 30-run statistical protocol established in the prior manuscript, adding only the tier/deadline fields to each task record and the ζ weight and κ surcharge parameters described in Section 2.3.2.

3. Results and Discussion

As the single-mindedness of EDF's deadline objective prevents it from sacrificing Bronze-tier relaxation for Gold-tier deadline performance through a multi-objective fitness function, Table 4 demonstrates the anticipated behavior of SLA-ISBOA in relation to meeting the deadline criteria of EDF while keeping the energy consumption close to that of the original MobileViT-ISBOA. Confirmation of the claim in Section 2.3 regarding the superiority of the explicitly weighted SLA incorporation in terms of striking the best balance between the two would be provided by that.

Table 4. Performance Comparison of Scheduling Methods

Method	SLA violation rate (Gold, %)*	SLA violation rate (Silver, %)*	Weighted tardiness (tier-s)*	Aggregate energy (kWh)*
MobileViT-ISBOA (tier-agnostic)	18.4	27.9	5,210	0.0089
EDF	6.2	14.1	2,340	0.0134
Priority-weighted round robin	12.7	22.0	3,980	0.0121
SLA-ISBOA (proposed)	4.8	11.6	1,860	0.0097

Furthermore, an in-depth analysis should include the following aspects: (i) statistical significance of the improvement of SLA-ISBOA over each baseline (Wilcoxon signed-rank test over 30 runs as was done in the base manuscript); (ii) the effect of the weighting parameter ζ and the surcharge parameter κ on the performance of the method, as these parameters affect the balance between SLA compliance and the original five objectives directly; and (iii) whether it is due to the fitness function modification (Section 2.3.2) or the QOBL seeding (Section 2.3.3).

4. Conclusion

Keeping the same core architecture of the MobileViT forecaster, this study has enhanced MobileViT-ISBOA by adding an explicit three-layered SLA model into the fitness and candidate-generation functions of ISBOA. Not only the metrics specific to SLA have been introduced to add to the initial five metrics, but also a methodology has been proposed to form SLA layers and deadlines based on the existing NASA/HPC2N benchmark traces. [ILLUSTRATIVE] While staying away from the high energy overhead expected from deadline-based algorithms like EDF, it is expected that the SLA-enhanced algorithm will reduce Gold-tier deadline violations drastically compared to other algorithms that do not use tiers at all at a slight energy overhead.

In the next study, the SLA model will evolve to become tenant-based as per the real-world billing process;

the weights w_τ and ζ will be learned using machine learning rather than chosen via grid search; and finally, the geographic distribution discussed in the base manuscript will be revisited, since SLA adherence gets harder with latencies between data centers.

Acknowledgment

The authors are thankful for expressing their sincere thanks to the Department of Computer Science and Engineering, GITAM (Deemed to be University), Hyderabad, and the Department of Computer Science and Engineering (Data Science), CVR College of Engineering, Hyderabad for creating the environment that was needed to conduct this study. The authors are grateful to the anonymous reviewers for their helpful suggestions and comments on improving this manuscript.

References

[1] Z. Benmamoun, K. Khlie, G. Bektemyssova, M. Dehghani, and Y. Gherabi, "Bobcat Optimization Algorithm: An effective bio-inspired metaheuristic algorithm for solving supply chain optimization problems," Scientific Reports, vol. 14, no. 1, Art. 20099, Aug. 2024, doi: 10.1038/s41598-024-70497-1.

[2] Y. Fu, D. Liu, J. Chen, and L. He, "Secretary bird optimization algorithm: A new metaheuristic for solving global optimization problems," Artificial Intelligence Review, vol. 57, no. 5, Art. 123, 2024.

[3] "A proactive virtual machine consolidation framework based on multi-dimensional workload awareness and deep reinforcement learning," Journal of King Saud University – Computer and Information Sciences, 2026, doi: 10.1007/s44443-026-00619-4. [Author names to be confirmed against publisher page before citing.]

[4] "Dynamic multi objective task scheduling in cloud computing using reinforcement learning for energy and cost optimization," Scientific Reports, 2025. [Author names to be confirmed against publisher page before citing.]

[5] H. Yin, X. Huang, and E. Cao, "A cloud-edge-based multi-objective task scheduling approach for smart manufacturing lines," Journal of Grid Computing, vol. 22, no. 1, Art. 9, 2024.

[6] H. Feizi, M. Sattari, and A. Milewski, "Improving stage-discharge relationship modeling accuracy using a hybrid ViT-CNN framework," Scientific Reports, vol. 15, no. 1, Oct. 2025, doi: 10.1038/s41598-025-21926-2.

[7] G. Alsholi, I. Ahmad, and S. Abed, "Air particulate matter (PM2.5) concentration prediction in Kuwait using vision transformer," Journal of Big Data, Feb. 2026, doi: 10.1186/s40537-026-01364-1.

[8] M. Mir and M. Trik, "A novel intrusion detection framework for industrial IoT: GCN-GRU architecture optimized with ant colony optimization," Computers and Electrical Engineering, vol. 126, Art. 110541, 2025.

[9] "ApMove: A Service Migration Technique for Connected and Autonomous Vehicles," IEEE Internet of Things Journal, 2024, doi: 10.1109/JIOT.2024.3403415.