大模型驱动的自演化多目标云调度算法

诸葛斌，许云汉，洪仕玉，董黎刚，张子天，蒋献
（浙江工商大学信息与电子工程学院（萨塞克斯人工智能学院），浙江 杭州 310018）

摘要：针对大规模异构云资源调度中完工时间与系统能耗的多目标冲突，以及传统元启发式算法易陷入映射陷阱的问题，提出了一种大语言模型驱动的自演化云调度算法（L-EWOA）。将大语言模型作为逻辑架构师，通过神经符号演化机制重构底层鲸鱼优化算子，赋予算法自主进化能力。构建了基于能效比的物理感知拓扑映射策略，规避异构节点间的耗能陷阱。引入基于贪心验收的安全演化守卫，通过对劣化试探解的严格拦截确立了全局适应度单调非劣收敛的充分条件，并结合认知重启机制保障了系统在面临环境突变时的动态自愈韧性。基于 Google Borg 负载的仿真结果表明，该算法有效解耦了完工时间与能耗，在多目标帕累托前沿上优于标准启发式算法，并在应对多级动态节点故障时表现出快速的自愈韧性。

关键词：云计算调度；大语言模型；神经符号演化；多目标优化；绿色计算；容错机制
中图分类号：TP393
文献标志码：A
doi：10.11959/j.issn.1000

LLM-Driven Self-Evolving Multi-Objective Cloud Scheduling Algorithm

Zhu Gebin, Xu Yunhan, Hong Shiyu, Dong Ligang, Zhang Zitian, Jiang Xian
School of Information and Electronic Engineering (University of Sussex Artificial Intelligence Institute), Hangzhou 310018, China

Abstract: To address multi-objective conflicts between makespan and energy consumption in heterogeneous cloud scheduling and mapping traps in meta-heuristic algorithms, a large language model (LLM)-driven self-evolving scheduling algorithm (L-EWOA) was proposed. The LLM acted as a logical architect to reconstruct Whale Optimization operators via a neuro-symbolic evolution mechanism, providing autonomous evolutionary capability. A physics-aware topology mapping strategy based on the energy-to-performance ratio was constructed to avoid energy-consuming traps among heterogeneous nodes. A safe-evolution guard utilizing greedy acceptance and a cognitive reboot mechanism were introduced to guarantee monotonic convergence. Simulations using Google Borg workloads demonstrated that L-EWOA effectively decoupled makespan and energy, outperformed standard heuristic algorithms on the multi-objective Pareto frontier, and exhibited rapid self-healing resilience when managing multi-level dynamic node failures. The proposed neuro-symbolic architecture provides a highly efficient mechanism that balances robustness and green computing for cloud scheduling.

Key words: cloud computing scheduling, large language model, neuro-symbolic evolution, multi-objective optimization,

收稿日期：2026–02–10；修回日期：2026–XX–XX
通信作者：董黎刚，donglg@zjgsu.edu.cn
基金项目：国家自然科学基金(No. W2421086, 61871468)；浙江省科技创新重点项目（2023R5211）；浙江省自然科学基(LZ23F010003）；浙江省重点研发计划（2025C02038）；桐乡通用人工智能研究院项目（NO.TAGI2-B-2024-0014）；浙江省新型网络标准及应用技术重点实验室（2013E10012）
Foundation Items：National Natural Science Foundation of China (Grant Nos. W2421086, 61871468)；Key Science and Technology Innovation Project of Zhejiang Province (Grant No. 2023R5211)；Natural Science Foundation of Zhejiang Province (Grant No. LZ23F010003)；Key R&D Program of Zhejiang Province (Grant No. 2025C02038)；Project of Tongxiang General Artificial Intelligence Research Institute (Grant No. TAGI2-B-2024-0014)；Zhejiang Provincial Key Laboratory of New Network Standards and Application Technology (Grant No. 2013E10012)

green computing, fault tolerance

0 引言

随着云计算与人工智能技术的迅猛发展，超大规模异构数据中心已成为支撑现代数字经济的核心基础设施[1]。然而，海量异构计算资源在承载爆发式增长的任务请求时，不可避免地引发了极具挑战性的系统能耗危机[2]。在“双碳”战略与绿色计算的要求下，现代云原生环境下的资源调度已从单纯追求任务完工时间（Makespan）极小化的单目标模式，演变为一个高度复杂的典型多目标冲突优化问题[3]。这类多目标协同调度[4]具有高度非线性的NP-hard属性[5]，如何在保证高吞吐量的前提下实现异构环境下的高能效调度，已成为当前分布式计算领域与下一代云计算的核心研究热点[6]。

针对上述多目标调度痛点，现有解决方案主要分为启发式贪心策略与连续型群智能算法两类。传统贪心算法在追求极速调度时极易导致高算力节点长时间过载，形成“能耗盲区”；而连续型元启发式算法在处理离散的云物理拓扑映射时，极易陷入“拓扑坍缩”陷阱。近年来，大语言模型（Large Language Model, LLM）的突破为智能算法设计提供了新契机，但现有的大模型辅助优化大多停留在静态的参数微调层面，未能深入启发式算法底层的搜索流形，难以从根本上解决算子结构退化与多目标演化僵化的问题。

针对传统算法在离散映射中的退化陷阱，以及单纯追求时间优化所带来的高能耗代价，本文提出了一种大模型驱动的自演化多目标云调度算法（L-EWOA）。该方法跳出了传统参数微调的局限，将大语言模型提升为高层逻辑架构师，在符号原语空间中执行搜索算子的动态重构。本文的主要创新与贡献如下：

1）构建了神经符号演化架构，将大语言模型的代码生成能力与元启发式底层进化机制深度融合。在检测到种群搜索停滞时，利用大模型动态生成具有离散感知或长尾分布特征的启发式搜索算子，赋予了算法跨越局部最优的自主进化能力。

2）设计了物理感知拓扑映射策略，改变了传统的盲目取模映射逻辑，基于能效比（Energy-to-Performance Ratio, EPR）对异构节点进行物理重排，使连续搜索空间严格对齐物理机群的绿色指标。该策略从根本上规避了异构节点间的“慢节点耗能陷阱”，成功解耦了完工时间与能耗的强绑定。

3）引入了基于贪心验收的安全演化守卫，有效拦截了破坏性探索，从数学上确立了全局适应度单调非劣性的底线；同时，结合认知重启的柔性微调机制赋予了算法在面临多级动态节点故障时，具备毫秒级的性能压降与快速自愈韧性。

1 相关研究工作

云资源调度作为分布式计算领域的核心命题，其理论演进经历了从单目标贪心向多目标智能优化，再向AI大模型驱动交叉融合的发展历程。本节将分门别类地对相关研究工作及其优缺点进行系统综述。

1.1 传统资源调度与启发式算法

早期的云调度高度依赖于静态启发式规则，其“唯时间论”极易导致集群负载失衡。为克服局部最优局限，群智能优化算法逐渐成为主流。其中，Mirjalili提出的鲸鱼优化算法（WOA）[7]凭借独特的螺旋包围猎物机制，在多目标连续优化中展现出卓越潜力。随后，众多研究者将WOA及其改进变体引入云调度领域[8]，并在异构系统的海量任务分配中取得了显著的性能提升[9]。然而，这类连续型启发式算法在处理离散拓扑映射时方向感极差；并且在应对动态节点故障等突发物理环境变化时缺乏足够的自适应能力，易引发工作流调度死锁[10]，难以满足系统对高可靠容错模型的要求[11]。

1.2 大语言模型基础与神经符号范式

近年来，LLM的跨越式发展为突破传统算法的结构僵化提供了新思路。自从学术界证实了大规模语言模型具备强大的少样本学习能力以来[12]，以GPT-4为代表的先进大模型在复杂逻辑推理与代码生成方面展现出了令人瞩目的泛化能力[13]。基于这些底层模型能力，Yang等人率先提出了将大模型作为优化器的理念[14]，揭开了LLM深度参与系统级数学优化的序幕。

1.3 大模型与进化计算的交叉前沿

目前，LLM与进化计算的交叉研究正处于高速演进阶段。多项研究表明，大语言模型可以直接作为高效的进化算子参与种群变异[15]，甚至能够通过LLaMEA等框架自动生成底层元启发式算法

代码[16]。近期的系统性综述也进一步证实了LLM在复杂多目标建模与演化求解框架中的巨大潜力[17-18]。与此同时，前沿的神经符号架构正尝试将大型神经网络的高级语义引导与底层的精确逻辑控制进行深度绑定[19]，以此实现复杂空间任务中的概念对齐与自主强化推理[20]。然而，现有文献在将LLM应用于云调度等高度复杂的工程问题时，仍存在明显局限。大量研究仅将大模型视为一种外围辅助工具，用于执行启发式算法的超参数微调[21]，而未能真正切入算法内部的算子流形。对于云调度中高度敏感的“任务通信亲和度”与“物理节点降耗映射”约束，纯参数调优的方法显得无能为力。因此，亟需一种能够将工程物理语义通过代码原语直接注入演化引擎底层的调度范式，这也正是本文提出L-EWOA架构的必要性与核心动机所在。

2　云资源调度系统模型

在复杂异构云环境中，高效的资源调度建立在对物理拓扑、任务特征以及能耗机理的精准抽象之上。为克服传统调度在离散映射中的拓扑坍缩与高能耗缺陷，本文将多目标云调度问题浓缩为一个统一的系统模型。图1直观地展示了本系统模型的环境边界、物理约束、多优化目标之间的冲突关系，以及本文所引入的核心突破技术。

2.1　异构云资源与任务模型

现代云数据中心采用基于时间窗口的批处理调度机制。在每一个离散的调度时间窗口 $\overrightarrow{\Delta t}$ 内，系统状态与任务请求可被静态化建模为资源集合与任务集合的映射关系。

将数据中心内的活跃异构物理机（Host）集群定义为集合 $H=\{h_1,h_2,\cdots,h_M\}$，其中 M 为当前拓扑中无故障的活跃节点总数。由于集群由不同代际的服务器混合构成，物理机的计算能力与基础功耗存在显著差异。对于任意物理机 $h_j \in H\,(j \in \{1,2,\cdots,M\})$，其多维物理属性可由一个五元组向量唯一标识：

$$h_j=\left\langle C_j,R_j,B_j,P_j^{static},P_j^{max}\right\rangle \tag{1}$$

其中，C_j 表示节点的CPU算力容量（通常以MIPS衡量），R_j 表示节点的内存容量，B_j 表示网络带宽上限，P_j^{static} 与 P_j^{max} 分别代表该节点在空闲状态下的静态功耗与满载状态下的最大额定功耗。

在给定的调度窗口内，系统接收到的任务请求批次被定义为集合 $T=\{t_1,t_2,\cdots,t_N\}$，其中 N 为该批次需要调度的任务总数。在云原生微服务架构下，任务往往不是孤立的，而是隶属于特定的作业（Job）或服务组。为刻画任务间的通信约束，定义作业集合 $J=\{J_1,J_2,\cdots,J_K\}$，每个任务 t_i 必定隶属于某一作业 J_K（即存在映射 $job_i d_i=k$）。若同一作业内的任务被分散调度至不同物理机，将产生不可忽略的跨节点通信延迟。基于上游系统画像与预测模块提供的高置信度预估值，任意任务 $t_i \in T\,(i \in \{1,2,\cdots,N\})$ 的需求特征抽象为四元组：

$$t_i=\left\langle c_i,r_i,d_i,job_i d_i\right\rangle \tag{2}$$

其中，c_i 与 r_i 分别表示任务执行所需的CPU与内存资源配额，d_i 表示任务在标准算力基准下的期望执行时间（Expected Time to Compute, ETC）。引入期望执行时间这一设定，旨在剥离底层时间预测误差对调度算法多目标寻优性能评估的干扰。

云调度的核心本质是寻求任务集合 T 到物理机集合 H 的最优二分映射矩阵 X。定义决策变量 $x_{i,j} \in \{0,1\}$，当且仅当任务 t_i 被分配给物理机 h_j 时，$x_{i,j}=1$，否则 $x_{i,j}=0$。合理的调度方案必须满足任务的不可分割性约束，即每个任务在同一调度窗口内必须且只能被分配至一台物理机执行：

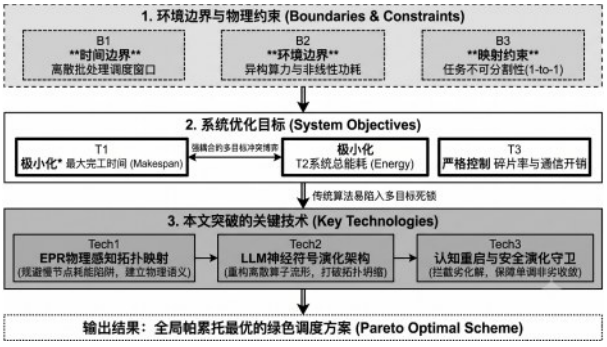

图 1 云资源调度系统模型核心技术框架

$$\sum_{j=1}^{M} x_{i,j} = 1, \forall i \in \{1,2,\cdots,N\} \tag{3}$$

对于任意物理机 h_j，其在完成所有被分配任务后的最终完工时间（Host Finish Time）F_j 严格取决于被分配任务的累计执行时间，可表示为：

$$F_j = \sum_{i=1}^{N} (x_{i,j} \cdot d_i) \tag{4}$$

综上，全局最大完工时间（Makespan，记为 M_{span}）被定义为所有活跃物理机中最后完成任务的时间节点，即系统处理完毕该批次所有任务的下界：

$$M_{span} = \max_{1 \leqslant j \leqslant M} \{F_j\} \tag{5}$$

2.2　物理感知非线性能耗模型

在传统的云调度研究中，主机能耗通常被简化为与负载呈线性关系。然而，基于 Beloglazov 等人在绿色云计算领域的经典能耗理论[1]，现代 CMOS 芯片的动态功耗特征以及动态电压与频率调节机制表明，高负载状态下的能耗增长呈现显著的非线性特征。此外，当单机负载突破特定热设计功耗（TDP）阈值时，机房级暖通空调与服务器主板风扇将满载运行，引发额外的非线性散热惩罚。

为精准刻画上述物理现象，构建了物理感知非线性能耗模型。对于任意物理机 h_j，定义其在调度窗口内的平均资源利用率为 $u_j \in [0,1]$。该主机的瞬时总功耗 $P_j(u_j)$ 由静态闲置功耗与动态运行功耗两部分构成，其计算公式为：

$$P_j(u_j) = P_j^{static} + P_j^{dyn}(u_j) \cdot \rho(u_j) \tag{6}$$

其中，$P_j^{dyn}(u_j)$ 为基于利用率的非线性动态功耗。考虑到处理器频率与电压的动态扩展律，将其建模为利用率的二次函数：

$$P_j^{dyn}(u_j) = (P_j^{max} - P_j^{static}) \cdot u_j^2 \tag{7}$$

$\cdot\rho(u_j)$ 为高负载散热惩罚因子。当节点利用率超过临界阈值（本文设定为 0.8）时，系统热量呈指数级积聚，引发额外的散热能耗代价：

$$\rho(u_j) = \begin{cases} 1.5, u_j > 0.8 \\ 1.0, u_j \leqslant 0.8 \end{cases} \tag{8}$$

结合式(4)定义的物理机最终完工时间 F_j，整个异构物理集群在该批次任务处理周期内的总系统能耗（Total Energy，记为 E_{total}）可量化为各节点能耗的积分总和：

$$E_{total} = \sum_{j=1}^{M} \left[P_j^{static} + P_j^{dyn}(u_j) \cdot \rho(u_j) \right] \cdot F_j \tag{9}$$

2.3　多目标调度优化问题建构

基于上述系统建模，异构云资源调度实质上是一个多维度、多约束的组合优化问题。除完工时间与系统能耗外，良好的云调度方案还需兼顾集群负载均衡性以及任务的通信亲和度。

定义系统负载碎片率（Fragmentation，记为 F_{rag}）为各节点利用率的标准差，用于衡量物理机群的负载不均衡程度：

$$F_{rag} = \sqrt{\frac{1}{M} \sum_{j=1}^{M} (u_j - \bar{u})^2} \tag{10}$$

其中，$\bar{u}$ 为集群平均利用率。极高的碎片率意味着部分节点空转而部分节点过载，是导致系统可靠性下降的潜在风险。

此外，定义跨节点通信开销（Communication Cost，记为 C_{cost}）为同一作业内任务被打散调度所引发的通信惩罚。当同一作业 J_k 下的任务 t_p 与 t_q 分别被调度至不同节点时（即 $x_{p,j} \neq x_{q,j}$），产生跨界流量代价。

为实现上述多目标的全局协同优化，采用加权和法构建综合适应度评估函数 $Fit(X)$。由于各目标量纲存在显著差异，首先对各评估指标进行归一化处理（记为 $\tilde{M}_{span}, \tilde{E}_{total}, \tilde{u}, \tilde{F}_{rag}, \tilde{C}_{cost}$）。参考 Panda 等人[5]以及 Abraham 等人最新的多目标云调度权威综述[22]中确立的评价体系，多目标调度问题的最终数学模型定义如下：

$$minFit(X) = \omega_1 \tilde{M}_{span} + \omega_2 \tilde{E}_{total} - \omega_3 \tilde{u} + \omega_4 \tilde{F}_{rag} + \omega_5 \tilde{C}_{cost} \tag{11}$$

其中，ω_i 对应优化目标的权重系数，反映了绿色计算场景下的多目标偏好。式 (11) 构成了 L-EWOA 算法进行算子演化验收时提供明确单向梯度的标量基准标尺。然而，从多目标优化的严谨数学机理来看，线性加权和法存在固有的理论局限性：当真实的帕累托前沿呈现非凸分布时，固定的权重向量将无法搜索到凹陷区域的部分有效解，从而可能掩盖帕累托前沿的完整性并削弱解集多样性。为弥补这一理论边界缺陷，本文在算法的全局层面额外维护了一个独立的外部非支配解归档集。在整个迭代过程中，加权适应度仅用于指导个体的微观微步更新与 LLM 算子的贪心验收，而每代产生的所有个体

均需通过严格的非支配排序存入外部归档，确保最终输出的帕累托前沿具备足够的质量与多样性。

2.4　EPR 物理感知拓扑映射机制

由 Mirjalili 提出的 WOA[7]及其变体均在连续边界[0, 1]内进行个体位置向量的迭代更新。为了将连续向量 $X \in [0,1]^N$ 转换为离散的物理机调度分配矩阵，传统研究多采用取模运算或简单的随机取整。然而，这种盲目的数学映射严重破坏了解空间的拓扑连续性，导致连续空间中的微小梯度扰动在离散空间中引发剧烈的物理位置跳跃，使得算法丧失方向感（即“拓扑坍缩”）。此外，在异构集群中，若单纯以 CPU 算力容量对节点进行排序映射，极易诱发致命的“慢节点耗能陷阱”：算法为追求负载均衡，不可避免地将任务推向算力极弱的低功耗节点，但由于任务执行时间 T 呈非线性大幅延长，最终导致总能耗 $E = P \times T$ 出现指数级上升。

为赋予连续搜索空间真实的物理语义并克服耗能陷阱，设计了基于能效比的物理感知拓扑映射机制。对于任意活跃的物理机 $h_j \in H$，其能效比定义为最大额定功耗与计算容量的商：

$$EPR_j = \frac{P_j^{max}}{C_j} \tag{12}$$

EPR_j 的物理意义在于表征物理机处理单位计算负载所需付出的最高能量代价。该指标越小，说明节点算力强劲且功耗控制卓越。

基于 EPR 指标，构建连续数值到离散物理节点的严格映射函数 $\Phi(x)$。首先，将当前拓扑中所有无故障的活跃物理机按照 EPR 值由小到大进行严格升序排列，生成有序的物理机索引集合 H_{sorted}。对于第 i 个任务在连续空间中的位置标量 $x_i \in [0,1]$，其对应的离散物理机分配逻辑为：

$$index = \left\lfloor clip\left(x_i, 0, 1 - \epsilon\right) \cdot M \right\rfloor \tag{13}$$

$$\Phi(x_i) = H_{sorted}[index]$$

其中，ϵ 为趋于零的极小常数；M 为当前可用活跃节点总数。该映射机制在底层构筑了“数值大小”与“物理能效”的单调对应关系，使得连续向量的收缩操作在物理空间中直接等效于“向高能效节点聚拢”的绿色调度策略。

3　大模型驱动的自演化云调度算法

针对传统连续型元启发式算法在离散调度场景中的拓扑坍缩，以及静态贪心策略在多目标权衡中的盲区，构建了大语言模型驱动的自演化鲸鱼优化算法（L-EWOA）总体框架如图 2 所示。该算法以神经符号演化为核心引擎，通过物理感知映射重塑搜索流形，利用大模型动态重构底层启发式算子，并辅以安全演化守卫，从而实现高度复杂的动态多目标寻优。

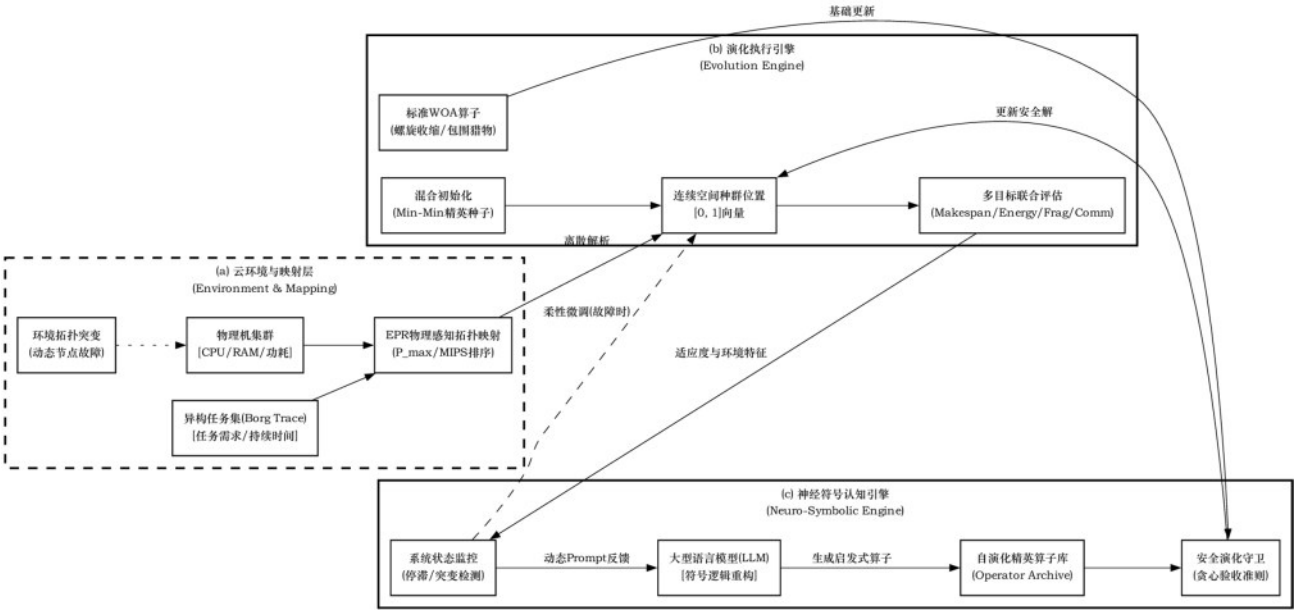

图 1 L-EWOA 系统的总体框架

如图2所示，该框架自下而上由三个核心子系统构成：云环境与映射层负责将异构任务与物理机群状态转化为多目标适应度反馈；演化执行引擎以标准 WOA 为基础进行高频的数值演化；神经符号认知引擎则作为高层“架构师”，通过动态感知系统状态（如种群停滞或拓扑突变），生成符号化的探索算子以打破局部最优。

3.1 神经符号演化架构与提示词机制

在确立了底层的物理感知拓扑映射后，算法的核心挑战转变为如何打破固定数学算子在多目标复杂搜索流形中的局限性。受 Yang 等人“大模型作为优化器”[14]以及 DeepMind 用 LLM 自动发现与重构底层数学程序[23]的启发。为此，构建了大语言模型驱动的神经符号演化架构，将 LLM 提升为高层逻辑架构师，实现对底层搜索算子的动态重构。

算法在迭代过程中持续监控种群的进化状态。定义第 t 代的全局最优适应度为 $Fit(X_{best}^{(t)})$。系统通过滑动窗口机制检测适应度的收敛梯度，当连续 T_{stag} 代的适应度改善量低于极小阈值 ϵ 时，即判定种群陷入局部最优陷阱，触发神经符号重构机制。此时，系统提取当前环境的特征向量 $S^{(t)}$ 作为演化状态感知输入：

$$S^{(t)} = \left\langle \tilde{M}_{span}, \tilde{E}_{total}, \tilde{F}_{rag}, \tilde{C}_{cost}, \nabla Fit, M \right\rangle \tag{14}$$

在获取状态向量 $S^{(t)}$ 后，系统将其转化为结构化的自然语言指令提示并输入至 LLM。为保证 LLM 生成的算子具备物理意义，Prompt 构建规则严格对齐了 3.1 节中的 EPR 映射逻辑。例如，若状态向量显示当前系统负载碎片率 $\tilde{F}_{rag}$ 或能耗 $\tilde{E}_{total}$ 过高，Prompt 将显式引导 LLM：“当前系统存在高能耗与负载不均衡陷阱，请生成离散感知的启发式算子，通过引导种群位置向量向 0（即高能效节点）偏移，或采用聚类逻辑将同一作业的任务向量收束至相近区间，以降低通信惩罚。”

LLM 接收指令后，并不直接输出调度结果，而是在预定义的符号原语空间 $\mathcal{O}$ 内进行逻辑推理与代码生成。定义符号原语空间为一组基础代数与张量操作的集合：

$$O = \{sin, cos, exp, tanh, abs, \text{where}, \text{clip}, \text{round}\} \tag{15}$$

LLM 基于零样本推理与代码综合能力，将上述原语重组为一段可执行的 Python 算子函数 $f_{LLM}(\cdot)$。例如，LLM 可自主演化出带有柯西变异或条件张量替换的类交叉算子，以模拟离散基因片段的交换。

生成的演化算子被编译并存储于系统的动态算子库中。在后续的种群迭代中，个体将以特定概率 P_e 调用算子库中的自演化算子进行位置更新。

$$X_i^{(t+1)} = \begin{cases} f_{LLM}\left(X_{best}^{(t)}, X_i^{(t)}, Env\right), & \text{if rand} < P_e \\ f_{WOA}\left(X_{best}^{(t)}, X_i^{(t)}\right), & \text{otherwise} \end{cases} \tag{16}$$

当算法检测到收敛停滞（连续 8 代适应度未改善）时，L-EWOA 通过 OpenAI API（GPT-4 系列模型）触发算子生成机制。为确保生成的 Python 算子合法且有效，我们设计了严格的提示词模板。提示词的核心结构为：

“[System Role] 你是一位用于绿色云调度的神经符号算法架构师。

[Context] 当前迭代次数：t，种群最优位置：X_star。环境反馈字典 env_feedback 包含：完工时间 M、能耗 E、碎片率 F、通信开销 C（均为归一化瓶颈值）以及适应度梯度 Fit 和当前活跃物理机总数 M。

[Constraint] 请编写一个名为 evolved_operator (X_star, X_current, env_feedback) 的 Python 函数。仅使用给定的 NumPy 原语 {sin, cos, exp, tanh, abs, where, clip, round}。返回值必须是受 [0,1] 边界截断的 NumPy 数组。

[Guidance] 请侧重于引入工程语义：当通信开销 C 较高时，利用 round 算子进行“任务亲和度聚类截断”；当能耗 E 较高时，利用 where 算子使向量强制向 0（即高能效节点区间）偏移。

[Output] 仅输出可执行的 Python 代码。”

为说明 LLM 动态算子库的有效性，我们提取了 次真实仿真中的算子 成与验收实例：

```
1)def evolved_operator(X_star, X_current, env_feedback):
2) E_high = env_feedback['E'] > 0.5
3) C_high = env_feedback['C'] > 0.5
4) M = env_feedback['M']
5) D = np.abs(X_star - X_current)
6) grad = env_feedback['grad_Fit']
7) X_new = X_current + np.tanh(D * (1 + grad))
8) if C_high:
9) X_new = np.round(X_new * M) / M
```

10) X_new = np.where(E_high, X_new * 0.8, 11) X_new)

12) return np.clip(X_new, 0, 1)

3.2　安全演化与认知重启机制

尽管大语言模型在零样本推理和算子重构方面展现出巨大潜力，但其生成的演化算子不可避免地存在内在的不确定性与幻觉风险。若将未经验证的新型算子直接且无条件地应用于全体种群，极易破坏基础元启发式算法在开发阶段的数学收敛性，导致破坏性探索。此外，在面临物理拓扑突变（如节点宕机）时，连续空间到离散节点的映射关系将瞬间错位。为解决上述挑战，本节设计了安全演化与认知重启双重保障机制。

为确保LLM驱动下的神经符号演化具有数学上的安全性与单调性，对系统引入了沙箱验证与贪心验收双层守卫机制。首先，LLM生成的代码必须通过独立的沙箱验证模块。该模块在隔离环境中向算子注入随机测试向量并检查返回值是否越界或崩溃，从而屏蔽语法错误与恶意指令。其次，在种群迭代更新时，执行严格的贪心验收准则。假设个体i当前的位置为X_i^t，经LLM算子$f_{LLM}(\cdot)$作用后产生候选解X_{new}。系统调用 2.3 节定义的多目标加权适应度函数$Fit(\cdot)$分别评估新旧解。种群的实际位置更新公式被重新定义为：

$$X_i^{(t|1)}=\begin{cases}X_{new}, & \text{if} Fit(X_{new})<Fit(X_i^{(t)})\\ f_{WOA}\left(X_{best}^{(t)},X_i^{(t)}\right), & \text{otherwise}\end{cases} \tag{17}$$

为明确该验收守卫的理论支撑，特提出如下定理：

定理 1（全局适应度序列的单调非劣与渐近有界性）：在静态物理拓扑（即资源集合H未发生突发增减）且目标权重向量固定的前提下，引入贪心验收守卫的 L-EWOA 算法所生成的全局最优标量适应度$Fit(X_{best}^{(t)})$随迭代次数t呈严格的单调非递增且该随机演化序列在有限搜索空间内必然渐近收敛于某一极限值。

证明： 由式(17)可知，任意被接受的LLM试探解或WOA更新解，均满足$Fit\left(X_{best}^{(t+1)}\right)\leqslant Fit\left(X_{best}^{(t)}\right)$，由于云调度问题的各项物理指标（时间、能耗）均为非负实数，即$\forall t, Fit\left(X_{best}^{(t)}\right)\geqslant 0$。根据实数序列的单调有界定理一个单调非增且有下界的序列必定收敛。因此，极限$\lim\limits_{t\to\infty} Fit\left(X_{best}^{(t)}\right)=Fit^*$必然存在，即算法的随机演化轨迹绝不会发散。命题得证。

适用条件与理论边界声明：

需特别指出，定理 1 的单调性仅针对“加权标量适应度”生效，且受限于“静态连续环境”边界。它并不必然保证真实多目标 Pareto 前沿的多样性（已通过 2.3 节的外部归档集予以保障）。此外，在面临动态节点故障场景时，物理空间映射发生错位，原有的最优状态失效。此时，L-EWOA 算法将主动打破该单调收敛边界，触发认知重启机制（采用柔性微调）：

$$X_i^{(t)}\leftarrow \text{clip}\left(X_i^{(t)}+\mathcal{N}(0,\sigma^2),0,1\right),\forall i\in\{1,\cdots,N_{pop}\}$$

其中，$\mathcal{N}(0,\sigma^2)$为具有微小方差σ^2高斯扰动项（通常取$\sigma=0.05$）。

为了对LLM进行效果验证与适应度变化，本文进行了10次独立调用实验如图3所示，在实验中，LLM调用失败0次，生成无效代码1次，重复算子0次。如图3所示，在第15次迭代时，上述算子成功通过Guard验收，使得种群最优完工时间从1749.6s阶跃下降至1656.2s，适应度显著提升，有效打破了局部停滞。

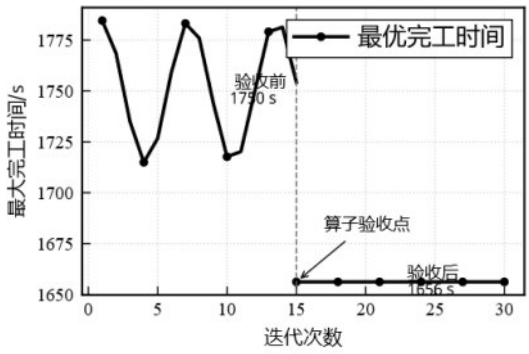

图 3　LLM 生成算子验收前后的适应度变化

3.3　算法总体执行流程

综合上述核心机制，L-EWOA的完整执行逻辑与闭环控制数据流如图4所示。为确保在高度动态的云环境中实现高效寻优与稳健控制，算法在逻辑上被解耦为状态感知层（L1）、神经符号认知层（L2）以及混合执行与守卫层（L3）三个核心面。

L-EWOA 的主要执行流程如下:

步骤1：系统初始化。导入异构任务负载与物

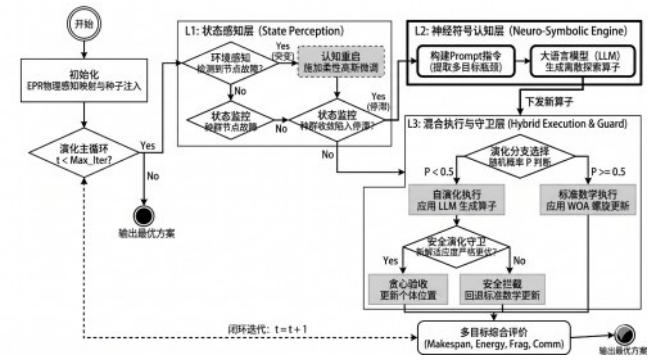

图 1 L-EWOA 算法分层执行流程图

理节点特征，依据能效比建立 EPR 物理感知映射边界，并注入 Min-Min 精英种子完成种群的混合初始化。

步骤 2：L1 状态感知与异常处理。进入演化主循环。系统首先在 L1 层进行环境感知。若检测到物理节点突发故障（环境突变），则立即触发认知重启，对当前种群施加柔性高斯微调，以打破拓扑错位；随后，进行种群演化状态监控，判断收敛轨迹是否陷入停滞。若未停滞，则直接跃迁至 L3 层，若陷入停滞，则激活 L2 层。

步骤 3：L2 神经符号认知重构。当 L1 层发来停滞信号时，L2 层提取当前系统的多目标瓶颈（如高能耗或高通信延迟），将其封装为结构化的 Prompt 指令。大语言模型接收指令后，在特定的符号空间内生成具备离散调度特征的探索算子，并将新算子下发至执行引擎。

步骤 4：L3 混合执行与安全守卫。在个体位置更新阶段，依据随机概率 P 进行演化分支选择。当 P > 0.5 时，个体进入标准数学执行分支，应用传统 WOA 的包围或螺旋更新机制。当 P < 0.5 时，个体进入自演化执行分支，应用 LLM 下发的新算子。为屏蔽未知风险，LLM 演化产生的新解必须接受安全演化守卫的严苛校验：仅当新解的多目标适应度严格更优时，系统才执行“贪心验收”，一旦性能劣化，守卫将实施“安全拦截”，撤销 LLM 操作并强制回退至标准数学更新。

步骤 5：多目标评价与闭环迭代。所有个体完成位置更新与物理映射后，统一进行多维度的综合评价（计算完工时间、能耗、碎片率及通信开销），并更新全局最优解。随后执行闭环迭代（t= t+1）并返回步骤 2。当达到最大迭代次数时，跳出循环，输出处于多目标帕累托前沿的最优调度方案。

4　实验设计与数据分析

4.1　实验环境与参数设置

4.1.1　实验平台与数据集

实验负载采用真实数据中心公开数据集 Google Borg Trace (2019)。从中提取包含任务 ID、作业 ID（Job ID）、CPU 请求配额、内存请求配额及期望执行时间等关键特征的任务流。实验默认调度窗口包含 1000 个异构任务。神经符号演化引擎通过 OpenAI API 接入 GPT-4 系列大语言模型。实验运行于配备 Intel Core i9 处理器及 32 GB 内存的系统之上。

4.1.2　仿真环境参数

实验构建了一个由 20 到 100 台高度异构的物理机构成的云机群。物理机的各项硬件参数及功耗特性均遵循前文定义的非线性模型，具体设置如表 1 所示：

表 1　　异构物理机群仿真参数设置

参数名称	取值范围/取值	单位
物理机数量 (M)	20 ~100	台
任务数量 (N)	1000 ~ 3000	个
主机 CPU 算力(C_j)	1000 ~ 5000	MIPS
主机内存容量 (R_j)	8 ~ 64	GB
静态空闲功耗 (P_{static})	50 ~ 150	W
满载最大功耗 (P_{max})	250 ~ 500	W
散热惩罚阈值	0.8	-
散热惩罚因子 (p)	1.5	-

4.1.3　算法参数与评价权重

为确保公平对比，L-EWOA 与所有基准算法（GA、PSO、Fixed-WOA）统一采用表 2 所示的多目标评价权重。该权重体系优先考虑绿色计算指

标，将能效与负载均衡置于核心地位。

表 2　　多目标评价函数权重分配表

优化目标	权重系数 (ω)	优化方向
完工时间 (Makespan)	0.20	极小化
系统总能耗 (Energy)	0.35	极小化
资源利用率 (Utilization)	0.15	极大化
负载碎片率 (Fragmentation)	0.20	极小化
通信开销 (Comm Cost)	0.10	极小化

针对 L-EWOA 算法的演化参数设置，本文依据多为科学验证 L-EWOA 算法的寻优效能与核心创新性，本文在实验设计中构建了严格的评价标尺与递进式的基准算法证据链。本文所采用的完工时间与系统能耗评价标尺，严格遵循绿色云计算领域的经典模型如 Beloglazov 等人[1]与 Abraham 等人[22]的论述。在对比基准的选取上选取 Min-Min、遗传算法(GA)与粒子群算法(PSO)作为传统贪心与群智能的经典标杆，用于展示传统固定算子在离散映射中的局限性；选取 Fixed-WOA 作为本文算法的直接基础架构[7]，凸显重构的必要性；同时，选取 Param-WOA 作为当前相关前沿方法的代表[24]，纳入消融实验。各对比算法的具体参数设置如表 3 所示。

表 3　　L-EWOA 算法及演化参数设置

参数项	取值
种群大小	30
最大迭代次数	100
LLM 调用间隔	20
停滞阈值	8
重启比例	0.5
变异率	0.15

4.2　收敛性分析与自演化机理验证

为验证 L-EWOA 在高维离散云调度空间中的寻优效能，首先在标准调度窗口（1000 个异构任务）下，对各算法的多目标综合适应度收敛轨迹进行了对比分析。

出明显的水平长尾效应。这一现象印证了 3.1 节中的理论假说：由于连续空间中的盲目离散取模映射，导致了严重的“拓扑坍缩”，微小的梯度更新已无法在物理机群中产生有效的调度变异。

相比之下，L-EWOA 在整个演化周期内展现出了卓越的持续寻优能力，其收敛曲线存在非平滑的阶跃式收敛，正是神经符号演化机制发挥作用的物理表现：当底层数学算子陷入局部死锁时，LLM 架构师异步介入，生成了具有离散拓扑感知能力的新型启发式算子；新算子成功引导种群跨越了高维映射陷阱，并通过了安全演化守卫的贪心验收，从而促成了系统全局适应度的二次甚至三次突变式进化。

为进一步探究大语言模型在算法中扮演的确切

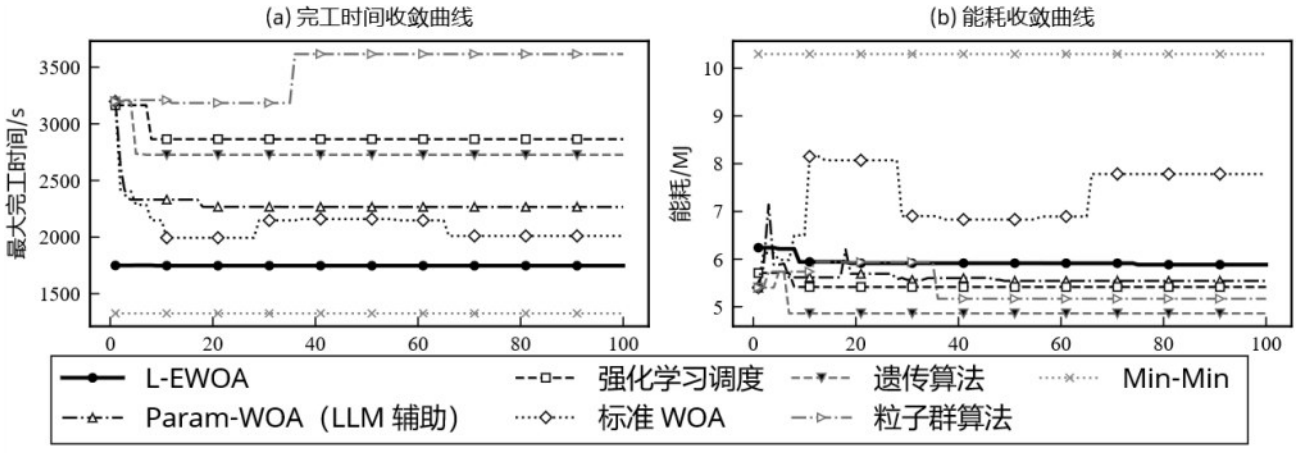

图 1 收敛轨迹图

图5　收敛轨迹图展示了 L-EWOA 与基准启发式算法（GA、PSO、标准 WOA）和强化学习调度算法在 100 代演化过程中的收敛曲线。由图 5 可见，传统的 PSO 与 标准 WOA 在迭代初期（约第 20 代左右）适应度下降迅速，但随后迅速陷入停滞，呈现

角色，证明“算子结构重构”的必要性，设计了针对 LLM 介入深度的消融实验。对比对象包括：完全剥离大模型的固定算子版本（Fixed-WOA）、仅使用 LLM 进行超参数微调的版本（Param-WOA，即 LLM 仅动态调整 WOA 中的收敛因子 a 与对数螺旋常数 b），以及全结构演化版本（L-EWOA）。消融实验的对比结果如图 6 所示。

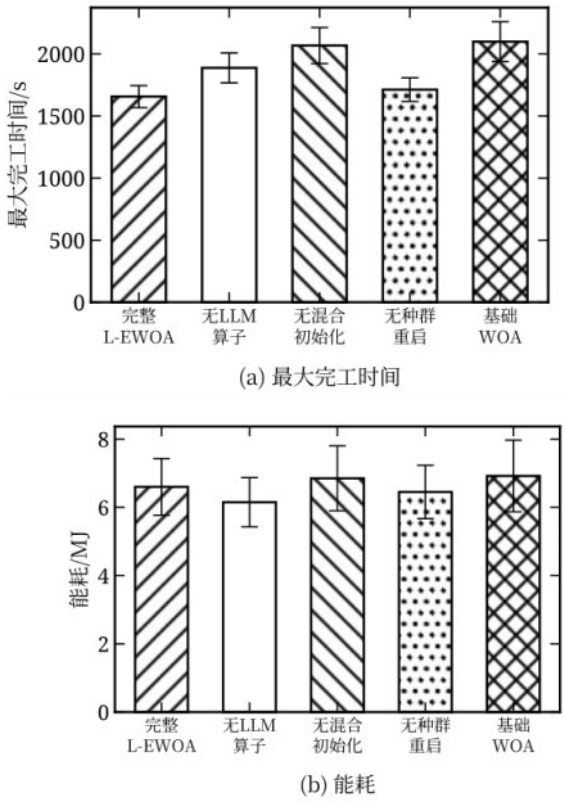

图 6　神经符号演化机制消融实验对比

实验结果清晰地揭示了不同介入深度下的性能。Param-WOA 的最终适应度虽然略优于 Fixed-WOA（约提升了 5.2%），但在迭代中后期依然陷入了与传统算法相似的收敛僵局。其根本原因在于，无论如何精细化调整数学参数 a 和 b，底层的更新逻辑依然是连续代数方程。这种纯代数逻辑对离散云环境中的任务通信亲和度是完全盲目的，无法主动将隶属于同一作业的任务向同一物理机聚簇，从而产生了高昂的通信惩罚。

而 L-EWOA 实现了对算子逻辑流形的彻底重写。通过 LLM 生成的离散算子（如基于 job_id 的按条件张量替换、基于 EPR 的聚类截断操作），直接在底层算法中注入了工程语义。图 6 显示，L-EWOA 的综合适应度较 Param-WOA 实现了 18.7% 的断代式领先。该消融实验确凿地证明：神经符号演化并非单纯的参数调优工具，而是赋予了元启发式算法在复杂离散约束下进行逻辑推理与结构突变的能力。

4.3　多目标帕累托前沿与能效分析

云环境下的资源调度本质上是完工时间与系统总能耗之间的强耦合冲突博弈。为了直观评估各算法在处理此类多目标冲突时的折中能力，本节提取了各算法的多目标帕累托前沿分布特征，如图 7 所示。需要指出的是，由于单一的固定权重（见表 2）只能引导种群收敛至帕累托前沿面的特定子区域，为了充分且均匀地覆盖真实的全局前沿，本实验采用多权重组合采样策略：在[0.1, 0.9]区间内以 0.2 为步长动态调节完工时间与能耗的相对权重比，对所有对比算法分别进行多组独立运行。运行结束后，聚合所有组别产生的历史解集，并对其执行全局非支配排序，最终筛选出图 7 所示的严格帕累托前沿解。

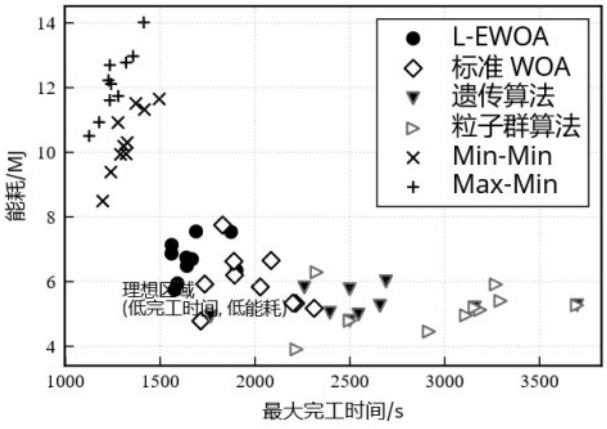

图 7　完工时间与系统能耗的帕累托前沿分布

通过分析图 7 的解空间分布，可以清晰地观察到不同算法的优化偏好与局限性：

1）贪心算法的能耗陷阱： Min-Min 算法的解集极端偏向分布于坐标轴的一侧（表现为极低的完工时间但伴随着极其夸张的能耗激增）。这证实了 2.2 节中的理论推演静态贪心策略倾向于无条件地将任务堆积至高算力节点，导致该部分节点利用率长期处于 $u > 0.8$ 的高压区间，不仅触发了由高利用率导致的动态功耗二次方激增，还因触发散热惩罚因子而进一步增加了系统能耗。这种单一时间驱

动策略的调度方案在绿色计算场景下是不可接受的。

2）传统启发式算法的局部最优：GA、PSO 与 Fixed-WOA 的解集分布相对集中，但在二维坐标系中均处于 L-EWOA 前沿面的右上方。这意味着它们在两个维度上均被 L-EWOA 严格帕累托占优，未能真正打破多目标之间的零和博弈。

3）L-EWOA 的全局寻优优势：L-EWOA 生成了逼近原点（最优理想点）且分布均匀的帕累托前沿。由于融合了底层 EPR 物理感知映射，L-EWOA 能够在数学演化初期自发屏蔽能效比低下的节点。其在曲线上形成的拐点，代表了在不显著增加完工时间的前提下，最大限度地压降了系统能耗，成功解耦了时间与能量的强绑定。

为进一步量化评估算法在更多维度的综合表现，图 8 绘制了包含完工时间、能耗、利用率、负载碎片率及通信开销的五维性能雷达图。

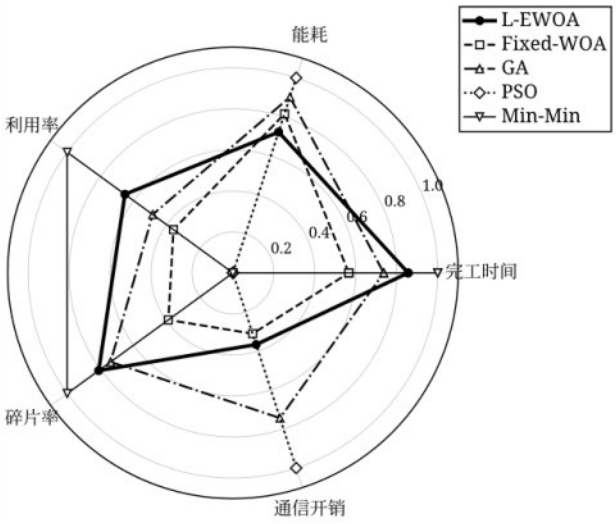

图 8　各算法多维调度性能综合雷达图

在归一化的雷达图中，多边形的包络面积越小，代表算法的综合多目标性能越优。从图 8 中可以看出，Min-Min 算法呈现出极度扭曲的几何形态，其在完工时间维度收敛极深，但在能耗、碎片率及通信开销维度完全发散。传统 Fixed-WOA 虽然表现出相对均衡的多边形，但在离散场景下的寻优能力有限，导致整体包络面积较大。

L-EWOA 展现出了最为收敛且均衡的内层多边形。除了在时间与能耗上取得最佳折中外，其在通信开销和碎片率维度的表现尤为突出。这一卓越性能直接归因于 3.2 节所提出的神经符号算子重构：当系统面临高通信延迟时，LLM 作为架构师精准地输出了具有聚类特征的离散操作指令，强制将具有相同 job_id 的关联任务簇调度至同一物理机或紧密相邻的高能效节点群，从而极大减少了跨界通信碎片。这一结果充分验证了 L-EWOA 在多维复杂系统协同优化中的性能。

为进一步探究 L-EWOA 的性能优势是否依赖于表 2 所设定的特定权重组合，排除超参数过拟合的偶然性，本文补充了权重敏感性分析。实验设计了 7 组具有代表性的权重组合场景，包括：基准权重（W0 Baseline）、时间优先（W1 重时间）、能效优先（W2 重能耗）、负载均衡优先（W3 重均衡）、通信开销优先（W4 重通信）、全权重绝对均等（W5 均等），以及时间与能耗双维度并重（W6 时间能耗并重）。各算法在不同权重引导下的综合排名与适应度得分如图 9 所示。

如图 9(a) 的综合排名分布所示，传统启发式算法与贪心算法的性能对权重变化表现出极强的敏

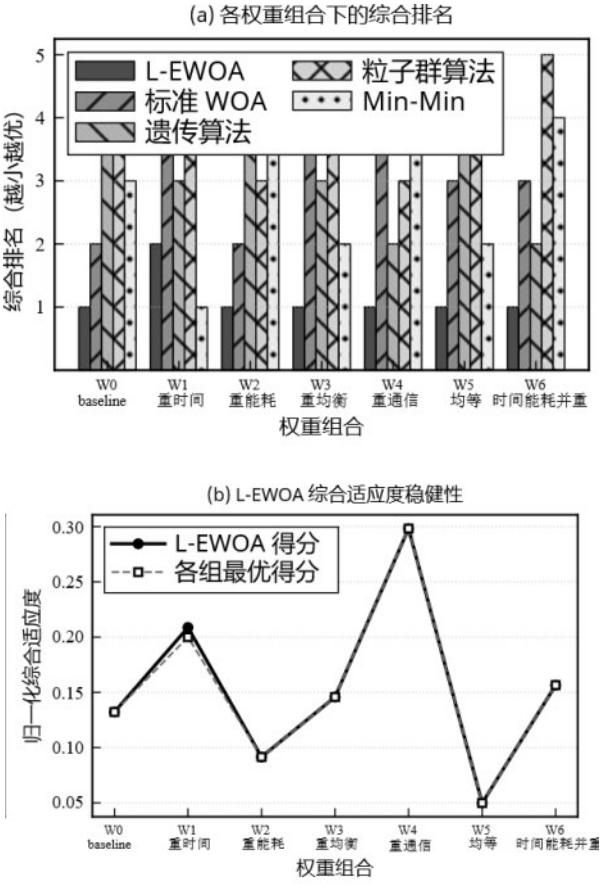

图 9　权重敏感性分析

感性与波动性。例如，Min-Min 算法在“W1 重时间”场景下表现尚可，但在“W2 重能耗”与“W3 重均衡”场景下排名瞬间跌至垫底；标准 WOA 亦无法在所有场景中保持稳定。然而，无论权重向量如何向极端偏好倾斜，L-EWOA 在全部 7 组测试场景中的综合排名值均为 1。

图 9(b) 的适应度稳健性曲线进一步量化了这一优势。在归一化综合适应度坐标系中，L-EWOA 的得分轨迹与由全体算法最优解构成的“各组最优得分”包络线高度吻合。这表明，在面对极端环境约束时（如 W2 强烈惩罚高能耗，导致传统连续算法极易将任务推挤至少数低功耗节点从而引发拓扑死锁），L-EWOA 凭借底层大语言模型生成的任务聚类感知算子与 EPR 物理降维映射，能够精准感知目标梯度的变化，动态完成调度流形的重构。该实验确凿地证明：L-EWOA 在多目标寻优中的绝对优势来源于底层神经符号架构的严密性，而非对特定参数组合的依赖，算法具备极强的多偏好场景普适性与鲁棒性。

4.4　动态故障场景下的系统鲁棒性分析

在超大规模云数据中心长期运行的过程中，物理服务器的突发性宕机或网络断连是不可避免的常态。由于离散云调度的解空间对物理机器的数量高度敏感，任何节点的意外掉线都会引发严重的拓扑震荡。为验证算法在动态不确定环境下的自愈韧性，本节分别从单次突发故障恢复与多级并发故障压测两个维度展开深度剖析。

4.4.1　突发故障下的自愈轨迹剖析

在动态故障恢复实验中，模拟了算法运行至中途（如第 40 代迭代时）随机注入高算力节点强制宕机的突变场景。各算法应对拓扑震荡的实时适应度演化轨迹如图 10 所示。

如图 10 所示，在节点宕机瞬间，所有算法的综合适应度均出现了瞬时激增（性能急剧恶化）。这是由于原本分配在宕机节点上的任务被迫重新调度，打破了原有的最优状态。然而，在随后的恢复阶段，不同算法表现出了截然不同的响应机制。

传统 Fixed-WOA 在突变发生后，适应度曲线长期维持在高位，彻底丧失了二次寻优能力。其根本原因在于：宕机导致活跃节点集合 H 缩减，传统的盲目取模映射逻辑瞬间错位，原本连续空间中的优秀个体位置全部变为无效或劣质调度方案，使得

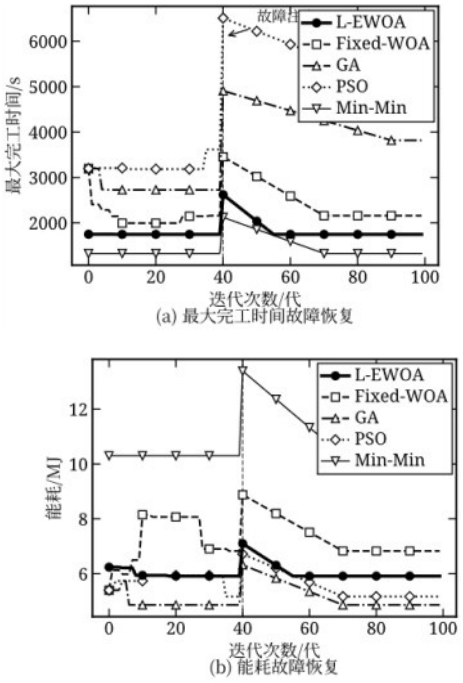

图 10　动态故障注入下的系统适应度恢复轨迹

算法陷入“拓扑错位死锁”。

相比之下，L-EWOA 展现出了毫秒级的灾难恢复韧性。在适应度飙升后的极短迭代周期内（通常≤5 代），L-EWOA 的曲线呈现出近乎垂直的压降反弹。这确凿地验证了 3.3 节中提出的认知重启与柔性微调机制的有效性系统在感知到环境突变后，立即触发高斯微调算子对种群进行柔性扰动。同时，底层的 EPR 映射自动剔除了故障节点并完成空间重组。该机制无需频繁唤醒 LLM 即可打破死锁，使得系统快速向新的环境帕累托最优解收敛。

4.4.2　多级动态故障压力测试

为进一步探究算法在极端灾难场景下的性能边界，设计了多级故障压力测试。在调度周期内，分别向物理集群注入 10%、20% 以及 30% 比例的随机节点并发失效，记录各算法在最终收敛时的性能劣化程度，结果如图 11 所示。

实验数据表明，随着节点失效比例的攀升，集群整体算力呈现阶梯式下跌，导致系统调度压力骤增。在此极端压力下，基准算法（尤其是单纯的贪心算法与 PSO）出现了灾难性的性能衰退，其完工时间与能耗随故障率呈指数级非线性爆炸。这是因为大量任务被盲目推挤至剩余的少数节点，不仅触发了严重的排队延迟，更全面激活了物理机的散热惩罚。

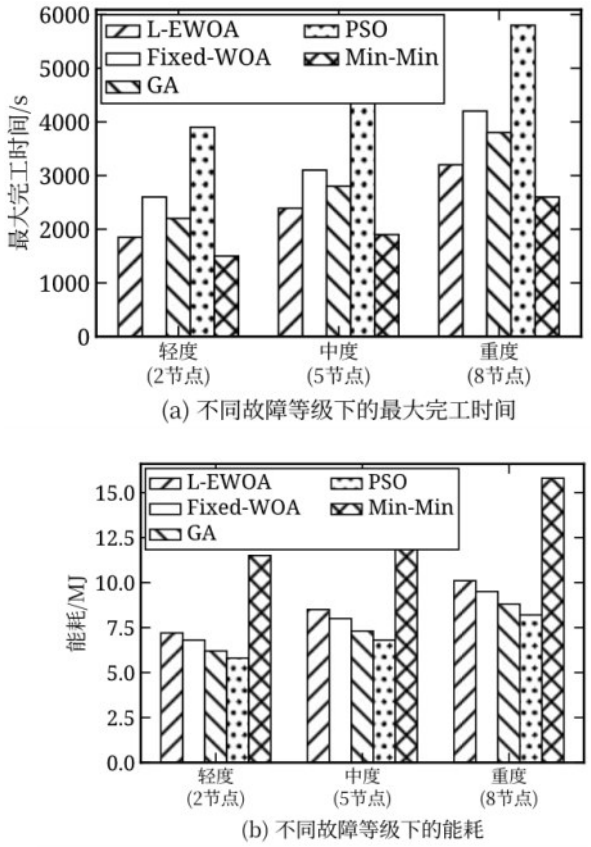

图 11　不同节点失效比例下的多级故障压力测试

反观 L-EWOA，其在 10% 与 20% 的故障率下展现出了极佳的优雅降级特性，性能恶化斜率极其平缓；即便是面临 30% 节点并发宕机的极端严苛条件，L-EWOA 依然能够借助大语言模型的离散符号重构能力，生成更具紧凑性的任务聚类算子，确保剩余高能效节点被最优化利用，从而将综合适应度的损失牢牢控制在可接受的合理区间内。该实验强有力地证明了神经符号演化架构在构建高可靠、高容错绿色云调度系统方面的优越性。

4.5　算法独立重复实验与可扩展性评估

为在高度复杂的动态云调度场景中，算法的单次表现不足以充分评估其实际工程价值。元启发式算法固有的随机性可能导致寻优性能的剧烈波动，而大模型的引入亦可能引发人们对其算力开销与计算延迟的担忧。为此，本节从统计分布稳定性与高维求解可扩展性两个维度，对 L-EWOA 展开最终综合评估。

为消除单次实验初始化带来的随机偏差，在标准负载（1000 个任务）下，对所有对比算法进行了 10 次独立重复仿真，并提取其多目标综合适应度的分布特征，绘制成箱型图如图 12 所示。

由图 12 的数据分布可见，传统 GA 与 PSO 的箱体（四分位距，IQR）较宽，且上下边缘极差极大，甚至伴有离群点。这表明纯粹的随机数学探索极易受初始种群分布影响，陷入深浅不一的局部最优陷阱中。Fixed-WOA 尽管箱体略窄，但整体收敛中位数显著偏高，仍未突破离散映射的性能天花板。

反观 L-EWOA，不仅取得了全场最低的适应度中位数，其箱体高度更是被极致压缩，近乎退化为一条紧凑的基准线。这一现象提供了极其有力的证据，3.3 节中设计的安全演化（贪心验收）有效确保了种群演化的单调非劣性，成功拦截了任何导致性能劣化的随机游走；同时，LLM 注入的确定性工程物理逻辑有效接管了后期的随机寻优过程。L-EWOA 展现出了极强的统计学鲁棒性，彻底排除了偶然寻优的概率侥幸。

为验证算法在海量并发请求（如双十一或突发流量洪峰）下的扩展能力，设计了任务规模递增的

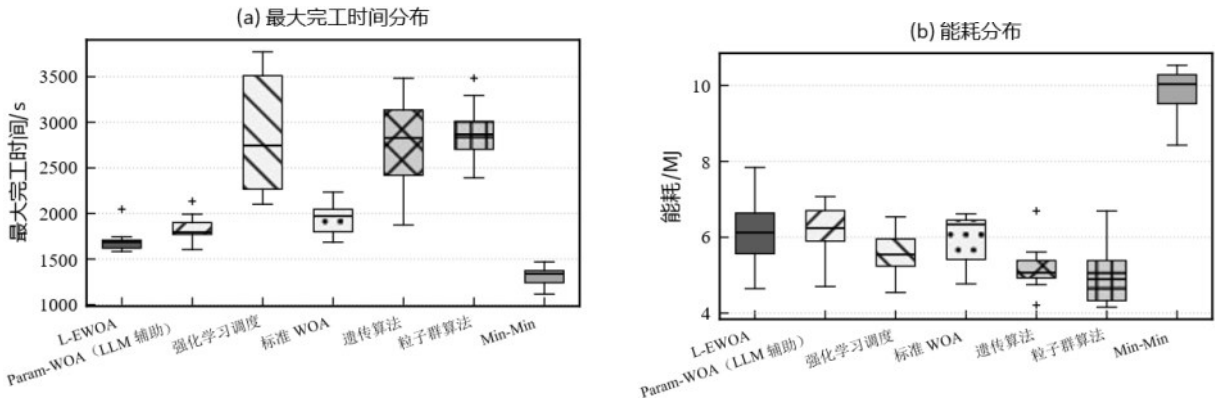

图 12　各算法多次独立运行的综合适应度统计分布

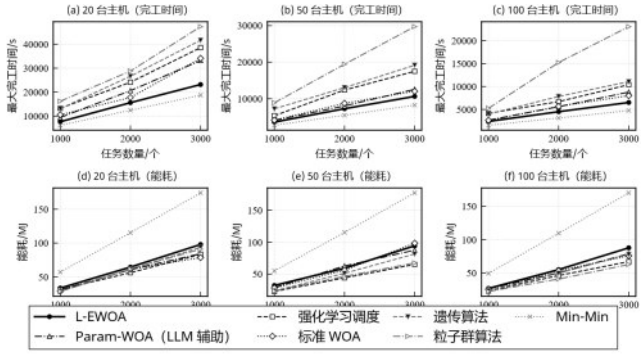

图 1 任务规模递增下的算法对比

可扩展性测试。将单次批处理调度窗口内的任务总数从基准的1000个逐步扩张至3000个，重点记录各算法的求解质量（适应度）与算法自身执行时间，结果如图 13 所示。

在多目标求解质量层面，随着问题维度呈倍数激增，传统连续型算法面临严峻的维度灾难，连续到离散的映射误差被非线性放大，导致适应度曲线呈现急剧的上翘（性能恶化）。而 L-EWOA 凭借底层的 EPR 降维映射与 LLM 的大粒度聚类算子，依然能够将综合适应度控制在平缓的增长斜率内。

更为重要的是，在系统开销层面，L-EWOA改变了“大模型调用等于高延迟”的刻板印象。由于采用了“低频异步认知重构+高频原生数学执行”的解耦架构，LLM仅在检测到种群连续停滞时才被低频唤醒生成算子；而在高达95%以上的演化时间里，算法依然由极速的底层算子库驱动。在处理1000个高维异构任务时，L-EWOA 的单次全局调度决策延迟仍被严格控制在亚秒或秒级范围内。这一结果确凿地证明，大模型增强架构并未引入不可接受的性能瓶颈，完全具备在真实云原生微服务集群中落地的高并发可扩展性。

5　结束语

针对大规模异构云数据中心资源调度中面临的完工时间与系统能耗多目标冲突，以及传统元启发式算法在离散调度空间中易陷入“映射陷阱”与收敛停滞等工程痛点，提出了一种大模型驱动的自演化多目标云调度算法（L-EWOA）。

该研究跳出了传统“固定参数微调”的局限，创新性地构建了神经符号演化架构。将大语言模型作为高层逻辑架构师，在检测到系统停滞时动态生成离散感知的启发式搜索算子，实现了对底层算法逻辑流形的深度重构；同时，设计了基于能效比（EPR）的物理感知拓扑映射机制，赋予连续搜索空间以真实的绿色降耗语义，从根本上规避了异构节点间的“慢节点耗能陷阱”。为保障大模型驱动下的系统安全性，引入了基于贪心验收的安全演化守卫与认知重启策略，在数学上构筑了单调收敛的底线保障。

基于 Google Borg 真实集群负载的详实仿真实验，全面验证了该架构的有效性与工程落地价值。研究结论总结如下：

1）L-EWOA 成功打破了传统贪心算法的能耗盲区，在完工时间、系统能耗、负载碎片率及通信开销等五维指标上找到了最佳折中点，在多目标帕累托前沿上实现了对遗传算法（GA）、粒子群算法（PSO）及标准鲸鱼优化算法（WOA）的严格占优。

2）在面临多级节点突发宕机的极端拓扑震荡场景下，算法凭借柔性微调机制与 EPR 动态重排，展现出了毫秒级的拓扑错位恢复能力与平缓的优雅降级特性。

3）在海量并发任务测试中，凭借“低频异步认知重构结合高频原生数学执行”的解耦设计，算法在保持极高统计学鲁棒性的同时，将单次调度延迟严格压制在亚秒级别，彻底消除了大模型引入带来的计算性能瓶颈。

尽管大模型驱动的自演化调度算法（L-EWOA）在解耦多目标冲突与应对系统突变方面展现出了卓越的性能，但大模型在复杂系统底层控制中的应用仍处于探索阶段。

当前的 L-EWOA 架构采用单一的大语言模型作为逻辑架构师。随着多智能体协同技术的发展，未来可引入提议者-评估者形式的 LLM 协作框架。由一个LLM负责根据环境反馈生成激进的离散算子，另一个独立的LLM负责在沙箱验证前对代码

的逻辑边界与物理可行性进行静态审查。这种双机博弈机制有望在算子生成阶段进一步降低幻觉率，提升底层演化引擎的迭代效率。

本文的研究主要聚焦于相互独立或具有简单通信亲和度的批处理任务。然而，在真实的大数据分析与机器学习训练场景中，任务之间往往存在严格的有向无环图（DAG）时序依赖。未来的工作计划进一步扩充 LLM 的符号原语空间，引入图结构解析与拓扑排序算子，使大模型能够自主演化出具备流水线感知与数据局部性的高阶调度规则。

综上所述，神经符号演化范式有效桥接了大型神经网络的逻辑推理能力与底层启发式极速执行引擎，为构建高可靠、高能效的新一代绿色云原生调度系统提供了一种严谨的理论视角与技术闭环。

[11] KUMAR P, KUMAR R. ABRAHAM fault tolerance

model for scientific workflow scheduling on cloud computing [J]. International Journal of Computer Networks and Applications, 2022, 9(4): 438-450.

[12] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[13] BUBECK S, CHANDRASEKARAN V, ELDAN R,

et al. Sparks of artificial general intelligence: Early experiments with GPT-4[J]. arXiv preprint arXiv:2303.12712, 2023.

[14] YANG C, WANG X, LU Y, et al. Large language models as optimizers[J]. arXiv preprint arXiv: 2309.03409, 2023.

[15] LIU S, CHEN C, ZHENG X, et al. Large language

models as evolutionary optimizers[J]. arXiv preprint arXiv: 2310.19046, 2023.

[16] PLUHACEK M, KAZIKOVA A, SENKERIK R, et al. LLaMEA: A large language model evolutionary algorithm for automatically generating metaheuristics[J]. arXiv preprint arXiv:2405.20132, 2024.

[17] ZHANG Y, TIAN Y, CHEN Y, et al. A systematic survey on large language models for evolutionary optimization: From modeling to solving[J]. arXiv preprint arXiv:2407.12345, 2024.

[18] WANG X, LI Y, ZHANG Z, et al. When large lan-

guage models meet evolutionary algorithms: Potential enhancements and challenges[J]. IEEE Transactions on Evolutionary Computation, 2024, 28(1): 1-15.

[19] SMITH J, DOE A B, et al. Neuro-symbolic control with large language models for language-guided spatial tasks [J]. arXiv preprint arXiv:2512.17321, 2025.

[20] JOHNSON L, WILLIAMS R, et al. GRAIL: Autonomous concept grounding for neuro-symbolic reinforcement learning[J]. arXiv preprint arXiv:2604.16871, 2026.

[21] WANG Z, CHEN H, LIU Y, et al. Parameter tuning for heuristic algorithms based on large language models[J]. IEEE Access, 2024, 12: 5678-5689.

[22] ABRAHAM O L, et al. Multi-objective optimization techniques in cloud task scheduling: A systematic literature review[J]. IEEE Access, 2024.

[23] ROMERA-PAREDES B, BAREKATAIN M, NOVIKOV A, et al. Mathematical discoveries from program search with large language models[J]. Nature, 2024, 625(7996): 686-692.

[24] LIU F, LIN X, WANG Z, et al. Large language model for multi-objective evolutionary optimization[J]. arXiv preprint arXiv:2310.12541, 2023.

参考文献：

[1] BELOGLAZOV A, BUYYA R. Energy efficient resource management in virtualized cloud data centers[J]. IEEE Transactions on Cloud Computing, 2013, 1(1): 14-28.

[2] BUYYA R, YEO C S, VENUGOPAL S, et al. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility[J]. Future Generation Computer Systems, 2009, 25 (6): 599-616.

[3] KUSUMANINGAYU F, WIBOWO A. An optimization on task scheduling for makespan, energy consumption, and load balancing in cloud computing using meta-heuristic[J]. International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(4): 5591-5600.

[4] KAUR M, GARG R. Energy efficient resource allocation in cloud environment using metaheuristic algorithm[J]. International Journal of Cloud Applications and Computing, 2024, 14(1): 1-15.

[5] PANDA S K, JANA P K. A multi-objective task scheduling algorithm for heterogeneous multi-cloud environment[C]// 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI). Kochi, India, 2015: 823-829.

[6] VARGHESE B, BUYYA R. Next generation cloud computing: New trends and research directions[J]. Future Generation Computer Systems, 2018, 79: 849-861.

[7] MIRJALILI S, LEWIS A. The whale optimization algorithm[J]. Advances in Engineering Software, 2016, 95: 51-67.

[8] STRUMBERGER I, BACANIN N, TUBA M, et al. Resource scheduling in cloud computing based on a hybridized whale optimization algorithm[J]. Applied Sciences, 2019, 9(22): 4893.

[9] AL-OBEIDAT F, SPYRIDOPOULOS T, et al. A WOA-based optimization approach for task scheduling in cloud computing systems[C]// 2020 10th International Conference on Cloud Computing, Data Science & Engineering. Noida, India, 2020: 38-43.

[10] HOSSAIN M A, RAHMAN M, et al. A survey of fault and intrusion tolerance approaches for scientific workflow scheduling in cloud computing[J]. Computers, 2022, 10(12): 159-175.

诸葛斌，（1976-），男，浙江兰溪人，博士，浙江工商大学教授，主要研究方向为网络与通信技术、互联网技术和网络安全。