(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2026 年 8 月 27 日 (27.08.2026)

WIPO | PCT

(10) 国际公布号
WO 2026/174669 A1

(51) 国际专利分类号:
G06F 9/50 (2006.01)

(21) 国际申请号: PCT/CN2025/094486

(22) 国际申请日: 2025 年 5 月 13 日 (13.05.2025)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202510188154.6 2025年2月20日 (20.02.2025) CN

(71) 申请人: 无限星途科技私人有限公司 (**X STAR TECHNOLOGY PTE. LTD.**) [SG/SG]; 新加坡, 新达城1号楼, 37-03, 淡马锡林荫道7号 038987 (SG)。

(72) 发明人: 肖鑫磊 (**XIAO, Xinlei**); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。 田立民 (**TIAN, Limin**); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。 唐守旭 (**TANG, Shouxu**); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。 张磊 (**ZHANG, Lei**); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。 李浩然 (**LI, Haoran**); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。 杨青相 (**YANG, Qingxiang**); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。

(74) 代理人: 北京品源专利代理有限公司 (**BEYOND ATTORNEYS AT LAW**); 中国北京市海淀区莲花池东路39号西金大厦6层 100036 (CN)。

(54) **Title:** MODEL INFERENCE SERVICE DEPLOYMENT METHOD AND APPARATUS, DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 模型推理服务部署方法、装置、设备及存储介质

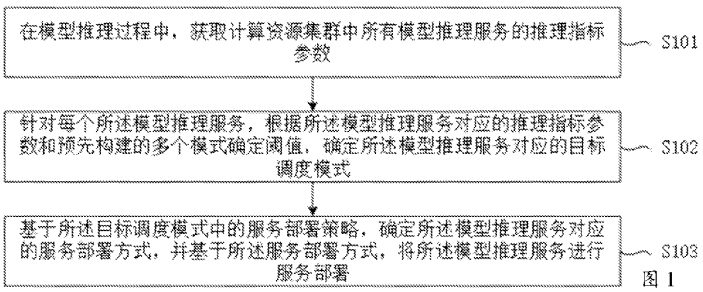


图 1

S101 In a model inference process, acquire inference metric parameters of all model inference services in a computing resource cluster

S102 For each model inference service, on the basis of the inference metric parameters corresponding to the model inference service and pre-constructed mode determination thresholds, determine a target scheduling mode corresponding to the model inference service

S103 On the basis of a service deployment policy in the target scheduling mode, determine a service deployment mode corresponding to the model inference service, and on the basis of the service deployment mode, perform service deployment on the model inference service

(57) **Abstract:** A model inference service deployment method and apparatus, a device, and a storage medium. The method comprises: in a model inference process, acquiring inference metric parameters of all model inference services in a computing resource cluster (S101); for each model inference service, on the basis of the inference metric parameters corresponding to the model inference service and pre-constructed mode determination thresholds, determining a target scheduling mode corresponding to the model inference service, wherein the target scheduling mode comprises a quality of service priority mode and a resource utilization priority mode (S102); and

WO 2026/174669 A1

[见续页]

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UY, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

on the basis of a service deployment policy in the target scheduling mode, determining a service deployment mode corresponding to the model inference service, and on the basis of the service deployment mode, performing service deployment on the model inference service (S103).

(57) 摘要: 一种模型推理服务部署方法、装置、设备及存储介质。该方法包括，在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数(S101)。针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的各个模式确定阈值，确定所述模型推理服务对应的目标调度模式，其中，所述目标调度模式包括服务质量优先模式和资源利用率优先模式(S102)。基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署(S103)。