

(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



(10) 国际公布号  
**WO 2025/086648 A1**

(43) 国际公布日  
2025 年 5 月 1 日 (01.05.2025)

(51) 国际专利分类号:  
**H04L 67/63** (2022.01)

(21) 国际申请号: PCT/CN2024/096924

(22) 国际申请日: 2024 年 6 月 3 日 (03.06.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:  
202311386911.8 2023 年 10 月 24 日 (24.10.2023) CN

(71) 申请人: 华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼 518129 (CN)。

(72) 发明人: 马翼鹏 (MA, Yipeng); 中国广东省深圳市龙岗区坂田华为总部办公楼 518129 (CN)。汪涛 (WANG, Tao); 中国广东省深圳市龙岗区坂田华为总部办公楼 518129 (CN)。彭勇 (PENG, Yong); 中国广东省深圳市龙岗区坂田

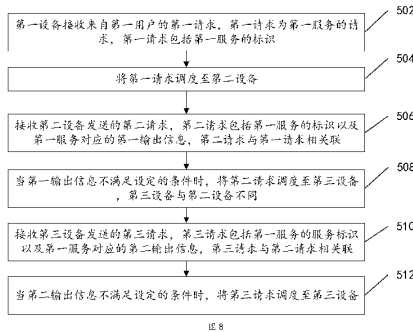
华为总部办公楼 518129 (CN)。周剑辉 (ZHOU, Jianhui); 中国广东省深圳市龙岗区坂田华为总部办公楼 518129 (CN)。

(74) 代理人: 北京汇思诚业知识产权代理有限公司 (UNI-INTEL PATENT AND TRADEMARK LAW FIRM); 中国北京市朝阳区建国门外大街永安东里甲3号通用国际中心A座3层 100022 (CN)。

(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, UZ, VC, VN, WS, ZA, ZM, ZW。

(54) Title: REQUEST SCHEDULING METHOD AND DEVICE

(54) 发明名称: 一种请求调度方法和设备



502 A first device receives a first request from a first user, the first request being a request for a first service, and the first request comprising an identifier of the first service  
504 Schedule the first request to a second device  
506 Receive a second request sent by the second device, the second request comprising the identifier of the first service and first output information corresponding to the first service, and the second request being associated with the first request  
508 If the first output information does not meet a preset condition, schedule the second request to a third device, the third device being different from the second device  
510 Receive a third request sent by the third device, the third request comprising the service identifier of the first service and second output information corresponding to the first service, and the third request being associated with the second request  
512 If the second output information does not meet the preset condition, schedule the third request to the third device

(57) Abstract: Disclosed are a request scheduling method and a device. The method comprises: a first device receiving a first request from a first user, the first request being a request for a first service, and the first request comprising an identifier of the first service; the first device scheduling the first request to a second device; the second device sending a second request to the first device, the second request comprising the identifier of the first service and first output information, and the second request being associated with the first request; if the first output information does not meet a preset condition, the first device scheduling the second request to a third device, the third device being different from the second device; the third device sending a third request to the first device, the third request comprising the identifier of the first service and second output information, and the third request being associated with the second request; and if the second output information does not meet the preset condition, the first device scheduling the third request to the third device, thereby improving the throughput of a service system in large-scale language model service deployment.

[见续页]

(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))。

(57) 摘要: 本发明公开了一种请求调度方法和设备, 所述方法包括: 第一设备接收第一用户的第一请求, 第一请求为第一服务的请求, 第一请求包括第一服务的标识; 第一设备将第一请求调度至第二设备; 第二设备向第一设备发送第二请求, 第二请求包括第一服务的标识和第一输出信息, 第二请求与第一请求相关联; 当第一输出信息不满足设定的条件时, 第一设备将第二请求调度至第三设备, 第三设备与第二设备不同; 第三设备向第一设备发送第三请求, 第三请求包括第一服务的标识和第二输出信息, 第三请求与第二请求相关联; 当第二输出信息不满足所述设定的条件时, 第一设备将第三请求调度至第三设备, 能够提升大规模语言模型服务部署中的服务系统的吞吐量。