(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
2025 年 8 月 7 日 (07.08.2025)

WIPO | PCT

(10) 国际公布号
WO 2025/161395 A1

(51) 国际专利分类号:
G06N 5/04 (2023.01)

(21) 国际申请号: PCT/CN2024/117525

(22) 国际申请日: 2024 年 9 月 6 日 (06.09.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202410137574.7 2024年1月31日 (31.01.2024) CN
202410223741.X 2024年2月28日 (28.02.2024) CN

(71) 申请人: 华为云计算技术有限公司 (**HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.**) [CN/CN]; 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心 550025 (CN)。

(72) 发明人: 高彬 (**GAO, Bin**); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心 550025 (CN)。左鹏飞 (**ZUO, Pengfei**); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心 550025 (CN)。杨幸坤 (**YANG, Xingkun**); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心 550025 (CN)。邓俊波 (**DENG, Junbo**); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心 550025 (CN)。

(74) 代理人: 北京三高永信知识产权代理有限责任公司 (**BEIJING SAN GAO YONG XIN INTELLECTUAL PROPERTY AGENCY CO., LTD.**); 中国北京市海淀区西直门北大街 32 号院 1号楼4层503 100088 (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,

(54) **Title:** COMPUTING SYSTEM, MULTI-ROUND SESSION INFERENCE METHOD, APPARATUS AND COMPUTING DEVICE CLUSTER

(54) 发明名称: 计算系统、多轮会话推理方法、装置和计算设备集群

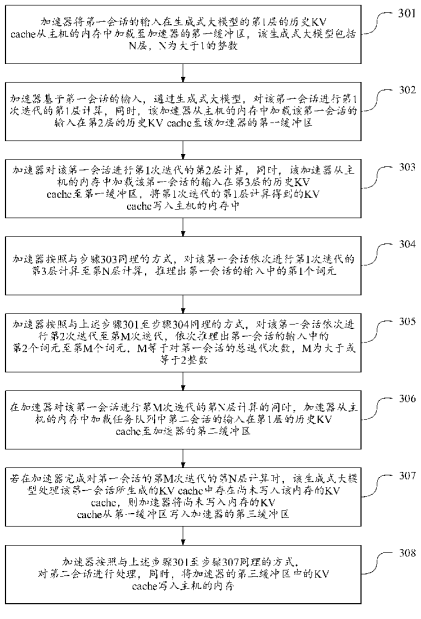

图3

301 An accelerator loads from an internal memory of a host to a first buffer area of the accelerator the historical KV cache of an input of a first session in a first layer of a generative large model, the generative large model comprising N layers, and N being an integer greater than 1

302 On the basis of the input of the first session and by means of the generative large model, the accelerator performs first-layer computing of a first iteration on the first session, and simultaneously the accelerator loads from the internal memory of the host to the first buffer area of the accelerator the historical KV cache of the input of the first session in a second layer

303 The accelerator performs second-layer computing of the first iteration on the first session, and simultaneously the accelerator loads from the internal memory of the host to the first buffer area the historical KV cache of the input of the first session in a third layer and write a KV cache obtained by the first-layer computing of the first iteration into the internal memory of the host

304 In the same way as that of step 303, the accelerator successively performs third-layer computing to N-th Layer computing of the first iteration on the first session, so as to infer a first token in the input of the first session

305 In the same way as that of step 301 to step 304, the accelerator successively performs a second iteration to an M-th iteration on the first session, so as to successively infer a second token to an M-th token in the input of the first session, M being equal to the total number of iterations on the first session, and M being an integer greater than or equal to 2

306 While performing N-th layer computing of the M-th iteration on the first session, the accelerator loads from the internal memory of the host to a second buffer area of the accelerator the historical KV cache of an input of a second session in a task queue in the first layer

307 When the accelerator completes the N-th layer computing of the M-th iteration on the first session, if there is a KV cache among the KV caches generated by the generative large model processing the session which has not yet been written into the internal memory, the accelerator writes the KV cache, which has not yet been written into the internal memory, from the first buffer area into a third buffer area of the accelerator

308 In the same way as that of step 301 to step 307, the accelerator processes the second session, and simultaneously writes the KV cache in the third buffer area of the accelerator into the internal memory of the host

(57) **Abstract:** The present application belongs to the technical field of cloud computing and provides a computing system, a multi-round session inference method, an apparatus and a computing device cluster. In the system, an external storage device stores a historical key value cache of a completed session. As the capacity of the external storage device is far greater than the capacity of an HBM in an accelerator, the hit rate of the historical key value cache can be increased, thereby avoiding recomputing the key value cache. In addition, when the accelerator processes a session, a host can preload a historical key value cache of a session to be processed in a task

[见续页]

WO 2025/161395 A1

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。

(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))。

queue from the external storage device to an internal memory of the host; when performing i-th layer computing on the session, the accelerator can preload from the internal memory of the host the historical key value cache required for the (i+1)-th layer computing. As the computing process and the data loading process are synchronously performed, the computing process does not need to wait for the completion of data loading, such that the time overheads of the accelerator accessing the external storage device can be hidden, improving the efficiency of multi-round session inference.

(57) 摘要: 本申请提供了一种计算系统、多轮会话推理方法、装置和计算设备集群，属于云计算技术领域。该系统中，外部存储设备存储了已完成的会话的历史键值缓存，由于外部存储设备的容量远大于加速器内的HBM的容量，因此，能够提高历史键值缓存的命中率，从而避免进行键值缓存重计算；并且，主机能够在加速器处理会话时，从外部存储设备中预加载任务队列中待处理会话的历史键值缓存至主机的内存中，加速器能够在对会话进行第i层计算时，从主机的内存中预加载第i+1层计算所需的历史键值缓存，由于计算过程与数据加载过程同步进行，计算过程无需等待数据加载完成，从而能够隐藏加速器访问外部存储设备的时间开销，提高多轮会话推理的效率。