

(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 2022/095815 A1

(43) 国际公布日
2022 年 5 月 12 日 (12.05.2022)

(51) 国际专利分类号:
G06F 9/50 (2006.01)

(21) 国际申请号: PCT/CN2021/127856

(22) 国际申请日: 2021 年 11 月 1 日 (01.11.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202011219652.6 2020 年 11 月 3 日 (03.11.2020) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人: 肖文聪 (XIAO, Wencong); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼, Zhejiang 311121 (CN)。任仕儒 (REN, Shiru); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼, Zhejiang 311121 (CN)。李永 (LI, Yong); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼, Zhejiang 311121 (CN)。

(74) 代理人: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,

(54) Title: GRAPHICS CARD MEMORY MANAGEMENT METHOD AND APPARATUS, DEVICE, AND SYSTEM

(54) 发明名称: 显存管理方法、装置、设备及系统

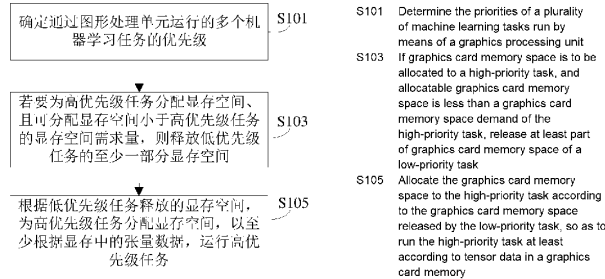


图 1

(57) Abstract: A graphics card memory management method and apparatus, a device, and a system. The method comprises: determining the priorities of a plurality of machine learning tasks run by means of a graphics processing unit (GPU); if graphics card memory resources are to be allocated to a high-priority task, and allocatable graphics card memory resources are less than a graphics card memory resource demand of the high-priority task, releasing at least some of graphics card memory resources occupied by a low-priority task; and allocating the graphics card memory resources to the high-priority task, so as to run the high-priority task at least according to tensor data of a graphics card memory space. By using such a processing method, when allocatable graphics card memory resources are insufficient, graphics card memory resources occupied by a low-priority task are allocated to a high-priority task for use, thereby performing dynamic scaling optimization on GPU graphics card memory resources occupied by a plurality of parallel machine learning tasks on a GPU. In this way, insofar as the performance of the high-priority task is guaranteed, the resource utilization rate of the whole cluster can be improved.

[见续页]

CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, 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：一种显存管理方法、装置、设备及系统。其中，所述方法包括：确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存资源；为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务。采用这种处理方式，使得在可分配显存资源不足时，将低优先级任务占用的显存资源分配给高优先级任务使用，由此实现对在一个GPU上并行的多个机器学习任务占用的GPU显存资源进行动态伸缩优化，这样就可以在保障高优先级任务性能的前提下，提升整体集群的资源利用率。