

(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 2024/244856 A1

(43) 国际公布日
2024 年 12 月 5 日 (05.12.2024)

(51) 国际专利分类号:
G06F 9/30 (2018.01)

(21) 国际申请号: PCT/CN2024/090157

(22) 国际申请日: 2024 年 4 月 26 日 (26.04.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202310639274.4 2023 年 5 月 31 日 (31.05.2023) CN
202311070464.5 2023 年 8 月 22 日 (22.08.2023) 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, Guangdong 518129 (CN)。

(72) 发明人: 杨玉超(YANG, Yuchao); 中国北京市海淀区颐和园路5号, Beijing 100871 (CN)。朱晓明(ZHU, Xiaoming); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。陶耀宇(TAO,

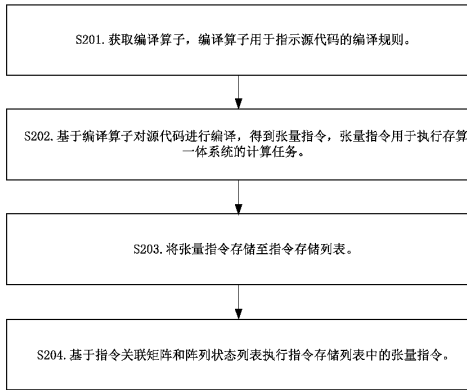
Yaoyu); 中国北京市海淀区颐和园路5号, Beijing 100871 (CN)。朱奕航(ZHU, Yihang); 中国北京市海淀区颐和园路5号, Beijing 100871 (CN)。董伟伟(DONG, Weiwei); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。王成旭(WANG, Chengxu); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。谭海波(TAN, Haibo); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所(普通合伙)(SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦B座18C2、18D、18E、18E2, Guangdong 518001 (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,

(54) Title: OPERATOR COMPILING METHOD AND APPARATUS

(54) 发明名称: 一种算子编译方法以及装置



S201 Acquire a compilation operator, which is used for indicating a compilation rule for source codes
S202 Compile the source codes on the basis of the compilation operator, so as to obtain tensor instructions, wherein the tensor instructions are used for executing computing tasks of an in-memory computing system
S203 Store the tensor instructions in an instruction storage list
S204 On the basis of an instruction association matrix and an array state list, execute the tensor instructions in the instruction storage list

图 2

(57) Abstract: An operator compiling method and apparatus, which are used for implementing tensor data processing of an in-memory computing architecture, and improving the computing task processing parallelism of the in-memory computing architecture. The method comprises: acquiring a compilation operator, which is used for indicating a compilation rule for source codes; compiling the source codes on the basis of the compilation operator, so as to obtain one or more tensor instructions, wherein the tensor instructions are used for executing computing tasks in an in-memory computing system, and each tensor instruction comprises one or more of the following instruction segments: an operation code, a tensor address, a data type, a data bit width and a tensor length; and storing the tensor instructions in an instruction storage list, which is used for storing compiled tensor instructions.

[见续页]

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)。

根据细则4.17的声明:
— 关于申请人有权要求在先申请的优先权(细则4.17(iii))

本国际公布:
— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种算子编译方法以及装置，用于实现存算一体架构的张量数据处理，提升存算一体架构的计算任务处理并行度。方法包括：获取编译算子，编译算子用于指示对源代码的编译规则。基于编译算子对源代码进行编译，得到一条或多条张量指令，张量指令用于执行存算一体系统中计算任务，张量指令包括以下一项或多项指令段：操作码、张量地址、数据类型、数据位宽和张量长度。将张量指令存储至指令存储列表，指令存储列表用于存储编译完成的张量指令。