

(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
2023 年 7 月 6 日 (06.07.2023)



(10) 国际公布号
WO 2023/124372 A1

(51) 国际专利分类号:
G06F 7/485 (2006.01)

(21) 国际申请号: PCT/CN2022/124517

(22) 国际申请日: 2022 年 10 月 11 日 (11.10.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111667694.0 2021 年 12 月 31 日 (31.12.2021) CN

(71) 申请人: 上海商汤智能科技有限公司 (SHANGHAI SENSETIME INTELLIGENT TECHNOLOGY CO., LTD.) [CN/CN]; 中国上海市徐汇区桂平路 391 号 3 号楼 1605A 室, Shanghai 200233 (CN)。

(72) 发明人: 霍冠廷 (HUO, Guanting); 中国上海市徐汇区桂平路 391 号 3 号楼 1605A 室, Shanghai 200233 (CN)。 王文强 (WANG, Wenqiang); 中国上海市徐汇区桂平路 391 号 3 号楼 1605A 室, Shanghai 200233 (CN)。 徐宁仪 (XU, Ningyi);

中国上海市徐汇区桂平路 391 号 3 号楼 1605A 室, Shanghai 200233 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区上地三街 9 号嘉华大厦 B 座 409, Beijing 100085 (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, 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,

(54) **Title:** FLOATING-POINT NUMBER PROCESSING APPARATUS AND METHOD, ELECTRONIC DEVICE, STORAGE MEDIUM, AND CHIP

(54) 发明名称: 浮点数处理装置、方法、电子设备、存储介质及芯片

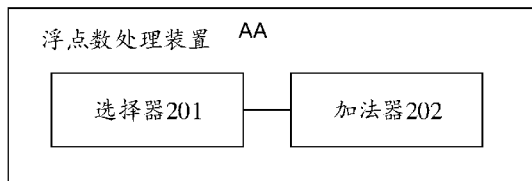


图 2

201 Selector
202 Adder
AA Floating-point number processing apparatus

(57) **Abstract:** The present disclosure provides a floating-point number processing apparatus and method, an electronic device, a storage medium, and a chip. The floating-point number processing method comprises: acquiring a plurality of floating-point numbers to be processed, the floating-point numbers to be processed being used to perform an operation in a target chip; extending a sign bit of each of the plurality of floating-point numbers to be processed to a target bit width, so as to obtain a plurality of extended floating-point numbers, wherein the target bit width matches the total number of the floating-point numbers to be processed; and performing accumulation processing on the plurality of extended floating-point numbers in the target chip, so as to obtain a target floating-point number.

(57) **摘要:** 本公开提供了一种浮点数处理装置、方法、电子设备、存储介质及芯片。该浮点数处理方法包括: 获取在目标芯片中进行运算的多个待处理浮点数; 将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽, 得到多个扩展后浮点数; 其中, 所述目标位宽与所述待处理浮点数的总数量匹配; 在所述目标芯片中将所述多个扩展后浮点数进行累加处理, 得到目标浮点数。

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, 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))。

浮点数处理装置、方法、电子设备、存储介质及芯片

交叉引用声明

本申请要求于 2021 年 12 月 31 日提交中国专利局的申请号为 202111667694.0 的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

[01] 本公开涉及集成电路技术领域，具体而言，涉及一种浮点数处理装置、方法、电子设备、计算机可读存储介质及芯片。

背景技术

[02] 随着半导体工艺、计算机体系结构、处理器设计架构的不断发展，处理器的功能越来越强大，结构也变得越来越复杂。其中，浮点数运算是处理器中运算步骤较多、延迟较大、功耗较大的一类计算，对处理器的性能指标有较大的影响。因此，提出一种优化浮点数处理过程的方法尤为重要。

发明内容

[03] 有鉴于此，本公开至少提供一种浮点数处理装置、方法、电子设备、计算机可读存储介质及芯片。

[04] 第一方面，本公开提供了一种浮点数处理装置，所述装置包括：选择器，用于将获取到的多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数；其中，所述目标位宽与所述待处理浮点数的总数量匹配；加法器，与所述选择器相连，用于将所述多个扩展后浮点数进行累加处理，得到目标浮点数。

[05] 为了保障浮点数累加过程的精度，满足加和运算后得到的中间浮点数的位宽要求，可以利用选择器将每个待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数，实现了待处理浮点数位宽的扩展，进而使得对多个扩展后浮点数进行累加处理过程中，无需对累加处理过程中得到的中间浮点数进行规格化处理，减少了规格化处理的次数，降低了芯片处理多个扩展后浮点数的累加过程的延迟和功耗；同时，减少了规格化处理的次数后，能够缓解规格化处理带来的数据丢失问题，提高了目标浮点数的精度。

[06] 在一种可能的实施方式中，所述装置还包括：比较器，与所述加法器相连，用于对从所述加法器输入的所述目标浮点数进行规格化处理，得到目标处理结果，其中，所述目标处理结果的格式与预设格式相匹配。

[07] 实施时，对多个扩展后浮点数进行累加处理过程中，无需对累加处理过程中得到的中间浮点数进行规格化处理；而是可以在累加得到目标浮点数后，对目标浮点数进行一次规格化处理，得到目标处理结果，减少了规格化处理的次数，降低了目标芯片处理多个浮点数的累加过程的延迟和功耗；同时，可以缓解规格化处理带来的数据丢失问题，提高了目标处理结果的精度。

[08] 在一种可能的实施方式中，所述装置还包括：对阶运算器，所述对阶运算器分别与所述加法器和所述选择器相连；所述对阶运算器包括减法器和移位器；所述减法器，用于确定每个所述扩展后浮点数的初始指数与目标指数之间的目标差值；其中，所述目标指数为基于所述多个扩展后浮点数的初始指数确定的；所述移位器用于针对每个所述扩展后浮点数，将所述扩展后浮点数的初始指数对齐至所述目标指数，并基于所述扩展后

浮点数对应的目标差值，对所述扩展后浮点数的尾数进行右移，得到处理后浮点数，并将所述处理后浮点数输入至所述加法器；其中，所述加法器用于将从所述多个扩展后浮点数分别得到的所述处理后浮点数的尾数进行累加处理，得到所述目标浮点数。

[09]在一种可能的实施方式中，所述选择器，在将获取到的多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数时，用于：从所述多个待处理浮点数中选取两个待处理浮点数，并将所述两个待处理浮点数的符号位扩展至所述目标位宽，得到两个扩展后浮点数；所述减法器用于将所述两个扩展后浮点数中的第一浮点数的初始指数作为所述目标指数；确定所述目标指数与所述两个扩展后浮点数中的第二浮点数的初始指数的第一差值；其中，所述第一浮点数的初始指数大于所述第二浮点数的初始指数；所述移位器用于将所述第二浮点数的初始指数对齐至所述目标指数，并基于所述第一差值，将所述第二浮点数的尾数进行右移，得到处理后的第二浮点数；将所述第一浮点数和所述处理后的第二浮点数输入至所述加法器；所述加法器用于将所述第一浮点数和所述处理后的第二浮点数进行加法处理，得到中间浮点数。

[10]在一种可能的实施方式中，所述装置还包括：第一寄存器，分别与所述加法器和所述对阶运算器中的所述减法器相连，用于存储由所述加法器得到并输入的所述中间浮点数和所述加法器发送的目标浮点数，并将所述中间浮点数发送给所述减法器。

[11]在一种可能的实施方式中，所述装置还包括：第二寄存器，与所述减法器相连，用于存储所述目标指数，并将所述目标指数发送给所述对阶运算器中的所述减法器。

[12]在一种可能的实施方式中，所述选择器还用于：从未被选取的待处理浮点数中选取一个待处理浮点数，并将所选取的待处理浮点数的符号位扩展至所述目标位宽，得到所选取的待处理浮点数对应的扩展后浮点数；响应于所选取的待处理浮点数对应的扩展后浮点数的初始指数大于所述第二寄存器存储的目标指数，将所选取的待处理浮点数对应的扩展后浮点数的初始指数作为新的目标指数，利用所述新的目标指数对所述第二寄存器进行更新；其中，所述第二寄存器还用于将所述新的目标指数发送给所述减法器；所述减法器用于基于接收到的新的目标指数，确定所述中间浮点数的指数与所述目标指数之间的第二差值；所述移位器用于将所述中间浮点数的指数对齐至所述目标指数，并基于所述第二差值，对所述中间浮点数的尾数进行右移，得到处理后中间浮点数，并将所述处理后中间浮点数和所选取的待处理浮点数对应的扩展后浮点数发送给所述加法器；所述加法器用于将所述中间浮点数的尾数和所选取的待处理浮点数对应的扩展后浮点数的尾数进行加法处理，得到新的中间浮点数，并将所述新的中间浮点数发送给所述第一寄存器；以及将最后一次加和处理后得到的中间浮点数作为所述目标浮点数，并将所述目标浮点数发送给所述第一寄存器。

[13]在一种可能的实施方式中，所述比较器，在对所述目标浮点数进行规格化处理，得到目标处理结果时，用于：利用移位方式、截尾方式和舍入方式中的一种或多种方式，对所述目标浮点数进行规格化处理，得到所述目标处理结果。

[14]在一种可能的实施方式中，所述目标位宽为基于所述多个待处理浮点数包括的初始尾数的初始位宽和所述多个待处理浮点数的总数量确定的。

[15]第二方面，本公开提供了一种浮点数处理方法，包括：获取在目标芯片中进行运算的多个待处理浮点数；将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数；其中，所述目标位宽与所述待处理浮点数的总数量匹配；在所述目标芯片中将所述多个扩展后浮点数进行累加处理，得到目标浮点数。

[16]上述方法中，为了保障浮点数累加过程的精度，满足加和运算后得到的中间浮点数的位宽要求，可以将每个待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点

数，实现了待处理浮点數位宽的扩展，进而使得在目标芯片中对多个扩展后浮点数进行累加处理过程中，无需对累加处理过程中得到的中间浮点数进行规格化处理，减少了规格化处理的次数，降低了目标芯片处理多个扩展后浮点数的累加过程的延迟和功耗；同时，减少了规格化处理的次数后，能够缓解规格化处理带来的数据丢失问题，提高了目标浮点数的精度。

[17]在一种可能的实施方式中，在所述得到目标浮点数之后，所述方法还包括：对所述目标浮点数进行规格化处理，得到目标处理结果，其中，所述目标处理结果的格式与预设格式相匹配。

[18]实施时，对多个扩展后浮点数进行累加处理过程中，无需对累加处理过程中得到的中间浮点数进行规格化处理；而是可以在累加得到目标浮点数后，对目标浮点数进行一次规格化处理，得到目标处理结果，减少了规格化处理的次数，降低了目标芯片处理多个浮点数的累加过程的延迟和功耗；同时，可以缓解规格化处理带来的数据丢失问题，提高了目标处理结果的精度。

[19]在一种可能的实施方式中，所述目标位宽为根据以下步骤确定的：基于所述多个待处理浮点数包括的初始尾数的初始位宽，确定所述多个待处理浮点数对应的调整位宽；基于所述调整位宽和所述待处理浮点数的总数量，确定所述目标位宽。

[20]上述实施方式中，根据调整位宽和待处理浮点数的总数量，较准确的确定目标位宽，后续将待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数后，可以使得扩展后浮点数的位宽能够满足累加处理的位宽要求，进而无需对累加处理过程得到的中间浮点数进行规格化处理。

[21]在一种可能的实施方式中，所述将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数，包括：从所述多个待处理浮点数中选取两个待处理浮点数，并将所述两个待处理浮点数的符号位扩展至所述目标位宽，得到两个扩展后浮点数；所述在所述目标芯片中将所述多个扩展后浮点数进行累加处理，得到目标浮点数，包括：将所述两个扩展后浮点数作为两个当前浮点数，在所述目标芯片中将所述两个当前浮点数进行加和处理，得到中间浮点数；从未被选取的待处理浮点数中选取一个待处理浮点数，并将所选取的待处理浮点数的符号位扩展至所述目标位宽，得到所选取的待处理浮点数对应的扩展后浮点数；将所选取的待处理浮点数对应的扩展后浮点数和所述中间浮点数作为更新后的两个当前浮点数，返回至在所述目标芯片中将所述两个当前浮点数进行加和处理的步骤，直至所述多个待处理浮点数中的每个待处理浮点数均进行过加和处理；将最后一次加和处理后得到的中间浮点数作为所述目标浮点数。

[22]这里，可以将多个待处理浮点数进行多次累加，得到目标浮点数，且在累加过程中无需对生成的中间浮点数进行规格化处理，缓解了对中间浮点数进行规格化处理时造成的数据丢失问题，保障了中间浮点数的精度，进而使得得到的目标浮点数的精度较高。

[23]在一种可能的实施方式中，所述在所述目标芯片中将所述两个当前浮点数进行加和处理，得到中间浮点数，包括：在所述目标芯片中将所述两个当前浮点数中的第一浮点数的初始指数对齐到第二浮点数的初始指数，并将所述第一浮点数的尾数右移目标位数，得到对阶处理后的第一浮点数；其中，所述两个浮点数中的所述第一浮点数的初始指数小于所述第二浮点数的初始指数；所述目标位数为所述第二浮点数的初始指数与所述第一浮点数的初始指数之间的差值；在所述目标芯片中将所述对阶处理后的第一浮点数与所述第二浮点数进行加和处理，得到所述中间浮点数。

[24]在一种可能的实施方式中，所述对所述目标浮点数进行规格化处理，得到目标处理

结果，包括：利用移位方式、截尾方式和舍入方式中的一种或多种方式，对所述目标浮点数进行规格化处理，得到所述目标处理结果。

[25] 这里，可以利用移位方式、截尾方式和舍入方式中的一种或多种方式，灵活的对目标浮点数进行规格化处理，得到目标处理结果。

[26] 以下装置、电子设备等的效果描述参见上述方法的说明，这里不再赘述。

[27] 第三方面，本公开提供了一种芯片，包括：存储器，用于存储待运算的多个待处理浮点数；如上述第一方面或任一实施方式所述的浮点数处理装置，用于对所述多个待处理浮点数进行处理，得到目标浮点数。

[28] 第四方面，本公开提供了一种浮点数处理装置，包括：获取模块，用于获取在目标芯片中进行运算的多个待处理浮点数；扩展模块，用于将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数；其中，所述目标位宽与所述待处理浮点数的总数量匹配；第一处理模块，用于在所述目标芯片中将所述多个扩展后浮点数进行累加处理，得到目标浮点数。

[29] 在一种可能的实施方式中，在所述得到目标浮点数之后，所述装置还包括：第二处理模块，用于对所述目标浮点数进行规格化处理，得到目标处理结果，其中，所述目标处理结果的格式与预设格式相匹配。

[30] 在一种可能的实施方式中，所述扩展模块用于根据以下步骤确定目标位宽：基于所述多个待处理浮点数包括的初始尾数的初始位宽，确定所述多个待处理浮点数对应的调整位宽；基于所述调整位宽和所述待处理浮点数的总数量，确定所述目标位宽。

[31] 在一种可能的实施方式中，所述扩展模块，在将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数时，用于：从所述多个待处理浮点数中选取两个待处理浮点数，将所述两个待处理浮点数的符号位扩展至所述目标位宽，得到两个扩展后浮点数；所述第一处理模块，在所述目标芯片中将所述多个扩展后浮点数进行累加处理，得到目标浮点数时，用于：将所述两个扩展后浮点数作为两个当前浮点数，在所述目标芯片中将所述两个当前浮点数进行加和处理，得到中间浮点数；从未被选取的待处理浮点数中选取一个待处理浮点数，并将所选取的待处理浮点数的符号位扩展至目标位宽，得到所选取的待处理浮点数对应的扩展后浮点数；将所选取的待处理浮点数对应的扩展后浮点数和所述中间浮点数作为更新后的两个当前浮点数，返回至在所述目标芯片中将所述两个当前浮点数进行加和处理的步骤，直至所述多个待处理浮点数中的每个待处理浮点数均进行过加和处理；将最后一次加和处理后得到的中间浮点数作为所述目标浮点数。

[32] 在一种可能的实施方式中，所述第一处理模块，在所述目标芯片中将所述两个当前浮点数进行加和处理，得到中间浮点数时，用于：在所述目标芯片中将所述两个当前浮点数中的第一浮点数的初始指数对齐到第二浮点数的初始指数，并将所述第一浮点数的尾数右移目标位数，得到对阶处理后的第一浮点数；其中，所述两个浮点数中的所述第一浮点数的初始指数小于所述第二浮点数的初始指数；所述目标位数为所述第二浮点数的初始指数与所述第一浮点数的初始指数之间的差值；在所述目标芯片中将所述对阶处理后的第一浮点数与所述第二浮点数进行加和处理，得到所述中间浮点数。

[33] 在一种可能的实施方式中，所述第二处理模块，在对所述目标浮点数进行规格化处理，得到目标处理结果时，用于：利用移位方式、截尾方式和舍入方式中的一种或多种方式，对所述目标浮点数进行规格化处理，得到所述目标处理结果。

[34] 第五方面，本公开提供一种电子设备，包括：处理器、存储器和总线，所述存储器

存储有所述处理器可执行的机器可读指令，当电子设备运行时，所述处理器与所述存储器之间通过总线通信，所述机器可读指令被所述处理器执行时实现如上述第二方面或任一实施方式所述的浮点数处理方法的步骤。

[35] 第六方面，本公开提供一种电子设备，包括如第三方面所述的芯片。

[36] 第七方面，本公开提供一种计算机可读存储介质，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器运行时执行如上述第二方面或任一实施方式所述的浮点数处理方法的步骤。

[37] 为使本公开的上述目的、特征和优点能更明显易懂，下文特举较佳实施例，并配合所附附图，作详细说明如下。

附图说明

[38] 为了更清楚地说明本公开实施例的技术方案，下面将对实施例中所需要使用的附图作简单地介绍。这些附图示出了符合本公开的实施例，并与说明书一起用于说明本公开的技术方案。应当理解，以下附图仅示出了本公开的某些实施例，因此不应被看作是对范围的限定，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他相关的附图。

[39] 图 1 示出了本公开实施例所提供的一种浮点数处理方法的流程示意图；

[40] 图 2 示出了本公开实施例所提供的一种浮点数处理装置的架构示意图；

[41] 图 3 示出了本公开实施例所提供的另一种浮点数处理装置的架构示意图；

[42] 图 4 示出了本公开实施例所提供的一种芯片的架构示意图；

[43] 图 5 示出了本公开实施例所提供的一种浮点数处理装置的架构示意图；

[44] 图 6 示出了本公开实施例所提供的一种电子设备的结构示意图。

具体实施方式

[45] 为使本公开实施例的目的、技术方案和优点更加清楚，下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行清楚、完整地描述。所描述的实施例仅仅是本公开一部分实施例，而不是全部的实施例。通常在此处附图中描述和示出的本公开实施例的组件可以以各种不同的配置来布置和设计。因此，以下对在附图中提供的本公开的实施例的详细描述无意限制要求保护的本公开的范围。基于本公开的实施例，本领域技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本公开保护的范围。

[46] 随着半导体工艺、计算机体系结构、处理器设计架构的不断发展，处理器的功能越来越强大，结构也变得越来越复杂。其中，浮点数运算是处理器中一类运算步骤较多、延迟较大、功耗较大的计算，对处理器的性能指标有较大的影响。

[47] 累加是一种常见的浮点数运算方式。一般的，在多个浮点数的累加运算中，可以先将两个浮点数进行累加得到中间结果，并将中间结果进行规格化处理；再将规格化处理后的中间结果与下一个浮点数进行累加，直至累加运算结束，得到目标结果。但是，每次累加后均对中间结果进行规格化处理的方式，造成执行浮点数累加运算的处理器功耗和延时较大。同时，由于规格化处理会造成数据丢失，每次对中间结果进行规格化处理，会造成规格化处理后中间结果的精度降低，进而在多次累加运算得到目标结果时，目标结果的精度较低。

[48] 为了缓解上述问题，本公开实施例提供了一种浮点数处理方法。

[49] 应注意到，相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步定义和解释。

[50] 为便于对本公开实施例进行理解，首先对本公开实施例所公开的一种浮点数处理方法进行详细介绍。本公开实施例所提供的浮点数处理方法的执行主体一般为具有一定计算能力的芯片，该芯片比如可以为处理器，该处理器可以包括中央处理器（central processing unit, CPU）、图形处理器（graphics processing unit, GPU）、张量处理单元（Tensor Processing Unit, TPU）、嵌入式神经网络处理器（Neural-network Processing Unit, NPU）等。在一些可能的实现方式中，该浮点数处理方法可以通过处理器调用存储器中存储的计算机可读指令的方式来实现。

[51] 参见图 1 所示，为本公开实施例所提供的浮点数处理方法的流程示意图，所述方法包括 S101-S103。

[52] S101，获取在目标芯片中进行运算的多个待处理浮点数。

[53] S102，将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数；其中，所述目标位宽与所述待处理浮点数的总数量匹配。

[54] S103，在所述目标芯片中将所述多个扩展后浮点数进行累加处理，得到目标浮点数。

[55] 上述方法中，为了保障浮点数累加过程的精度，满足加和运算后得到的中间浮点数的位宽要求，可以将每个待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数，实现了待处理浮点数位宽的扩展，进而使得在目标芯片中对多个扩展后浮点数进行累加处理过程中，无需对累加处理过程中得到的中间浮点数进行规格化处理，减少了规格化处理的次数，降低了目标芯片处理多个扩展后浮点数的累加过程的延迟和功耗；同时，减少了规格化处理的次数后，能够缓解规格化处理带来的数据丢失问题，提高了目标浮点数的精度。

[56] 下述对 S101-S103 进行具体说明。

[57] 在 S101 中，目标芯片可以为任一需要进行浮点数运算的芯片，比如，目标芯片可以为 CPU、GPU、NPU 等。浮点数一般包括符号位、尾数和指数。比如，浮点数 $(0.3)_{10} = (0011\ 1110\ 1001\ 1001\ 1001\ 1001\ 1010)_2$ ，其中，符号位 $Sa=0$ ，指数 $Ea=011\ 1110\ 1$ ，尾数 $Ma=1.001\ 1001\ 1001\ 1001\ 1010$ 。

[58] 在 S102 中，考虑到在进行多个浮点数的累加时，会将进行累加的多个浮点数进行对阶处理，再将对阶处理后的多个浮点数的包含符号位的尾数进行加和处理。在将多个浮点数的包含符号位的尾数进行加和时，加和后得到的和值的位宽大于或等于加和前的每个浮点数的包含符号位的尾数的位宽。比如，若浮点数 1 的包含符号位的尾数的位宽为 12，浮点数 2 的包含符号位的尾数的位宽为 13，则浮点数 1 和浮点数 2 加和后的位宽可能为 14。

[59] 为了保障加和后得到的浮点数的精度，满足加和运算后得到的和值的位宽要求，可以将每个待处理浮点数的符号位进行扩展。其中，扩展后的位宽可以为目标位宽。该目标位宽可以为用户设置的位宽值，该目标位宽大于任一待处理浮点数的尾数对应的初始位宽。或者，该目标位宽也可以为根据各个待处理浮点数的尾数对应的初始位宽和待处理浮点数的总数量确定的。

[60] 实施时，可以确定目标位宽，并将每个待处理浮点数的符号位扩展至目标位宽，得到扩展后浮点数。比如，上述案例中符号位 $Sa=0$ ，扩展后的符号位可以为 $Sa=0000$ 。

[61] 在一种可选实施方式中，所述目标位宽为根据以下步骤确定的：基于所述多个待处理浮点数包括的初始尾数的初始位宽，确定所述多个待处理浮点数对应的调整位宽；基于所述调整位宽和所述待处理浮点数的总数量，确定所述目标位宽。

[62] 实施时，可以确定每个待处理浮点数的初始尾数的初始位宽，并基于多个初始尾数的初始位宽，确定调整位宽。该调整位宽可以为多个初始位宽中的最大值，也可以为大于该最大值的任一整数。下述以调整位宽为多个初始位宽中的最大位宽为例进行说明。

[63] 比如，若待处理浮点数的数量为 5 个，则确定 5 个初始尾数的初始位宽中的最大位宽，将该最大位宽确定为调整位宽；再根据调整位宽和待处理浮点数的总数量，确定目标位宽；即，在最大位宽的基础上，根据待处理浮点数的总数量和待处理浮点数的进制，确定和值的溢出位数，再根据溢出位数和最大位宽，确定目标位宽。比如，目标位宽可以为 $n + \log_a m$ ，其中， n 为调整位宽， m 为待处理浮点数的总数量， a 为待处理浮点数的数制，比如， $a=2$ 时，表示待处理浮点数为二进制数。

[64] 其中，在 $n + \log_a m$ 的值为小数时可以向上取整，得到目标位宽。比如，在 $n + \log_a m = 12 + \log_2 5 = 14.32$ 时，确定该目标位宽可以为 15。若待处理浮点数为 $a=(0.3)_{10}=0\ 011\ 1110\ 1\ 100\ 1100\ 1100\ 1100\ 1100\ 1101$ ，其中，0 为符号位，011 1110 1 为初始指数，100 1100 1100 1100 1101 为初始尾数，则将待处理浮点数的符号位扩展至目标位宽后，该待处理浮点数可以为 0000 011 1110 1 100 1100 1100 1100 1100 1101。

[65] 上述实施方式中，根据调整位宽和待处理浮点数的总数量，较准确的确定目标位宽，后续将待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数后，可以使得扩展后浮点数的位宽能够满足累加处理的位宽要求，进而无需对累加处理过程得到的中间浮点数进行规格化处理。

[66] 在 S103 中，可以将多个扩展后浮点数进行对阶处理，使得多个扩展后浮点数的初始指数一致，再将对阶处理后的多个扩展后浮点数的初始尾数进行累加，得到目标浮点数。实施时，对阶处理可以包括：确定多个扩展后浮点数对应的目标指数，比如，该目标指数可以为多个扩展后浮点数分别对应的初始指数中的最大指数，将每个扩展后浮点数的初始指数对齐至目标指数，并将该扩展后浮点数的尾数右移 x 位，得到对阶处理后的扩展后浮点数。其中， x 为正整数， x 的值为目标指数与该扩展后浮点数的初始指数之差。

[67] 实施时，还可以先将多个扩展后浮点数中的两个扩展后浮点数进行累加处理，将得到的中间浮点数与下一个扩展后浮点数进行累加处理，直至每个扩展后浮点数均被累加，得到目标浮点数。

[68] 在一种可选实施方式中，在 S102 中，所述将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数，可以包括：从所述多个待处理浮点数中选取两个待处理浮点数，并将所述两个待处理浮点数的符号位扩展至所述目标位宽，得到两个扩展后浮点数。

[69] 实施时，可以从多个待处理浮点数中随机选取两个待处理浮点数，或者，也可以设置多个待处理浮点数的处理顺序，按照处理顺序选取两个待处理浮点数；再将选取的两个待处理浮点数的符号位扩展至目标位宽，得到两个扩展后浮点数。

[70] 在得到两个扩展后浮点数之后，S103 可以包括：

[71] S1031，将所述两个扩展后浮点数作为两个当前浮点数，在所述目标芯片中将所述两个当前浮点数进行加和处理，得到中间浮点数；

[72] S1032，从未被选取的待处理浮点数中选取一个待处理浮点数，并将所选取的待处

理浮点数的符号位扩展至所述目标位宽，得到所选取的待处理浮点数对应的扩展后浮点数；

[73] S1033，将所选取的待处理浮点数对应的扩展后浮点数和所述中间浮点数作为更新后的两个当前浮点数，返回至在所述目标芯片中将所述两个当前浮点数进行加和处理，得到所述中间浮点数的步骤，直至所述多个待处理浮点数中的每个待处理浮点数均进行过加和处理；

[74] S1034，将最后一次加和处理后得到的中间浮点数作为所述目标浮点数。

[75] 在 S1031 中，可以将两个扩展后浮点数作为两个当前浮点数，在目标芯片中将两个当前浮点数进行加和处理，得到中间浮点数。

[76] 在一种可选实施方式中，所述在所述目标芯片中将所述两个当前浮点数进行加和处理，得到中间浮点数，可以包括：

[77] 步骤 A1，在所述目标芯片中将所述两个当前浮点数中的第一浮点数的初始指数对齐到第二浮点数的初始指数，并将所述第一浮点数的尾数右移目标位数，得到对阶处理后的第一浮点数；其中，所述两个浮点数中的所述第一浮点数的初始指数小于所述第二浮点数的初始指数；所述目标位数为所述第二浮点数的初始指数与所述第一浮点数的初始指数之间的差值；

[78] 步骤 A2，在所述目标芯片中将所述对阶处理后的第一浮点数与所述第二浮点数进行加和处理，得到中间浮点数。

[79] 实施时，第一浮点数可以为两个当前浮点数中初始指数较小的浮点数；第二浮点数可以为两个当前浮点数中初始指数较大的浮点数。在将两个当前浮点数进行加和之前，可以先对第一浮点数进行对阶处理，得到对阶处理后的第一浮点数，使得对阶处理后的第一浮点数的指数与第二浮点数的初始指数一致。

[80] 第一浮点数的对阶处理的过程可以包括：在目标芯片中将第一浮点数的初始指数对齐至第二浮点数的初始指数，并将第一浮点数的初始尾数右移目标位数，该目标位数为第二浮点数的初始指数与第一浮点数的初始指数之间的差值。比如，若第一浮点数的初始指数为 10，第二浮点数的初始指数为 15，则将第一浮点数的初始指数对齐至 15，并将第一浮点数的初始尾数右移 5 位，得到对阶处理后的第一浮点数，即，对阶处理后的第一浮点数的指数为 15。

[81] 在目标芯片中将对阶处理后的第一浮点数的尾数与第二浮点数的初始尾数进行加和处理，得到中间浮点数。

[82] 在 S1032 中，在得到中间浮点数之后，可以从未被选取的待处理浮点数中选取一个待处理浮点数，所选取的待处理浮点数可以为随机选取的待处理浮点数，也可以为按照处理顺序选取的待处理浮点数；再将所选取的待处理浮点数的符号位扩展至目标位宽，得到所选取的待处理浮点数对应的扩展后浮点数。

[83] 在 S1033 中，可以将所选取的待处理浮点数对应的扩展后浮点数和中间浮点数作为更新后的两个当前浮点数，返回至 S1031 的步骤，直至多个待处理浮点数中的每个待处理浮点数均进行过加和处理为止。

[84] 在 S1034 中，可以将最后一次加和处理后得到的中间浮点数作为目标浮点数。

[85] 这里，可以将多个待处理浮点数进行多次累加，得到目标浮点数，由于在将待处理浮点数进行累加之前，将每个待处理浮点数的符号位扩展至了目标位宽，使得累加过程中得到的中间浮点数的尾数位宽不会超出该目标位宽，进而在累加过程中无需对生成

的中间浮点数进行规格化处理,缓解了对中间浮点数进行规格化处理时造成的数据丢失问题,保障了中间浮点数的精度,进而使得得到的目标浮点数的精度较高。

[86] 在一种可能的实施方式中,在所述得到目标浮点数之后,所述方法还包括:对所述目标浮点数进行规格化处理,得到目标处理结果,其中,所述目标处理结果的格式与预设格式相匹配。

[87] 实施时,对多个扩展后浮点数进行累加处理过程中,无需对累加处理过程中得到的中间浮点数进行规格化处理;而是可以在累加得到目标浮点数后,对目标浮点数进行一次规格化处理,得到目标处理结果,减少了规格化处理的次数,降低了目标芯片处理多个浮点数的累加过程的延迟和功耗;同时,可以缓解规格化处理带来的数据丢失问题,提高了目标处理结果的精度。

[88] 由于在目标芯片中为了保证浮点数表示的唯一性,浮点数在目标芯片中都是以规格化形式存储的,因此,可以在将多个待处理浮点数进行累加处理得到目标浮点数后,对得到的目标浮点数进行规格化处理,得到目标处理结果,该目标处理结果的格式与预设格式相匹配。其中,该预设格式可以为基于设置的浮点数标准确定的。

[89] 在一种可选实施方式中,所述对所述目标浮点数进行规格化处理,得到目标处理结果,可以包括:利用移位方式、截尾方式和舍入方式中的一种或多种方式,对所述目标浮点数进行规格化处理,得到目标处理结果。

[90] 实施时,可以利用移位方式、截尾方式和舍入方式中的一种或多种方式,对目标浮点数进行规格化处理,得到目标处理结果。移位方式可以包括左移、右移等;截尾方式可以为将超出特定位数的数据进行删除的方式等。该目标处理结果可以为规格化处理后的浮点数。

[91] 这里,可以利用移位方式、截尾方式和舍入方式中的一种或多种方式,灵活的对目标浮点数进行规格化处理,得到目标处理结果。

[92] 本领域技术人员可以理解,在具体实施方式的上述方法中,各步骤的撰写顺序并不意味着严格的执行顺序而对实施过程构成任何限定,各步骤的具体执行顺序应当以其功能和可能的内在逻辑确定。

[93] 基于相同的构思,本公开实施例还提供了一种浮点数处理装置。参见图 2 所示,为本公开实施例提供的浮点数处理装置的架构示意图,该浮点数处理装置包括选择器 201 和加法器 202;其中,所述加法器 202 与所述选择器 201 相连。

[94] 所述选择器 201 用于将获取到的多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽,得到多个扩展后浮点数;其中,所述目标位宽与所述待处理浮点数的总数量匹配。

[95] 所述加法器 202 用于将所述多个扩展后浮点数进行累加处理,得到目标浮点数。

[96] 为了保障浮点数累加过程的精度,满足加和运算后得到的中间浮点数的位宽要求,可以利用选择器将每个待处理浮点数的符号位扩展至目标位宽,得到多个扩展后浮点数,实现了待处理浮点数位宽的扩展,进而使得对多个扩展后浮点数进行累加处理过程中,无需对累加处理过程中得到的中间浮点数进行规格化处理,减少了规格化处理的次数,降低了芯片处理多个扩展后浮点数的累加过程的延迟和功耗;同时,减少了规格化处理的次数后,能够缓解规格化处理带来的数据丢失问题,提高了目标浮点数的精度。

[97] 在一种可能的实施方式中,参见图 3 所示,所述装置还包括比较器 203;所述比较器 203 与所述加法器 202 相连;所述加法器 202 还用于将所述目标浮点数输入至所述比较器;所述比较器 203 用于对所述目标浮点数进行规格化处理,得到目标处理结果,其

中, 所述目标处理结果的格式与预设格式相匹配。

[98] 实施时, 对多个扩展后浮点数进行累加处理过程中, 无需对累加处理过程中得到的中间浮点数进行规格化处理; 而是可以在累加得到目标浮点数后, 对目标浮点数进行一次规格化处理, 得到目标处理结果, 减少了规格化处理的次数, 降低了目标芯片处理多个浮点数的累加过程的延迟和功耗; 同时, 可以缓解规格化处理带来的数据丢失问题, 提高了目标处理结果的精度。

[99] 在一种可能的实施方式中, 所述装置还包括对阶运算器 204。所述对阶运算器 204 分别与所述加法器 202 和所述选择器 201 相连; 所述对阶运算器包括减法器 241 和移位器 242。其中, 所述减法器 241 用于确定所述多个扩展后浮点数中的每个扩展后浮点数的初始指数与目标指数之间的目标差值, 所述目标指数为基于多个扩展后浮点数的初始指数确定的; 所述移位器 242 用于针对每个所述扩展后浮点数, 将所述扩展后浮点数的初始指数对齐至所述目标指数, 并基于所述扩展后浮点数对应的目标差值, 对所述扩展后浮点数的尾数进行右移, 得到处理后浮点数, 并将所述处理后浮点数输入至所述加法器; 所述加法器 202 用于将所述多个处理后浮点数的尾数进行累加处理, 得到目标浮点数。

[100] 在一种可能的实施方式中, 所述选择器 201 在将获取到的多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽, 得到多个扩展后浮点数时, 用于: 从所述多个待处理浮点数中选取两个待处理浮点数, 将所述两个待处理浮点数的符号位扩展至目标位宽, 得到两个扩展后浮点数; 所述减法器 241 用于将所述两个扩展后浮点数中的第一浮点数的初始指数作为所述目标指数; 确定所述目标指数与所述两个扩展后浮点数中的第二浮点数的初始指数的第一差值; 其中, 所述第一浮点数的初始指数大于所述第二浮点数的初始指数; 所述移位器 242 用于将所述第二浮点数的初始指数对齐至所述目标指数, 并基于所述第一差值, 将所述第二浮点数的尾数进行右移, 得到处理后的第二浮点数; 将所述第一浮点数和所述处理后的第二浮点数输入至所述加法器; 所述加法器 202, 用于将所述第一浮点数和所述处理后的第二浮点数进行加法处理, 得到中间浮点数。

[101] 在一种可能的实施方式中, 所述装置还包括第一寄存器 205。所述第一寄存器 205 分别与所述加法器 202 和所述对阶运算器的减法器 241 相连; 所述加法器 202 用于将得到的所述中间浮点数输入至所述第一寄存器 205; 所述第一寄存器 205 用于存储所述中间浮点数和所述加法器发送的目标浮点数, 并将所述中间浮点数发送给所述减法器 241。

[102] 在一种可能的实施方式中, 所述装置还包括第二寄存器 206。所述第二寄存器 206 与所述减法器 241 相连; 所述第二寄存器 206 用于存储目标指数, 并将所述目标指数发送给所述对阶运算器中的所述减法器 241。

[103] 在一种可能的实施方式中, 所述选择器 201 还用于: 从未被选取的待处理浮点数中选取一个待处理浮点数, 并将所选取的待处理浮点数的符号位扩展至目标位宽, 得到所选取的待处理浮点数对应的扩展后浮点数; 响应于所选取的待处理浮点数对应的扩展后浮点数的初始指数大于所述第二寄存器存储的目标指数, 将所选取的待处理浮点数对应的扩展后浮点数的初始指数作为新的目标指数, 利用所述新的目标指数对所述第二寄存器进行更新; 所述第二寄存器 206 还用于将新的目标指数发送给减法器; 所述减法器 241 用于基于接收到的新的目标指数, 确定所述中间浮点数的指数与所述目标指数之间的第二差值; 所述移位器 242 用于将所述中间浮点数的指数对齐至所述目标指数, 并基于所述第二差值, 对所述中间浮点数的尾数进行右移, 得到处理后中间浮点数, 并将所述处理后中间浮点数和选取的该待处理浮点数对应的扩展后浮点数发送给所述加法器 202; 所述加法器 202 用于将所述中间浮点数的尾数和所选取的待处理浮点数对应的扩

展后浮点数的尾数进行加法处理，得到新的中间浮点数，并将所述新的中间浮点数发送给所述第一寄存器 205；以及将最后一次加和处理后得到的中间浮点数作为目标浮点数，将所述目标浮点数发送给所述第一寄存器 205。

[104] 在一种可能的实施方式中，所述比较器 203，在对所述目标浮点数进行规格化处理，得到目标处理结果时，用于：利用移位方式、截尾方式和舍入方式中的一种或多种方式，对所述目标浮点数进行规格化处理，得到目标处理结果。

[105] 在一种可能的实施方式中，所述目标位宽为基于所述多个待处理浮点数包括的初始尾数的初始位宽和所述多个待处理浮点数的总数量确定的。

[106] 其中，加法器和第一寄存器的位宽至少为目标位宽。第二寄存器的位宽可以为多个待处理浮点数包括的初始指数的指数位宽中的最大值；即，可以确定多个待处理浮点数中的每个待处理浮点数的初始指数的指数位宽，并将多个指数位宽中的最大值确定为第二寄存器的位宽。

[107] 结合图 3，对上述实施方式提出的浮点数处理方法进行示例性说明。假设多个待处理浮点数包括 a1、a2、a3、a4、a5，该浮点数处理方法可以包括如下步骤。

[108] 步骤一、将进行累加运算的两个待处理浮点数（a1 和 a2）输入至选择器中，选择器将两个待处理浮点数的符号位扩展至目标位宽，得到两个扩展后浮点数，并将两个扩展后浮点数输入至对阶运算器。

[109] 步骤二、对阶运算器对两个扩展后浮点数进行对阶处理，得到指数对齐的两个扩展后浮点数，并将指数对齐的两个扩展后浮点数输入至加法器中；以及将两个扩展后浮点数包括的初始指数中的最大指数作为目标指数，并将该目标指数存储至第二寄存器中。

[110] 实施时，对阶运算器中的减法器确定第一浮点数的初始指数与目标指数之间的目标差值；移位器可以将第一浮点数的初始指数对齐到目标指数，并将第一浮点数的尾数右移目标位数，得到对阶处理后的第一浮点数。该对阶处理后的第一浮点数和第二浮点数的指数一致。其中，两个扩展后浮点数中，第一浮点数的初始指数小于第二浮点数的初始指数；目标位数为第二浮点数的初始指数（目标指数）与第一浮点数的初始指数之间的目标差值。这里，可以将第二浮点数的初始指数作为目标指数存储至第二寄存器中。

[111] 步骤三、加法器将指数对齐的两个扩展后浮点数的尾数进行加和运算，得到中间浮点数，并将该中间浮点数寄存在第一寄存器。

[112] 步骤四、将待处理浮点数 a3 输入至选择器中，选择器将待处理浮点数 a3 的符号位扩展至目标位宽，得到扩展后浮点数，并将扩展后浮点数输入至对阶运算器；以及第一寄存器将中间浮点数输入至对阶运算器。

[113] 步骤五、对阶运算器对待处理浮点数 a3 对应的扩展后浮点数和中间浮点数进行对阶处理，得到指数对齐的两个扩展后浮点数，并将指数对齐的两个扩展后浮点数输入至加法器中。

[114] 实施时，判断待处理浮点数 a3 对应的扩展后浮点数中初始指数与中间浮点数的指数（即，第二寄存器中存储的目标指数）之间的大小，若扩展后浮点数中初始指数小于中间浮点数的指数，则将扩展后浮点数的初始指数对齐到中间浮点数的指数，并对扩展后浮点数的尾数右移目标位数（即，中间浮点数与扩展后浮点数的指数之差），得到对阶处理后的扩展后浮点数。该对阶处理后的扩展后浮点数和中间浮点数的指数一致。

[115] 若扩展后浮点数中初始指数大于中间浮点数的指数，则将中间浮点数的初始指数对阶到扩展后浮点数的指数，并对中间浮点数的尾数右移目标位数（即，扩展后浮点数与中间浮点数的指数之差），得到对阶处理后的中间浮点数。该对阶处理后的中间浮点

数和扩展后浮点数为指数对齐的两个扩展后浮点数。

[116] 在扩展后浮点数中初始指数与中间浮点数的指数相同时，可以无需进行对阶处理，加法器直接将扩展后浮点数的尾数与中间浮点数的尾数进行相加，得到中间浮点数。

[117] 步骤六、在待处理浮点数 a3 对应的扩展后浮点数中初始指数大于中间浮点数的指数时，可以将扩展后浮点数中的初始指数作为更新后的目标指数，并将该更新后的目标指数更新至第二寄存器。

[118] 步骤七、加法器将指数对齐的两个扩展后浮点数的尾数进行加和运算，得到中间浮点数，并将该中间浮点数寄存在第一寄存器。

[119] 分别获取待处理浮点数 a4、a5，并针对获取到的待处理浮点数 a4 和 a5 分别重复步骤四至步骤七，得到 a1、a2、a3、a4、a5 累加后的目标浮点数。

[120] 步骤八，在得到目标浮点数之后，第一寄存器可以将目标浮点数输入至比较器。比较器对目标浮点数进行规格化处理，得到目标处理结果，其中，目标处理结果的格式与预设格式相匹配。

[121] 基于相同的构思，本公开实施例提供了一种芯片。参见图 4 所示，该芯片包括：上述任一实施方式所述的浮点数处理装置 401 和存储器 402。所述存储器 402 用于存储待运算的多个待处理浮点数；所述浮点数处理装置 401 用于对所述多个待处理浮点数进行处理，得到目标浮点数。

[122] 基于相同的构思，本公开实施例还提供了一种浮点数处理装置。参见图 5 所示，为本公开实施例提供的浮点数处理装置的架构示意图，该浮点数处理装置包括获取模块 501、扩展模块 502 和第一处理模块 503。

[123] 获取模块 501 用于获取在目标芯片中进行运算的多个待处理浮点数。

[124] 扩展模块 502 用于将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数；其中，所述目标位宽与所述待处理浮点数的总数量匹配。

[125] 第一处理模块 503 用于在所述目标芯片中将所述多个扩展后浮点数进行累加处理，得到目标浮点数。

[126] 在一种可能的实施方式中，在所述得到目标浮点数之后，所述装置还包括：第二处理模块 504，用于对所述目标浮点数进行规格化处理，得到目标处理结果，其中，所述目标处理结果的格式与预设格式相匹配。

[127] 在一种可能的实施方式中，所述扩展模块 502 用于根据以下步骤确定目标位宽：基于所述多个待处理浮点数包括的初始尾数的初始位宽，确定所述多个待处理浮点数对应的调整位宽；基于所述调整位宽和所述待处理浮点数的总数量，确定所述目标位宽。

[128] 在一种可能的实施方式中，所述扩展模块 502，在将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数时，用于：从多个待处理浮点数中选取两个待处理浮点数，将所述两个待处理浮点数的符号位扩展至目标位宽，得到两个扩展后浮点数；所述第一处理模块 503，在所述目标芯片中将所述多个扩展后浮点数进行累加处理，得到目标浮点数时，用于：将所述两个扩展后浮点数作为两个当前浮点数，在所述目标芯片中将所述两个当前浮点数进行加和处理，得到中间浮点数；从未被选取的待处理浮点数中选取一个待处理浮点数，并将所选取的待处理浮点数的符号位扩展至目标位宽，得到所选取的待处理浮点数对应的扩展后浮点数；将所选取的待处理浮点数对应的扩展后浮点数和所述中间浮点数，作为更新后的两个当前浮

点数，返回至在所述目标芯片中将所述两个当前浮点数进行加和处理，得到中间浮点数的步骤，直至所述多个待处理浮点数中的每个待处理浮点数均进行过加和处理；将最后一次加和处理后得到的中间浮点数作为所述目标浮点数。

[129] 在一种可能的实施方式中，所述第一处理模块 503，在所述目标芯片中将所述两个当前浮点数进行加和处理，得到中间浮点数时，用于：在所述目标芯片中将所述两个当前浮点数中的第一浮点数的初始指数对齐到第二浮点数的初始指数，并将所述第一浮点数的尾数右移目标位数，得到对阶处理后的第一浮点数；其中，所述两个浮点数中的所述第一浮点数的初始指数小于所述第二浮点数的初始指数；所述目标位数为所述第二浮点数的初始指数与所述第一浮点数的初始指数之间的差值；在所述目标芯片中将所述对阶处理后的第一浮点数与所述第二浮点数进行加和处理，得到中间浮点数。

[130] 在一种可能的实施方式中，所述第二处理模块 504，在对所述目标浮点数进行规格化处理，得到目标处理结果时，用于：利用移位方式、截尾方式和舍入方式中的一种或多种方式，对所述目标浮点数进行规格化处理，得到目标处理结果。

[131] 在一些实施例中，本公开实施例提供的装置具有的功能或包含的模板可以用于执行上文方法实施例描述的方法，其具体实现可以参照上文方法实施例的描述，为了简洁，这里不再赘述。

[132] 基于同一技术构思，本公开实施例还提供了一种电子设备。参照图 6 所示，为本公开实施例提供的电子设备的结构示意图，该电子设备包括处理器 601、存储器 602 和总线 603。

[133] 其中，存储器 602 用于存储执行指令，包括内存 6021 和外部存储器 6022。这里的内存 6021 也称内存储器，用于暂时存放处理器 601 中的运算数据，以及与硬盘等外部存储器 6022 交换的数据，处理器 601 通过内存 6021 与外部存储器 6022 进行数据交换，当电子设备 600 运行时，处理器 601 与存储器 602 之间通过总线 603 通信，使得处理器 601 在执行以下指令：

[134] 获取在目标芯片中进行运算的多个待处理浮点数；将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数；其中，所述目标位宽与所述待处理浮点数的总数量匹配；在所述目标芯片中将所述多个扩展后浮点数进行累加处理，得到目标浮点数。

[135] 其中，处理器 601 的具体处理流程可以参照上述方法实施例的记载，这里不再赘述。

[136] 此外，本公开实施例还提供一种计算机可读存储介质，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器运行时执行上述方法实施例中所述的浮点数处理方法的步骤。其中，该存储介质可以是易失性或非易失的计算机可读取存储介质。

[137] 本公开实施例还提供一种计算机程序产品，该计算机程序产品承载有程序代码，所述程序代码包括的指令可用于执行上述方法实施例中所述的浮点数处理方法的步骤，具体可参见上述方法实施例，在此不再赘述。

[138] 其中，上述计算机程序产品可以具体通过硬件、软件或其结合的方式实现。在一个可选实施例中，所述计算机程序产品具体体现为计算机存储介质，在另一个可选实施例中，计算机程序产品具体体现为软件产品，例如软件开发包(Software Development Kit, SDK) 等等。

[139] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的装置和装置的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。在本

公开所提供的几个实施例中，应该理解到，所揭露的装置、装置和方法，可以通过其它的方式实现。以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，又例如，多个单元或组件可以结合或者可以集成到另一个装置，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些通信接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[140] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即，可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[141] 另外，在本公开各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

[142] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个处理器可执行的非易失的计算机可读取存储介质中。基于这样的理解，本公开的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本公开各个实施例所述方法的全部或部分步骤。而前述的存储介质包括U盘、移动硬盘、只读存储器（Read-Only Memory, ROM）、随机存取存储器（Random Access Memory, RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

[143] 以上仅为本公开的具体实施方式，但本公开的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本公开揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本公开的保护范围之内。因此，本公开的保护范围应以权利要求的保护范围为准。

权利要求书

1、一种浮点数处理装置，其特征在于，所述装置包括：

选择器，用于将获取到的多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数；其中，所述目标位宽与所述待处理浮点数的总数量匹配；

加法器，与所述选择器相连，用于将所述多个扩展后浮点数进行累加处理，得到目标浮点数。

2、根据权利要求1所述的装置，其特征在于，所述装置还包括：

比较器，与所述加法器相连，用于对从所述加法器输入的所述目标浮点数进行规格化处理，得到目标处理结果，其中，所述目标处理结果的格式与预设格式相匹配。

3、根据权利要求1或2所述的装置，其特征在于，所述装置还包括对阶运算器，所述对阶运算器分别与所述加法器和所述选择器相连；所述对阶运算器包括：

减法器，用于确定每个所述扩展后浮点数的初始指数与目标指数之间的目标差值；其中，所述目标指数为基于所述多个扩展后浮点数的初始指数确定的；

移位器，用于针对每个所述扩展后浮点数，

将所述扩展后浮点数的初始指数对齐至所述目标指数；

基于所述扩展后浮点数对应的目标差值，对所述扩展后浮点数的尾数进行右移，得到处理后浮点数；并

将所述处理后浮点数输入至所述加法器；

其中，所述加法器用于将从所述多个扩展后浮点数分别得到的所述处理后浮点数的尾数进行累加处理，得到所述目标浮点数。

4、根据权利要求3所述的装置，其特征在于，

所述选择器，在将获取到的多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数时，用于：

从所述多个待处理浮点数中选取两个待处理浮点数，并将所述两个待处理浮点数的符号位扩展至所述目标位宽，得到两个扩展后浮点数；

所述减法器用于：

将所述两个扩展后浮点数中的第一浮点数的初始指数作为所述目标指数；

确定所述目标指数与所述两个扩展后浮点数中的第二浮点数的初始指数的第一差值；其中，所述第一浮点数的初始指数大于所述第二浮点数的初始指数；

所述移位器用于：

将所述第二浮点数的初始指数对齐至所述目标指数；

基于所述第一差值，将所述第二浮点数的尾数进行右移，得到处理后的第二浮点数；

将所述第一浮点数和所述处理后的第二浮点数输入至所述加法器；

所述加法器用于：

将所述第一浮点数和所述处理后的第二浮点数进行加法处理，得到中间浮点数。

5、根据权利要求4所述的装置，其特征在于，所述装置还包括：

第一寄存器，分别与所述加法器和所述对阶运算器中的所述减法器相连，用于存储由所述加法器得到并输入的所述中间浮点数和所述加法器发送的目标浮点数，并将所述中间浮点数发送给所述减法器。

6、根据权利要求5所述的装置，其特征在于，所述装置还包括：

第二寄存器，与所述减法器相连，用于存储所述目标指数，并将所述目标指数发送给所述对阶运算器中的所述减法器。

7、根据权利要求6所述的装置，其特征在于，

所述选择器还用于：

从未被选取的待处理浮点数中选取一个待处理浮点数，并将所选取的待处理浮点数的符号位扩展至所述目标位宽，得到所选取的待处理浮点数对应的扩展后浮点数；

响应于所选取的待处理浮点数对应的扩展后浮点数的初始指数大于所述第二寄存器存储的目标指数，将所选取的待处理浮点数对应的扩展后浮点数的初始指数作为新的目标指数；

利用所述新的目标指数对所述第二寄存器进行更新；其中，所述第二寄存器还用于将所述新的目标指数发送给所述减法器；

所述减法器用于：

基于接收到的新的目标指数，确定所述中间浮点数的指数与所述目标指数之间的第二差值；

所述移位器用于：

将所述中间浮点数的指数对齐至所述目标指数，并基于所述第二差值对所述中间浮点数的尾数进行右移，得到处理后中间浮点数；

将所述处理后中间浮点数和所选取的待处理浮点数对应的扩展后浮点数发送给所述加法器；

所述加法器用于：

将所述中间浮点数的尾数和所选取的待处理浮点数对应的扩展后浮点数的尾数进行加法处理，得到新的中间浮点数，并将所述新的中间浮点数发送给所述第一寄存器；以及

将最后一次加和处理后得到的中间浮点数作为所述目标浮点数，并将所述目标浮点数发送给所述第一寄存器。

8、根据权利要求2所述的装置，其特征在于，所述比较器，在对所述目标浮点数进行规格化处理，得到目标处理结果时，用于：

利用移位方式、截尾方式和舍入方式中的一种或多种方式，对所述目标浮点数进行规格化处理，得到所述目标处理结果。

9、根据权利要求1至8任一项所述的装置，其特征在于，所述目标位宽为基于所述多个待处理浮点数包括的初始尾数的初始位宽和所述多个待处理浮点数的总数量确定的。

10、一种浮点数处理方法，其特征在于，包括：

获取在目标芯片中进行运算的多个待处理浮点数；

将所述多个待处理浮点数中的每个所述待处理浮点数的符号位扩展至目标位宽，得到多个扩展后浮点数；其中，所述目标位宽与所述待处理浮点数的总数量匹配；

在所述目标芯片中将所述多个扩展后浮点数进行累加处理，得到目标浮点数。

11、一种芯片，其特征在于，包括：

存储器，用于存储待运算的多个待处理浮点数；

如权利要求 1 至 9 任一项所述的浮点数处理装置，用于对所述多个待处理浮点数进行处理，得到目标浮点数。

12、一种电子设备，其特征在于，包括：处理器、存储器和总线，所述存储器存储有所述处理器可执行的机器可读指令，当所述电子设备运行时，所述处理器与所述存储器之间通过所述总线通信，所述机器可读指令被所述处理器执行时实现如权利要求 10 所述的浮点数处理方法的步骤。

13、一种电子设备，其特征在于，包括如权利要求 11 所述的芯片。

14、一种计算机可读存储介质，其特征在于，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器运行时执行如权利要求 10 所述的浮点数处理方法的步骤。

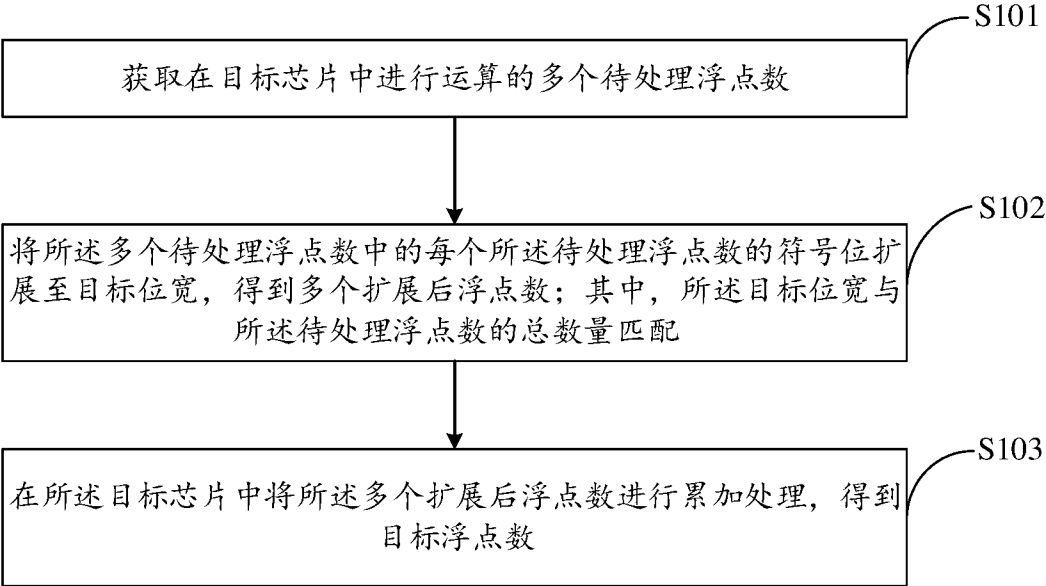


图 1

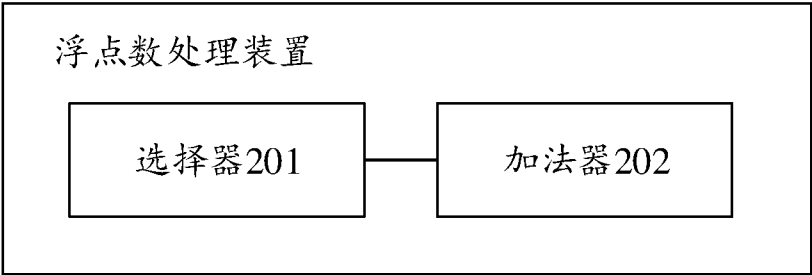


图 2

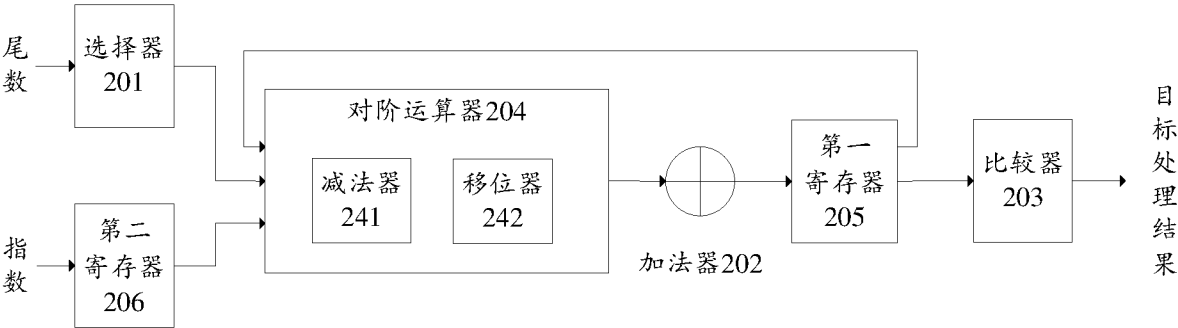


图 3

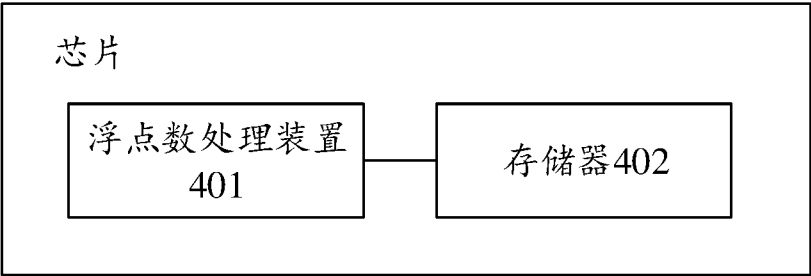


图 4

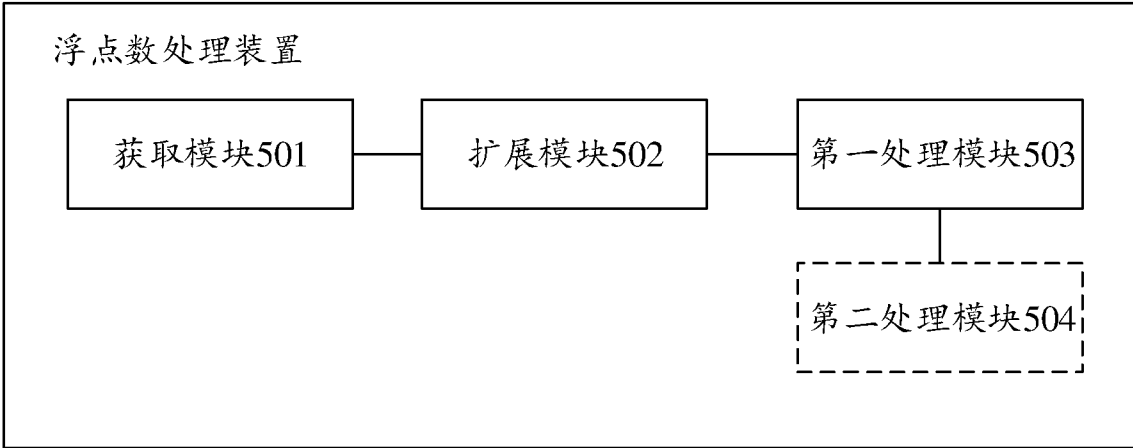


图 5

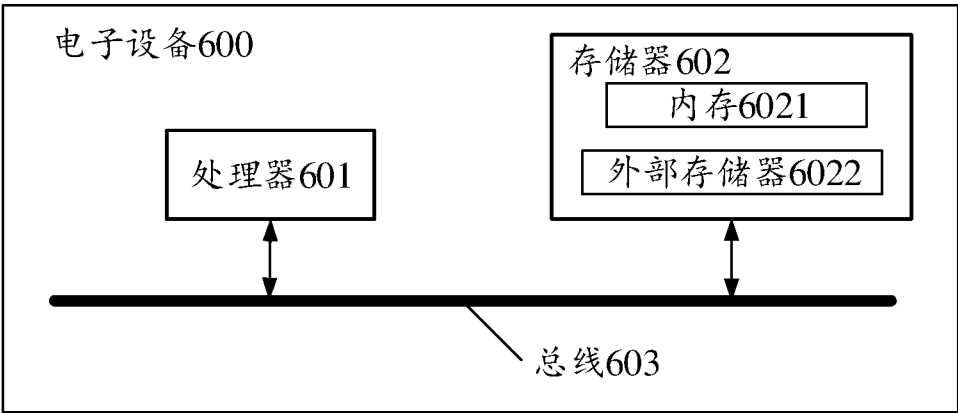


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/124517

A. CLASSIFICATION OF SUBJECT MATTER

G06F 7/485(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, CNABS, DWPI, ENTXT, ENTXTC: 浮点数处理装置, 选择器, 加法器, 扩展, 目标位宽, 累加处理, floating point number process device, selector, adder, expand, target bit width, accumulation, process

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 108255777 A (INSTITUTE OF ELECTRONICS, CHINESE ACADEMY OF SCIENCES) 06 July 2018 (2018-07-06) description, paragraphs [0035]-[0053]	1-14
PX	CN 114296682 A (SHANGHAI ZHENLIANG INTELLIGENT TECHNOLOGY CO., LTD.) 08 April 2022 (2022-04-08) claims, and entire description	1-14
A	US 2018067721 A1 (ADVANCED RISC MACH LTD.) 08 March 2018 (2018-03-08) entire description	1-14



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

06 December 2022

Date of mailing of the international search report

28 December 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/124517

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	108255777	A	06 July 2018	CN	108255777	B	06 August 2021
CN	114296682	A	08 April 2022	None			
US	2018067721	A1	08 March 2018	US	10061561	B2	28 August 2018

国际检索报告

国际申请号

PCT/CN2022/124517

A. 主题的分类

G06F 7/485 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNTXT, CNABS, DWPI, ENTXT, ENTXTC:浮点数处理装置, 选择器, 加法器, 扩展, 目标位宽, 累加处理, floating point number process device, selector, adder, expand, target bit width, accumulation, process

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 108255777 A (中国科学院电子学研究所) 2018年7月6日 (2018 - 07 - 06) 说明书第[0035]段—第[0053]段	1-14
PX	CN 114296682 A (上海阵量智能科技有限公司) 2022年4月8日 (2022 - 04 - 08) 权利要求书, 说明书全文	1-14
A	US 2018067721 A1 (ADVANCED RISC MACH LTD) 2018年3月8日 (2018 - 03 - 08) 说明书全文	1-14

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2022年12月6日

国际检索报告邮寄日期

2022年12月28日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

唐田田

电话号码 62412065

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/124517

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	108255777	A	2018年7月6日	CN 108255777 B	2021年8月6日
CN	114296682	A	2022年4月8日	无	
US	2018067721	A1	2018年3月8日	US 10061561 B2	2018年8月28日