(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
2024 年 4 月 25 日 (25.04.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/082674 A1

(51) 国际专利分类号:
G06F 7/483 (2006.01)

(21) 国际申请号: PCT/CN2023/102089

(22) 国际申请日: 2023 年 6 月 25 日 (25.06.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202211281416.6 2022年10月19日 (19.10.2022) 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: 罗元勇 **(LUO, Yuanyong)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 陈敏琪**(CHEN, Minqi)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 张忠星**(ZHANG, Zhongxing)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 伍玮翔 **(WU, Wei Hsiang)**; 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: 北京中博世达专利商标代理有限公司 **(BEIJING ZBSD PATENT & TRADEMARK AGENT LTD.)**; 中国北京市海淀区交大东路 31 号 11 号楼8层, Beijing 100044 (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,

(54) **Title:** FLOATING-POINT DATA PRECISION CONVERSION METHOD AND APPARATUS

(54) 发明名称: 浮点数据精度转换方法和装置

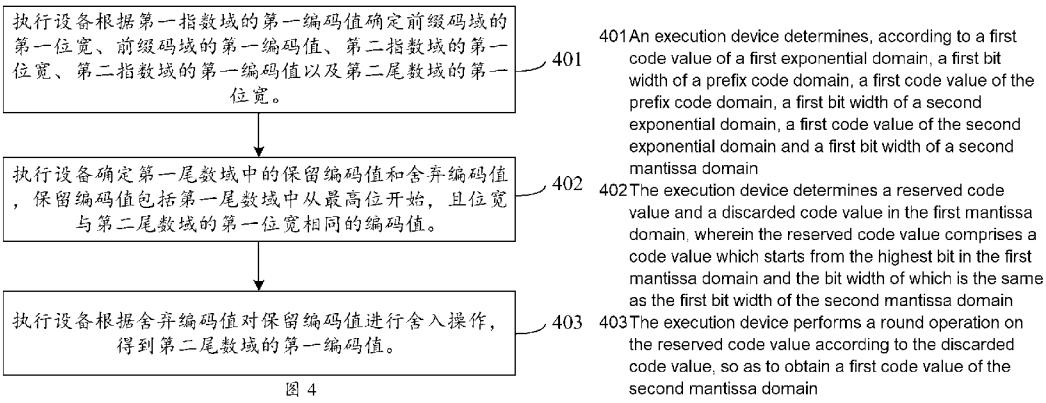

图 4

(57) **Abstract:** The embodiments of the present application relate to the technical field of chips. Provided are a floating-point data precision conversion method and apparatus, which improve the invariance of the overall mean value when high-precision data is converted into low-precision data. The specific solution is: according to a first code value of a first exponential domain, determining a first bit width of a prefix code domain, a first code value of the prefix code domain, a first bit width of a second exponential domain, a first code value of the second exponential domain and a first bit width of a second mantissa domain; determining a reserved code value and a discarded code value in the first mantissa domain, wherein the reserved code value comprises a code value which starts from the highest bit in the first mantissa domain and the bit width of which is the same as the first bit width of the second mantissa domain; and performing a round operation on the reserved code value according to the discarded code value, so as to obtain a first code value of the second mantissa domain. The embodiments of the present application are applied to the process of converting high-precision data into low-precision data.

WO 2024/082674 A1

[见续页]

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))。

(57) 摘要: 本申请实施例提供一种浮点数据精度转换方法和装置，涉及芯片技术领域，提高了高精度数据向低精度数据转换时的整体均值不变性。具体方案为：根据第一指数域的第一编码值确定前缀码域的第一位宽、前缀码域的第一编码值、第二指数域的第一位宽、第二指数域的第一编码值以及第二尾数域的第一位宽；确定第一尾数域中的保留编码值和舍弃编码值，保留编码值包括第一尾数域中从最高位开始，且位宽与第二尾数域的第一位宽相同的编码值；根据舍弃编码值对保留编码值进行舍入操作，得到第二尾数域的第一编码值。本申请实施例用于高精度数据向低精度数据转换的过程。

说 明 书

浮点数据精度转换方法和装置

本申请要求于 2022 年 10 月 19 日提交国家知识产权局、申请号为 202211281416.6、申请名称为“浮点数据精度转换方法和装置”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

本申请涉及芯片技术领域，尤其涉及一种浮点数据精度转换方法和装置。

背景技术

随着混合精度计算的快速发展，开始大规模的部署低精度浮点数据格式的计算资源。比如在人工智能（Artificial Intelligence，AI）领域部署有浮点（Floating Point，FP）16 和 FP32 混合训练模式，脑浮点（Brain Floating Point，BF）16 和 FP32 混合训练模式，以及高性能计算（High Performance Computing，HPC）业务领域的 FP32 和 FP64 混合精度等。目前，学界和业界提出了多种 8 比特浮点数据格式，比如共享指数位（Shard Exponent Bias，SEB）、混合 FP8（Hybrid FP8，HFP8）以及可配置的 FP8（Configurable FP8，CFP8）等。对于精度要求较高的 HPC 业务领域也想部署大规模的低精度算力，于是提出了多种混合精度的求解器算法。这些算法中先利用低精度算力，如 FP16/BF16，计算出低精度的初始计算结果，然后使用迭代算法和高精度数据格式 FP32/FP64，求解出高精度的计算结果。针对混合精度计算场景，需要涉及到不同精度数据间的相互转换。低精度数据向高精度数据进行格式转换时，可以实现无误差的转换。高精度数据向低精度数据进行格式转换时，需要对高精度数据进行舍入（round）操作，由此会产生转换误差，降低了高精度数据向低精度数据转换时的整体均值不变性。不同的舍入方式对应的转换误差也不相同，特别是 AI 训练场景，对高精度数据进行舍入操作会出现误差的累积效应，从而影响 AI 模型的训练精度。

发明内容

本申请实施例提供一种浮点数据精度转换方法和装置，实现了高精度数据向低精度数据的转换，第二浮点数据采用前缀码域指示第二指数域的位宽，有效平衡了第二浮点数据位宽、范围和精度之间的关系。并通过提供简单的舍入方式，根据第一尾数域中的舍弃编码值对保留编码值进行舍入操作，无需其他设备的支持，提高了高精度数据向低精度数据转换的转换效率，降低了硬件开销。

为达到上述目的，本申请实施例采用如下技术方案。

第一方面，本申请实施例提供一种浮点数据精度转换方法，第一浮点数据包括符号域、第一指数域和第一尾数域，第二浮点数据包括符号域、前缀码域、第二指数域和第二尾数域，前缀码域用于指示第二指数域的位宽，第一浮点数据的精度高于第二浮点数据的精度，该方法包括：根据第一指数域的第一编码值确定前缀码域的第一位宽、前缀码域的第一编码值、第二指数域的第一位宽、第二指数域的第一编码值以及第二尾数域的第一位宽；确定第一尾数域中的保留编码值和舍弃编码值，保留编码值包括第一尾数域中从最高位开始，且位宽域第二尾数域的第一位宽相同的编码值；根据舍弃编码值对保留编码值进行舍入操作，得到第二尾数域的第一编码值。

本申请实施例提供的浮点数据精度转换方法，实现了将高精度数据转换为低精度数据。在数据格式转换中，基于第一浮点数据的符号域可以得到第二浮点数据的符号域，基于第一浮点数据的第一指数域可以得到第二浮点数据的前缀码域和第二指数域，以及基于第一浮点数据的第一尾数域可以得到第二浮点数据的第二尾数域。第二浮点数据中通过较短的前缀码域指示第二指数域的位宽，可以有效提升第二浮点数据的尾数的精度或位宽，同时对于只提供 1 位尾数的精度的第二浮点数据可以表示较大的数值范围，有效平衡了第二浮点数据位宽、范围和精度之间的关系。且前缀码域可以采用前缀码编码方式，占用位宽少，解析第二指数域和第二尾数域便捷。并通过提供简单的舍入方式，根据第一尾数域中的舍弃编码值对保留编码值进行舍入操作，无需其他设备的支持，提高了高精度数据向低精度数据转换的转换效率，降低了硬件开销。

在一种可能的设计中，舍入操作包括进位操作和舍弃操作，根据舍弃编码值对保留编码值进行舍入操作，得到第二尾数域的第一编码值包括：舍弃编码值中从最高位开始，且位宽为预设位宽的编码值大于或等于第二预设阈值时，对保留编码值的最低位进行进位操作，对舍弃编码值进行舍弃操作，保留编码值进位后的编码值为第二尾数域的第一编码值；舍弃编码值中从最高位开始，且位宽为预设位宽的编码值小于第二预设阈值时，对舍弃编码值进行舍弃操作，保留编码值为第二尾数

域的第一编码值；其中，第二预设阈值为舍弃编码值中从最低位开始，且位宽为预设位宽的编码值。

这种设计中，对于随机舍入方式，用于比较的第二预设阈值为舍弃编码值中从最低位开始，且位宽为预设位宽的编码值，第二预设阈值的生成无需额外的随机数生成器，不存在随机数生成的性能瓶颈，提高了高精度数据向低精度数据的转换效率，同时硬件开销更低。

在一种可能的设计中，舍入操作包括进位操作或舍弃操作，根据舍弃编码值对保留编码值进行舍入操作，得到第二尾数域的第一编码值包括：舍弃编码值的最高位大于或等于第一预设阈值时，对保留编码值的最低位进行进位操作，并对舍弃编码值进行舍弃操作，保留编码值进行进位操作后得到的编码值为第二尾数域的第一编码值；舍弃编码值的最高位小于第一预设阈值时，对舍弃编码值进行舍弃操作，保留编码值为第二尾数域的第一编码值。

这种设计中，第一预设阈值可以为 0 或 1，将舍弃编码值的最高位和第一预设阈值进行比较，属于远离 0 进位舍入方式。除了远离 0 进位舍入方式，还可以包括远离偶数进位舍入方式和远离奇数进位舍入方式等。但远离 0 进位舍入方式对于其他舍入方式硬件实现面积更小、功耗开销更小，且具有更高的数据分辨率。

在一种可能的设计中，本申请实施例提供的浮点数据精度转换方法还包括：判断进位操作后的保留编码值是否溢出；若进位操作后的保留编码值溢出，则对第一指数域的第一编码值的最低位进行加 1 操作，得到第一指数域的第二编码值；若前缀码域的第二位宽和前缀码域的第一位宽不同，根据第一指数域的第二编码值确定前缀码域的第二编码值、第二指数域的第二编码值、第二尾数域的第二位宽和第二尾数域的第二编码值；若前缀码域的第二位宽和前缀码域的第一位宽相同，判断第二指数域的第一位宽和第二指数域的第二位宽是否相同；若第二指数域的第二位宽小于第二指数域的第一位宽，对保留编码值的位宽进行加 1 操作，得到第二尾数域的第二位宽和第二尾数域的第二编码值；若第二指数域的第二位宽大于或等于第二指数域的第一位宽，对保留编码值的最低位进行舍弃操作，得到第二尾数域的第二位宽和第二尾数域的第二编码值。

这种设计中，在溢出的情况下，对第一指数域的第一编码值的最低位进行加 1 操作之后得到前缀码域的第二位宽和第二指数域的位宽，如果前缀码域的第二位宽和前缀码域的第一位宽相同时，如果第二指数域的位宽变化，可以得到第二尾数域的第二位宽，可以解决保留编码值进行进位操作后产生的溢出的问题。

在一种可能的设计中，根据第一指数域的第一编码值确定前缀码域的第一位宽、前缀码域的第一编码值、第二指数域的第一位宽和第二指数域的第一编码值包括：根据第一指数域的第一编码值确定指示值，通过查表确定与指示值对应的前缀码域的第一位宽和前缀码域的第一编码值，指示值还用于指示第二指数域的第一位宽；根据第一指数域的第一编码值确定第二指数域的第一位宽对应的第一编码值。

在一种可能的设计中，根据第一指数域的第一编码值确定第二尾数域的第一位宽包括：根据第二浮点数据的总位宽、前缀码域的第一位宽、第二指数域的第一位宽确定第二尾数域的第一位宽。

这种设计中，例如，对于 HiFloat8 数据格式的第二浮点数据，符号域的位宽为 1，前缀码域的位宽为 2 或 3，第二指数域的第一位宽为 0 至 4 中的一个整数，第二尾数域的第一位宽为 1 至 4 中的一个整数。由此，第二浮点数据中采用较短的前缀码域指示第二指数域的第一位宽，使得第二浮点数据最高可以提供 4 位尾数的精度，同时对于只提供 1 位尾数的精度的第二浮点数据可以表示较大的数值范围，有效平衡了第二浮点数据位宽、范围和精度之间的关系。且第二指数域存储时隐藏最高位，减少了第二指数域需要存储的第一位宽，有效避免了不同前缀码域的指示值对应的第二指数域的第一编码值出现数值重叠的问题，使得 HiFloat8 数据格式中无冗余编码。

在一种可能的设计中，第一浮点数据超出第二浮点数据的数据范围的上限时，基于饱和方式或无穷大方式确定第二浮点数据；第一浮点数据超出第二浮点数据的数据范围的下限时，第二浮点数据为零；第一浮点数据为非数字值时，第二浮点数据为非数字值。

这种设计中，第一浮点数据超出第二浮点数据的数据范围的上限和下限时，第二浮点数据可以通过特殊值表示第一浮点数据，例如饱和值、无穷大值和零值。当第一浮点数据为非数字值时，第二浮点数据也为非数字值表示。

第二方面，本申请实施例提供一种浮点数据精度转换装置，第一浮点数据包括符号域、第一指数域和第一尾数域，第二浮点数据包括符号域、前缀码域、第二指数域和第二尾数域，前缀码域用于指示第二指数域的位宽，第一浮点数据的精度高于第二浮点数据的精度，该装置包括：位宽计算

单元，用于根据第一指数域的第一编码值确定前缀码域的第一位宽、前缀码域的第一编码值、第二指数域的第一位宽、第二指数域的第一编码值以及第二尾数域的第一位宽；尾数域计算单元，用于确定第一尾数域中的保留编码值和舍弃编码值，保留编码值包括第一尾数域中从最高位开始，且位宽与第二尾数域的第一位宽相同的编码值；舍入操作单元，用于根据舍弃编码值对保留编码值进行舍入操作，得到第二尾数域的第一编码值。

第二方面的有益效果可参见第一方面的说明。

在一种可能的设计中，舍入操作包括进位操作或舍弃操作，舍入操作单元还用于：舍弃编码值中从最高位开始，且位宽为预设位宽的编码值大于或等于第二预设阈值时，对保留编码值的最低位进行进位操作，对舍弃编码值进行舍弃操作，保留编码值进位后的编码值为第二尾数域的第一编码值；舍弃编码值中从最高位开始，且位宽为预设位宽的编码值小于第二预设阈值时，对舍弃编码值进行舍弃操作，保留编码值为第二尾数域的第一编码值；其中，第二预设阈值为舍弃编码值中从最低位开始，且位宽为预设位宽的编码值。

在一种可能的设计中，舍入操作包括进位操作或舍弃操作，舍入操作单元还用于：舍弃编码值的最高位大于或等于第一预设阈值时，对保留编码值的最低位进行进位操作，并对舍弃编码值进行舍弃操作，保留编码值进行进位操作后得到的编码值为第二尾数域的第一编码值；舍弃编码值的最高位小于第一预设阈值时，对舍弃编码值进行舍弃操作，保留编码值为第二尾数域的第一编码值。

在一种可能的设计中，装置还包括：溢出单元，用于判断进位操作后的保留编码值是否溢出；位宽计算单元还用于若进位操作后的保留编码值溢出，则对第一指数域的第一编码值进行加 1 操作，得到第一指数域的第二编码值；根据第一指数域的第二编码值确定第二指数域的第二位宽和前缀码域的第二位宽；若前缀码域的第二位宽和前缀码域的第一位宽不同，根据第一指数域的第二编码值确定前缀码域的第二编码值、第二指数域的第二编码值、第二尾数域的第二位宽和第二尾数域的第二编码值；若所述前缀码域的第二位宽和所述前缀码域的第一位宽相同，判断所述第二指数域的第一位宽和第二指数域的第二位宽是否相同；若第二指数域的第二位宽小于第二指数域的第一位宽，对保留编码值的位宽进行加 1 操作，得到第二尾数域的第二位宽和第二尾数域的第二编码值；若第二指数域的第二位宽大于或等于第二指数域的第一位宽，对保留编码值的最低位进行舍弃操作，得到第二尾数域的第二位宽和第二尾数域的第二编码值。

在一种可能的设计中，位宽计算单元还用于：根据第一指数域的第一编码值确定指示值，通过查表确定与指示值对应的前缀码域的第一位宽和前缀码域的第一编码值，指示值还用于指示第二指数域的第一位宽；根据第一指数域的第一编码值确定第二指数域的第一位宽对应的第一编码值。

在一种可能的设计中，位宽计算单元还用于：根据第二浮点数据的总位宽、前缀码域的第一位宽、第二指数域的第一位宽确定第二尾数域的第一位宽。

在一种可能的设计中，位宽计算单元还用于：第一浮点数据超出第二浮点数据的转换范围的上限时，基于饱和方式或无穷大方式确定第二浮点数据；第一浮点数据超出第二浮点数据的转换范围的下限时，第二浮点数据为零；第一浮点数据为非数字值时，第二浮点数据为非数字值。

第三方面，提供一种通信装置，包括至少一个处理器，所述至少一个处理器与存储器相连，所述至少一个处理器用于读取并执行所述存储器中存储的程序，以使得所述装置执行如上述第一方面或第一方面的任一项所述的方法。

第四方面，提供一种芯片，所述芯片与存储器耦合，用于读取并执行所述存储器中存储的程序指令，以实现如上述第一方面或第一方面的任一项所述的方法。

第五方面，本申请提供一种芯片系统，该芯片系统应用于云中心。该芯片系统包括一个或多个接口电路，以及一个或多个处理器。该接口电路和该处理器通过线路互联；该接口电路用于从云中心的存储器接收信号，并向处理器发送该信号，该信号包括该存储器中存储的计算机指令。当该处理器执行该计算机指令时，云中心执行如第一方面或其相应的可能的设计提供的浮点数据精度转换方法。

第六方面，本申请实施例提供了一种计算机可读存储介质，包括计算机指令，当计算机指令在电子设备上运行时，使得电子设备执行上述任一方面及任一项可能的实现方式中的浮点数据精度转换方法。

第七方面，本申请实施例提供了一种计算机程序产品，当计算机程序产品在计算机或处理器上运行时，使得计算机或处理器执行上述任一方面及任一项可能的实现方式中的浮点数据精度转换方

法。

可以理解的是，上述提供的任一种浮点数据精度转换装置、芯片系统、计算机可读存储介质或计算机程序产品等均可以应用于上文所提供的对应的方法，因此，其所能达到的有益效果可参考对应的方法中的有益效果，此处不再赘述。

本申请的这些方面或其他方面在以下的描述中会更加简明易懂。

附图说明

图 1 为本申请实施例提供的一种 IEEE754 浮点数据格式图；

图 2 为本申请实施例提供的一种浮点数据精度转换方法或装置应用的系统或设备示意图；

图 3 为本申请实施例提供的一种 SoC 的结构示意图；

图 4 为本申请实施例提供的一种浮点数据精度转换方法的流程图；

图 5 为本申请实施例提供的一种随机舍入方式的结构示意图；

图 6 为本申请实施例提供的一种远离 0 进位舍入方式的流程图；

图 7 为本申请实施例提供的另一种浮点数据精度转换方法的流程图；

图 8 为本申请实施例提供的另一种浮点数据精度转换方法的流程图；

图 9 为本申请实施例提供的一种 FP32 数据转换为 HiFloat8 数据的流程图；

图 10 为本申请实施例提供的一种电子设备的结构示意图。

具体实施方式

为了便于理解，示例的给出了部分与本申请实施例相关概念的说明以供参考。如下所示：

标量计算单元，针对标量计算的电路称为标量计算单元，其中，标量又称纯量，只有大小，没有方向。标量计算多用于通用计算，在中央处理器（Central Processing Unit，CPU）多级流水线的执行单元（Execution Unit，EXU）部分和其他类似功能的处理器的标量计算部分，可以嵌入基于 HiFloat 数据格式的算数逻辑单元（Arithmetic Logic Unit，ALU）。

向量计算单元，针对向量计算而特殊设计的具有一定并行度的计算单元，如单指令多数据流（Single Instruction Multiple Data，SIMD）处理机，其中，向量又称矢量，通常指长度大于 1 的一维数组。向量计算单元多用于 HPC 和 AI 机器学习等领域，包括如线性规划、傅里叶变换、滤波计算以及线性代数、偏微分方程、积分等数学问题的求解。在向量计算加速单元或向量处理机中，可以嵌入基于 HiFloat 数据格式的算数执行单元（Vector Unit）。

矩阵计算单元，针对矩阵计算而特殊设计的具有相应并行度的计算单元，如脉动阵列（systolic array）处理机，其中，矩阵是一个按照长方阵列排列的 2 维数组。矩阵计算单元多用于 HPC 和 AI 机器学习等领域的矩阵计算，包括如矩阵乘、矩阵求逆和矩阵分解等。在矩阵计算加速单元中，可以嵌入基于 HiFloat 数据格式的矩阵单元（Matrix Unit）。

张量计算单元，针对张量计算而特殊设计的具有相应并行度的计算单元称为张量计算单元，如立方（Cube）计算单元，其中，张量是维数超过 2 维的多维数组，常见的为 3 维数组。张量计算单元多用于 AI 机器学习领域，如卷积操作。在张量计算加速单元中，可以嵌入基于 HiFloat 数据格式的张量单元（Tensor Unit）。

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行描述。其中，在本申请实施例的描述中，除非另有说明，“/”表示或的意思，例如，A/B 可以表示 A 或 B；本文中的“和/或”仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。另外，在本申请实施例的描述中，“多个”是指两个或多于两个。

以下，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。在本实施例的描述中，除非另有说明，“多个”的含义是两个或两个以上。

电气电子工程师协会（Institute of Electrical and Electronics Engineers，IEEE）制定了 IEEE 754 为二进制浮点数算数标准，分别定义了双精度 FP64、单精度 FP32 以及半精度 FP16 等浮点数据表示方法。其中，双精度 FP64 数据和单精度 FP32 数据适用于 CPU 和浮点运算器环境，半精度 FP16 数据适用于计算机图形环境。如图 1 所示，图 1 为本申请实施例提供的 IEEE754 浮点数据格式图。IEEE 754 浮点数据包括符号域（bit sign，S）、指数域（bits exponent，E）和尾数域（bits mantissa，M），

其中，对于 FP64 数据，符号域为 1 比特，指数域为 11 比特，尾数域为 52 比特；对于 FP32 数据，符号域为 1 比特，指数域为 8 比特，尾数域为 23 比特；对于 FP16 数据，符号域为 1 比特，指数域为 5 比特，尾数域为 10 比特。IEEE 754 表示的浮点数据与十进制的值（Value）的转换公式为：$Value=(-1)^{sign}*2^{exponent-bias}*(1+mantissa)$，其中，bias 对于 FP64、FP32 和 FP16 等浮点数据为不同的常数。

本申请实施例提供的第一浮点数据为 FP32 数据，下面先对 FP32 数据的数据格式进行介绍，例如表 1 所示，表 1 为 FP32 数据的数据格式。

表 1 FP32 数据的数据格式表

S	E								M																					
0	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FP32 数据包括符号域 S、指数域 E 和尾数域 M，FP32 数据的符号域决定 FP32 数据是正数或是负数，其中，0 表示正数，1 表示负数。FP32 数据的指数域为 2 的幂，可以对 FP32 数据进行加权处理。FP32 数据的尾数域是二进制小数。FP32 数据转十进制数据的转换步骤如下：（1）FP32 数据的符号域为 0，则该 FP32 数据为正数。（2）FP32 数据的指数域的编码值为 01111100，表示十进制的 124，去除 FP32 数据中的偏置（bias）（对于 FP32 数据，bias=127），则该 FP32 数据的指数域为 2 的-3 次方。（3）FP32 数据的尾数域为 01000000000000000000000，由于指数域不为全“0”或全“1”，则尾数域表示的十进制数据为 1.25。则基于浮点数据与十进制的值的转换公式该 FP32 数据对应的十进制数据为 0.15625。

对于 AI 的混合训练模式和 HPC 业务领域，现有的数据格式，例如 FP32 数据，数据的位宽较大，导致数据存储和转移开销大。且越来越多的应用，不需要例如 FP64 数据等高精度数据格式。因此，需要将高精度的数据转换为低精度的数据，在转换过程中，需要对高精度数据进行舍入操作，由此会产生转换误差。

现有技术提供了一种 FP32 向 FP16 或 BF16 数据格式转换的一种简化的随机舍入方式，随机舍入的阈值采用数据本身的特定的比特进行计算后完成。该方式中通过 FP32 数据的尾数域（共 23 位，最高有效位（Most Significant Bit，MSB）向最低有效位（Least Significant Bit，LSB）编号为 1-23）中的第 11-18 比特与第 16-23 比特相加，溢出的比特作为随机输入的阈值。但该方式仅描述了单一的随机舍入转化方式，仅涉及到 FP32 到 FP16 或 BF16 数据格式的转换，无法满足其他格式的高精度数据向低精度数据的转换。且该方式的阈值生成方式涉及到较多尾数域的比特，计算较为复杂，硬件开销较大。

由此，本申请实施例提供的浮点数据精度转换方法，实现了将高精度数据转换为低精度数据。在数据格式转换中，基于第一浮点数据的符号域可以得到第二浮点数据的符号域，基于第一浮点数据的第一指数域可以得到第二浮点数据的前缀码域和第二指数域，以及基于第一浮点数据的第一尾数域可以得到第二浮点数据的第二尾数域。第二浮点数据中通过较短的前缀码域指示第二指数域的位宽，可以有效提升第二浮点数据的尾数的精度或位宽，同时对于只提供 1 位尾数的精度的第二浮点数据可以表示较大的数值范围，有效平衡了第二浮点数据位宽、范围和精度之间的关系。且前缀码域可以采用前缀码编码方式，占用位宽少，解析第二指数域和第二尾数域便捷。并通过提供简单的舍入方式，根据第一尾数域中的舍弃编码值对保留编码值进行舍入操作，无需其他设备的支持，提高了高精度数据向低精度数据转换的转换效率，降低了硬件开销。

本申请实施例提供的第二浮点数据为 HiFloat8 数据，如表 2 所示，表 2 为 HiFloat8 数据的编码方式。

表 2 HiFloat8 数据的编码方式

HiFloat8 数据	符号域（S）	前缀码域(D)：{值}	指数域（E）	尾数域（M）
位宽	1	2：{2,3,4}	D	8-3-D
	1	3：{0,1}	D	8-4-D

其中，8 为 HiFloat8 数据的总位宽，总位宽可以变化。符号域占据一个比特位，0 表示正数，1 表示负数，或是 1 表示负数，0 表示正数。前缀码域占据 2 或 3 个比特位，前缀码域可以表达 5 个不同信息，D 的值（Value）可以是 0、1、2、3、4。指数域的位宽根据 D 的值变化，以及尾数域占据剩余位宽。

其中，前缀码域可以采用整数（integer）编码，此时 D 为固定值。前缀码域也可以采用前缀码（prefix code）编码，此时 D 为有限值集合。前缀码编码中用 2 个比特位编码值 2、3 和 4，用 3 个

比特位编码值 0 和 1。前缀码域的前缀码编码方式如表 3 所示，表 3 为前缀码域的编码方式。

表 3 前缀码域的前缀码编码方式

值	编码	位宽
4	11	2
3	10	
2	01	
1	001	3
0	000	

本申请实施例提供的 HiFloat8 数据和十进制的值（X）转换公式为：$X=(-1)^{S}*2^{E_i+E_c}*(1+M)$。

其中，Ec 为阶码对称中心，也是 FP32 数据中的 bias。

当 D 为 0 时，表示指数域的值为 0。当 D 不为 0 时，指数域采用符号量值（signed magnitude）编码，即符号位尾随原码（True Form，TF），指数域的编码 Ei={Se，1'b1，TF[2：end]},Se 为指数位符号位。TF 的最高位 1'b1 隐藏，不存储，则指数域的编码值为 Es={Se，TF[2：end]}。指数域对应于十进制的编码值为 Ev=Ei+Ec。

HiFloat(N, 5, Ec)可配置为 HiFloat(8, 5, 0)，简写为 HiF8，也可以配置为其他情况。HiFloat8 编码数值分布如表 4 所示。

表 4 HiFloat8 编码数值分布表

D	0	1	2	3	4
Es	None	Se	Se, TF[2]	Se, TF[2:3]	Se, TF[2:4]
Ei	0	Se,1	Se, 1, TF[2]	Se, 1, TF[2:3]	Se, 1, TF[2:4]
Ev	0	±1	±[2,3]	±[4,7]	±[8,15]
M 位宽	4	3	3	2	1

在上述场景中，本申请的浮点数据精度转换方法和装置可以应用于不同的系统或设备中，如应用于图 2 所示的执行设备 20，图 2 为本申请实施例提供的一种浮点数据精度转换方法和装置应用的系统或设备示意图。该执行设备可以是终端，如手机终端，平板电脑，笔记本电脑，AR 设备（图 2 中未示出），VR 设备（图 2 中未示出），车载终端（图 2 中未示出）等，还可以是服务器等。本申请提供的浮点数据精度转换方法可以应用于执行设备 20 中涉及到 CPU、HPC 和 AI 等关于混合精度计算的场景中，例如标量计算单元、向量计算单元、矩阵计算单元和张量计算单元等。

在一些实施例中，本申请提出的浮点数据精度转换的装置可以为芯片，例如该芯片为系统级芯片（System-on-a-Chip，SoC）。如图 3 所示，图 3 为本申请实施例提供的一种 SoC 的结构示意图。SoC 包括处理器，该处理器可以为单核处理器或多核处理器、存储器和 I/O 接口等。处理器可加载存储器中的数据和应用程序后，对数据进行处理，例如进行本申请中的计算处理。例如数据为 FP32 数据时，可通过读取 FP32 数据中的符号域、第一指数域和第一尾数域确定第二浮点数据的符号域、前缀码域、第二指数域和第二尾数域。

本申请实施例提供一种浮点数据精度转换方法，应用于将第一浮点数据转换为第二浮点数据，第一浮点数据包括符号域、第一指数域和第一尾数域，第二浮点数据包括符号域、前缀码域、第二指数域和第二尾数域，前缀码域用于指示第二指数域的位宽，第一浮点数据的精度高于第二浮点数据的精度。如图 4 所示，图 4 为本申请实施例提供的一种浮点数据精度转换方法的流程图，该方法包括：

步骤 401、执行设备根据第一指数域的第一编码值确定前缀码域的第一位宽、前缀码域的第一编码值、第二指数域的第一位宽、第二指数域的第一编码值以及第二尾数域的第一位宽。

示例性的，第一浮点数据的精度高于第二浮点数据的精度，其中，第一浮点数据可以是 FP32 数据，第二浮点数据可以是 HiFloat8 数据。

在转换过程中，由于数据格式转换并不会影响浮点数据的正负，因此，第一浮点数据的符号域和第二浮点数据的符号域相同。第一浮点数据或第二浮点数据包括二进制整数部分和二进制小数部分，其中，第一指数域和第二指数域分别决定第一浮点数据和第二浮点数据的二进制整数部分，第一尾数域决定第一浮点数据的二进制小数部分，前缀码域和第二尾数域决定第二浮点数据的二进制小数部分。基于第一指数域的第一编码值进行计算操作，可以得到前缀码域的第一位宽和前缀码域的第一编码值。基于第二浮点数据的数据格式，由此也可以得到第二指数域的第一位宽和第二指数

域的第一编码值。在得到前缀码域的第一位宽、第二指数域的第一位宽后，可以确定第二尾数域的第一位宽。

步骤 402、执行设备确定第一尾数域中的保留编码值和舍弃编码值，保留编码值包括第一尾数域中从最高位开始，且位宽与第二尾数域的第一位宽相同的编码值。

示例性的，由于第一浮点数据的精度高于第二浮点数据的精度，第一浮点数据的第一尾数域的位宽大于第二浮点数据的第二尾数域的位宽。将第一浮点数据转换为第二浮点数据，由于第二浮点数据的第二尾数域的位宽有限，需要对第一尾数域中的编码值进行取舍。将第一尾数域中从最高位开始，且位宽域第二尾数域的第一位宽相同的编码值确定为保留编码值，以及将第一尾数域中除保留编码值外剩余的编码值确定为舍弃编码值。

步骤 403、执行设备根据舍弃编码值对保留编码值进行舍入操作，得到第二尾数域的第一编码值。

示例性的，舍入操作可以是进位操作和舍弃操作，根据舍弃编码值判断对保留编码值是进行进位操作还是对舍弃编码值进行舍弃操作。其中一种判断方式可以是通过将舍弃编码值和阈值进行比较，若舍弃编码值大于阈值，对保留编码值进行进位操作，若舍弃编码值小于阈值，对舍弃编码值进行舍弃操作。

示例性的，本申请实施例提供的浮点数据精度转换方法，实现了将高精度数据转换为低精度数据。在数据格式转换中，基于第一浮点数据的符号域可以得到第二浮点数据的符号域，基于第一浮点数据的第一指数域可以得到第二浮点数据的前缀码域和第二指数域，以及基于第一浮点数据的第一尾数域可以得到第二浮点数据的第二尾数域。第二浮点数据中通过较短的前缀码域指示第二指数域的位宽，使得第二浮点数据最高可以提供 4 位尾数的精度，同时对于只提供 1 位尾数的精度的第二浮点数据可以表示较大的数值范围，有效平衡了第二浮点数据位宽、范围和精度之间的关系。且前缀码域可以采用前缀码编码方式，占用位宽少，解析第二指数域和第二尾数域便捷。并通过提供简单的舍入方式，根据第一尾数域中的舍弃编码值对保留编码值进行舍入操作，无需其他设备的支持，提高了高精度数据向低精度数据转换的转换效率，降低了硬件开销。

可选的，本申请实施例还提供一种随机舍入（Stochastic Round，SR）方式。如图 5 所示，图 5 为本申请实施例提供的一种随机舍入方式的结构示意图。步骤 403 还可以包括：

步骤 4033、舍弃编码值中从最高位开始，且位宽为预设位宽的编码值大于或等于第二预设阈值时，执行设备对保留编码值的最低位进行进位操作，对舍弃编码值进行舍弃操作，保留编码值进位后的编码值为第二尾数域的第一编码值。

步骤 4034、舍弃编码值中从最高位开始，且位宽为预设位宽的编码值小于第二预设阈值时，执行设备对舍弃编码值进行舍弃操作，保留编码值为第二尾数域的第一编码值。

其中，第二预设阈值为舍弃编码值中从最低位开始，且位宽为预设位宽的编码值。

示例性的，对于 SR 舍入方式，预设位宽可以为 10 至 14 中的整数。以预设位宽为 14 为例，第二预设阈值可以为舍弃编码值中从最低位开始，且位宽为 14 的编码值，则舍弃编码值中用于与第二预设阈值比较的部分舍弃编码值为从最高位开始，且位宽为 14 的编码值。

在一个实例中，对于表 1 中的第一尾数域 23’b01000000000000000000000，若第二尾数域的第一位宽为 2，则第一尾数域中的保留编码值为 2’b01，舍弃编码值为 21’b00000000000000000000，则部分舍弃编码值为 14’b00000000000000，第二预设阈值为 14’b00000000000000。由于部分舍弃编码值等于第二预设阈值，则对舍弃编码值进行舍弃操作，保留编码值进位后的编码值为第二尾数域的第一编码值，即第二尾数域的第一编码值为 2’b01。

在另一个实例中，对于表 1 中的第一尾数域 23’b01000000000000000000000，若第二尾数域的第一位宽为 1，则第一尾数域中的保留编码值为 1’b0，舍弃编码值为 22’b1000000000000000000000，则部分舍弃编码值为 14’b10000000000000，第二预设阈值为 14’b00000000000000。由于部分舍弃编码值大于第二预设阈值，则对保留编码值的最低位进行进位操作，并对舍弃编码值进行舍弃操作，保留编码值进行进位操作后得到的编码值为第二尾数域的第一编码值，即第二尾数域的第一编码值为 1’b1。

示例性的，对于 SR 舍入方式，用于比较的第二预设阈值为舍弃编码值中从最低位开始，且位宽为预设位宽的编码值，第二预设阈值的生成无需额外的随机数生成器，不存在随机数生成的性能瓶颈，提高了高精度数据向低精度数据的转换效率，同时硬件开销更低。

可选的，舍入操作包括进位操作或舍弃操作，本申请实施例提供一种远离 0 进位（Round Half To

Away，TA）的舍入方式，如图 6 所示，图 6 为本申请实施例提供的一种远离 0 进位舍入方式的流程图，步骤 403 可以包括：

步骤 4031、舍弃编码值的最高位大于或等于第一预设阈值时，执行设备对保留编码值的最低位进行进位操作，并对舍弃编码值进行舍弃操作，保留编码值进行进位操作后得到的编码值为第二尾数域的第一编码值。

步骤 4032、舍弃编码值的最高位小于第一预设阈值时，执行设备对舍弃编码值进行舍弃操作，保留编码值为第二尾数域的第一编码值。

示例性的，对于 TA 舍入方式，第一预设阈值可以为 1。舍弃编码值的最高位大于或等于预设阈值时，即舍弃编码值的最高位为 1，则对保留编码值的最低位进行进位操作，并对舍弃编码值进行舍弃操作。舍弃编码值的最高位小于第一预设阈值时，即舍弃编码值的最高位为 0 时，对舍弃编码值进行舍弃操作。

在一个实例中，对于表 1 中的第一尾数域 23’b01000000000000000000000，若第二尾数域的第一位宽为 2，则第一尾数域中的保留编码值为 2’b01，舍弃编码值为 21’b00000000000000000000，此时舍弃编码值的最高位为 0。由于舍弃编码值的最高位小于第一预设阈值，则对舍弃编码值进行舍弃操作，保留编码值为第二尾数域的第一编码值，即第二尾数域的第一编码值为 2’b01。

在另一个实例中，对于表 1 中的第一尾数域 23’b01000000000000000000000，若第二尾数域的第一位宽为 1，则第一尾数域中的保留编码值为 1’b0，舍弃编码值为 22’b10000000000000000000000，此时舍弃编码值的最高为 1。由于舍弃编码值的最高位大于第一预设阈值，则对保留编码值的最低位进行进位操作，并对舍弃编码值进行舍弃操作，保留编码值进行进位操作后得到的编码值为第二尾数域的第一编码值，即第二尾数域的第一编码值为 1’b1。

示例性的，对于 TA 舍入方式，其预设阈值也可以为 0。舍弃编码值的最高位大于第一预设阈值时，对保留编码值的最低位进行进位操作，并对舍弃编码值进行舍弃操作，保留编码值进行进位操作后得到的编码值为第二尾数域的第一编码值。舍弃编码值的最高位小于或等于第一预设阈值时，对舍弃编码值进行舍弃操作，保留编码值为第二尾数域的第一编码值。

示例性的，除了 TA 舍入方式，还可以包括远离偶数进位（round half to even）舍入方式和远离奇数进位（round half to odd）舍入方式等。本申请实施例提供的 TA 舍入方式对于其他舍入方式硬件实现面积更小、功耗开销更小，且具有更高的数据分辨率。

可选的，如图 7 所示，图 7 为本申请实施例提供的另一种浮点数据精度转换方法的流程图。本申请实施例提供的浮点数据精度转换方法还可以包括：

步骤 404、执行设备判断进位操作后的保留编码值是否溢出。

示例性的，若对保留编码值进行进位操作，保留编码值有可能产生溢出。在一个实例中，若保留编码值为 3’b111，对保留编码值的最低位进行进位操作后，会出现溢出现象。

步骤 405、若进位操作后的保留编码值溢出，执行设备则对第一指数域的第一编码值的最低位进行加 1 操作，得到第一指数域的第二编码值。

示例性的，如表 1 中的第一指数域，其第一编码值为 8’b01111100，若进位操作后的保留编码值溢出，则对第一指数域的第一编码值的最低位进行加 1 操作，得到第一指数域的第二编码值，即第一指数域的第二编码值为 8’b01111101。

步骤 406、执行设备根据第一指数域的第二编码值确定第二指数域的第二位宽和前缀码域的第二位宽。

示例性的，根据第一指数域的第二编码值 8’b01111101 可以得到第二指数域的第二位宽为 1 和前缀码域的第二位宽为 3。

步骤 407、若前缀码域的第二位宽和前缀码域的第一位宽不同，执行设备根据第一指数域的第二编码值确定前缀码域的第二编码值、第二指数域的第二编码值第二尾数域的第二位宽和第二尾数域的第二编码值。

示例性的，请参看表 4，若前缀码域的第二位宽和前缀码域的第一位宽不同，由于前缀码域用于指示第二指数域的位宽，则第二指数域的第一位宽和第二指数域的第二位宽不同。若前缀码域的第二位宽大于前缀码域的第一位宽，则第二指数域的第二位宽小于第二指数域的第一位宽，此时前缀码域增加的位宽数和第二指数域减少的位宽数相同，因此第二尾数域的第一位宽不变。若前缀码域的第二位宽小于前缀码域的第一位宽，则第二指数域的第二位宽大于第二指数域的第一位宽，此时

前缀码域减少的位宽数和第二指数域减少的位宽数相同，因此第二尾数域的第一位宽不变。由此，若前缀码域的第二位宽和前缀码域的第一位宽不同，此时的第二指数域和前缀码域的位宽发生变化，而第二尾数域的第二位宽不变，执行设备根据第一指数域的第二编码值确定前缀码域的第二编码值和第二指数域的第二编码值。其中，第二尾数域的第二编码值为全 0，例如若第二尾数域的第二位宽为 3，则第二尾数域的第二编码值为 3'b000。

在一个实例中，基于第一指数域的第一比特 8'b01111100 确定的前缀码域的第一位宽为 2，第二指数域的第二位宽为 2，此时前缀码域减少的位宽数和第二指数域增加的位宽数相同，因此第二尾数域的第一位宽不变。

步骤 408、若前缀码域的第二位宽和前缀码域的第一位宽相同，执行设备判断第二指数域的第一位宽和第二指数域的第二位宽是否相同。

步骤 409、若第二指数域的第二位宽小于第二指数域的第一位宽，执行设备对保留编码值的位宽进行加 1 操作，得到第二尾数域的第二位宽和第二尾数域的第二编码值。

示例性的， 前缀码域的第二位宽和前缀码域的第一位宽相同，若第二指数域的第二位宽小于第二指数域的第一位宽，则第二尾数域的第一位宽会增加。对保留编码值的位宽进行加 1 操作得到第二尾数域的第二位宽和第二尾数域的第二编码值。在一个实例中，若保留编码值为 2'b01，保留编码值的位宽为 2，对保留编码值的位宽进行加 1 操作后，保留编码值的位宽为 3，保留编码值为 3'b010。

步骤 4010、若第二指数域的第二位宽大于第二指数域的第一位宽，执行设备对保留编码值的最低位进行舍弃操作，得到所述第二尾数域的第二位宽和所述第二尾数域的第二编码值。

示例性的，前缀码域的第二位宽和前缀码域的第一位宽相同，若第二指数域的第二位宽大于第二指数域的第一位宽，则第二尾数域的第一位宽会减少。对保留编码值的最低位进行舍弃操作得到第二尾数域的第二位宽和第二尾数域的第二编码值。在一个实例中，若保留编码值为 2'b01，保留编码值的位宽为 2，对保留编码值的最低位进行舍弃操作后，保留编码值为 1'b0，保留编码值的位宽为 1。

下面对本申请提供的浮点数据精度转换方法进一步进行说明，如图 8 所示，图 8 为本申请实施例提供的另一种浮点数据精度转换方法的流程图，步骤 401 包括：

步骤 4011、执行设备根据第一指数域的第一编码值确定指示值，通过查表确定与指示值对应的前缀码域的第一位宽和前缀码域的第一编码值，指示值还用于指示第二指数域的第一位宽。

示例性的，基于第一指数域的第一编码值可以确定第一指数域的指数值 N，基于第一指数域的指数值可以确定指示值。此处的查表为查看表 3，指示值为 D 的值，指示值可以是 0、1、2、3、4。在一个实例中，例如表 1 中的第一指数域，其第一编码值为 8'b01111100，表示十进制的 124，去除对于 FP32 数据的偏置 127 后，得到十进制的-3，其中，-3 为第一指数域的指数值 N，利用公式 $D=INT[\log_2|N|]$，可以得到 D 为 2。通过查表 3 可以确定与 2 对应的前缀码域的第一位宽为 2 和前缀码域的第一编码值为 01。指示值还用于指示第二指数域的第一位宽，即第二指数域的第一位宽为 2。

步骤 4012、执行设备根据第一指数域的第一编码值确定第二指数域的第一位宽对应的第一编码值。

示例性的，对于表 1 中的第一指数域，请参看表 4，当 D 为 2 时，且第一指数域的指数值为-3，即指数位符号位 Se 为 1，由于指示值确定的第二指数域的第一位宽为 2，则确定的第二指数域的第一编码值为 11。

步骤 4013、执行设备根据第二浮点数据的总位宽、前缀码域的第一位宽、第二指数域的第一位宽确定第二尾数域的第一位宽。

示例性的，第二浮点数据的总位宽为 Nb，前缀码域的第一位宽为 Db，第二指数域的第一位宽为 Eb，符号域的位宽为 1，第二尾数域的第一位宽为 Mb，则 Mb=Nb-Db-Eb-1。在一个实例中，对于 HiFloat8 数据，Nb=8，Db=2 或 3，则第二尾数域的第一位宽 Mb=Nb-3-Eb 或 Mb=Nb -Eb-4。

示例性的，对于为 HiFloat8 数据格式的第二浮点数据，符号域的位宽为 1，前缀码域的位宽为 2 或 3，第二指数域的第一位宽为 0 至 4 中的一个整数，第二尾数域的第一位宽为 1 至 4 中的一个整数。由此，第二浮点数据中采用较短的前缀码域指示第二指数域的第一位宽，使得第二浮点数据最高可以提供 4 位尾数的精度，同时对于只提供 1 位尾数的精度的第二浮点数据可以表示较大的数值范围，有效平衡了第二浮点数据位宽、范围和精度之间的关系。且第二指数域存储时隐藏最高位，减少了第二指数域需要存储的第一位宽，有效避免了不同前缀码域的指示值对应的第二指数域的第一编码

值出现数值重叠的问题，使得 HiFloat8 数据格式中无冗余编码。

可选的，第一浮点数据超出第二浮点数据的数据范围的上限时，基于饱和方式或无穷大方式确定第二浮点数据。

示例性的，饱和方式可以为用低精度浮点数据能表示的最大浮点数据作为第一浮点数据。无穷大方式可以为用低精度浮点数据的无穷大数据作为第一浮点数据。在一个实例中，对于 HiFloat8 数据，若第一浮点数据超出 HiFloat8 数据的数据范围的上限时，第一浮点数据精度转换后的第二浮点数据可以表示为 8'b01101111。

可选的，第一浮点数据超出第二浮点数据的数据范围的下限时，第二浮点数据为零。

示例性的，对于 HiFloat8 数据，若第一浮点数据超出 HiFloat8 数据的数据范围的下限时，第二浮点数据为零，第二浮点数据可以表示为 8'b01111110。

可选的，第一浮点数据为非数字值时，第二浮点数据为非数字值。

示例性的，对于 HiFloat8 数据，若第一浮点数据为非数字值（Not a Number，NAN）时，第二浮点数据可以表示为 8'b11111110。

示例性的，如图 9 所示，图 9 为本申请实施例提供的一种 FP32 数据转换为 HiFloat8 数据的流程图。以 FP32 数据转换为 HiFloat8 数据为例，应用于转换模块，该转换过程包括以下流程。

（1）转换模块接收 FP32 数据，FP32 数据包括符号域 S，指数域 E[0:7]和尾数域 M[0:22];

（2）对 FP32 数据进行判断是否为特殊值（零值、非数字值（Not a Number，NAN）、正无穷大和负无穷大），若 FP32 数据为特殊值，转（21）；

（3）若 FP32 数据不为特殊值，获取 FP32 数据的符号域，转（21）；

（4）对于 FP32 数据的指数域进行去除偏置操作，即 E=E+bias;

（5）计算 HiFloat8 数据中前缀码域、指数域和尾数域对应的位宽，分别用 db、eb 和 mb 表示;

（6）根据 FP32 数据的指数域和 HiFloat8 数据的指数域的位宽，确定 HiFloat8 数据的指数域的编码值 e=E[0:k]，其中，k 为对应于 HiFloat8 数据的指数域的位宽的值;

（7）配置舍入方式;

（8）生成阈值;

（9）获取 FP32 数据中的尾数域的舍弃编码值的最高位，作为舍弃位;

（10）舍弃位和阈值比较;

（11）进行舍入判断，若舍弃位大于阈值，进行进位操作，若舍弃位小于阈值，进行舍弃操作;

（12）若进行舍弃操作，FP32 数据中尾数域的保留编码值为 HiFloat8 数据的尾数域 m=M[0:N-db-eb-1]，转（21）；

（13）若进行进位操作，FP32 数据中尾数域的保留编码值进行加 1 操作，m=M[0:N-db-eb-1]+1;

（14）判断 FP32 数据中尾数域的保留编码值是否溢出，若 FP32 数据中尾数域的保留编码值未溢出，转（21）；

（15）若 FP32 数据中尾数域的保留编码值溢出，FP32 数据的指数域进行加 1 操作 E=E+1;

（16）基于进行加 1 操作后的 FP32 的指数域计算得到 HiFloat8 数据的新的前缀码域的位宽和指数域的位宽，分别用 db1 和 eb1 表示;

（17）判断 db1 是否等于 db，若 db1 不等于 db，转（21）；

（18）若 db1 等于 db，判断 eb1 是否大于 eb;

（19）若 eb1 小于 eb，HiFloat8 数据的尾数域的位宽进行加 1 操作，即；mb=mb+1;

（20）若 eb1 大于 eb，HiFloat8 数据的尾数域的位宽进行减 1 操作，即 mb=mb-1;

（21）进行 HiFloat8 数据编码;

（22）得到 HiFloat8 数据;

可以理解的是，为了实现上述功能，电子设备包含了执行各个功能相应的硬件和/或软件模块。结合本文中所公开的实施例描述的各示例的算法步骤，本申请能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行，取决于技术方案的特定应用和设计约束条件。本领域技术人员可以结合实施例对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

本实施例可以根据上述方法示例对电子设备进行功能模块的划分，例如，可以对应各个功能划分各个功能模块，也可以将两个或两个以上的功能集成在一个处理模块中。上述集成的模块可以采

用硬件的形式实现。需要说明的是，本实施例中对模块的划分是示意性的，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。

在采用对应各个功能划分各个功能模块的情况下，图 10 示出了上述实施例中涉及的电子设备 100 的一种可能的组成示意图，如图 10 所示，图 10 为本申请实施例提供的一种电子设备的结构示意图。该电子设备 100 可以包括：位宽计算单元 101、尾数域计算单元 102 和舍入操作单元 103。

其中，位宽计算单元 101 可以用于支持电子设备 100 执行上述步骤 401、步骤 4011、步骤 4012、步骤 4013 等，和/或用于本文所描述的技术的其他过程。

尾数域计算单元 102 可以用于支持电子设备 100 执行上述步骤 402 等，和/或用于本文所描述的技术的其他过程。

舍入操作单元 103 可以用于支持电子设备 100 执行上述步骤 403、步骤 4031、步骤 4032 等，和/或用于本文所描述的技术的其他过程。

需要说明的是，上述方法实施例涉及的各步骤的所有相关内容均可以援引到对应功能模块的功能描述，在此不再赘述。

本实施例提供的电子设备 100，用于执行上述浮点数据精度转换方法，因此可以达到与上述实现方法相同的效果。

在采用集成的单元的情况下，电子设备 100 可以包括处理模块、存储模块和通信模块。其中，处理模块可以用于对电子设备 100 的动作进行控制管理，例如，可以用于支持电子设备 100 执行上述位宽计算单元 101、尾数域计算单元 102 和舍入操作单元 103 执行的步骤。存储模块可以用于支持电子设备 100 存储程序代码和数据等。通信模块，可以用于支持电子设备 100 与其他设备的通信，例如与无线接入设备的通信。

其中，处理模块可以是处理器或控制器。其可以实现或执行结合本申请公开内容所描述的各种示例性的逻辑方框，模块和电路。处理器也可以是实现计算功能的组合，例如包含一个或多个微处理器组合，数字信号处理（digital signal processing，DSP）和微处理器的组合等等。存储模块可以是存储器。通信模块具体可以为射频电路、蓝牙芯片、Wi-Fi 芯片等与其他电子设备交互的设备。

在一个实施例中，当处理模块为处理器，存储模块为存储器时，本实施例所涉及的电子设备可以为服务器和电脑等。

本申请实施例还提供一种电子设备，包括一个或多个处理器以及一个或多个存储器。该一个或多个存储器与一个或多个处理器耦合，一个或多个存储器用于存储计算机程序代码，计算机程序代码包括计算机指令，当一个或多个处理器执行计算机指令时，使得电子设备执行上述相关方法步骤实现上述实施例中的浮点数据精度转换方法。

本申请的实施例还提供一种计算机存储介质，该计算机存储介质中存储有计算机指令，当该计算机指令在电子设备上运行时，使得电子设备执行上述相关方法步骤实现上述实施例中的浮点数据精度转换方法。

本申请的实施例还提供了一种计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机执行上述相关步骤，以实现上述实施例中电子设备执行的浮点数据精度转换方法。

另外，本申请的实施例还提供一种装置，这个装置具体可以是芯片，组件或模块，该装置可包括相连的处理器和存储器；其中，存储器用于存储计算机执行指令，当装置运行时，处理器可执行存储器存储的计算机执行指令，以使芯片执行上述各方法实施例中电子设备执行的浮点数据精度转换方法。

其中，本实施例提供的电子设备、计算机存储介质、计算机程序产品或芯片均用于执行上文所提供的对应的方法，因此，其所能达到的有益效果可参考上文所提供的对应的方法中的有益效果，此处不再赘述。

通过以上实施方式的描述，所属领域的技术人员可以了解到，为描述的方便和简洁，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将装置的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。

在本申请所提供的几个实施例中，应该理解到，所揭露的装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块或单元的划分，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，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。

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一个物理单元或多个物理单元，即可以位于一个地方，或者也可以分布到多个不同地方。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。

另外，在本申请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。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元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时，可以存储在一个可读取存储介质中。基于这样的理解，本申请实施例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全部或部分可以以软件产品的形式体现出来，该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个设备（可以是单片机，芯片等）或处理器（processor）执行本申请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。

以上内容，仅为本申请的具体实施方式，但本申请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申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

权 利 要 求 书

1.一种浮点数据精度转换方法，其特征在于，第一浮点数据包括符号域、第一指数域和第一尾数域，第二浮点数据包括所述符号域、前缀码域、第二指数域和第二尾数域，所述前缀码域用于指示所述第二指数域的位宽，所述第一浮点数据的精度高于所述第二浮点数据的精度，所述方法包括：

根据所述第一指数域的第一编码值确定所述前缀码域的第一位宽、所述前缀码域的第一编码值、所述第二指数域的第一位宽、所述第二指数域的第一编码值以及所述第二尾数域的第一位宽；

确定所述第一尾数域中的保留编码值和舍弃编码值，所述保留编码值包括所述第一尾数域中从最高位开始，且位宽与所述第二尾数域的第一位宽相同的编码值；

根据所述舍弃编码值对所述保留编码值进行舍入操作，得到所述第二尾数域的第一编码值。

2. 根据权利要求 1 所述的方法，其特征在于，所述舍入操作包括进位操作和舍弃操作，所述根据所述舍弃编码值对所述保留编码值进行舍入操作，得到所述第二尾数域的第一编码值包括：

所述舍弃编码值中从最高位开始，且位宽为预设位宽的编码值大于或等于第二预设阈值时，对所述保留编码值的最低位进行进位操作，对所述舍弃编码值进行舍弃操作，所述保留编码值进位后的编码值为所述第二尾数域的第一编码值；

所述舍弃编码值中从最高位开始，且位宽为预设位宽的编码值小于所述第二预设阈值时，对所述舍弃编码值进行舍弃操作，所述保留编码值为所述第二尾数域的第一编码值；

其中，所述第二预设阈值为所述舍弃编码值中从最低位开始，且位宽为预设位宽的编码值。

3.根据权利要求 1 所述的方法，其特征在于，所述舍入操作包括进位操作或舍弃操作，所述根据所述舍弃编码值对所述保留编码值进行舍入操作，得到所述第二尾数域的第一编码值包括：

所述舍弃编码值的最高位大于或等于第一预设阈值时，对所述保留编码值的最低位进行进位操作，并对所述舍弃编码值进行舍弃操作，所述保留编码值进行进位操作后得到的编码值为所述第二尾数域的第一编码值；

所述舍弃编码值的最高位小于所述第一预设阈值时，对所述舍弃编码值进行舍弃操作，所述保留编码值为所述第二尾数域的第一编码值。

4.根据权利要求 2 或 3 所述的方法，其特征在于，所述方法还包括：

判断进位操作后的所述保留编码值是否溢出；

若进位操作后的所述保留编码值溢出，则对所述第一指数域的第一编码值的最低位执行加 1 操作，得到所述第一指数域的第二编码值；

根据所述第一指数域的第二编码值确定所述第二指数域的第二位宽和所述前缀码域的第二位宽；

若所述前缀码域的第二位宽和所述前缀码域的第一位宽不同，根据所述第一指数域的第二编码值确定所述前缀码域的第二编码值、所述第二指数域的第二编码值、所述第二尾数域的第二位宽和所述第二尾数域的第二编码值；

若所述前缀码域的第二位宽和所述前缀码域的第一位宽相同，判断所述第二指数域的第一位宽和所述第二指数域的第二位宽是否相同；

若所述第二指数域的第二位宽小于所述第二指数域的第一位宽，对所述保留编码值的位宽进行加 1 操作，得到所述第二尾数域的第二位宽和所述第二尾数域的第二编码值；

若所述第二指数域的第二位宽大于或等于所述第二指数域的第一位宽，对所述保留编码值的最低位进行舍弃操作，得到所述第二尾数域的第二位宽和所述第二尾数域的第二编码值。

5.根据权利要求 1 所述的方法，其特征在于，所述根据所述第一指数域的第一编码值确定所述前缀码域的第一位宽、所述前缀码域的第一编码值、所述第二指数域的第一位宽和所述第二指数域的第一编码值包括：

根据所述第一指数域的第一编码值确定指示值，通过查表确定与所述指示值对应的所述前缀码域的第一位宽和所述前缀码域的第一编码值，所述指示值还用于指示所述第二指数域的第一位宽；

根据所述第一指数域的第一编码值确定所述第二指数域的第一位宽对应的第一编码值。

6.根据权利要求 1 或 5 所述的方法，其特征在于，所述根据所述第一指数域的第一编码值确定所述第二尾数域的第一位宽包括：

根据所述第二浮点数据的总位宽、所述前缀码域的第一位宽、所述第二指数域的第一位宽确定所述第二尾数域的第一位宽。

7.根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

所述第一浮点数据超出所述第二浮点数据的数据范围的上限时，基于饱和方式或无穷大方式确定第二浮点数据；

所述第一浮点数据超出所述第二浮点数据的数据范围的下限时，所述第二浮点数据为零；

所述第一浮点数据为非数字值时，所述第二浮点数据为非数字值。

8.一种浮点数据精度转换装置，其特征在于，第一浮点数据包括符号域、第一指数域和第一尾数域，第二浮点数据包括所述符号域、前缀码域、第二指数域和第二尾数域，所述前缀码域用于指示所述第二指数域的位宽，所述第一浮点数据的精度高于所述第二浮点数据的精度，所述装置包括：

位宽计算单元，用于根据所述第一指数域的第一编码值确定所述前缀码域的第一位宽、所述前缀码域的第一编码值、所述第二指数域的第一位宽、所述第二指数域的第一编码值以及所述第二尾数域的第一位宽；

尾数域计算单元，用于确定所述第一尾数域中的保留编码值和舍弃编码值，所述保留编码值包括所述第一尾数域中从最高位开始，且位宽与所述第二尾数域的第一位宽相同的编码值；

舍入操作单元，用于根据所述舍弃编码值对所述保留编码值进行舍入操作，得到所述第二尾数域的第一编码值。

9.根据权利要求 8 所述的装置，其特征在于，所述舍入操作包括进位操作或舍弃操作，所述舍入操作单元还用于：

所述舍弃编码值中从最高位开始，且位宽为预设位宽的编码值大于或等于第二预设阈值时，对所述保留编码值的最低位进行进位操作，对所述舍弃编码值进行舍弃操作，所述保留编码值进位后的编码值为所述第二尾数域的第一编码值；

所述舍弃编码值中从最高位开始，且位宽为预设位宽的编码值小于所述第二预设阈值时，对所述舍弃编码值进行舍弃操作，所述保留编码值为所述第二尾数域的第一编码值；

其中，所述第二预设阈值为所述舍弃编码值中从最低位开始，且位宽为预设位宽的编码值。

10.根据权利要求 8 所述的装置，其特征在于，所述舍入操作包括进位操作或舍弃操作，所述舍入操作单元还用于：

所述舍弃编码值的最高位大于或等于第一预设阈值时，对所述保留编码值的最低位进行进位操作，并对所述舍弃编码值进行舍弃操作，所述保留编码值进行进位操作后得到的编码值为所述第二尾数域的第一编码值；

所述舍弃编码值的最高位小于所述第一预设阈值时，对所述舍弃编码值进行舍弃操作，所述保留编码值为所述第二尾数域的第一编码值。

11.根据权利要求 9 或 10 所述的装置，其特征在于，所述装置还包括：

溢出单元，用于判断进位操作后的所述保留编码值是否溢出；

所述位宽计算单元还用于若进位操作后的所述保留编码值溢出，则对所述第一指数域的第一编码值进行加 1 操作，得到所述第一指数域的第二编码值；

根据所述第一指数域的第二编码值确定所述第二指数域的第二位宽和所述前缀码域的第二位宽；

若所述前缀码域的第二位宽和所述前缀码域的第一位宽不同，根据所述第一指数域的第二编码值确定所述前缀码域的第二编码值、所述第二指数域的第二编码值、所述第二尾数域的第二位宽和所述第二尾数域的第二编码值；

若所述前缀码域的第二位宽和所述前缀码域的第一位宽相同，判断所述第二指数域的第一位宽和所述第二指数域的第二位宽是否相同；

若所述第二指数域的第二位宽小于所述第二指数域的第一位宽，对所述保留编码值的位宽进行加 1 操作，得到所述第二尾数域的第二位宽和所述第二尾数域的第二编码值；

若所述第二指数域的第二位宽大于或等于所述第二指数域的第一位宽，对所述保留编码值的最低位进行舍弃操作，得到所述第二尾数域的第二位宽和所述第二尾数域的第二编码值。

12.根据权利要求 8 所述的装置，其特征在于，所述位宽计算单元还用于：

根据所述第一指数域的第一编码值确定指示值，通过查表确定与所述指示值对应的所述前缀码域的第一位宽和所述前缀码域的第一编码值，所述指示值还用于指示所述第二指数域的第一位宽；

根据所述第一指数域的第一编码值确定所述第二指数域的第一位宽对应的第一编码值。

13.根据权利要求 8 至 12 任一项所述的装置，其特征在于，所述位宽计算单元还用于：

根据所述第二浮点数据的总位宽、所述前缀码域的第一位宽、所述第二指数域的第一位宽确定

所述第二尾数域的第一位宽。

14.根据权利要求 8 所述的装置，其特征在于，所述位宽计算单元还用于：

所述第一浮点数据超出所述第二浮点数据的转换范围的上限时，基于饱和方式或无穷大方式确定所述第二浮点数据；

所述第一浮点数据超出所述第二浮点数据的转换范围的下限时，所述第二浮点数据为零；

所述第一浮点数据为非数字值时，所述第二浮点数据为非数字值。

15.一种计算机可读存储介质，其特征在于，包括计算机指令，当计算机指令在电子设备上运行时，使得电子设备执行上述权利要求 1-7 中的任一项所述的方法。

16.一种计算机程序产品，其特征在于，当计算机程序产品在计算机或处理器上运行时，使得所述计算机或所述处理器执行上述权利要求 1-7 中的任一项所述的方法。

符号位：1比特	指数位：FP16 5比特,FP32 8比特，FP64 11比特	尾数位：FP16 10比特,FP32 23比特，FP64 52比特

图 1

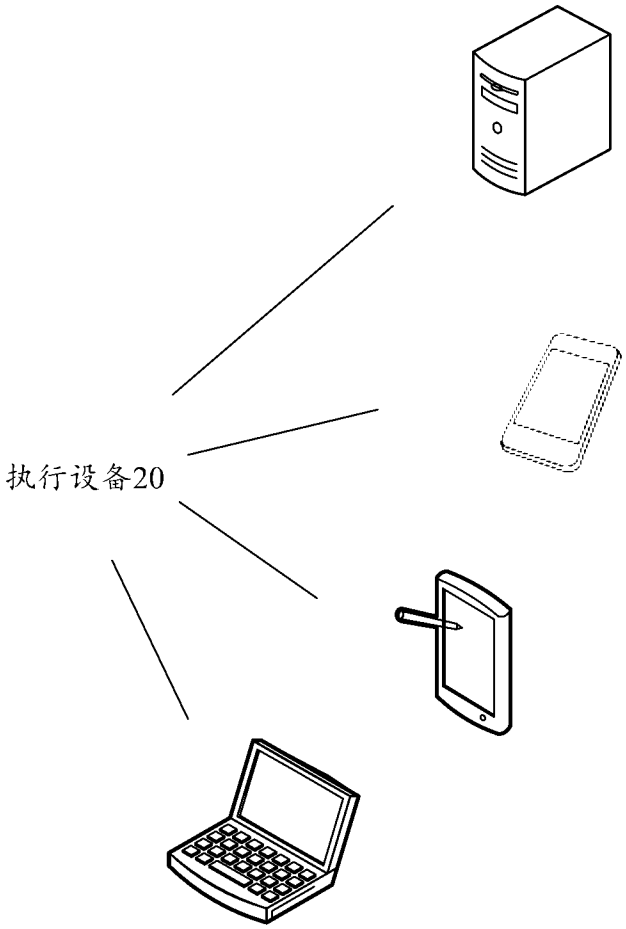

图 2

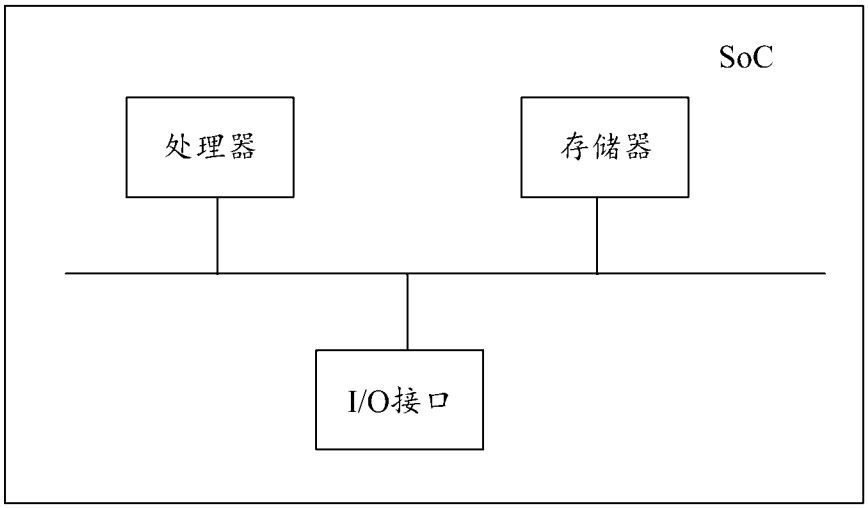

图 3

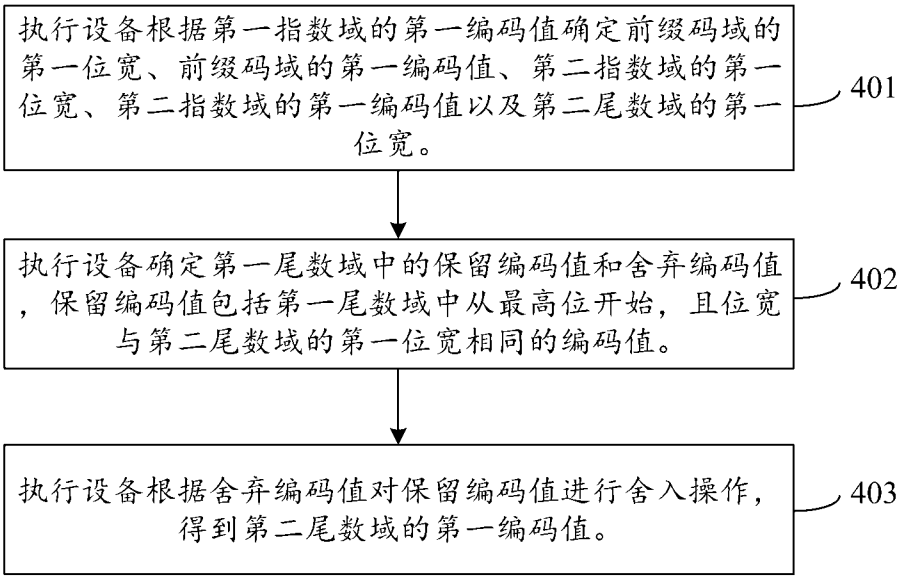
执行设备根据第一指数域的第一编码值确定前缀码域的第一位宽、前缀码域的第一编码值、第二指数域的第一位宽、第二指数域的第一编码值以及第二尾数域的第一位宽。
401
执行设备确定第一尾数域中的保留编码值和舍弃编码值，保留编码值包括第一尾数域中从最高位开始，且位宽与第二尾数域的第一位宽相同的编码值。
402
执行设备根据舍弃编码值对保留编码值进行舍入操作，得到第二尾数域的第一编码值。
403

图 4

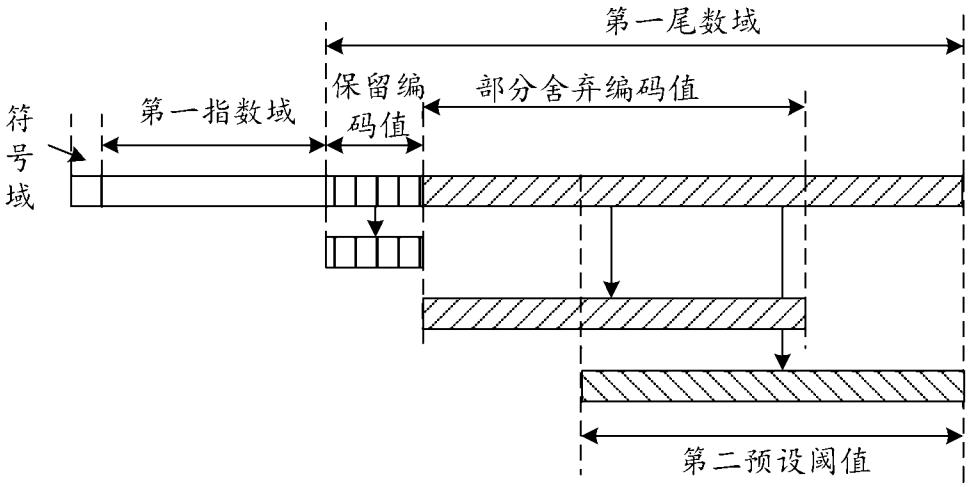
第一尾数域
保留编码值
部分舍弃编码值
符号域
第一指数域
第二预设阈值

图 5

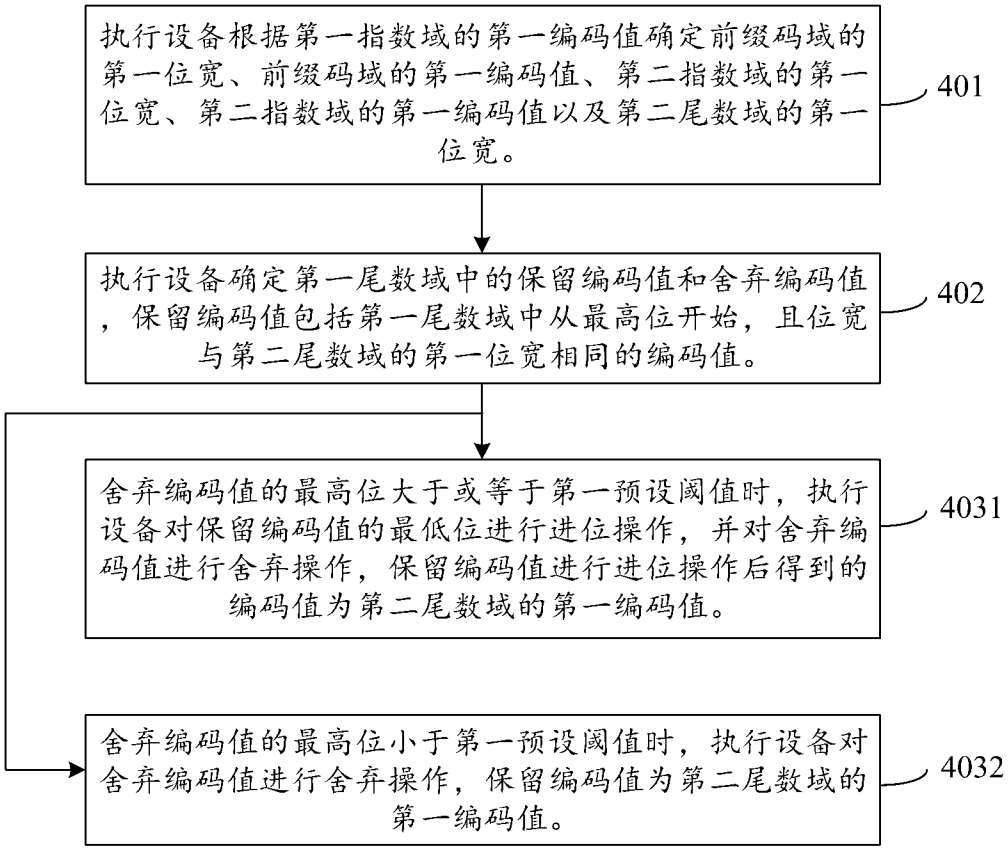
执行设备根据第一指数域的第一编码值确定前缀码域的第一位宽、前缀码域的第一编码值、第二指数域的第一位宽、第二指数域的第一编码值以及第二尾数域的第一位宽。
401
执行设备确定第一尾数域中的保留编码值和舍弃编码值，保留编码值包括第一尾数域中从最高位开始，且位宽与第二尾数域的第一位宽相同的编码值。
402
舍弃编码值的最高位大于或等于第一预设阈值时，执行设备对保留编码值的最低位进行进位操作，并对舍弃编码值进行舍弃操作，保留编码值进行进位操作后得到的编码值为第二尾数域的第一编码值。
4031
舍弃编码值的最高位小于第一预设阈值时，执行设备对舍弃编码值进行舍弃操作，保留编码值为第二尾数域的第一编码值。
4032

图 6

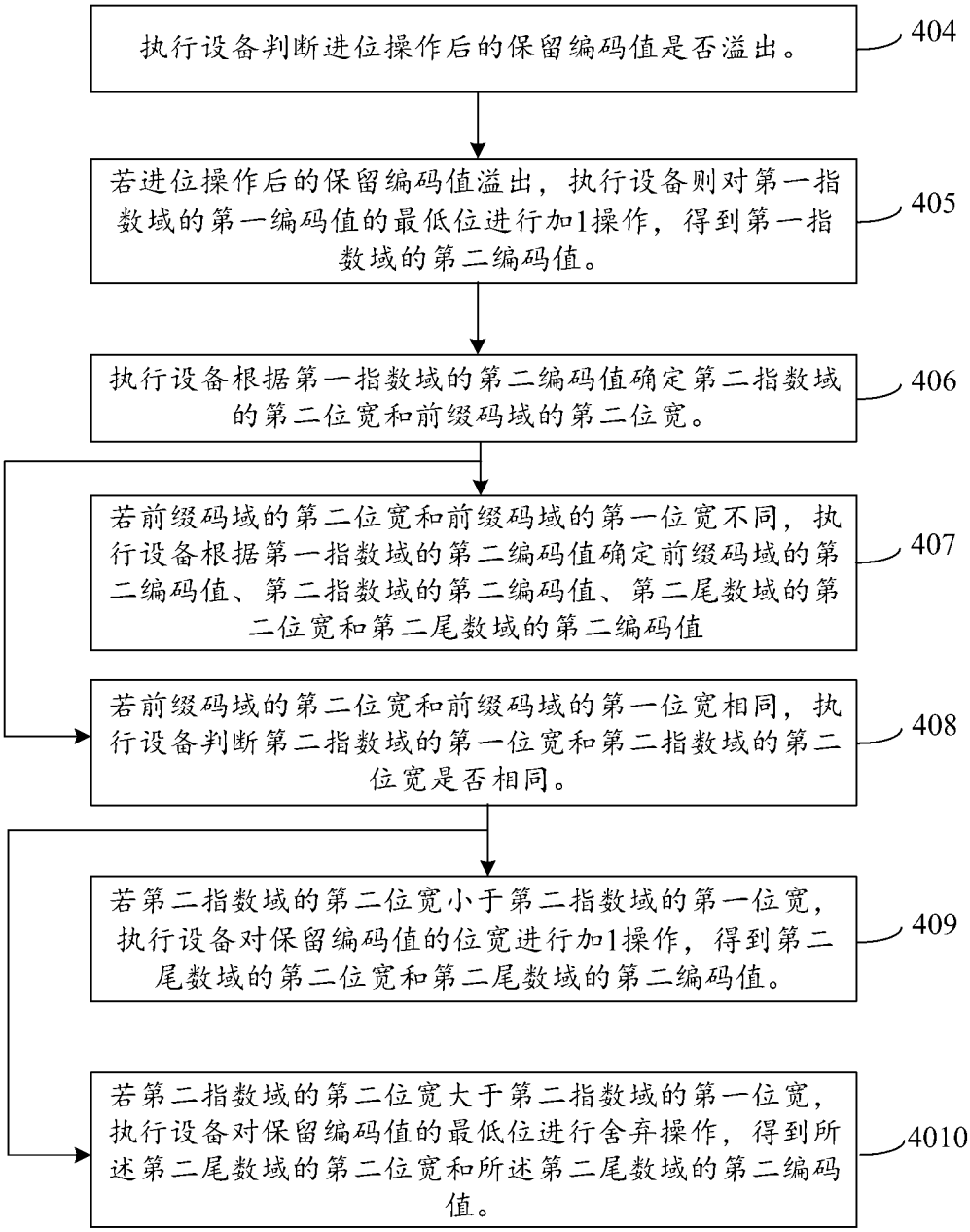
执行设备判断进位操作后的保留编码值是否溢出。
404
若进位操作后的保留编码值溢出，执行设备则对第一指数域的第一编码值的最低位进行加1操作，得到第一指数域的第二编码值。
405
执行设备根据第一指数域的第二编码值确定第二指数域的第二位宽和前缀码域的第二位宽。
406
若前缀码域的第二位宽和前缀码域的第一位宽不同，执行设备根据第一指数域的第二编码值确定前缀码域的第二编码值、第二指数域的第二编码值、第二尾数域的第二位宽和第二尾数域的第二编码值
407
若前缀码域的第二位宽和前缀码域的第一位宽相同，执行设备判断第二指数域的第一位宽和第二指数域的第二位宽是否相同。
408
若第二指数域的第二位宽小于第二指数域的第一位宽，执行设备对保留编码值的位宽进行加1操作，得到第二尾数域的第二位宽和第二尾数域的第二编码值。
409
若第二指数域的第二位宽大于第二指数域的第一位宽，执行设备对保留编码值的最低位进行舍弃操作，得到所述第二尾数域的第二位宽和所述第二尾数域的第二编码值。
4010

图 7

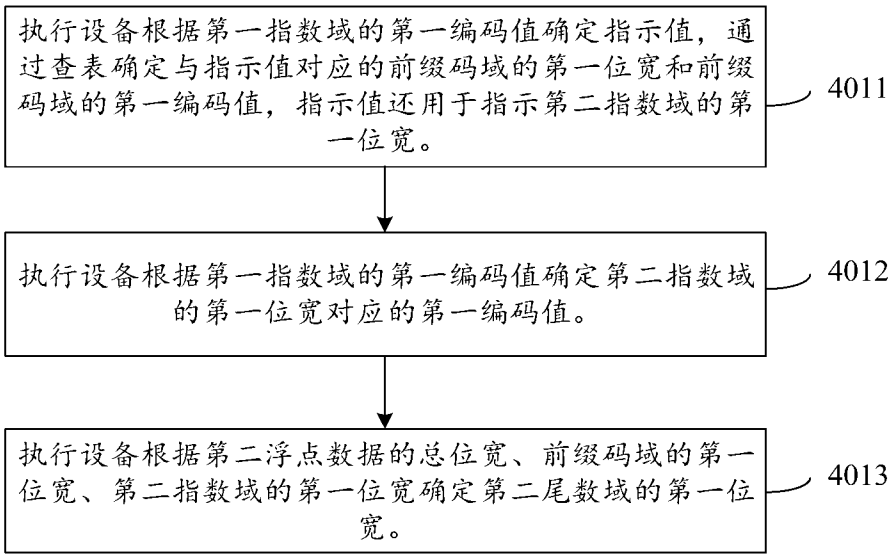
执行设备根据第一指数域的第一编码值确定指示值，通过查表确定与指示值对应的前缀码域的第一位宽和前缀码域的第一编码值，指示值还用于指示第二指数域的第一位宽。
4011
执行设备根据第一指数域的第一编码值确定第二指数域的第一位宽对应的第一编码值。
4012
执行设备根据第二浮点数据的总位宽、前缀码域的第一位宽、第二指数域的第一位宽确定第二尾数域的第一位宽。
4013

图 8

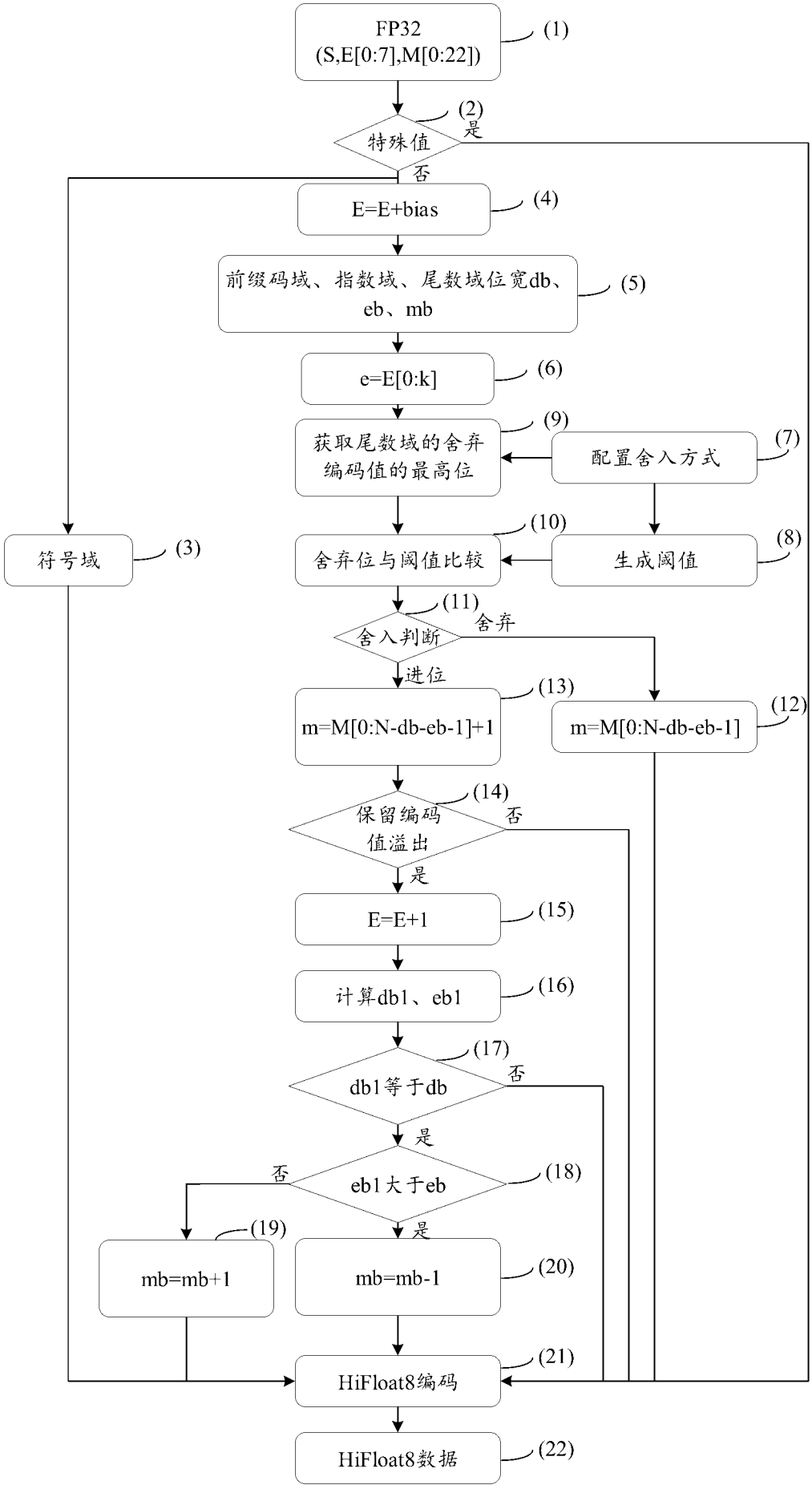
FP32
(S,E[0:7],M[0:22])
(1)
特殊值
(2)
是
否
E=E+bias
(4)
前缀码域、指数域、尾数域位宽db、eb、mb
(5)
e=E[0:k]
(6)
获取尾数域的舍弃编码值的最高位
(9)
配置舍入方式
(7)
符号域
(3)
舍弃位与阈值比较
(10)
生成阈值
(8)
舍入判断
(11)
舍弃
进位
m=M[0:N-db-eb-1]+1
(13)
m=M[0:N-db-eb-1]
(12)
保留编码值溢出
(14)
否
是
E=E+1
(15)
计算db1、eb1
(16)
db1等于db
(17)
否
是
eb1大于eb
(18)
否
是
mb=mb+1
(19)
mb=mb-1
(20)
HiFloat8编码
(21)
HiFloat8数据
(22)

图 9

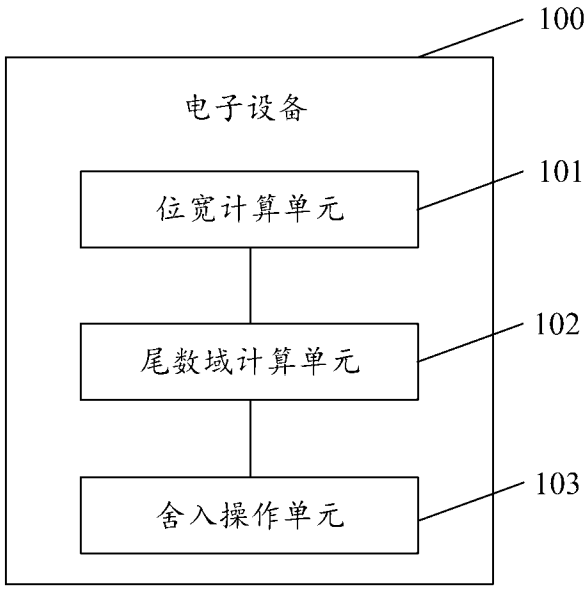

图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/102089

A. CLASSIFICATION OF SUBJECT MATTER

G06F7/483(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)

IPC: 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)

CNKI, CNTXT, ENTXT, ENTXTC, DWPI: 浮点数, 转换, 符号段, 指数段, 尾数段, floating-point number, conversion, symbol segment, exponent segment, mantissa segment

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 111340207 A (NANJING UNIVERSITY) 26 June 2020 (2020-06-26) description, paragraphs 0005-0153	1, 8, 15-16
A	CN 104778026 A (INSPUR ELECTRONIC INFORMATION INDUSTRY CO., LTD.) 15 July 2015 (2015-07-15) entire document	1-16
A	US 2013282780 A1 (LSI CORP.) 24 October 2013 (2013-10-24) entire document	1-16

☐ 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
"D" document cited by the applicant in the international application
"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	Date of mailing of the international search report
15 September 2023	**18 September 2023**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/102089

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 111340207 A	26 June 2020	None	
CN 104778026 A	15 July 2015	None	
US 2013282780 A1	24 October 2013	US 8898214 B2	25 November 2014

国际检索报告

国际申请号
PCT/CN2023/102089

A. 主题的分类

G06F7/483(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)

IPC：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））

CNKI、CNTXT，ENTXT，ENTXTC，DWPI：浮点数，转换，符号段，指数段，尾数段，floating-point number, conversion, symbol segment, exponent segment, mantissa segment

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 111340207 A (南京大学) 2020年6月26日 (2020 - 06 - 26) 说明书第0005-0153段	1、8、15-16
A	CN 104778026 A (浪潮电子信息产业股份有限公司) 2015年7月15日 (2015 - 07 - 15) 全文	1-16
A	US 2013282780 A1 (LSI CORPORATION) 2013年10月24日 (2013 - 10 - 24) 全文	1-16

☐ 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

“D” 申请人在国际申请中引证的文件

“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	国际检索报告邮寄日期
2023年9月15日	2023年9月18日
ISA/CN的名称和邮寄地址 中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	受权官员 刘瑛 电话号码 (+86) 62412234

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/102089

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 111340207 A	2020年6月26日	无	
CN 104778026 A	2015年7月15日	无	
US 2013282780 A1	2013年10月24日	US 8898214 B2	2014年11月25日