[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 7/57 (2006.01)
G06F 7/499 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200780006490.4

[43] 公开日 2009 年 3 月 18 日

[11] 公开号 CN 101390045A

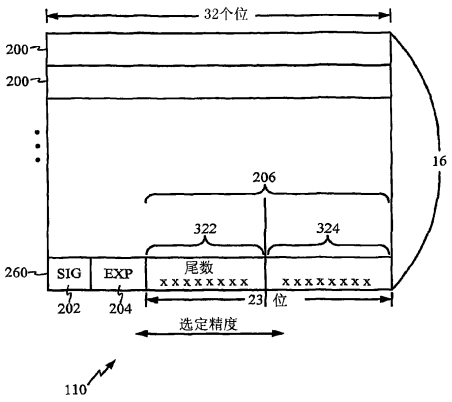
[22] 申请日 2007.2.27
[21] 申请号 200780006490.4
[30] 优先权
[32] 2006. 2.27 [33] US [31] 11/363,118
[86] 国际申请 PCT/US2007/062908 2007.2.27
[87] 国际公布 WO2007/101216 英 2007.9.7
[85] 进入国家阶段日期 2008.8.22
[71] 申请人 高通股份有限公司
地址 美国加利福尼亚州
[72] 发明人 肯尼思·艾伦·多克塞尔
[74] 专利代理机构 北京律盟知识产权代理有限责任公司
代理人 刘国伟
权利要求书 3 页 说明书 8 页 附图 4 页

[54] 发明名称

对可选次精度的功率要求减少的浮点处理器

[57] 摘要

本发明揭示一种用于用具有给定精度的浮点处理器来执行浮点运算的方法和设备。 为对一个或一个以上浮点数字进行所述浮点运算选择次精度。 所述次精度的选择针对所述一个或一个以上浮点数字中的每一者产生一个或一个以上过剩位。 可从所述浮点处理器中另外将用于存储或处理所述一个或一个以上过剩位的一个或一个以上组件移除功率，且用从所述一个或一个以上组件移除的功率来执行所述浮点运算。



1. 一种用具有最大精度的浮点处理器来执行浮点运算的方法，其包括：

为对一个或一个以上浮点数字进行所述浮点运算选择小于所述最大精度的次精度，所述次精度的所述选择针对所述一个或一个以上浮点数字中的每一者产生一个或一个以上过剩位；

从所述浮点处理器中另外将用于存储或处理所述一个或一个以上过剩位的一个或一个以上组件移除功率；以及

用从所述一个或一个以上组件移除的功率来执行所述浮点运算。

2. 根据权利要求 1 所述的方法，其进一步包括使用具有多个存储元件的浮点寄存器，所述一个或一个以上过剩位存储在所述存储元件中的一者或一者以上中，且其中所述从中移除功率的一个或一个以上组件包含用于所述一个或一个以上过剩位的所述存储元件。

3. 根据权利要求 2 所述的方法，其进一步包括使用具有用以执行所述浮点运算的逻辑的浮点运算器，且其中所述从中移除功率的一个或一个以上组件包含所述逻辑的另外将用于处理所述一个或一个以上过剩位的部分。

4. 根据权利要求 1 所述的方法，其进一步包括使用具有用以执行所述浮点运算的逻辑的浮点运算器，且其中所述从中移除功率的一个或一个以上组件包含所述逻辑的另外将用于处理所述一个或一个以上过剩位的部分。

5. 根据权利要求 4 所述的方法，其中所述浮点运算包括加法。

6. 根据权利要求 5 所述的方法，其进一步包括将来自所述逻辑的所述部分的进位输出强制为零。

7. 根据权利要求 4 所述的方法，其中所述浮点运算包括乘法。

8. 一种具有最大精度的浮点处理器，其包括：

浮点控制器，其经配置以为对一个或一个以上浮点数字进行浮点运算选择小于所述最大精度的次精度，所述次精度的所述选择针对所述一个或一个以上浮点数字中的每一者产生一个或一个以上过剩位，所述浮点控制器进一步经配置以从所述浮点处理器中另外将用于存储或处理所述一个或一个以上过剩位的一个或一个以上组件移除功率；以及

浮点运算器，其经配置以执行所述浮点运算。

9. 根据权利要求 8 所述的浮点处理器，其进一步包括浮点寄存器，所述浮点寄存器具有多个存储元件，所述一个或一个以上过剩位存储在所述存储元件中的一者或一者以上中，且其中所述可从中移除功率的一个或一个以上组件包含用于所述一个或一个以上过剩位的所述存储元件。

10. 根据权利要求 9 所述的浮点处理器，其中所述浮点运算器包括用以执行所述浮点运算的逻辑，且其中所述可从中移除功率的一个或一个以上组件包含所述逻辑的另外将用于处理所述一个或一个以上过剩位的部分。

11. 根据权利要求 8 所述的浮点处理器，其中所述浮点运算器包括用以执行所述浮点运算的逻辑，且其中所述可从中移除功率的一个或一个以上组件包含所述逻辑的另外将用于处理所述一个或一个以上过剩位的部分。

12. 根据权利要求 11 所述的浮点处理器，其中所述浮点运算器包含浮点加法器。

13. 根据权利要求 12 所述的浮点处理器，其中所述浮点运算器进一步经配置以在所述功率被移除时，将来自所述逻辑的所述部分的进位输出强制为零。

14. 根据权利要求 11 所述的浮点处理器，其中所述浮点运算器包含浮点乘法器。

15. 一种具有最大精度的浮点处理器，其包括：

浮点寄存器，其具有多个存储元件，所述存储元件经配置以存储多个浮点数字；

浮点运算器，其经配置以对存储在所述浮点寄存器中的所述浮点数字中的一者或一者以上执行浮点运算；以及

浮点控制器，其经配置以为对所述浮点数字中的所述一者或一者以上进行浮点运算选择小于所述最大精度的次精度，所述次精度的所述选择针对所述浮点数字中的所述一者或一者以上中的每一者产生一个或一个以上过剩位，所述一个或一个以上过剩位存储在所述浮点寄存器的所述存储元件中的一者或一者以上中，且其中所述浮点控制器进一步经配置以从用于所述一个或一个以上过剩位的所述存储元件移除功率。

16. 根据权利要求 15 所述的浮点处理器，其中所述浮点运算器包括经配置以执行所述浮点运算的逻辑，且其中所述浮点控制器进一步经配置以从所述逻辑的另外将用于处理所述一个或一个以上过剩位的部分移除功率。

17. 根据权利要求 16 所述的浮点处理器，其中所述浮点运算器包含浮点加法器。

18. 根据权利要求 17 所述的浮点处理器，其中所述浮点运算器进一步经配置以在所述功率被移除时，将来自所述逻辑的所述部分的进位输出强制为零。

19. 根据权利要求 16 所述的浮点处理器，其中所述浮点运算器包含浮点乘法器。

20. 一种具有最大精度的浮点处理器，其包括：

浮点寄存器，其经配置以存储多个浮点数字；

浮点运算器，其具有经配置以对存储在所述浮点寄存器中的所述浮点数字中的一者或一者以上执行浮点运算的逻辑；以及

浮点控制器，其经配置以为对所述浮点数字中的所述一者或一者以上进行浮点运算选择小于所述最大精度的次精度，所述次精度的所述选择针对所述浮点数字中的所述一者或一者以上中的每一者产生一个或一个以上过剩位，且其中所述浮点控制器进一步经配置以从所述逻辑的另外将用于处理所述一个或一个以上过剩位的部分移除功率。

21. 根据权利要求 20 所述的浮点处理器，其中所述浮动寄存器包括多个存储元件，所述存储元件经配置以存储所述浮点数字，所述一个或一个以上过剩位存储在所述存储元件中的一者或一者以上中，且其中所述浮点控制器进一步经配置以从用于所述一个或一个以上过剩位的所述存储元件移除功率。

22. 根据权利要求 20 所述的浮点处理器，其中所述浮点运算器包含浮点加法器。

23. 根据权利要求 22 所述的浮点处理器，其中所述浮点运算器进一步经配置以在所述功率被移除时将来自所述逻辑的所述部分的进位输出强制为零。

24. 根据权利要求 20 所述的浮点处理器，其中所述浮点运算器包含浮点乘法器，且功率被从所述浮点乘法器内包括部分乘积的元件的部分移除。

对可选次精度的功率要求减少的浮点处理器

技术领域

无

背景技术

浮点处理器是以较高速度执行某些数学运算（例如，乘法、除法、三角函数和指数函数）的专用计算单元。因此，强大的计算系统通常并入有浮点处理器，作为主处理器的一部分或作为协处理器。数字的浮点表示通常包含符号分量、指数和尾数。为了找出浮点数字的值，使尾数乘以升高为指数的幂的基数（在计算机中通常为 2）。将符号应用于所得值。

浮点处理器的精度由用于表示尾数的位的数目界定。尾数中的位越多，精度越大。浮点处理器的精度通常视特定应用而定。举例来说，ANSI/IEEE-754 标准（现代计算机通常遵循所述标准）指定 32 位单一格式，其具有 1 位符号、8 位指数和 23 位尾数。在 32 位编码中只存储尾数的 23 个分数位，紧接在二进制小数点左方的整数位是隐含的。IEEE-754 还指定 64 位双重格式，其具有 1 位符号、11 位指数和 53 位尾数。与单一编码类似，在 64 位编码中只存储尾数的 52 个分数位，紧接在二进制小数点左方的整数位是隐含的。较高精度可达成较高准确度，但通常导致功率消耗增加。

浮点算术运算的执行可能必然伴有计算低效，因为浮点处理器通常局限于单一格式或单一格式和双重格式两者所提供的精度。虽然一些应用可能需要这些类型的精度，但其它应用可能不需要。举例来说，一些图形应用可能只需要 16 位尾数。对于这些图形应用，任何超过 16 位精度的准确度都趋向于导致不必要的功率消耗。这在功率特别受到重视的电池操作的装置中尤其受到关注，所述装置例如是无线电话、个人数字助理（PDA）、膝上型计算机、游戏控制台、寻呼机和相机（只列举几种）。如果已知应用总是需要某一减小的精度，那么可根据所述减小的精度来设计和构造浮点处理器。然而，对于通用处理器，常见的情形是，对于某些应用（例如产生 3D 图形），减小的精度可能是可接受的，且对于其它应用（例如实施全球定位系统（GPS）功能），可能需要较大的精度。因此，此项技术中需要一种浮点格式的减小的精度或次精度为可选择的浮点处理器。还可使用功率管理技术来确保浮点处理器不会消耗比支持选定次精度所必需的功率多的功率。

发明内容

揭示一种用具有精度格式的浮点处理器来执行浮点运算的方法的一方面。所述方法包含为对一个或一个以上浮点数字的浮点运算选择次精度，所述次精度的选择针对一个或一个以上浮点数字中的每一者产生一个或一个以上过剩位。所述方法进一步包含：从浮点处理器中另外将用于存储或处理一个或一个以上过剩位的一个或一个以上组件移除功率；以及用从所述一个或一个以上组件移除的功率来执行浮点运算。

揭示具有精度格式的浮点处理器的一个方面。所述浮点处理器包含浮点控制器，其经配置以为对一个或一个以上浮点数字的浮点运算选择次精度，所述次精度的选择针对一个或一个以上浮点数字中的每一者产生一个或一个以上过剩位，所述浮点控制器进一步经配置以从浮点处理器中另外将用于存储或处理一个或一个以上过剩位的一个或一个以上组件移除功率。浮点处理器进一步包含浮点运算器，其经配置以执行浮点运算。

揭示具有精度格式的浮点处理器的另一方面。所述浮点处理器包含：浮点寄存器，其具有多个存储元件，所述存储元件经配置以存储多个浮点数字；以及浮点运算器，其经配置以对存储在浮点寄存器中的浮点数字中的一者或一者以上执行浮点运算。浮点处理器进一步包含浮点控制器，其经配置以为对浮点数字中的所述一者或一者以上的浮点运算选择次精度，所述次精度的选择针对浮点数字中的所述一者或一者以上中的每一者产生一个或一个以上过剩位，所述一个或一个以上过剩位存储在浮点寄存器的存储元件中的一者或一者以上中，且其中浮点控制器进一步经配置以从用于所述一个或一个以上过剩位的存储元件移除功率。

揭示具有精度格式的浮点处理器的又一方面。所述浮点处理器包含：浮点寄存器，其经配置以存储多个浮点数字；以及浮点运算器，其具有经配置以对存储在浮点寄存器中的浮点数字中的一者或一者以上执行浮点运算的逻辑。浮点处理器进一步包含浮点控制器，其经配置以为对浮点数字中的所述一者或一者以上的浮点运算选择次精度，所述次精度的选择针对浮点数字中的所述一者或一者以上中的每一者产生一个或一个以上过剩位，且其中浮点控制器进一步经配置以从所述逻辑的另外将用于处理所述一个或一个以上过剩位的部分移除功率。

应了解，所属领域的技术人员通过以下具体实施方式将容易明白浮点处理器和执行浮点运算的方法的其它实施例，以下具体实施方式中以说明的方式展示和描述浮点处理器和执行浮点运算的方法的各个实施例。将认识到，浮点处理器和执行浮点运算的方法的其它和不同实施例是可能的，且用于描述这些实施例的细节能够在许多方面进行修改。因此，附图和具体实施方式将被视为本质上是说明性的而非限制性的。

附图说明

图 1 是说明具有可选次精度的浮点处理器的实例的功能框图；

图 2 是在具有可选次精度的浮点处理器中使用的浮点寄存器堆的实例的图解说明；

图 3A 是说明使用具有可选次精度的浮点处理器执行的浮点加法的实例的概念图；以及

图 3B 是说明使用具有可选次精度的浮点处理器执行的浮点乘法的实例的概念图。

具体实施方式

下文结合附图而陈述的具体实施方式意在描述本发明的各个实施例，而无意表示可实践本发明的仅有实施例。为了便于全面理解本发明，具体实施方式包含特定细节。然而，所属领域的技术人员将了解，可在没有这些特定细节的情况下实践本发明。在有些情况下，为了更清楚地说明本发明的概念，以框图形式展示众所周知的结构和组件。

在浮点处理器的至少一个实施例中，一个或一个以上浮点运算的精度可相对于指定格式的精度而减小。另外，可使用功率管理技术来确保浮点处理器不会消耗比支持选定次精度所必需的功率多的功率。向浮点处理器提供的用以执行数学运算的指令可包含可编程控制字段。所述控制字段可用于选择浮点格式的次精度，且管理功率消耗。通过将浮点格式的次精度选择为特定运算所需的次精度，从而减少浮点处理器支持选定次精度的功率消耗，可实现较大的效率以及显著的功率节省。

图 1 是说明具有可选次精度的浮点处理器（FPP）100 的实例的功能框图。浮点处理器 100 包含浮点寄存器堆（FPR）110；浮点控制器（CTL）130；以及浮点数学运算器（FPO）140。浮点处理器 100 可实施为主处理器的一部分、协处理器或通过总线或其它信道连接到主处理器的单独实体。

浮点寄存器堆 110 可以是任何合适的存储媒体。在图 1 所示的实施例中，浮点寄存器堆 110 包含若干可寻址寄存器位置 115-1（REG1）、115-2（REG2）、…、115-N（REGN），其每一者经配置以存储浮点运算的运算数。所述运算数可包含（例如）来自存储器的数据和/或先前浮点运算的结果。向浮点处理器提供的指令可用于将运算数移动到主存储器或从主存储器移动运算数。

图 2 示意性地说明在具有可选次精度的浮点处理器 100 中使用的浮点寄存器堆 110（如结合图 1 所述）的数据结构的实例。在图 2 所说明的实施例中，浮点寄存器堆 110 包含十六个可寻址寄存器位置，为了方便起见，在图 2 中用参考标号 200 来指代每一寄存器位置。每一寄存器位置 200 经配置以存储 32 位二进制浮点数字（采用 IEEE-754 32 位

单一格式)。具体地说，每一寄存器位置 200 含有 1 位符号 202、8 位指数 204 和 24 位分数 206。然而，当然应了解，浮点处理器 100 的其它实施例可包含浮点寄存器堆 110，其格式与 IEEE 32 位单一格式不同（包含但不限于 IEEE 64 位双重格式)，且/或可含有不同数目的寄存器位置。

返回参看图 1，浮点控制器 130 可用于使用控制信号 133 来选择浮点运算的次精度。控制寄存器（CRG）137 可加载有（例如）在一个或一个以上指令的控制字段中传输的次精度选择位。以稍后将更详细描述的方式，浮点控制器 130 可使用次精度选择位来减小运算数的精度。次精度选择位还可用于断开浮点处理器 100 的一部分。举例来说，次精度选择位可用于从用于所选次精度不需要的位的浮点寄存器元件移除功率。次精度选择位还可用于从浮点运算器 FPO 140 中当选定次精度减小时不使用的逻辑移除功率。一连串开关可用于移除功率和将功率施加到浮点寄存器元件和浮点运算器 140 中的逻辑。所述开关可在浮点寄存器 110 和浮点运算器 140 的内部或外部，所述开关可以是场效应晶体管或任何其它类型的开关。

浮点运算器 140 可包含一个或一个以上经配置以执行浮点运算的组件。这些组件可包含（但不限于）计算单元，例如：浮点加法器（ADD）142，其经配置以执行浮点相加和相减指令；及浮点乘法器（MUL）144，其经配置以执行浮点相乘指令。如图 1 中所见，浮点运算器 140 中的计算单元 ADD 142 和 MUL 144 中的每一者以允许在计算单元之间且在每一计算单元与浮点寄存器堆 110 之间转移运算数的方式彼此耦合且耦合到浮点寄存器堆 110。浮点运算器可通过各个连接件 134、135、136、137、138 和 139 耦合到浮点寄存器（如所说明的)，或可经由总线或任何其它合适耦合件而耦合。在浮点处理器 100 的至少一个实施例中，计算单元（ADD 142 和 MUL 144）中的任何一者的输出可以是任何其它计算单元的输入。浮点寄存器堆 110 可用于存储中间结果以及从浮点运算器 140 输出的结果。

加法器 142 可以是常规浮点加法器，其经配置以便以浮点格式执行标准算术运算。乘法器 144 可以是常规浮点乘法器，其经配置以执行浮点乘法。乘法器 144 可与（例如）布斯（Booth）或经修改的布斯算法一起实施，且可包含：部分乘积产生逻辑，其产生部分乘积；以及许多进位保留加法器（carry-save adder)，其将所述部分乘积相加。

虽然为了简单起见，图 1 中只展示了加法器 142 和乘法器 144，但浮点运算器 140 还可包含其它计算单元（未图示)，所述计算单元是此项技术中已知的，且经配置以执行其它类型的浮点数学运算。这些计算单元可包含（但不限于)：浮点除法器，其经配置以

执行浮点相除指令；浮点开方器，其经配置以执行浮点开方指令；浮点指数运算器，其经配置以执行浮点指数指令；浮点对数运算器，其经配置以执行用于计算对数函数的指令；以及浮点三角运算器，其经配置以执行用于计算三角函数的指令。

浮点处理器 100 的不同实施例可包含上文所列计算单元中的仅一个、或一些或所有。举例来说，加法器 142 和乘法器 144 每一者可包含一个或一个以上众所周知的常规子单元，例如：对准器，其使输入运算数对准；规范器(normalizer)，其将结果转换成标准格式；以及舍入器，其基于指定舍入模式而对结果进行舍入。加法器 142 和乘法器 144 中还包含众所周知的电路元件，例如位反相器、多路复用器、计数器和组合逻辑电路。

如图 1 中所说明，浮点运算器 140 耦合到浮点寄存器堆 110，使得对于所请求的浮点运算的每一指令，相关计算单元（即，加法器 142 或乘法器 144）可从浮点寄存器堆 110 接收存储在寄存器位置 REG1、…、REGN 中的一者或一者以上中的一个或一个以上运算数。

在从浮点寄存器堆 110 接收到运算数之后，浮点运算器 140 中的一个或一个以上计算单元可以浮点控制器 130 所选择的次精度，对接收到的运算数执行所请求的浮点运算的指令。可将输出发送回到浮点寄存器 110 以供存储，如图 1 中所示。

在浮点处理器 110 的实施例中，可在程序控制下使用软件可选模式来减小浮点运算的精度，或者如上文所阐释，向浮点处理器 100 提供的指令可包含含有次精度选择位的可编程控制字段。将次精度选择位写入到控制寄存器 137，控制寄存器 137 又在浮点运算期间控制每一运算数的尾数的位长度。或者，可将次精度选择位直接从任何合适的用户界面写入到控制寄存器 137，所述用户界面包含（例如，但不限于）图 1 中所示的监视器屏幕/键盘或鼠标 150。在浮点处理器 100 的另一实施例中，可将次精度选择位直接从主处理器或其操作系统写入到控制寄存器 137。控制寄存器 137（其图示于浮点控制器 130 中）可作为独立实体驻存在其它位置、并入到另一实体中或分布在多个实体上。

次精度选择位可用于减小浮点运算的精度。这可以多种方式来实现。在一个实施例中，浮点控制器 130 可使功率从用于分数的满足次精度选择位所指定精度所不需要的过剩位的浮点寄存器元件被移除。举例来说，如果浮点寄存器文件中的每一位置含有 23 位分数，且浮点运算所需的次精度为 10 位，那么只需要分数的 9 个普通有效位（MSB）；隐藏位或整数位构成第十个位。可从用于其余 14 个分数位的浮点寄存器元件移除功率。如果一个或一个以上指令的次精度增加到 16 位，那么将需要尾数的 15 个 MSB。在后一种情况下，可从用于分数的 8 个最低有效位（LSB）的浮点寄存器元件移除功率。

另外，浮点运算器 140 中对应于过剩尾数位的逻辑不需要功率。因此，可通过将功率移除到浮点运算器 140 中由于所选择的次精度而保持未使用的逻辑。

图 3A 是说明浮点加法运算的实例的概念图，其中选择性地将功率施加到浮点运算器中的逻辑。具体地说，图 3A 概念性地说明浮点加法运算，其中两个输入浮点数字 302 和 304（每一者由选定次精度表征）相加在一起。出于简单起见，假定两个数字 302 和 304 已经对准，从而不需要进行移位。通过一连串级（参看图 3A，具有参考标号 310_1、301_2、310_i、…、310_n）来执行完整精度模式下的浮点加法运算。根据标准惯例，浮点寄存器依序存储组成每个数字的位，从最右边的 LSB 到最左边的 MSB。跨越图 3A 从右向左移动的级中的每一连续级均涉及与先前级中所涉及的位相比有效性增加的位。

在图 3A 所说明的实例中，用线 305 来表示选定次精度。可从用于实施线 305 右边的每一级的逻辑移除功率。将来自最后断电的级 310_i 的进位输出 C 强制成零。功率只供应到用于实施线 305 左边的每一级的逻辑。在图 3A 中，向浮点运算器的现用级提供的带电的位使用参考标号 322 展示为 X，而向功率被移除的级提供的不带电的位使用参考标号 324 展示为圆圈。

图 3B 是说明浮点乘法运算的实例的概念图，其中将功率选择性地施加到浮点运算器中的逻辑。在浮点乘法器 MUL（图 1 中以参考标号 144 展示）中执行浮点乘法运算。在乘法器中，可使大量逻辑断电，从而提供显著的功率节省。如图 3B 中所说明的二进制乘法基本上是经移位浮点数字的一连串加法。在所说明的实施例中，使用移位与相加技术，在 k 位被乘数 402 与 k 位乘数 404 之间执行二进制乘法。移位与相加技术可由布斯算法或经修改的布斯算法乘法器代替。

与在浮点相加的情况下一样，在一连串级（在图 3B 中说明为 410-1、…、410-m）中执行浮点相乘。出于简单起见假定使用布斯算法，可针对乘数 404 中的每个位产生一个部分乘积，部分乘积 420-i 在对应的级 410-i 期间产生。如果乘数的值为 0，那么其对应的部分乘积只由 0 组成；如果所述位的值为 1，那么其对应的部分乘积是被乘数的拷贝。每一部分乘积 420-i 以随与之相关联的乘数位而变的形式向左移位，此后运算继续移到下一个级。每一部分乘积因此可被视为经移位的数字。与乘数中的位 0 相关联的部分乘积向左移位零个位，且与位 1 相关联的部分乘积向左移位一个位。将部分乘积或经移位的浮点数字 420-i 相加在一起，以便为相乘产生输出值 430。

在图 3B 所说明的实施例中，用线 405 来指示控制器 130 对所需减小精度的选择。与在结合图 3A 而描述的浮点加法的情况下一样，可从用于实施线 405 右边的级的逻辑移除

功率。只将功率施加到支持选定次精度实际需要的级，即在线 405 左边的级。在图 3B 中，向已加电的逻辑提供的位展示为 X，而向已断电的级提供的位展示为圆圈。

如从图 3B 所见，对于第一部分乘积 420-1，用于许多（N）个位（使用参考标号 402 展示）的逻辑不带电。对于第二部分乘积，用于 N-1 个位的逻辑不带电，依此类推。对于第 m 个部分乘积或经移位的浮点数字 420-m，用于许多（N-m+1）个位（使用参考标号 414 展示）的逻辑不带电。位的数目 N 经选择以使得其余级的精度不会受到不利影响。

从上文所述的浮点乘法得出的输出值的宽度（即位的数目）等于相乘在一起的两个输入值 402 和 404 的宽度的总和。输出值 430 可舍位到选定次精度，即可舍去输出值 430 中精度小于选定精度的位中的任何一者，以产生由选定精度表征的经舍位的输出数字。或者，可将输出值 430 舍入到选定精度。在任一种情况下，有效性小于选定精度的输出位也可不带电。

结合本文所揭示的实施例而描述的各种说明性逻辑单元、区块、模块、电路、元件和/或组件可在浮点处理器中实施或执行，浮点处理器是通用处理器、数字信号处理器（DSP）、专用集成电路（ASIC）、现场可编程门阵列（FPGA）或其它可编程逻辑组件、离散门或晶体管逻辑、离散硬件组件或其经设计以执行本文所述功能的任何组合的一部分。通用处理器可以是微处理器，但在替代方案中，处理器可以是任何常规处理器、控制器、微控制器或状态机。处理器还可实施为计算组件的组合，例如 DSP 与微处理器的组合、多个微处理器、一个或一个以上微处理器与 DSP 核心的结合或任何其它此配置。

结合本文所揭示的实施例而描述的方法或算法可直接在硬件中、在由处理器执行的软件模块中或在所述两者的组合中实施。软件模块可驻存在 RAM 存储器、快闪存储器、ROM 存储器、EPROM 存储器、EEPROM 存储器、寄存器、硬盘、可拆式盘、CD-ROM 或此项技术中已知的任何其它形式的存储媒体中。存储媒体可耦合到处理器，使得处理器可从存储媒体读取信息和向存储媒体写入信息。在替代方案中，存储媒体可与处理器形成一体。

提供对所揭示实施例的先前描述是为了使所属领域的技术人员能够制作或使用本发明。所属领域的技术人员将容易了解对这些实施例的各种修改，且在不脱离本发明精神或范围的情况下，本文所界定的一般原理可应用于其它实施例。因此，不希望本发明限于本文展示的实施例，而是希望赋予本发明与权利要求书一致的完整范围，其中以单数形式对元件的参考不希望意味着“一个且只有一个”（除非明确这样陈述），而是意味着“一个或一个以上”。所属领域的技术人员已知或以后将知道的在整个本发明中描述的各

种实施例的元件的所有结构和功能等效物都特意以引用的方式并入本文中，且希望权利要求书涵盖所述结构和功能等效物。此外，本文所揭示的内容都无意奉献给公众，不管权利要求书中是否明确陈述此揭示内容。所有权利要求要件都不应根据 35 U.S.C. §112 第六段的规定来解释，除非特意使用短语“用于……的装置”来陈述所述要件，或在方法项的情况下使用短语“用于……的步骤”来陈述所述要件。

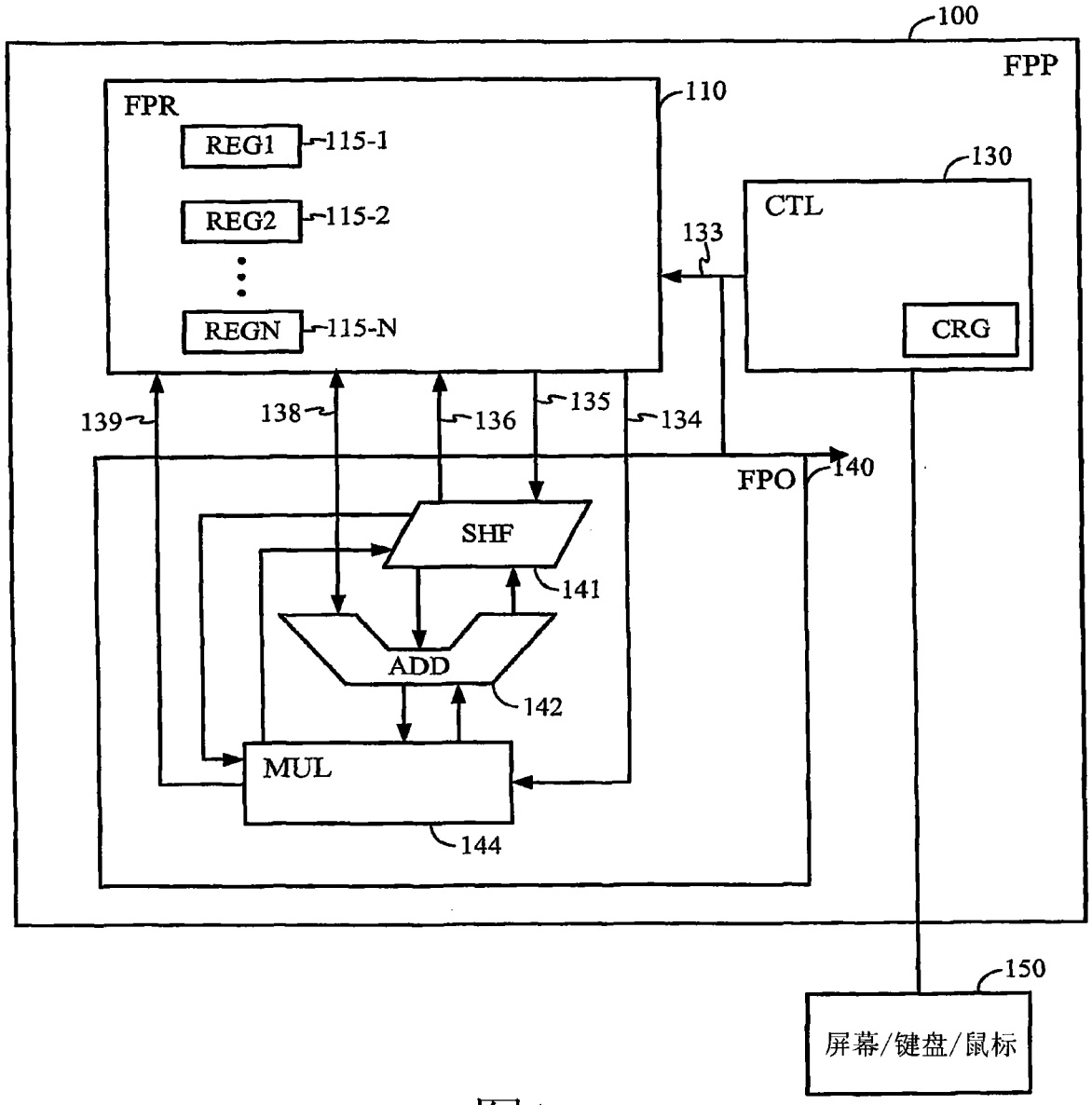

100
FPP
FPR
110
REG1
115-1
REG2
115-2
REGN
115-N
130
CTL
133
CRG
139
138
136
135
134
FPO
140
SHF
141
ADD
142
MUL
144
150
屏幕/键盘/鼠标

图1

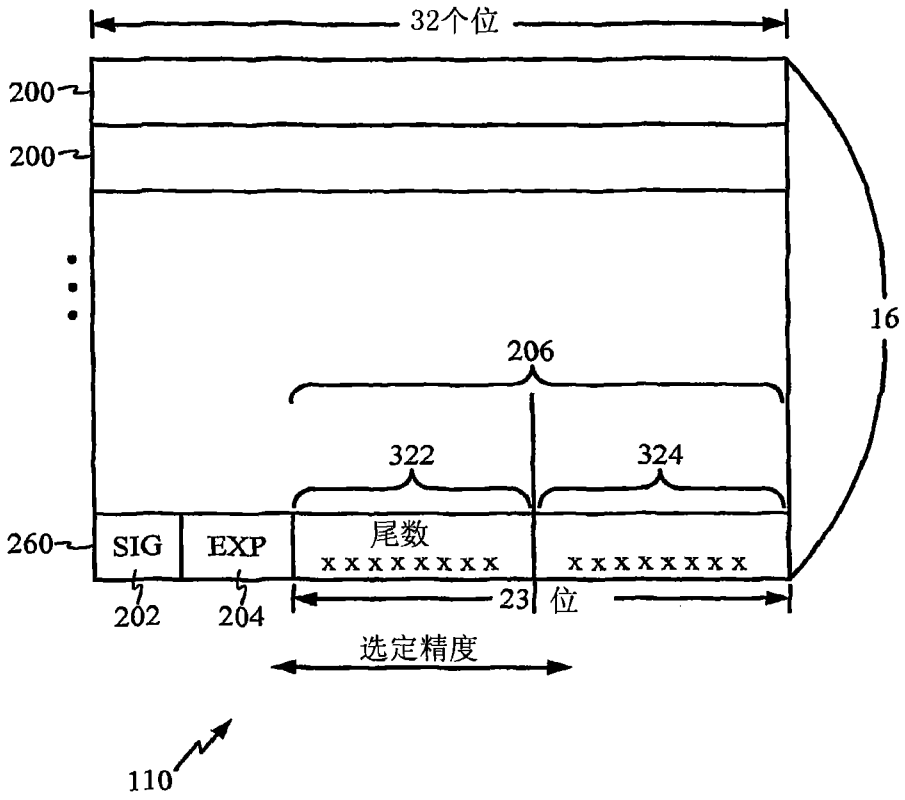

32个位
200
200
16
206
322
324
260
SIG
EXP
尾数
xxxxxxxx
xxxxxxxx
202
204
23 位
选定精度
110

图2

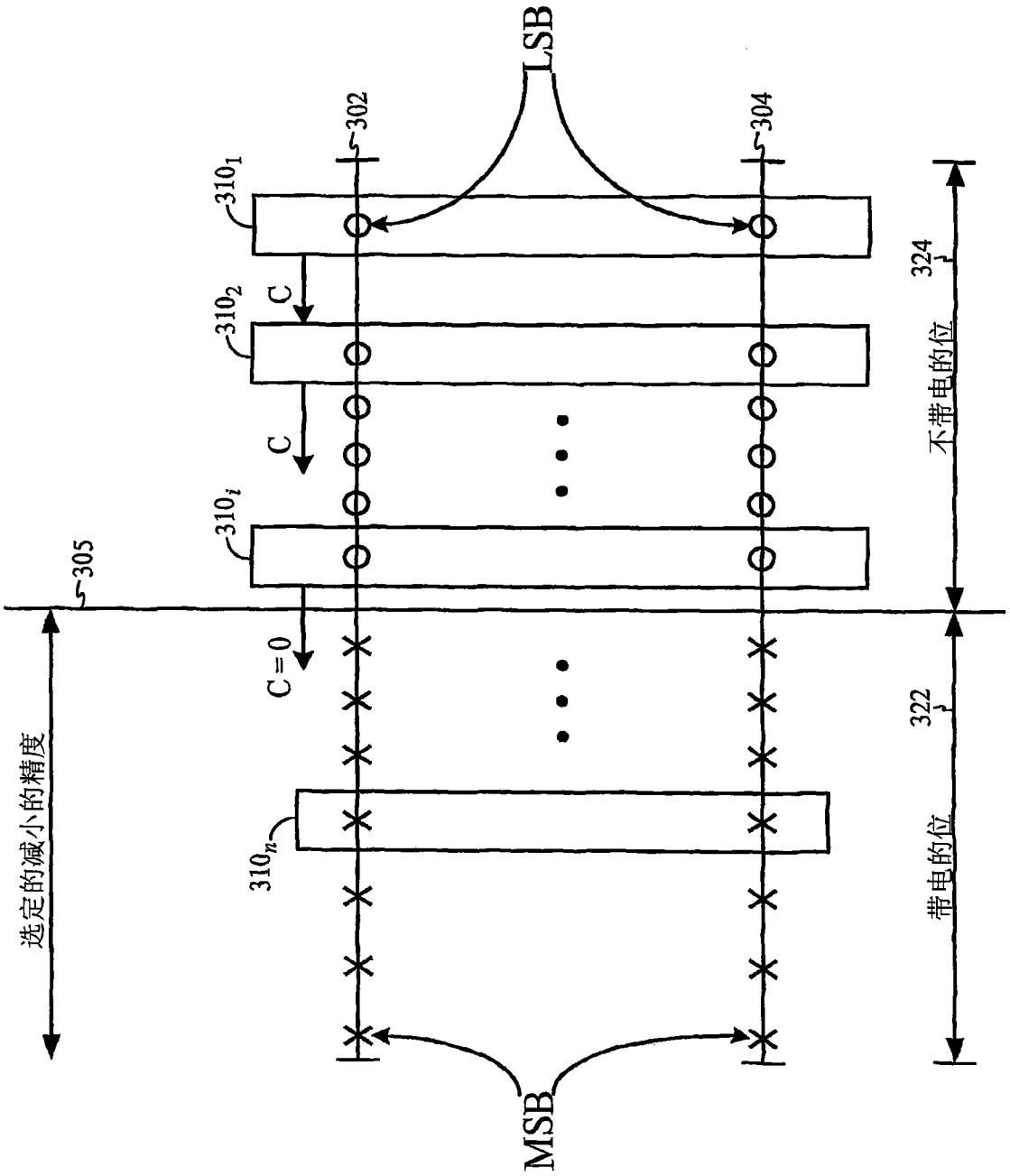

LSB
302
304
310_1
310_2
310_i
310_n
C
C
C=0
305
选定的减小的精度
324
不带电的位
322
带电的位
MSB

图3A

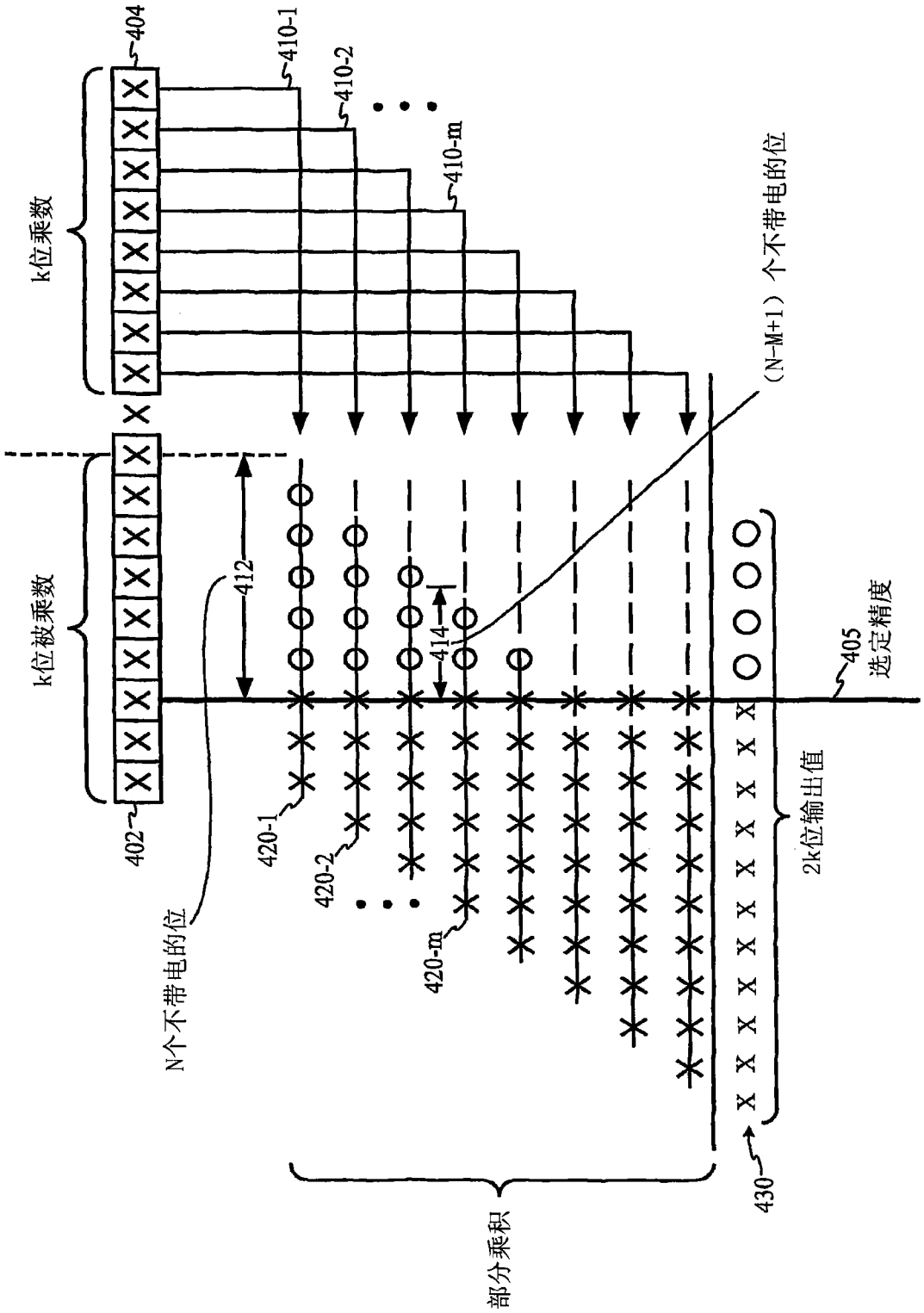

k位乘数
404
410-1
410-2
410-m
k位被乘数
402
412
414
420-1
420-2
420-m
N个不带电的位
(N-M+1)个不带电的位
405
选定精度
2k位输出值
430
部分乘积

图3B