República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) **PI0708284-3 A2**

BRPI0708284A2

(22) Data de Depósito: 27/02/2007
(43) Data da Publicação: 24/05/2011
(RPI 2107)

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

(54) Título: **PROCESSADOR DE PONTO FLUTUANTE COM EXIGÊNCIAS DE POTÊNCIA REDUZIDA PARA SUB-PRECISÃO SELECIONÁVEL**

(30) Prioridade Unionista: 27/02/2006 US 11/363,118

(73) Titular(es): Qualcomm Incorporated

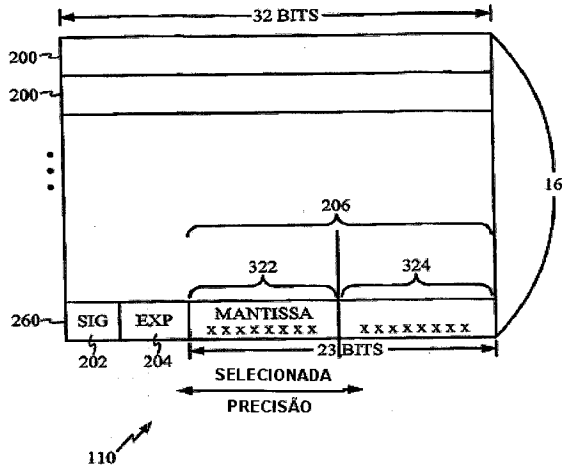
(72) Inventor(es): Kenneth Alan Dockser

(74) Procurador(es): Montaury Pimenta, Machado & Lioce

(86) Pedido Internacional: PCT US2007062908 de 27/02/2007

(87) Publicação Internacional: WO 2007/101216de 07/09/2007

(57) Resumo: PROCESSADOR DE PONTO FLUTUANTE COM EXIGÊNCIAS DE POTÊNCIA REDUZIDA PARA SUB-PRECISÁO SELECIONÁVEL. Um método e aparelho para a realização de uma operação de ponto flutuante com um processador de ponto flutuante possuindo uma determinada precisão são descritos. Uma sub-precisão para a operação de ponto flutuante em um ou mais dos números de ponto flutuante é selecionada. A seleção da sub-precisão resulta em um ou mais bits excessivos para cada um dos um ou mais números de ponto flutuante. A potência pode ser removida de um ou mais componentes no processador de ponto flutuante que, do contrário, seria utilizado para armazenar ou processar os um ou mais bits excessivos, e a operação de ponto flutuante é realizada com a potência removida dos um ou mais componentes.



"PROCESSADOR DE PONTO FLUTUANTE COM EXIGÊNCIAS DE POTÊNCIA REDUZIDA PARA SUB-PRECISÃO SELECIONÁVEL"

FUNDAMENTOS

Os processadores de ponto flutuante são unidades computacionais especializadas que realizam determinadas operações matemáticas, por exemplo, multiplicação, divisão, funções trigonométricas, e funções exponenciais, a altas velocidades. Portanto, sistemas computacionais poderosos freqüentemente incorporam processadores de ponto flutuante, como parte do processador principal ou como um co-processador. Uma representação de ponto flutuante de um número inclui comumente um componente de sinal, um expoente e uma mantissa. Para se encontrar o valor de um número de ponto flutuante, a mantissa é multiplicada por uma base (comumente 2 em computadores) elevada à potência do expoente. O sinal é aplicado ao valor resultante.

A precisão do processador de ponto flutuante é definida pelo número de bits utilizado para representar a mantissa. Quanto mais bits na mantissa, maior a precisão. A precisão do processador de ponto flutuante depende geralmente da aplicação em particular. Por exemplo, o padrão ANSI/IEEE-754 (comumente seguido pelos computadores modernos) especifica um formato único de 32 bits possuindo um sinal de 1 bit, um expoente de 8 bits, e uma mantissa de 23 bits. Apenas a fração de 23 bits da mantissa é armazenada na codificação de 32 bits, um bit inteiro, imediatamente à esquerda do ponto binário, é implicado. O IEEE-754 também especifica um formato duplo de 64 bits possuindo um sinal de 1 bit, um expoente de 11 bits, e uma mantissa de 53 bits. Análogo à codificação única, apenas a fração de 52 bits da mantissa é armazenada na codificação de 64 bits, um bit inteiro, imediatamente à esquerda do ponto binário, é implicado. A maior precisão resulta em uma

maior exatidão, mas comumente resulta em consumo de energia aumentado.

O desempenho das operações aritméticas de ponto flutuante pode resultar em ineficiência computacional visto que os processadores de ponto flutuante são comumente limitados à precisão fornecida pelo formato único, ou ambos os formatos, único e duplo. Enquanto algumas aplicações podem exigir esses tipos de precisão, outras aplicações podem não exigir. Por exemplo, algumas aplicações gráficas podem apenas exigir uma mantissa de 16 bits. Para essas aplicações gráficas, qualquer precisão além de 16 bits de precisão tende a resultar em consumo de energia desnecessário. Essa é uma preocupação particular para dispositivos operados por bateria onde a potência se torna crítica, tal como telefones sem fio, assistentes digitais pessoais (PDA), laptops, consoles de jogos, pagers, e câmeras, apenas para citar alguns. Se for sabido que um aplicativo exige sempre uma determinada precisão reduzida, o processador de ponto flutuante pode ser projetado e construído de acordo com essa precisão reduzida. Para os processadores de finalidade geral, no entanto, a situação comum é que para determinados aplicativos, por exemplo, geração de gráficos em 3D, uma precisão reduzida pode ser aceitável, e para outros aplicativos, por exemplo, implementação de funções de Sistema de Posicionamento Global (GPS), uma maior precisão pode ser necessária. Portanto, existe uma necessidade na técnica para se criar um processador de ponto flutuante no qual a precisão reduzida, ou sub-precisão, do formato de ponto flutuante é selecionável. As técnicas de gerenciamento de energia também podem ser empregadas para garantir que esse processador de ponto flutuante não consuma mais energia do que o necessário para suportar a sub-precisão selecionada.

SUMÁRIO

Um aspecto de um método de realização de uma operação de ponto flutuante com um processador de ponto flutuante possuindo um formato de precisão é descrito. O método inclui a seleção de uma sub-precisão para a operação de ponto flutuante em um ou mais números de ponto flutuante, a seleção da sub-precisão resultando em um ou mais bits excessivos para cada um dos um ou mais números de ponto flutuante. O método inclui adicionalmente a remoção da potência de um ou mais componentes no processador de ponto flutuante que, do contrário, seria utilizado para armazenar ou processar os um ou mais bits excessivos, e realizar a operação de ponto flutuante com potência removida dos um ou mais componentes.

Um aspecto de um processador de ponto flutuante possuindo um formato de precisão é descrito. O processador de ponto flutuante inclui um controlador de ponto flutuante configurado para selecionar uma sub-precisão para uma operação de ponto flutuante em um ou mais números de ponto flutuante, a seleção da sub-precisão resultando em um ou mais bits excessivos para cada um dos um ou mais números de ponto flutuante, o controlador de ponto flutuante sendo configurado adicionalmente para remover potência de um ou mais componentes no processador de ponto flutuante que, do contrário, seria utilizado para armazenar ou processar um ou mais dos bits excessivos. O processador de ponto flutuante inclui adicionalmente um operador de ponto flutuante configurado para realizar a operação de ponto flutuante.

Outro aspecto de um processador de ponto flutuante possuindo um formato de precisão é descrito. O processador de ponto flutuante inclui um registro de ponto flutuante possuindo uma pluralidade de elementos de

armazenamento configurados para armazenar uma pluralidade de números de ponto flutuante, e um operador de ponto flutuante configurado para realizar uma operação de ponto flutuante em um ou mais dos números de ponto flutuante armazenados no registro de ponto flutuante. O processador de ponto flutuante inclui adicionalmente um controlador de ponto flutuante configurado para selecionar uma sub-precisão para uma operação de ponto flutuante em um ou mais números de ponto flutuante, a seleção da sub-precisão resultando em um ou mais bits excessivos para cada um dos um ou mais dos números de ponto flutuante, os um ou mais bits excessivos sendo armazenados em um ou mais dos elementos de armazenamento do registro de ponto flutuante, e onde o controlador de ponto flutuante é adicionalmente configurado para remover a potência dos elementos de armazenamento para os um ou mais bits excessivos.

Um aspecto adicional de um processador de ponto flutuante possuindo um formato de precisão é descrito. O processador de ponto flutuante inclui um registro de ponto flutuante configurado para armazenar uma pluralidade de números de ponto flutuante, e um operador de ponto flutuante possuindo lógica configurada para realizar uma operação de ponto flutuante em um ou mais dos números de ponto flutuante armazenados no registro de ponto flutuante. O processador de ponto flutuante inclui adicionalmente um controlador de ponto flutuante configurado para selecionar uma sub-precisão para uma operação de ponto flutuante nos um ou mais dos números de ponto flutuante, a seleção da sub-precisão resultando em um ou mais bits excessivos para cada um dos ditos um ou mais números de ponto flutuante, e onde o controlador de ponto flutuante é adicionalmente configurado para remover energia de uma parte da lógica

que, do contrário, seria utilizada para processar os um ou mais bits excessivos.

Deve-se entender que outras modalidades do processador de ponto flutuante, e do método de realização das operações de ponto flutuante, se tornarão prontamente aparentes aos versados na técnica a partir da descrição detalhada a seguir, na qual várias modalidades do processador de ponto flutuante e do método de realização das operações de ponto flutuante são apresentadas e descritas por meio de ilustração. Como será realizado, outras modalidades diferentes do processador de ponto flutuante e do método de realização das operações de ponto flutuante são possíveis, e os detalhes utilizados para descrever essas modalidades podem ser modificados em muitos aspectos. Portanto, os desenhos e a descrição detalhada devem ser considerados como ilustrativos por natureza e não como restritivos.

BREVE DESCRIÇÃO DOS DESENHOS

A figura 1 é um diagrama de blocos funcional ilustrando um exemplo de um processador de ponto flutuante com sub-precisão selecionável;

A figura 2 é uma ilustração gráfica de um exemplo de um arquivo de registro de ponto flutuante utilizado em um processador de ponto flutuante com sub-precisão selecionável;

A figura 3a é um diagrama conceitual ilustrando um exemplo de uma adição de ponto flutuante realizada utilizando-se um processador de ponto flutuante com sub-precisão escalonável; e

A figura 3b é um diagrama conceitual ilustrando um exemplo de uma multiplicação de ponto flutuante que é realizada utilizando-se um processador de ponto flutuante com sub-precisão selecionável.

DESCRIÇÃO DETALHADA

A descrição detalhada apresentada abaixo com relação aos desenhos em anexo deve descrever as várias modalidades da presente descrição, mas não deve representar a única modalidade na qual a presente descrição pode ser praticada. A descrição detalhada inclui detalhes específicos, a fim de permitir uma compreensão profunda da presente descrição. Será apreciado pelos versados na técnica, no entanto, que a presente descrição pode ser praticada sem esses detalhes específicos. Em alguns casos, as estruturas e componentes bem conhecidos são apresentados sob a forma de diagrama de blocos, a fim de ilustrar mais claramente os conceitos da presente descrição.

Em pelo menos uma modalidade de um processador de ponto flutuante, a precisão para uma ou mais operações de ponto flutuante podem ser reduzidas a partir do formato especificado. Adicionalmente, as técnicas de gerenciamento de potência podem ser empregadas para garantir que o processador de ponto flutuante não consuma mais energia do que o necessário para suportar a sub-precisão selecionada. As instruções fornecidas para o processador de ponto flutuante para realização das operações matemáticas podem incluir um campo de controle programável. O campo de controle pode ser utilizado para selecionar a sub-precisão do formato de ponto flutuante e gerenciar o consumo de energia. Pela seleção da sub-precisão do formato de ponto flutuante, de acordo com a necessária para uma operação particular, reduzindo, assim, o consumo de energia do processador de ponto flutuante para suportar a sub-precisão selecionada, uma maior eficiência além de uma economia energética significativa podem ser alcançadas.

A figura 1 é um diagrama de blocos funcional ilustrando um exemplo de um processador de ponto flutuante

(FPP) 100 com sub-precisão selecionável. O processador de ponto flutuante 100 inclui um arquivo de registro de ponto flutuante (FPR) 110; um controlador de ponto flutuante (CTL) 130; e um operador matemático de ponto flutuante (FPO) 140. O processador de ponto flutuante 100 pode ser implementado como parte do processador principal, um co-processador, ou uma entidade separada conectada ao processador principal através de um barramento ou outro canal.

O arquivo de registro de ponto flutuante 100 pode ser qualquer meio de armazenamento adequado. Na modalidade ilustrada na figura 1, o arquivo de registro de ponto flutuante 110 inclui várias localizações de registro endereçável 115-1 (REG1), 115-2 (REG2), ..., 115-N (REGN), cada uma configurada para armazenar um operando para uma operação de ponto flutuante. Os operandos podem incluir, por exemplo, dados de uma memória e/ou resultados de operações de ponto flutuante anteriores. Instruções fornecidas para o processador de ponto flutuante podem ser utilizadas para mover os operandos para e da memória principal.

A figura 2 ilustra de forma esquemática um exemplo da estrutura de dados para um arquivo de registro de ponto flutuante 110 utilizado em um processador de ponto flutuante 100 com sub-precisão selecionável, como descrito em conjunto com a figura 1. Na modalidade ilustrada na figura 2, o arquivo de registro de ponto flutuante 110 inclui dezesseis localizações de registro endereçáveis, cada localização de registro sendo referida com referência numérica 200 na figura 2, por motivos de conveniência. Cada localização de registro 200 é configurada para armazenar um número de ponto flutuante binário de 32 bits, em um formato único de 32 bits IEEE-754. Em particular, cada localização

de registro 200 contém um sinal de 1 bit 202, um expoente de 8 bits 204, e uma fração de 24 bits 206. Deve-se, obviamente, compreender, no entanto, que outras modalidades do processador de ponto flutuante 100 podem incluir um arquivo de registro de ponto flutuante 110 que é formatado diferentemente do formato único de 32 bits IEEE (incluindo, mas não limitado, ao formato duplo de 64 bits IEEE), e/ou pode conter um número diferente de localizações de registro.

Com referência novamente à figura 1, o controlador de ponto flutuante 130 pode ser utilizado para selecionar a sub-precisão das operações de ponto flutuante utilizando um sinal de controle 133. Um registro de controle (CRG) 137 pode ser carregado com bits de seleção de sub-precisão, por exemplo, transmitidos no campo de controle de uma ou mais instruções. De uma forma a ser descrita em maiores detalhes posteriormente, os bits de seleção de sub-precisão podem ser utilizados pelo controlador de ponto flutuante 130 para reduzir a precisão dos operandos. Os bits de seleção de sub-precisão podem ser utilizados também para desligar as partes do processador de ponto flutuante 100. Por meio de exemplo, os bits de seleção de sub-precisão podem ser utilizados para remover a potência dos elementos de registro de ponto flutuante para os bits que não são necessários para a sub-precisão selecionada. Os bits de seleção de sub-precisão também podem ser utilizados para remover energia da lógica no operador de ponto flutuante FPO 140 que não é utilizado quando a sub-precisão selecionada é reduzida. Uma série de comutadores pode ser utilizada para remover e aplicar potência aos elementos de registro de ponto flutuante e lógica no operador de ponto flutuante 140. Os comutadores, que podem ser internos ou externos ao registro de ponto

flutuante 110 e ao operador de ponto flutuante 140, podem ser transistores de efeito de campo ou qualquer outro tipo de comutador.

O operador de ponto flutuante 140 pode incluir um ou mais componentes configurados para realizar as operações de ponto flutuante. Esses componentes podem incluir, mas não estão limitados a, unidades computacionais como um adicionador de ponto flutuante (ADD) 142 configurado para executar instruções de adição e subtração de ponto flutuante, e um multiplicador de ponto flutuante (MUL) 144 configurado para executar as instruções de multiplicação de ponto flutuante. Como observado na figura 1, cada uma das unidades de computação ADD 142 e MUL 144 no operador de ponto flutuante 140 é acoplado um ao outro e ao arquivo de registro de ponto flutuante 110 de uma forma a permitir que os operandos sejam transferidos entre as unidades de computação, assim como entre cada unidade de computação e o arquivo de registro de ponto flutuante 110. O operador de ponto flutuante pode ser acoplado ao registro de ponto flutuante através de conexões individuais 134, 135, 136, 137, 138 e 139, como ilustrado ou pode ser acoplado através de um barramento ou qualquer outro acoplamento adequado. Em pelo menos uma modalidade do processador de ponto flutuante 100, a saída de qualquer uma das unidades computacionais (ADD 142 e MUL 144) pode ser a entrada de qualquer outra unidade computacional. O arquivo de registro de ponto flutuante 110 pode ser utilizado para armazenar resultados intermediários, além dos resultados que são enviados do operador de ponto flutuante 140.

O adicionador 142 pode ser um adicionador de ponto flutuante convencional, configurado para realizar as operações aritméticas padrão em um formato de ponto flutuante. O multiplicador 144 pode ser um multiplicador de

ponto flutuante convencional, configurado para realizar a multiplicação de ponto flutuante. O multiplicador 144 pode implementar com, por meio de exemplo, um algoritmo Booth ou Booth modificado, e pode incluir a lógica de geração de produto parcial que gera os produtos parciais, e um número de adicionadores de economia de onda que adicionam os produtos parciais.

Enquanto por motivos de simplicidade apenas um adicionador 142 e um multiplicador 144 são ilustrados na figura 1, o operador de ponto flutuante 140 também pode incluir outras unidades computacionais (não ilustradas), que são conhecidas da técnica, e que são configuradas para executar outros tipos de operações matemáticas de ponto flutuante. Essas unidades computacionais podem incluir, mas não estão limitadas a: um divisor de ponto flutuante configurado para realizar as instruções de divisão de ponto flutuante; um extrator de raiz quadrada de ponto flutuante configurado para realizar as instruções de extração de raiz quadrada de ponto flutuante; um operador exponencial de ponto flutuante configurado para executar as instruções exponenciais de ponto flutuante; um operador logarítmico de ponto flutuante configurado para realizar as instruções para o cálculo das funções logarítmicas; e um operador trigonométrico de ponto flutuante configurado para realizar as instruções para o cálculo de funções trigonométricas.

Diferentes modalidades do processador de ponto flutuante 100 podem incluir apenas uma, ou algumas, ou todas as unidades computacionais listadas acima. Por exemplo, o adicionador 142 e o multiplicador 144 podem, cada um, incluir uma ou mais sub-unidades convencionais bem conhecidas tal como alinhadores que alinham os operandos de entrada, normalizadores que mudam o resultado para um formato padrão, e arredondadores que arredondam os

resultados baseados em um modo de arredondamento especificado. Os elementos de circuito bem conhecidos tal como inversores de bit, multiplexadores, contadores e circuitos lógicos combinatoriais também são incluídos no adicionador 142 e no multiplicador 144.

Como ilustrado na figura 1, o operador de ponto flutuante 140 é acoplado ao arquivo de registro de ponto flutuante 110 de forma que para cada instrução de uma operação de ponto flutuante solicitada, a unidade computacional relevante, isso é, o adicionador 142 ou o multiplicador 144, possa receber do arquivo de registro de ponto flutuante 110 um ou mais operandos armazenados em uma ou mais das localizações de registro REG1,...,REGN.

Depois do recebimento dos operandos do arquivo de registro de ponto flutuante 110, uma ou mais unidades computacionais no operador de ponto flutuante 140 podem executar as instruções da operação de ponto flutuante solicitada nos operandos recebidos, na sub-precisão selecionada pelo controlador de ponto flutuante 130. A saída pode ser enviada de volta para o registro de ponto flutuante 110 para armazenamento, como ilustrado na figura 1.

Em uma modalidade do processador de ponto flutuante 100, um modo selecionável de software pode ser utilizado para reduzir a precisão das operações de ponto flutuante sob um controle de programa ou como explicado acima, as instruções fornecidas para o processador de ponto flutuante 100 podem incluir um campo de controle programável contendo os bits de seleção de sub-precisão. Os bits de seleção de sub-precisão são escritos no registro de controle 137, que, por sua vez, controla o comprimento de bit da mantissa para cada operando durante a operação de ponto flutuante. Alternativamente, os bits de seleção de sub-precisão podem ser escritos no registro de controle 137

diretamente a partir de qualquer interface de usuário adequada, incluindo, por exemplo, mas não limitado a uma tela de monitor/teclado ou mouse 150, ilustrados na figura 1. Em outra modalidade do processador de ponto flutuante 100, os bits de seleção de sub-precisão podem ser escritos no registro de controle 137 diretamente a partir do processador principal, ou seu sistema operacional. O registro de controle 137, que é ilustrado no controlador de ponto flutuante 130, pode residir em outro local como uma entidade independente, integrada a outra entidade, ou distribuída através de múltiplas entidades.

Os bits de seleção de sub-precisão podem ser utilizados para reduzir a precisão da operação de ponto flutuante. Isso pode ser alcançado de várias formas. Em uma modalidade, o controlador de ponto flutuante 130 pode fazer com que a potência seja removida dos elementos de registro de ponto flutuante para a fração de bits excessivos que não são necessários para se encontrar à precisão especificada pelos bits de seleção de sub-precisão. Por meio de exemplo, se cada localização no arquivo de registro de ponto flutuante contiver uma fração de 23 bits, e a sub-precisão necessária para a operação de ponto flutuante for de 10 bits, apenas os 9 bits comumente significativos (MSBs) da fração serão necessários; o bit escondido ou inteiro é o décimo. A potência pode ser removida dos elementos de registro de ponto flutuante para a fração de 14 bits restantes. Se a sub-precisão para uma ou mais instruções for aumentada para 16 bits, então os 15 MSBs da mantissa serão necessários. No último caso, a potência pode ser removida dos elementos de registro de ponto flutuante para os 8 bits menos significativos (LSBs) da fração.

Adicionalmente, a lógica no operador de ponto flutuante 140 correspondente aos bits de mantissa

excedentes não exigem maior potência. Dessa forma, a economia energética pode ser alcançada pela remoção de energia para a lógica no operador de ponto flutuante 140 que permanece não utilizado como resultado da sub-precisão selecionada.

A figura 3a é um diagrama conceitual ilustrando um exemplo de uma operação de adição de ponto flutuante com uma energia sendo seletivamente aplicada à lógica no operador de ponto flutuante. Em particular, a figura 3a ilustra de forma conceitual uma operação de adição de ponto flutuante com dois números de ponto flutuante de entrada 302 e 304, cada um caracterizado pela sub-precisão selecionada, somados juntos. Por motivos de simplicidade, considera-se que os dois números 302 e 304 já tenham sido alinhados, de forma que nenhuma mudança precise ser feita. A operação de adição de ponto flutuante no modo de precisão total é realizada através de uma sucessão de estágios, referidos na figura 3a pelas referências numéricas 310_1, 310_2,..., 301_i,..., 310_n. De acordo com a convenção padrão, o registro de ponto flutuante armazena em ordem os bits que criam cada número, variando de um LSB mais a direita para um MSB mais a esquerda. Cada estágio sucessivo dentre os estágios se movendo da direita para a esquerda através da figura 3 envolve bits que possuem um aumento significativo em comparação com os bits envolvidos nos estágios anteriores.

No exemplo ilustrado na figura 3a, a sub-precisão selecionada é representada por uma linha 305. A energia pode ser removida da lógica utilizada para implementar cada estágio à direita da linha 305. A realização C do último estágio descendente energizado 310_i é forçada para zero. A energia é suprida apenas para a lógica utilizada para implementar cada estágio à esquerda da linha 305. Na figura

3a, os bits energizados fornecidos para os estágios ativos do operador de ponto flutuante são representados por Xs, utilizando a referência numérica 322, enquanto os bits não energizados fornecidos para os estágios com energia removida são representados por círculos, utilizando-se as referências numéricas 324.

A figura 3b é um diagrama conceitual ilustrando um exemplo de uma operação de multiplicação de ponto flutuante com uma energia sendo seletivamente aplicada à lógica no operador de ponto flutuante. A operação de multiplicação de ponto flutuante é realizada no multiplicador de ponto flutuante MUL, ilustrado na figura 1 com referência numérica 144. É no multiplicador que uma quantidade substancial da lógica pode ser desenergizada, fornecendo economias energéticas significativas. A multiplicação binária como ilustrado na figura 3b é basicamente uma série de adições de números de ponto flutuante alterados. Na modalidade ilustrada, a multiplicação binária é realizada entre um multiplicando de k-bit 402 e um multiplicador de k-bit 404, utilizando uma técnica de alterar e adicionar. A técnica de alterar e adicionar pode ser substituída por um algoritmo Booth, ou um multiplicador de algoritmo Booth modificado.

Como no caso de adição de ponto flutuante, a multiplicação de ponto flutuante é realizada em uma série de estágios, ilustrados na figura 3b como 410-1,...410-m. Assumindo-se, por motivos de simplicidade, que o algoritmo Booth seja utilizado, um produto parcial é gerado para cada bit no multiplicador 404, um produto parcial 420-i sendo gerado durante um estágio correspondente 410-i. Se o valor do multiplicador for igual a 0, seu produto parcial correspondente consiste apenas em zeros; se o valor do bit for igual a 1, seu produto parcial correspondente é uma

cópia do multiplicando. Cada produto parcial 420-i é alterado para a esquerda, como uma função do bit multiplicador com o qual está associado, depois que a operação move para o próximo estágio. Cada produto parcial pode, dessa forma, ser observado como um número alterado. O produto parcial associado com o bit 0 no multiplicador é de bits zero mudados para a esquerda, e o produto parcial associado com o bit 1 é um bit mudado para a esquerda. Os produtos parciais ou números de ponto flutuante alterados 420-i são adicionados juntos para gerar o valor de saída 430 da multiplicação.

Na modalidade ilustrada na figura 3b, a seleção de uma precisão reduzida desejada pelo controlador 130 é indicada com uma linha 405. Como no caso da adição de ponto flutuante, descrita em conjunto com a figura 3a, a energia pode ser removida da lógica utilizada para implementar os estágios à direita da linha 405. A energia só é aplicada aos estágios que são na verdade necessários para suportar a sub-precisão selecionada, isso é, os estágios à esquerda da linha 405. Na figura 3b, os bits fornecidos para a lógica energizada são representados por Xs, enquanto os bits fornecidos para os estágios desenergizados são representados por círculos.

Como observado a partir da figura 3b, para o primeiro produto parcial 420-1, a lógica para um número de N bits, ilustrada utilizando-se a referência numérica 402, é desenergizada. Para o segundo produto parcial, a lógica para os N-1 bits é desenergizada, e assim por diante. Para o produto parcial m ou número de ponto flutuante alterado 420-m, a lógica para um número (N-m+1) de bits, ilustrado utilizando a referência numérica 414, é desenergizada. O número de N bits é escolhido de modo que a precisão dos estágios restantes não seja afetada de forma adversa.

O valor de saída, resultando da multiplicação de ponto flutuante descrita acima, possui uma largura (isto é, número de bits) que é igual à soma das larguras dos dois valores de entrada 402 e 404 que estão sendo multiplicados. O valor de saída 430 pode ser truncado para a sub-precisão selecionada, isto é, qualquer um dos bits do valor de saída 430 que seja inferior à precisão selecionada pode ser truncado, para gerar um número de saída truncado caracterizado pela precisão selecionada. Alternativamente, o valor de saída 430 pode ser arredondado para a precisão selecionada. Em qualquer caso os bits de saída menos significativos que a precisão selecionada também podem ser desenergizados.

As várias unidades lógicas, blocos, módulos, circuitos, elementos e/ou componentes ilustrativos descritos com relação às modalidades descritas aqui podem ser implementados ou realizados em um processador de ponto flutuante que é parte de um processador de finalidade geral, um processador de sinal digital (DSP), um circuito integrado específico de aplicativo (ASIC), um campo de conjunto de portal programável (FPGA), ou outro componente lógico programável, portal discreto ou lógica de transistor, componentes discretos de hardware, ou qualquer combinação dos mesmos projetada para realizar as funções descritas aqui. Um processador de finalidade geral pode ser um microprocessador, mas na alternativa, o processador pode ser qualquer processador convencional, controlador, micro controlador ou máquina de estado. O processador também pode ser implementado como uma combinação de componentes de computação, por exemplo, uma combinação de um DSP e um microprocessador, uma pluralidade de microprocessadores, um ou mais microprocessadores em conjunto com um núcleo DSP, ou qualquer outra configuração.

Os métodos ou algoritmos descritos com relação às modalidades descritas aqui podem ser consubstanciados diretamente em hardware, em um módulo de software executado por um processador, ou em uma combinação dos dois. Um módulo de software pode residir na memória RAM, memória flash, memória ROM, memória EPROM, memória EEPROM, registros, disco rígido, um disco removível, CD-ROM, ou qualquer outra forma de meio de armazenamento conhecido da técnica. Um meio de armazenamento pode ser acoplado ao processador de forma que o processador possa ler informação a partir de, e escrever informação no meio de armazenamento. Na alternativa, o meio de armazenamento pode ser integral ao processador.

A descrição anterior das modalidades descritas é fornecida para permitir que qualquer pessoa versada na técnica crie ou faça uso da presente descrição. Várias modificações a essas modalidades serão prontamente aparentes aos versados na técnica, e os princípios genéricos definidos aqui podem ser aplicados a outras modalidades sem se distanciar do espírito ou escopo da descrição. Dessa forma, a presente descrição não deve ser limitada às modalidades ilustradas aqui, mas deve ser acordado o escopo total consistente com as reivindicações, onde a referência a um elemento no singular não significa "um e apenas um" a menos que especificamente mencionados, mas ao invés disso "um ou mais um". Todas as equivalências estruturais e funcionais aos elementos de várias modalidades descritas por toda essa descrição que são conhecidas ou se tornarão conhecidas posteriormente dos versados na técnica são expressamente incorporadas aqui por referência e devem ser englobadas pelas reivindicações. Ademais, nada do que foi descrito aqui deve ser dedicado ao público a menos que tal descrição seja explicitamente

recitada nas reivindicações. Nenhum elemento de reivindicação deve ser considerado sob 35 U.S.C. § 112, sexto parágrafo, a menos que o elemento seja expressamente recitado utilizando a frase "meios para", ou no caso de uma reivindicação de método, o elemento seja recitado utilizando-se a frase "etapa para".

REIVINDICAÇÕES

1. Método para realizar uma operação de ponto flutuante com um processador de ponto flutuante possuindo uma precisão máxima, compreendendo:

selecionar uma sub-precisão inferior à precisão máxima para a operação de ponto flutuante em um ou mais números de ponto flutuante, a seleção da sub-precisão resultando em um ou mais bits excessivos para cada um ou mais dos números de ponto flutuante;

remover energia de um ou mais componentes no processador de ponto flutuante que, do contrário, seria utilizada para armazenar ou processar os um ou mais bits excessivos; e

realizar operação de ponto flutuante com energia removida de um ou mais componentes.

2. Método, de acordo com a reivindicação 1, compreendendo adicionalmente a utilização de um registro de ponto flutuante com uma pluralidade de elementos de armazenamento, os um ou mais bits excessivos sendo armazenados em um ou mais elementos de armazenamento, e onde os um ou mais componentes dos quais a potência é removida inclui os elementos de armazenamento para os um ou mais bits excessivos.

3. Método, de acordo com a reivindicação 2, compreendendo adicionalmente a utilização de um operador de ponto flutuante com lógica para realizar a operação de ponto flutuante, e onde os um ou mais componentes dos quais a potência é removida incluem uma parte da lógica que, do contrário, seria utilizada para processar os um ou mais bits excessivos.

4. Método, de acordo com a reivindicação 1, compreendendo adicionalmente a utilização de um operador de ponto flutuante com lógica para realizar a operação de

ponto flutuante, e onde os um ou mais componentes dos quais a potência é removida incluem uma parte da lógica que, do contrário, seria utilizada para processar os um ou mais bits excessivos.

5. Método, de acordo com a reivindicação 4, no qual a operação de ponto flutuante compreende adição.

6. Método, de acordo com a reivindicação 5, compreendendo adicionalmente uma realização forçada da parte da lógica para zero.

7. Método, de acordo com a reivindicação 4, no qual a operação de ponto flutuante compreende multiplicação.

8. Processador de ponto flutuante possuindo uma precisão máxima, compreendendo:

um controlador de ponto flutuante configurado para selecionar uma sub-precisão inferior à precisão máxima para uma operação de ponto flutuante em um ou mais números de ponto flutuante, a seleção da sub-precisão resultando em um ou mais bits excessivos para cada um dos um ou mais números de ponto flutuante, o controlador de ponto flutuante sendo configurado adicionalmente para remover potência de um ou mais componentes no processador de ponto flutuante que, do contrário, seria utilizada para armazenar ou processar os um ou mais bits excessivos; e

um operador de ponto flutuante configurado para realizar a operação de ponto flutuante.

9. Processador de ponto flutuante, de acordo com a reivindicação 8, compreendendo adicionalmente um registro de ponto flutuante possuindo uma pluralidade de elementos de armazenamento, os um ou mais bits excessivos sendo armazenados em um ou mais dos elementos de armazenamento, e onde os um ou mais componentes dos quais a potência pode

ser removida inclui os elementos de armazenamento para os um ou mais bits excessivos.

10. Processador de ponto flutuante, de acordo com a reivindicação 9, no qual o operador de ponto flutuante compreende lógica para realizar a operação de ponto flutuante, e onde um ou mais dos componentes dos quais a potência pode ser removida incluem uma parte da lógica que, do contrário, seria utilizada para processar os um ou mais bits excessivos.

11. Processador de ponto flutuante, de acordo com a reivindicação 8, no qual o operador de ponto flutuante compreende a lógica para realizar a operação de ponto flutuante, e onde os um ou mais dos componentes dos quais a potência pode ser removida incluem uma parte da lógica que, do contrário, seria utilizada para processar os um ou mais bits excessivos.

12. Processador de ponto flutuante, de acordo com a reivindicação 11, no qual o operador de ponto flutuante inclui um adicionador de ponto flutuante.

13. Processador de ponto flutuante, de acordo com a reivindicação 12, no qual o operador de ponto flutuante é adicionalmente configurado para forçar uma realização da parte da lógica para zero quando a energia é removida.

14. Processador de ponto flutuante, de acordo com a reivindicação 11, no qual o operador de ponto flutuante inclui um multiplicador de ponto flutuante.

15. Processador de ponto flutuante possuindo uma precisão máxima, compreendendo:

um registro de ponto flutuante possuindo uma pluralidade de elementos de armazenamento configurada para armazenar uma pluralidade de números de ponto flutuante;

um operador de ponto flutuante configurado para realizar uma operação de ponto flutuante no um ou mais

números de ponto flutuante armazenados no registro de ponto flutuante; e

um controlador de ponto flutuante configurado para selecionar uma sub-precisão inferior à precisão máxima para uma operação de ponto flutuante nos um ou mais números de ponto flutuante, a seleção da sub-precisão resultando em um ou mais bits excessivos para cada um dos um ou mais números de ponto flutuante, os um ou mais bits excessivos sendo armazenados em um ou mais elementos de armazenamento do registro de ponto flutuante, e onde o controlador de ponto flutuante é adicionalmente configurado para remover energia dos elementos de armazenamento para os um ou mais bits excessivos.

16. Processador de ponto flutuante, de acordo com a reivindicação 15, no qual o operador de ponto flutuante compreende a lógica configurada para realizar a operação de ponto flutuante, e onde o controlador de ponto flutuante é adicionalmente configurado para remover energia de uma parte da lógica que, do contrário, seria utilizada para processar os um ou mais bits excessivos.

17. Processador de ponto flutuante, de acordo com a reivindicação 16, no qual o operador de ponto flutuante inclui um adicionador de ponto flutuante.

18. Processador de ponto flutuante, de acordo com a reivindicação 17, no qual o operador de ponto flutuante é adicionalmente configurado para forçar uma realização da parte da lógica para zero quando energia é removida.

19. Processador de ponto flutuante, de acordo com a reivindicação 16, no qual o operador de ponto flutuante inclui um multiplicador de ponto flutuante.

20. Processador de ponto flutuante possuindo uma precisão máxima, compreendendo:

um registro de ponto flutuante configurado para armazenar uma pluralidade de números de pontos flutuantes;

um operador de ponto flutuante possuindo lógica configurada para realizar uma operação de ponto flutuante nos um ou mais números de ponto flutuante armazenados no registro de ponto flutuante; e

um controlador de ponto flutuante configurado para selecionar uma sub-precisão inferior à precisão máxima para uma operação de ponto flutuante nos um ou mais números de ponto flutuante, a seleção da sub-precisão resultando em um ou mais bits excessivos para cada um dos um ou mais números de pontos flutuantes, e onde o controlador de ponto flutuante é adicionalmente configurado para remover a energia de uma parte da lógica que, do contrário, seria utilizada para processar um ou mais dos bits excessivos.

21. Processador de ponto flutuante, de acordo com a reivindicação 20, no qual o registro flutuante compreende uma pluralidade de elementos de armazenamento configurada para armazenar o número de ponto flutuante, os um ou mais bits excessivos sendo armazenados em um ou mais elementos de armazenamento, e onde o controlador de ponto flutuante é adicionalmente configurado para remover energia dos elementos de armazenamento para os um ou mais bits excessivos.

22. Processador de ponto flutuante, de acordo com a reivindicação 20, no qual o operador de ponto flutuante inclui um adicionador de ponto flutuante.

23. Processador de ponto flutuante, de acordo com a reivindicação 22, no qual o operador de ponto flutuante é adicionalmente configurado para forçar uma realização da parte da lógica para zero quando a energia é removida.

24. Processador de ponto flutuante, de acordo com a reivindicação 20, no qual o operador de ponto flutuante

inclui um multiplicador de ponto flutuante, e onde a energia é removida de partes dos elementos compreendendo produtos parciais dentro do multiplicador de ponto flutuante.

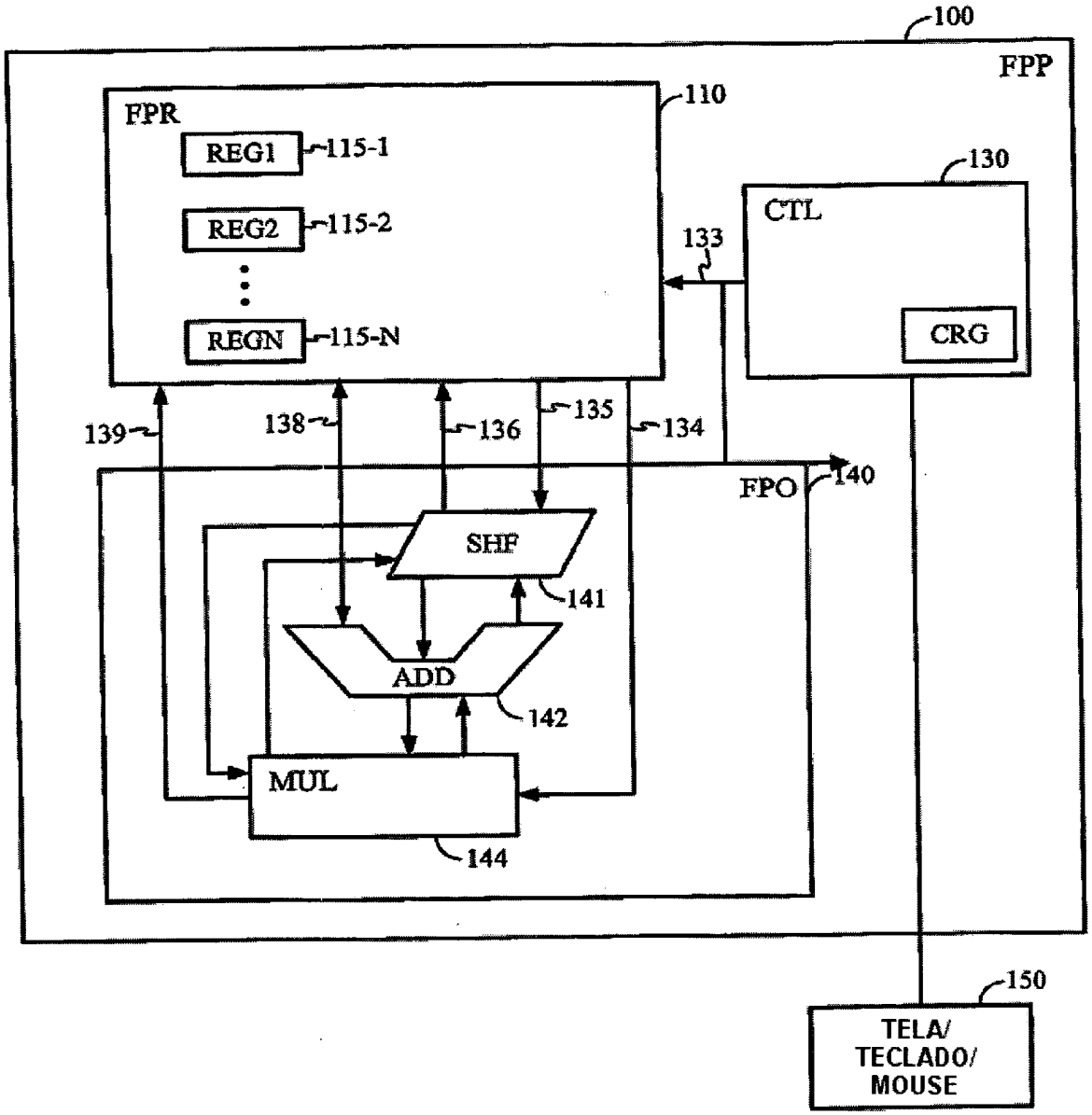

100
FPP
FPR
110
REG1
115-1
REG2
115-2
REGN
115-N
130
CTL
133
CRG
139
138
136
135
134
FPO
140
SHF
141
ADD
142
MUL
144
150
TELA/
TECLADO/
MOUSE

FIGURA 1

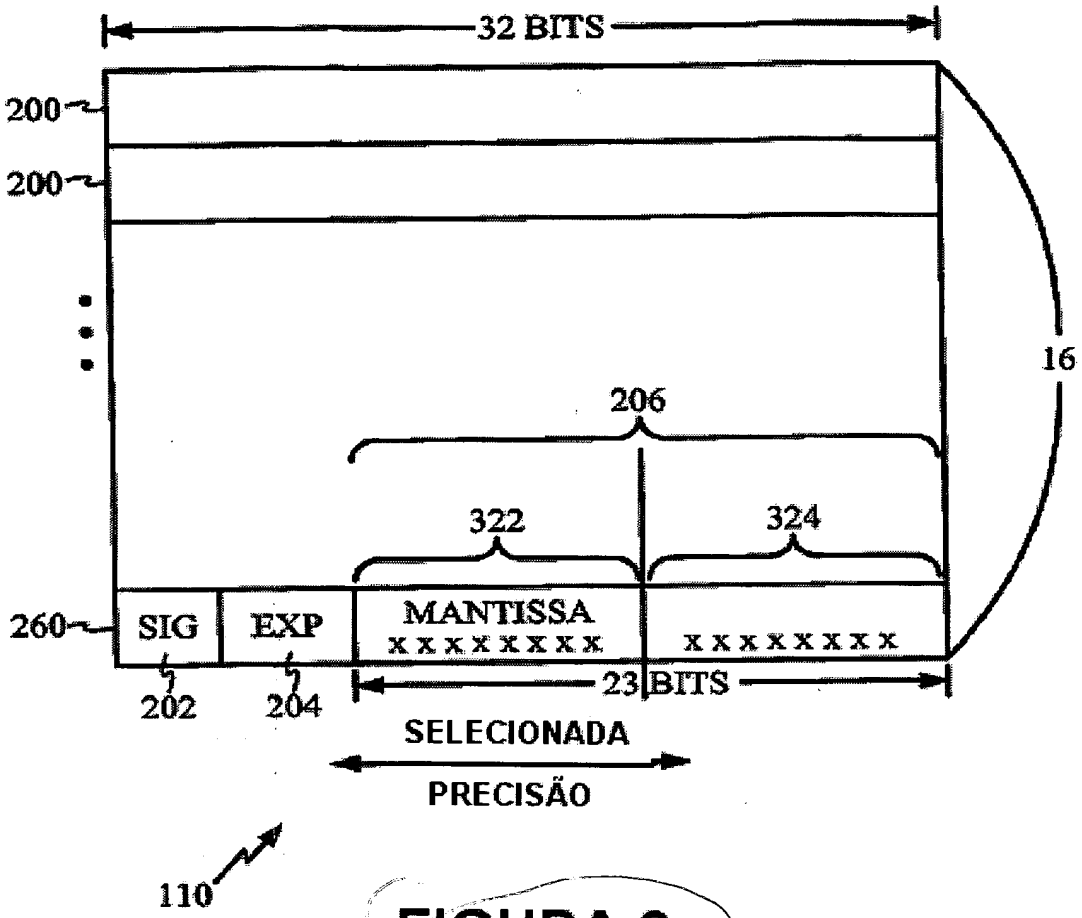

32 BITS
200
200
16
206
322
324
260
SIG
EXP
MANTISSA
xxxxxxxx
xxxxxxxx
202
204
23 BITS
SELECIONADA
PRECISÃO
110

FIGURA 2

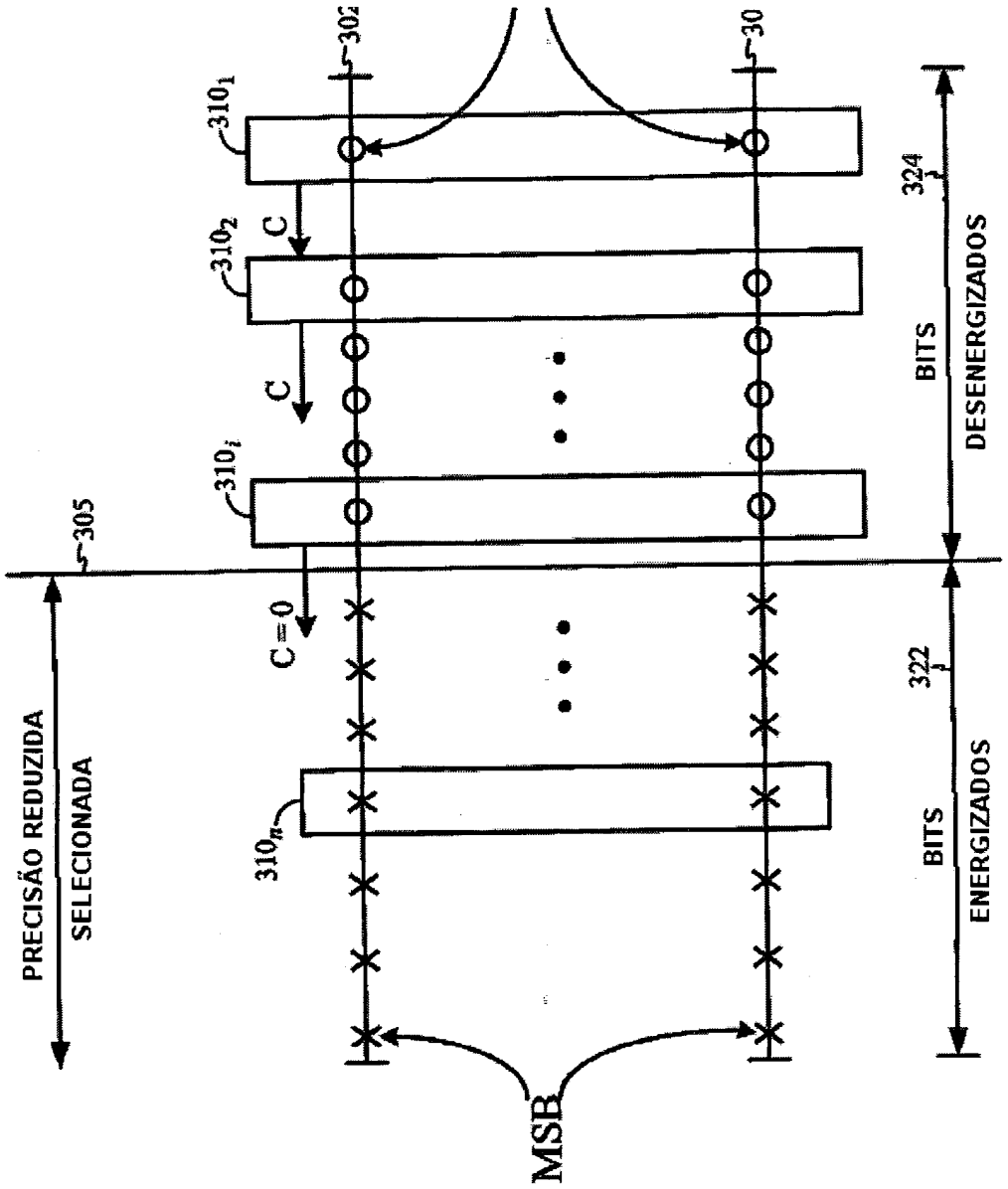

PRECISÃO REDUZIDA
SELECIONADA
305
310_1
310_2
310_i
310_n
C
C
C=0
302
30
324
BITS
DESENERGIZADOS
322
BITS
ENERGIZADOS
MSB

FIGURA 3A

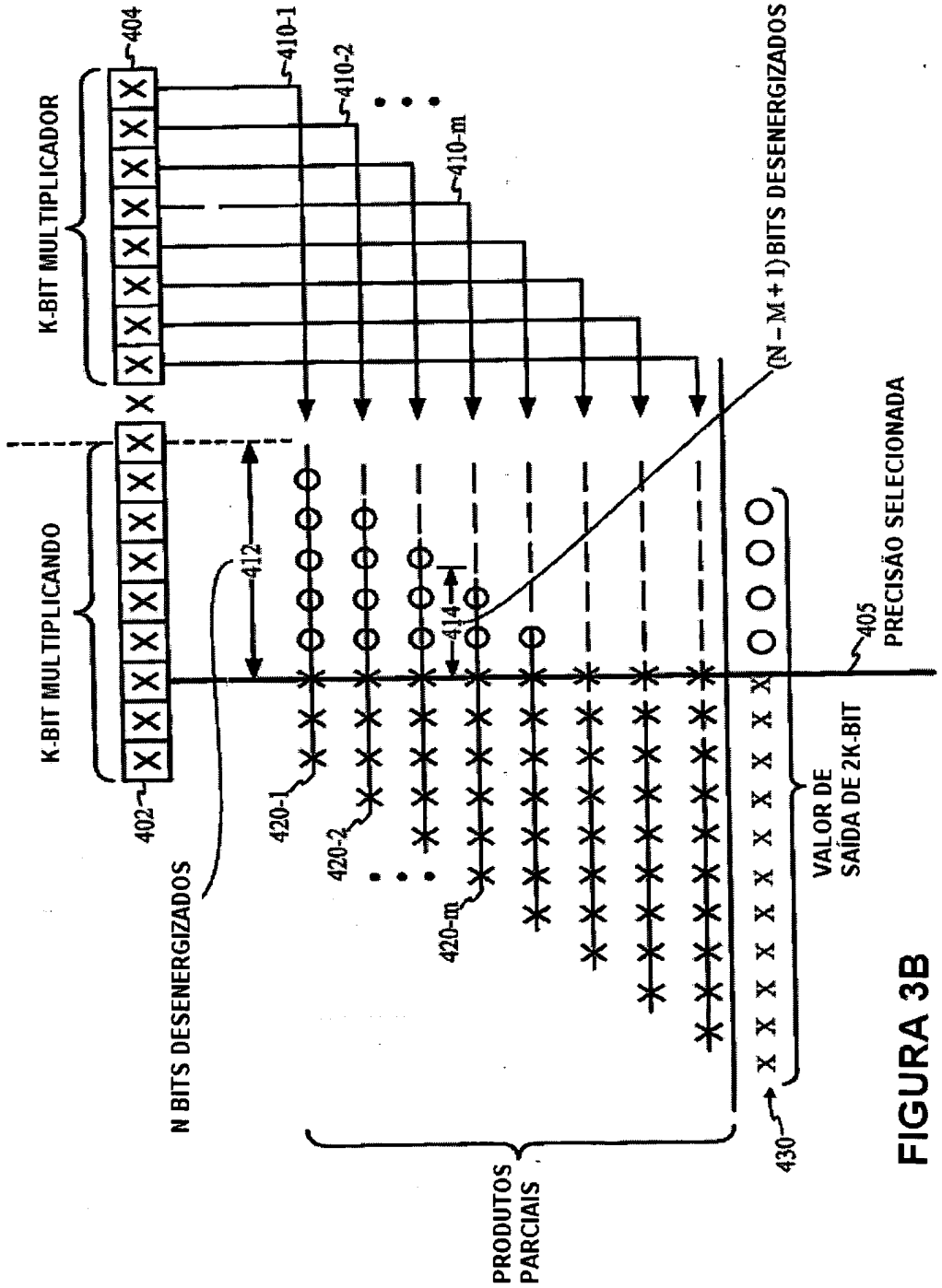

K-BIT MULTIPLICADOR
404
410-1
410-2
410-m
K-BIT MULTIPLICANDO
402
412
414
N BITS DESENERGIZADOS
420-1
420-2
420-m
(N – M + 1) BITS DESENERGIZADOS
405
PRECISÃO SELECIONADA
VALOR DE
SAÍDA DE 2K-BIT
430
PRODUTOS
PARCIAIS

FIGURA 3B

PI0708284-3

RESUMO

"PROCESSADOR DE PONTO FLUTUANTE COM EXIGÊNCIAS DE POTÊNCIA REDUZIDA PARA SUB-PRECISÃO SELECIONÁVEL"

Um método e aparelho para a realização de uma operação de ponto flutuante com um processador de ponto flutuante possuindo uma determinada precisão são descritos. Uma sub-precisão para a operação de ponto flutuante em um ou mais dos números de ponto flutuante é selecionada. A seleção da sub-precisão resulta em um ou mais bits excessivos para cada um dos um ou mais números de ponto flutuante. A potência pode ser removida de um ou mais componentes no processador de ponto flutuante que, do contrário, seria utilizado para armazenar ou processar os um ou mais bits excessivos, e a operação de ponto flutuante é realizada com a potência removida dos um ou mais componentes.