公告本

申請日期	90-11-30
案　　號	90129656
類　　別	G06F 7/544

（以上各欄由本局填註）

A4
C4

528986

發明 ~~新型~~ 專利說明書

一、發明 ~~新型~~ 名稱	中　文	用以計算倒數的方法和裝置
	英　文	METHOD AND APPARATUS FOR CALCULATING A RECIPROCAL
二、發明 ~~創作~~ 人	姓　名	1.亞力克西 克魯格羅夫 ALEXEI KROUGLOV 2.傑 羅 JIE ZHOU 3.丹尼爾 葛門森 DANIEL GUDMUNSON
	國　籍	均加拿大
	住、居所	1.加拿大安大略省伊多比可市#310皇冠區9號 2.加拿大安大略省北約克市低堤區29號 3.加拿大安大略省新市場市傑里路200號
三、申請人	姓　名（名稱）	加拿大商斯康視訊公司 SICON VIDEO CORPORATION
	國　籍	加拿大
	住、居所（事務所）	加拿大安大略省理奇蒙市F區第16大道1550號2樓
	代表人姓　名	艾德 海克 ED HACKER

（由本局填寫）

承辦人代碼：	
大　　類：	
ＩＰＣ分類：	

A6
B6

本案已向：

國（地區）　申請專利，申請日期：　　案號：　　，□有　□無主張優先權

加拿大	2000年12月20日	2,329,104	□有	☑無主張優先權

有關微生物已寄存於：　　寄存日期：　　，寄存號碼：

五、發明說明（ 1 ）

發明範圍

本發明乃關於信號處理。質言之，本發明係關於計算倒數或反數之方法與裝置。

發明背景

計算倒數乃屬除法運算中重要之一環，特別是於浮點小數為然。利用倒數於兩數相除，其答案可由以除數之反數乘被除數而獲得。此種相除方法可用以增進處理如電腦等數位處理裝置中繁複計算以及如數位信號處理器等特殊用途積體電路中之處理速度。

根據美國電機電子工程師協會之二進位浮點演算標準0754P-1985，茲引介於此以供參考，浮點格式中之浮點標準數字乃封入32位元內，以24有效位元長度為單一精準度，或封入64位元內以53有效位元長度為加倍精準度。

有數種插入及疊代方法廣泛為開發者採用以計算倒數，包括直接近似值，線性插入，平方插入，立方插入等。

在直接近似值方法以獲得數字之倒數中，倒數之一切可能性尾數儲存於唯讀記憶器中。採用此法可迅速獲得答案，但其需極大之記憶容量。例如，依據美國電機電子工程師協會標準754單一精準度浮點格式以獲得一倒數，需要$2^{23}\times 23=184$百萬位元之記憶器。

線性插入法係根據微積分中之均值原理，可適用於下列之倒數計算：

$$\frac{1}{x}=\frac{1}{x_0}-\frac{1}{\xi^2}(x-x_0) \qquad (1)$$

五、發明說明（ 2 ）

其中 $\xi \in [x_0, x]$ 及 $x \geqq x_0$，

亦可採用平方插入，立方插入或其它插入法以獲得所要求精準度之倒數。不過，所有此等方法皆需另加乘法運算，並另需記憶器以儲存改正係數。插入法之主要缺點在於所要之精準度愈高，則所需儲存必要資料之記憶器量就愈大。

牛頓拉弗生疊代法廣泛用於數位電腦中以計算倒數。此法提供方程式之解法

$$f(z)=0 \qquad (2)$$

根據採用循環公式

$$z_{i+1}=z_i-\frac{f(z_i)}{f'(z_i)} \qquad (3)$$

疊代i後所獲得之值z_i乃屬二次收斂至z，故疊代i及疊代i+1後之對應誤差ε以下式表達：

$$\varepsilon(z_{i+1}) \leqq \varepsilon^2(z_i) \qquad (4)$$

利用牛頓拉弗生法計算倒數 $x=\frac{1}{a}$ 產生以下表達式：

$$x_{i+1}=x_i*(2-a*x_i) \qquad (5)$$

五、發明說明（ 3 ）

自式(5)中所見，此法之每一疊代步驟涉及依序所做之兩次乘法運算及一次"2之補數"運算。故每一疊代步驟後之倒數精準度皆獲倍增。牛頓拉弗生疊代法本身之缺點在於其需多次疊代步驟以獲得所要求精準度之倒數。

為克服上述缺點，業已發展出採用某種插入法獲得倒數之初期近似值，然後根據此近似值採用疊代法。例如，經建議使用反數表以獲得連續疊代之初期值。

發明概述

本發明提供一種方法和裝置，用以快速精準地除可遞送一數之倒數的值。

依據此方法和裝置，採用線性插入法以獲得一數倒數之近似值。然後此近似值用做牛頓拉弗生疊代之輸入值，以計算高精準度之倒數。

本發明之方法與以前工技方法不同者在於提供一種計算檢查表中最少數項目之公式，以獲所要求精準度之倒數的近似值。本發明之方法並提供計算初期近似值及組成檢查表中項目之改正係數的公式。實施本發明方法之裝置包括儲存此等值之檢查表記憶器，整數乘法器及減法器。

準此，本發明提供一種產生輸出信號之方法，該信號代表具規格化尾數M($1 \leqq M < 2$)輸入值D倒數之近似輸出值，該尾數表示輸入信號，輸入信號包含一組最有效位元之N_0，且近似所要精準度$\varepsilon = 2^{-N}$倒數之輸出信號，其中$N \leqq N_0$，包含步驟：a，就輸入信號之一組最有效位元P，產生檢查表中項目之數n，其中$n = 2^{1'}$，包括次步驟：I，在第一

五、發明說明（ 4 ）

一檢查表中產生一組含有一群有效位元N之輸入項目y_i，其中i=0 ... n-1；及ii，在第二檢查表中產生一組含有一群有效位元(N-P)之輸入項目K_i，其中i=0 ... .n-1；b，查出對應於輸入信號最有效位元P組之檢查表中之諸項目y_i及K_i；c，以輸入信號有效位元P組後繼之含有效位元(N-P)群位元之信號乘K_i；d，自項目y_i之有效位元N組減去最有效位元群(N-P)。

在本發明方法之另一方面，產生檢查表n項目之步驟包含次步驟：iii，計算必要之檢查表項目之最小數1以獲得較所要精準度更高之精準度，其中

$$\frac{2l+1}{2l+2}-\sqrt{\frac{l}{l+1}}<\varepsilon \text{ 及 } \left|\frac{2l-1}{2l}-\sqrt{\frac{l-1}{l}}\right|\geq\varepsilon$$

及iv，就$n=2^P$查出檢查表項目之最小數，其中$2^{P-1}<1$及$2^P\geqq 1$；於第一檢查表中產生一組輸入項目之步驟包含各次步驟：A，計算

$$\hat{y}_i=\frac{\sqrt{x_i(x_i+\frac{1}{n})}+\frac{1}{2n}}{x_i(x_i+\frac{1}{n})}$$

其中i=0 ... n-1，$x_0=1$及$x_{i+1}=x_i+1/n$，及B，就i=0 ... n-1查出含一組N有效位元及$\hat{y}_i$近似尾數之項目y_i；及/或第二檢查表中產生一組輸入項目之步驟包含次步驟：計算

五、發明說明（ 5 ）

$$K_i = \frac{2^{N-P}}{x_i(x_i + \frac{1}{n})} \text{ 其中 } i = 0,\ldots, n-1,\ x_0 = 1, \text{ 及 } x_{i+1} = x_i + \frac{1}{n}$$

以及就i=0 ... n-1查出含一組(N-P)有效位元及K_i之近似整數部分之項目K_i。

本發明並提供一種計算倒數1之裝置，倒數含具有規格化尾數M之輸入值D精準度 $\varepsilon = 2^{-N}$ (其中 $1 \leqq M < 2$) 包含一組有效位元N_0其中$N_0 \geqq N$，此裝置至少含有一處理器及下述各項：第一記憶器，構成檢查表稱為尾數M有效位元P之函數，並具含一組有效位元N之輸出I_0；第二記憶器，構成檢查表稱為尾數M有效位元P之函數，並具含一組有效位元(N-P)之輸出K；(N-P)×(N-P)容量之乘法器，其具有尾數M有效位元P之後繼有效位元(N-P)組及輸出K之兩輸入，以及含有效位元(N-P)×(N-P)組之輸出MU；一個加法器/減法器，其具有輸出I並具兩輸入分別連接以接受輸入I_0和輸出MU之有效位元(N-P)組。

在本發明裝置之另一方面，第一和第二兩記憶器併入一儲存裝置內，此裝置儲存I_0和K並稱為尾數M最有效位元P之函數；本發明裝置另含一種裝置，用以依據I而執行程式計劃之牛頓拉弗生疊代法；第一記憶器含一唯讀記憶器；第二記憶器亦含一唯讀記憶器；儲存裝置至少含一唯讀記憶器；且或本發明裝置可含數位信號處理器中。

圖式之簡單說明

五、發明說明（ 6 ）

圖式中僅以範例方式顯示本發明之較佳具體實例。

圖1為一顯示本發明所採用線性插入法之曲線圖；

圖2為一方塊圖，顯示本發明之線性插入法裝置，用以獲得倒數尾數之N-位元精準度；以及

圖3為一方塊圖，顯示一牛頓拉弗生疊代法裝置，用以獲得倒數尾數之2N-位元精準度。

發明說明

本發明適用於利用浮點格式中二進位數D之尾數M的計算。輸入數之尾數M認定業經先期規格化，即等於或大於一而小於二，亦即$1 \leqq M < 2$。

圖1顯示線性插入法之較佳具體實例，本發明方法中用之約計一數之倒數。關於直接近似值法，線性插入法大為減少檢查表中所需儲存之項目數。在採用直接插入法以達成倒數尾數M中N-位元精準度$\varepsilon = 2^{-N}$時，檢查表或需要$2^N - 1$平均分隔之項目：緣於使用線性插入法以獲得倒數尾數之相同N-位元精準度$\varepsilon = 2^{-N}$，檢查表中具有2^P項目即足，其中$P \leqq N$。每一項目與其前一項目之區別為2^{-P}，故尾數M之最有效位元P構成檢查表位址。

圖1中x_i及x_{i+1}之代表檢查表中有兩連續項目。y_i值代表儲存於檢查表中$1/x_i$值之近似值。$(x - x_i)$乃由第一最有效位元後繼尾數位元代表。

所需倒數之近似值由下式表達：

$$y = y_i - k_i(x - x_i) \qquad (6)$$

五、發明說明（ 7 ）

其中k_i為改正係數，且$x_i \leqq x < x_{i+1}$

為獲得具有N-位元精準度之倒數尾數M，尾數M之最有效位元P之至少(N-P)位元必須與圖1中所示改正係數k_i之至少(N-P)有效位元相匹配。此等係數以整數形式$K_i = 2^{N-P} \times K_i$儲存於檢查表中。因此，獲得具有N-位元精準度之倒數尾數M都涉及在(N-P)×(N-P)容量之整數乘法器中，以係數K_i之(N-P)位元乘尾數M之(N-P)位元，以2^{N-P}除其結果，再自y_i減去其商等步驟。

y_i之尾數M的N位元及整數K_i之(N-P)位元皆儲存於檢查表中。y_i及K_i諸值乃依下式以N及(N-P)精準度計算：

$$y_i \cong \frac{\sqrt{x_i(x_i + \frac{1}{n})} + \frac{1}{2n}}{x_i(x_i + \frac{1}{n})} \qquad (7)$$

$$K_i \cong \frac{2^{N-I'}}{x_i(x_i + \frac{1}{n})} \qquad (8)$$

其中n為檢查表中項目數。

找出1與2兩者間規格化尾數M之倒數方面，線性插入法之最大誤差ε乃由n決定，並以下式表示：

$$\varepsilon = \frac{2n+1}{2n+2} - \sqrt{\frac{n}{n+1}} \qquad (9)$$

五、發明說明（ 8 ）

必需之檢查表中項目之數n可由公式(9)決定，由之獲知所要求之最大誤差ε。

例如，就n=64而言，獲得倒數之線性插入法最大誤差ε≈ $2.98*10^{-5}>2^{-16}$，而n=128時，依據線性插入法，最大誤差ε≈ $7.54*10^{-6}<2^{-16}$。

圖2顯示本發明計算倒數之裝置(10)，其實施上述之方法。輸入數D尾數M之最有效位元P構成含2^P項目之唯讀記憶器(12)之位址線。此唯讀記憶器(12)最好儲存N位元以供尾數最有效位元P之反數y_i(儲存(N-1)位元固已足，因反數之首位元永遠為"0"故也)及修正係數K_i之(N-P)位元，以進行線性插入法。

改正係數K之(N-P)位元供至具有(N-P)×(N-P)容量之整數乘法器(14)之一輸入。供至乘法器(14)另一輸入者為輸入尾數M之最有效位元P後繼之(N-P)位元值。(N-P)×(N-P)位元相乘之乘積(N-P)最有效位元MU供至具N容量之整數減法器(16)之輸入。輸入MU之最有效位元P均為"0"而乘積之最低有效位元(N-P)予以拋棄。供至減法器(16)另一輸入者為自唯讀記憶器(12)之倒數近似值y_i之N位元(如圖2中所示之I)。相減後之答案構成裝置(10)之N位元輸出。

請注意，若輸入尾數M之最有效位元P等於"1"(即最有效位元為"1"而其它(P-1)位元為"0")，則輸出I可表示為MU之(N-P)位元之"1"的補數，從而簡化了計算。

圖3顯示一裝置(20)，用以實施線性插入法結果之牛頓拉

五、發明說明（ 9 ）

弗生疊代法，以提升N-位元至2N-位元結果之精準度。自插入法裝置(10)之輸出I之N位元供至N×2N容量之整數乘法器(22)。供至乘法器(22)之另一輸入為多工器(24)之輸出2N位元。多工器(24)交互選定輸入尾數M之2N最有效位元及2之補充裝置(26)輸出之2N位元。

乘法器(22)產生一3N位元長之答案。此乘積之最低有效組N捨棄。乘積之最有效位元MU1之2N供至補充裝置(26)。此2之補充裝置(26)之輸出2N位元長供至乘法器(24)。在第二次通過乘法器(22)，乘積之最有效位元2N形成裝置(20)之2N位元輸出(如圖3中所示之I_1)。

固然已以範例方式說明本發明之較佳具體實例，惟對嫻於技藝者而言，可做變化與修改而不離本發明所附申請專利範圍者。

A5
B5

四、中文發明摘要（發明之名稱：用以計算倒數的方法和裝置）

一種用以計算浮點輸入數"D"中規格化尾數"M"之倒數的方法和裝置。提供根據所要求精準度以決定檢查表中最小量之公式以及計算檢查表項目之諸公式。檢查表儲存初期近似值及改正係數，其由尾數最有效位元之對應數定位，並用於以減法運算及乘法運算之線性插入法而獲得倒數之初期近似值。線性插入法之答案可供至牛頓拉弗生(Newton-Raphson)疊代法裝置，其每一疊代需二次插入運算及一次二之補數運算，從而倒數之精準度可獲倍增。

英文發明摘要（發明之名稱：METHOD AND APPARATUS FOR CALCULATING A RECIPROCAL）

A method and apparatus for the calculation of the reciprocal of a normalized mantissa M for a floating-point input number D. A formula for determining the minimum size for the look-up table in accordance with the required precision is provided, as well as formulas for calculating look-up table entries. The look-up table stores the initiation approximations and the correction coefficients, which are addressed by the corresponding number of the mantissa's most significant bits and used to obtain the initial approximation of the reciprocal by means of linear interpolation requiring one subtraction operation and one multiplication operation. The result of the linear interpolation may be fed to a Newton-Raphson iteration device requiring, for each iteration, two multiplication operations and one two's complement operation, thereby doubling the precision of the reciprocal.

（請先閱讀背面之注意事項再填寫本頁各欄）

裝 訂 線

六、申請專利範圍

1. 一種產生輸出信號之方法，輸出信號代表近似輸入值D倒數之輸出值，輸入值含有由輸入信號所表示之規格化尾數M(其中$1 \leqq M < 2$)，輸入信號包含一組最有效位元N_0且輸出信號以所要精準度$\varepsilon = 2^{-N}$(其中$N \leqq N_0$)而近似於該倒數，此方法包含之步驟為：

   a，就輸入信號之一組最有效位元P，於多個檢查表中產生n數項目，其中$n=2^P$，本步驟含各分步驟：

      i，在第一檢查表中產生含有效位元組N之一組輸入項目y_i，其中$i=0 \ldots n-1$；及

      ii，在第二檢查表中產生含有效位元組(N-P)之一組輸入項目K_i，其中$i=0 \ldots n-1$；

   b，查出檢查表中對應於輸入信號最有效位元組P之項目y_i及K_i；

   c，以輸入信號最有效位元組P後繼之含有效位元組(N-P)之信號與K_i相乘；及

   d，自項目y_i有效位元組N減去最有效位元組(N-P)。

2. 如申請專利範圍第1項之方法，其中於檢查表中產生n項目之步驟包含分步驟：

   iii，為獲得所要精準度更高精準，計算其必要之檢查表項目之最小數1，其中

$$\frac{2l+1}{2l+2} - \sqrt{\frac{l}{l+1}} < \varepsilon \text{ 及 } \frac{2l-1}{2l} - \sqrt{\frac{l-1}{l}} \geq \varepsilon; \text{ 以及}$$

   iv，查出$n=2^P$所需之檢查表項目最小數n其中$2^{P-1}<1$及

六、申請專利範圍

$2^P \geqq 1$。

3. 如申請專利範圍第1項之方法，其中在第一檢查表中產生一組輸入項目之步驟含有分步驟：

A，計算

$$\hat{y}_i = \frac{\sqrt{x_i(x_i + \frac{1}{n}) + \frac{1}{2n}}}{x_i(x_i + \frac{1}{n})}$$

其中i=0，...，n-1，$x_0=1$及$x_{i+1}=x_i+1/n$；及

B，查出含有效位元組N之項目$\hat{y}_i$及就i=0 ... n-1求y_i尾數之近似值。

4. 如申請專利範圍第1項之方法，其中在第二檢查表中產生一組輸入項目之步驟含有分步驟：

A，計算

$$\hat{K}_i = \frac{2^{N-P}}{x_i(x_i + \frac{1}{n})}$$

其中i=0，...，n-1，$x_0=1$及$x_{i+1}=x_i+1/n$；及

B，查出含有效位元組(N-P)之項目K_i及就i=0 ... n-1求$\hat{K}_i$整數部分之近似值。

5. 一種計算反數I之裝置，其具規格化尾數M($1 \leqq M < 2$)之輸入值D之精準度$\varepsilon = 2^{-N}$，尾數含一組最有效位元N_0，其中$N_0 \geqq N$，此裝置包含：

至少一個處理器；

第一記憶器構成檢查表，稱為尾數M最有效位元P之函

六、申請專利範圍

數，並具有含一組有效位元N之輸出I_0；

第二記憶器構成檢查表，稱為尾數M最有效位元P之函數，並具有含一組有效位元(N-P)之輸出K；

(N-P)×(N-P)容量之乘法器，其具有兩個輸入，其一輸入為尾數M最有效位元組P後繼之有效位元組(N-P)及另一輸入為輸出K，並具有含(N-P)×(N-P)有效位元組之輸出MU；以及

加法器/減法器，其具以輸出I並具有兩個輸入，分別連接以接受輸出I_0及輸出MU之最有效位元組(N-P)。

6. 如申請專利範圍第5項之裝置，其中第一及第二兩記憶器併入一儲存裝置，其儲存I_0及K並稱為尾數M最有效位元P之函數。

7. 如申請專利範圍第5項之裝置，其另含根據I以執行牛頓拉弗生疊代法程式規劃之裝置。

8. 如申請專利範圍第5項之裝置，其中第一記憶器含有一唯讀記憶器(ROM)。

9. 如申請專利範圍第5項之裝置，其中第二記憶器含有一唯讀記憶器。

10. 如申請專利範圍第6項之裝置，其中儲存裝置至少含一個唯讀記憶器。

11. 一種含有申請專利範圍第5項裝置之數位化信號處理裝置。

圖 1

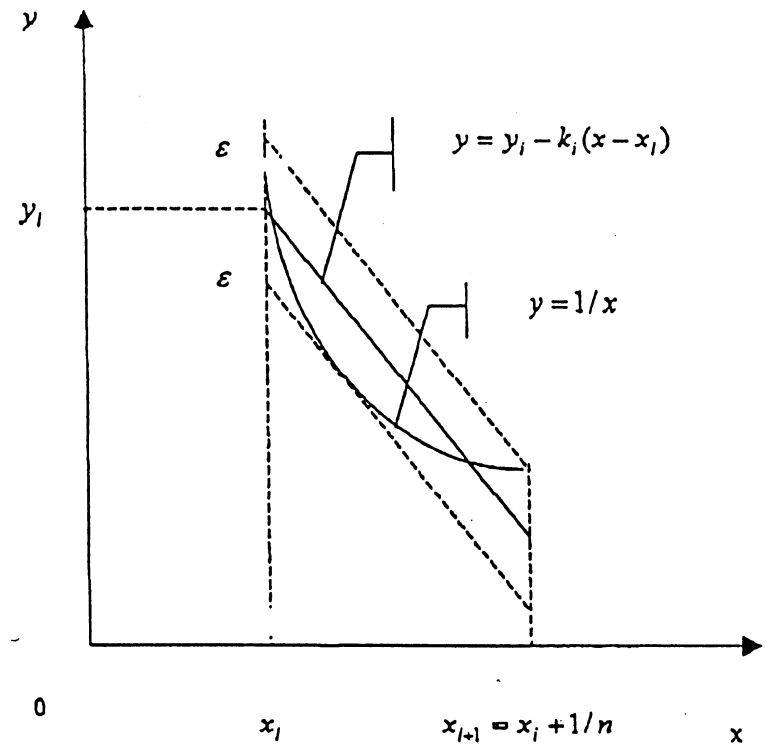

圖 2

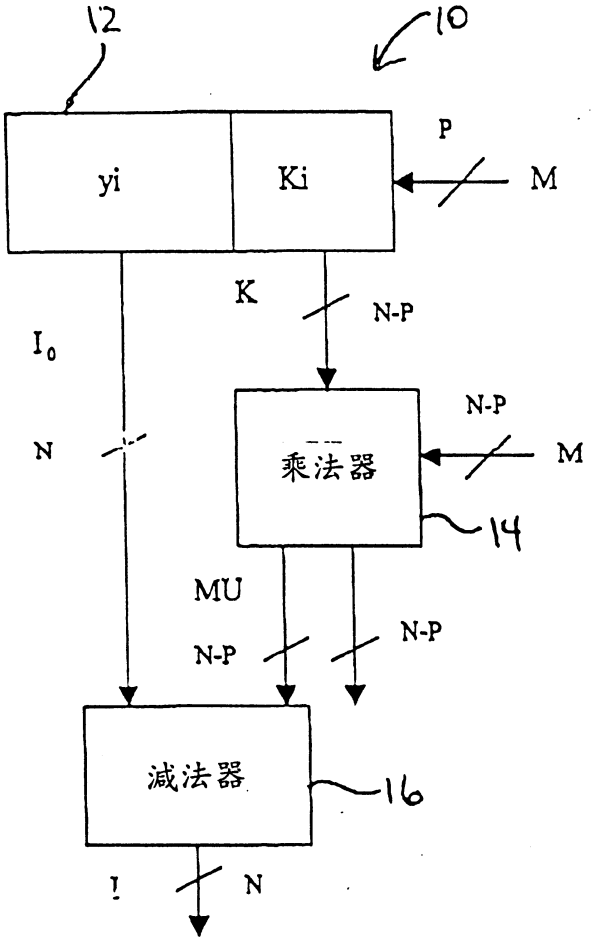

圖 3