拓展算法

对于$\frac{1}{x}$如果能够快速地取得一个近似值a,$ax \approx 1$,可以令$y = ax - 1$,只要y很小,那么做如下变化就可以使用$\frac{1}{1+y}$进行展开了

$$\frac{1}{x} = \frac{a}{ax} = \frac{a}{1+y}$$

查表法其实就是用取整、查表的方法来估计a,但是在MIC上查表的效率不尽人意.

获取近似值a,有以下的一些方法

一次多项式

记$a = f(1+y)$,多项式逼近的思路是在$[0, 1]$上获得一个多项式估计使得$max|(1+y)*f(y) - 1|$最小.然后即可对任意的x取出尾数部分作为$1+y$计算,最后把指数部分添加上去即可.

$$f(x) = -\frac{8}{17}x + \frac{16}{17} \qquad error = \frac{1}{17}$$

二次多项式

$$f(x) = \frac{245}{796}x^2 - \frac{935}{1194}x + \frac{2345}{2388} \qquad error = \frac{43}{2388} = \frac{1}{55.534\cdots}$$
$$f(x) = \frac{32}{99}x^2 - \frac{80}{99}x + \frac{98}{99} \qquad error = \frac{1}{99}$$

三次多项式

$$f(x) = -\frac{642}{3035}x^3 + \frac{1954}{3035}x^2 - \frac{8468}{9105}x + \frac{63508}{63735} \qquad error = 0.00141655$$
$$f(x) = -0.2115x^3 + 0.6438x^2 - 0.93x + 0.9964 \qquad error = 0.00137882$$

类似卡马克快速平方根倒数的方法

IEEE754的双精度浮点数可以表示为 $2^e(1+m)$.对$\frac{1}{x}=y$等号左右取底为2的对数

$$
\begin{aligned}
\frac{1}{x} &= y \\
-\log x &= logy \\
-\log(2^{e_x}(1+m_x)) &= \log(2^{e_y}(1+m_y)) \\
-e_x - \log(1+m_x) &= e_y + \log(1+m_y)
\end{aligned}
$$

由于$m \in [0,1)$, 所以可以做一个近似$\log(1+m) = m + \Delta$.

$$
\begin{aligned}
-e_x - \log(1+m_x) &= e_y + \log(1+m_y) \\
-(e_x + m_x + \Delta_x) &= e_y + m_y + \Delta_y
\end{aligned}
$$

如果把浮点数看做是64位整数$2^{52}E + M$, 有如下的对应关系

$$
\begin{aligned}
e &= E - 1023 \\
m &= \frac{M}{2^{52}}
\end{aligned}
$$

于是又得到

$$
\begin{aligned}
-(e_x + m_x + \Delta_x\ ) & = e_y + m_y + \Delta_y \\
-(E_x - 1023 + \frac{M_x}{2^{52}} + \Delta_x\ ) & = (E_y - 1023 + \frac{M_y}{2^{52}}) + \Delta_y \\
-(2^{52}E_x - 2^{52}\cdot 1023 + M_x + 2^{52}\Delta_x) & = (2^{52}E_y - 2^{52}\cdot 1023 + M_y) + 2^{52}\Delta_y \\
-(2^{52}E_x + M_x) + 2^{52}(1023 - \Delta_x) & = (2^{52}E_y + M_y) + 2^{52}(\Delta_y - 1023)
\end{aligned}
$$

$2^{52}E_x + M_x$和$2^{52}E_y + M_y$正好是x和y对应的64位整数, 整理一下

$$
\begin{aligned}
(2^{52}E_y + M_y) &= -(2^{52}E_x + M_x) + 2^{52}(1023 - \Delta_x) - 2^{52}(\Delta_y - 1023) \\
&= 2^{52}(2046 - (\Delta_x + \Delta_y)) - (2^{52}E_x + M_x) \\
y_{int64} &= 2^{52}(2046 - (\Delta_x + \Delta_y)) - x_{int64}
\end{aligned}
$$

把$\Delta_x + \Delta_y$使用一个常数进行近似.由于$\Delta = log(1+m) - m \in [0, 0.0860...], m \in [0,1]$, 于是可以用$\Delta_{max}$来近似表示$\Delta_x + \Delta_y$, 这样, 误差为

$$
\epsilon = \Delta_x + \Delta_y - \Delta_{max} \in [-\Delta_{max}, \Delta_{max}]
$$

于是算得所谓MAGIC NUMBER

$$
2^{52}(2046 - \Delta_{max}) = 0x7fde9f73aabb2400
$$

所以得到了快速的倒数近似

```
union {
    long long i;
    double y;
} p;
p.y = x;
p.i = 0x7fde5f73aabb2400 - p.i;
rec = p.y;
```

使用暴力验证的方法可以算得这样做出来的误差$ax-1$大约是0.05084,精度虽然不很高,但是仅仅需要一条指令即可完成估计,且不需要单独处理尾数(拆分尾数和合并指数至少需要4条指令),可以让$\dfrac{1}{1+y}$多迭代1次.