Représentation des nombres en virgule fixe
Exercice supplémentaire

Sébastien Piérard

Contexte

Balthazar Picsou souhaite disposer d'une application pour gérer ses avoirs. Toutes les valeurs sont exprimées en sous, avec une précision d'un centième. L'application doit permettre d'effectuer des additions et des soustractions. On sait que l'ordinateur sur lequel l'application sera exécutée manipule des signaux binaires. On vous demande de faire une analyse détaillée de la faisabilité d'une application basée sur une représentation en virgule fixe des valeurs.

Question 1

Est-ce que les valeurs de Balthazar Picsou sont exactement représentables en virgule fixe sur cet ordinateur ?

Solution

Étant donné les caractéristiques techniques de cet ordinateur, on sait que l'on travaillera en binaire (notation positionnelle en base 2).

Commençons par s'intéresser au nombre de bits nécessaires à gauche du séparateur. Les avoirs d'Oncle Picsou étant finis, on sait que ce nombre est fini (borné), il n'y a donc aucune difficulté de ce coté là.

Intéressons-nous à présent au nombre de bits nécessaires à droite du séparateur, pour pouvoir représenter exactement la richesse d'Oncle Picsou. Pour que k bits suffisent, il faut qu'en décalant le séparateur de k positions vers la droite, soit en multipliant par 2^k, on obtienne toujours un entier. Donc, existe-t-il un entier k tel que $2^k \,/_{100}$ est un entier ? Voici deux façons de répondre à cette question, l'une plus courte que l'autre, mais toutes les deux correctes.

— $^{2^k}/_{100}$ ne peut pas être entier, comme on le remarque en factorisant le numérateur et le dénominateur :

$$\frac{2 \times 2 \times \ldots \times 2}{2 \times 2 \times 5 \times 5}$$

(100 a 5 dans ses facteurs premiers, que 2^k n'a pas).

— On se demande s'il existe un entier k tel que $2^k \bmod 100 = 0$. Pour $k = 0$, on a $2^k \bmod 100 = 1$. Puis, on peut utiliser la propriété

$$\begin{aligned} (ab) \bmod c &= ((a \bmod c)\,(b \bmod c)) \bmod c \\ \Rightarrow 2^k \bmod 100 &= \left(2\,2^{k-1}\right) \bmod 100 = \left(2\left(2^{k-1} \bmod 100\right)\right) \bmod 100 \end{aligned}$$

pour obtenir la série $2^k \bmod 100$ pour $k = 0, 1, 2, 3, \ldots$. On obtient :

$$1, 2, 4, 8, 16, 32, 64, 28, 56, 12, 24, 48, 96, 92, 84, 68, 36, 72, 44, 88, 76, 52, 4, \ldots$$

Étant donné le procédé de calcul, on sait que cette suite est périodique (chaque élément dépend uniquement du précédent, et il ne peut y avoir que 100 valeurs possibles pour ses éléments). Nous pouvons nous arrêter de calculer les éléments de la suite, car nous avons trouvé sa période. Il n'existe donc aucun k fini tel que $2^k \bmod 100 = 0$.

On en conclut qu'il faudrait une infinité de bits à droite du séparateur pour pouvoir représenter exactement des centièmes de sous.

Question 2

Est-ce que les valeurs de Balthazar Picsou sont approximativement représentables en virgule fixe sur cet ordinateur, avec une précision de $1/400$?

Solution

Soit k le nombre de bits à droite du séparateur. Il faut qu'il y ait au moins un rationnel $\frac{j}{2^k}$ dans tout intervalle $\left[\frac{i}{100} - \frac{1}{400}, \frac{i}{100} + \frac{1}{400}\right]$ (avec i et j entiers). Ce sera le cas si l'écart entre les nombres représentables, $1/2^k$, est au maximum égal à la largeur de cet intervalle, $\left(\frac{i}{100} + \frac{1}{400}\right) - \left(\frac{i}{100} - \frac{1}{400}\right) = \frac{2}{400}$. On a

$$\arg\min_k \left(\frac{1}{2^k} \leq \frac{2}{400}\right) = \arg\min_k \left(200 \leq 2^k\right) = \lceil \log_2 200 \rceil = 8$$

ce qui signifie que les valeurs de Balthazar Picsou sont représentables en virgule fixe, sur cet ordinateur, avec une précision de $1/400$ à condition d'utiliser au moins 8 bits à droite du séparateur.

Question 3

Sur base des réponses données aux questions précédentes, proposez des solutions pour réaliser une application permettant d'effectuer des additions et des soustractions exactes de centièmes de sous.

Solution

Beaucoup de solutions différentes peuvent être proposées. Voyons-en quelques-unes. On distingue le problème de la représentation des données du problème d'implémentation des opérations nécessaires.

1. Manipuler les valeurs à un facteur d'échelle près, en les multipliant toutes par 100. En effet, $100\,v_1 \pm 100\,v_2 = 100\,(v_1 \pm v_2)$. Les calculs se font donc sur des entiers.
2. Travailler avec des rationnels. On a alors deux variables entières, l'une pour le numérateur et l'autre pour le dénominateur. Les calculs se font donc sur des entiers.
3. Utiliser une représentation en base 10, chaque chiffre étant bien évidemment stocké de façon individuelle en binaire, et implémenter des algorithmes permettant d'effectuer des additions et soustractions avec cette représentation.
   — Une possibilité est alors de travailler en BCD (n'hésitez pas à être curieux et à vous renseigner là-dessus … on n'en dira pas plus ici !).
   — Une autre solution consiste à stocker les valeurs comme des chaînes de caractères dans lesquelles elles sont écrites en base 10.
4. Utiliser une représentation approximative, entachée d'une erreur maximale de $1/400$. A partir de cette représentation, il est possible de retrouver la vraie valeur. Suite à une addition ou une soustraction, l'erreur sur le résultat sera au maximum de $1/200$, ce qui permet de retrouver de façon univoque la vraie valeur. Pour éviter que les erreurs ne se cumulent et que l'incertitude devienne telle que la vraie valeur ne puisse plus être retrouvée, il est nécessaire de corriger le résultat de l'opération pour ramener à chaque fois l'erreur maximale à $1/400$.

Parmi ces solutions, seule la dernière correspond à une approche orientée virgule fixe. Notez que si les opérations à effectuer étaient autres que des additions et soustractions, d'autres solutions auraient peut-être dû être envisagées.