Première NSI - types et valeurs de base : flottants

Travaux dirigés

qkzk

2020/08/01

1. Écriture en base 2 des nombres à virgule

1. Parmi les nombres à virgule binaires suivants, lesqules sont strictement supérieurs à $\frac{1}{2}$

   a. $0,011111$
   b. $0,100001$
   c. $1,000001$
   d. $0,11$

2. L'écriture binaire de 6.625_{10} est : _

   a. $110,101$
   b. $0,101$
   c. $110,101101...$

3. Parmi les nomres suivants, écrits en base 10, quels sont ceux qui ont une écriture infinie en base 2 ?

   a. 1,25
   b. 0,75
   c. 1,7

2. Représenter des nombres en virgule fixe

Le microcontrôleur de l'antimissile *Patriot* stocke la valeur $\frac{1}{10}$ en ne conservant que 23 bits pour la partie décimale (codage en virgule fixe).

Il calcule le temps écoulé depuis démarrage en multiples de $1/10^{\text{eme}}$ de seconde.

1. Écrire 1/10 en binaire, en conservant au moins 30 chiffres binaires après la virgule.
2. Sachant que les registres du *Patriot* ne conservent que 23 bits après la virgule, quelle est, en base 10, la valeur qui est codée effectivement à la place de 1/10 ?
3. Quelle est l'erreur approximative commise sur la représentation de 1/10 ?
4. Combien de signaux d'horloge le *Patriot* reçoit-il en 100 h de fonctionnement ?
5. En tenant compte de l'erreur calculée à la question c., quel est le décalage de l'horloge du *Patriot* par rapport à l'heure réelle au bout de 100h.
6. Sachant qu'un missile se déplace à une vitesse d'environ 1 676 m/s, à quelle erreur de position en mètres correspond le décalage d'horloge d'un *Patriot* ayant fonctionné 100 h sans interruption ?
7. Conclure, sachant que, pour atteindre sa cible un *Patriot* doit l'approcher à moins de 500 m.

3. Nombres à virgule flottante

1. Soit le nombre dyadique $x = +1,10101 \times 2^{11}$. Convertir ce nombre en décimal :

   a. à l'aide de la calculatrice,
   b. à la main. Commencer par écrire $1,10101_2$ comme un quotient d'entiers.

2. On considère le décimal $x = 0.62890625$. Nous allons le convertir en dyadique.

a. x s'écrit exactement comme la somme de puissances négatives de 2.

   Décomposer x sous la forme $x = \frac{1}{2^{p_1}} + \frac{1}{2^{p_2}} + \cdots + \frac{1}{2^{p_k}}$

b. En déduire la *mantisse* de la réprésentation dyadique de x.

c. Donner la représentation dyadique complète.

3. On considère $x = 23.2578125$.

   a. Décomposer x en partie entière et partie décimale. Écrire la partie entière comme somme de puissance de 2 et la partie décimale comme somme de puissance de $1/2$. Ce calcul est exact.
   b. En déduire la mantisse de la réprésentation dyadique de ce nombre.
   c. Déterminer la représentation dyadique complète.

4. On considère $x = 14.04688$. Ce nombre n'est pas représentable exactement en dyadique.

   a. Écrire 14 comme la somme de puissances de 2.
   b. Encadrer la partie décimale de x entre deux sommes de puissances de $\frac{1}{2}$. On ira jusqu'à la puissance 6.
   c. Choisir, parmi les deux bornes, la meilleure approximation.
   d. Donner la mantisse de x dans sa représentation dyadique.
   e. En déduire la représentation dyadique de x.

4. Comprendre la norme IEEE-754

Le codage des nombres en virgule flottantes est la manière la plus répandue de coder les nombres à virgule. Elle est employée dans les ordinateurs, les smartphones, les cartes graphiques.

Le sujet

La norme IEEE-754 (virgule flottante) présente plusieurs variantes, dont une est nommée double précision. En double précision, les nombres sont codés sur 64 bits.

Remarque : *Le codage en double précision est, par exemple, celui utilisé par Python.*

Le premier bit est un bit de **signe** (1 pour un nombre négatif, 0 pour un nombre positif).

Les 11 bits suivants codent **l'exposant décalé** : il vaut $n + 1023$.

À noter : *Le décalage sert à ne manipuler que des nombres positifs (même si n est négatif).*

Les 52 bits suivants sont les bits de **mantisse** : c'est la valeur de M.

Dans toute la suite, on ne s'intéresse qu'aux nombres positifs.

1. Écrire le nombre 49.78125 en binaire.
2. Écrire ce résultat en notation scientifique binaire, sous la forme :

$$1, M \times 2^n$$

   Déterminer les valeurs de M (une suite de chiffres binaires) et la valeur de n (un nombre entier).

3. Sachant q'un nombre flottant est codé en mettant bout à bout le signe, les 11 bits d'exposant décalé et les 52 bits de mantisse, donner dans l'ordre les 64 bits de l'écriture en virgule flottante double précision de 40.78125.
4. Les nombres normalisés sont ceux pour lesquels l'exposant décalé ne contient ni que des 0 ni que des 1 (en binaire). Quels sont le plus petit nombre et le plus grand nombre normalisé ?
5. Quelle est la différence entre le plus grand nombre normalisé et celui qui lui est immédiatement inférieur ?
6. Les nombres dénormalisés sont pour lesquels l'exposant est nul et la mantisse est non nulle. Ces nombres ont un codage à part, et leur valeur est donnée par $0, M \times 2^{-1022}$. Quel est le plus petit nombre dénormalisé ? Quel est le plus grand nombre dénormalisé ?
7. Quelles sont les valeurs approchées, en utilisant les puissances de 10, du plus petit et du plus grand nombre dénormalisé ainsi que celles du plus petit et du plus grand nombre normalisé ?

8. Le zéro est obtenu avec un exposant décalé nul et une mantisse nulle. représentez le 0 et tous les nombres déjà calculés sur l'axe réel.

La feuille de route

1. **Convertir un nombre à virgule en binaire**
   Convertissez d'abord la partie entière, puis par multiplications successives pour la partie décimale.

2. **Décaler la virgule en binaire**
   Vous avez l'habitude de manipuler la notation scientifique en base 10 :

   $$156,23 = 1,5623 \times 10^2$$

   On fait la même chose en binaire en tenant compte que décaler la virgule correspond à un multiplication par 2 plutôt que 10.

3. **Coder un nombre**
   Calculez la valeur de l'exposant décalé et convertissez la en binaire. Puis écrivez simplement le bit de signe, les 11 bits de l'exposant décalé et les 52 bits de mantisse (en complétant par des 0 sur la gauche de l'exposant et de la mantisse si nécessaire).

4. **Trouver la plage de nombres représentables**
   Déduisez la plage de variation de l'exposant réel. Le plus petit nombre normalisé est obtenu avec une mantisse de comportant que des 0 et le plus petit exposant réel. Le plus grand nombre normalisé est obtenu avec une mantisse ne contenant que des 1 et le plus grand exposant réel.

5. **Trouver deux valeurs consécutives**
   Le nombre immédiatement inférieur au plus grand nombre normalisé est obtenu avec le même exposant et une mantisse ne comportant que des 1, sauf en toute dernière position.

6. **Trouver la plage de nombres représentables**
   Le plus petit nombre dénormalisé est obtenu avec la plus petite mantisse non nulle. Le plus grand nombre dénormalisé est obtenu avec une mantisse ne comportant que des 1.

7. **Écrire un nombre en notation scientifique**
   Écrivez simplement les nombres sous la forme $a \times 10^b$ avec $1 \leq a < 10$ et b entier.

8. **Représenter des nombres sur l'axe réel**
   Vérifiez que le plus petit nombre normalisé se trouve juste après le plus grand nombre dénormalisé.

Remarque: *Les nombres dénormalisés constituent un cas particulier de la norme, ils permettent de représenter des nombres plus proches de 0 que ceux qu'on pourrait avoir en utilisant uniquement le codage des nombres normalisés.*