



VTF

Design de systèmes de traitement vidéo temps réel par FPGA

Eléments de traitement du signal en virgule fixe

appliqué au traitement du signal vidéo

Michel Starkier

27/01/2011 V1.4

Format virgule fixe (non-signé)



 La représentation d'un nombre binaire non-signé avec virgule fixe est :

$$x_{(10)} = \sum_{i=-m}^{n-1} b_i 2^i$$

- n nombre de bit de la partie entière
 m nombre de bit de la partie fractionnaire
- avec $0 \le x \le 2^n 2^{-m}$
- Exemple:

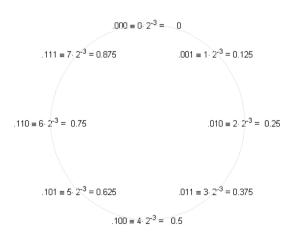
$$n=2, m=2$$

10. $11_{(2)} = 1 \cdot 2^{1} + 0 \cdot 2^{0} + 1 \cdot 2^{-1} + 1 \cdot 2^{-2} = 2.75$

27/01/2011

VTF - MSR

Cercle des nombres non-signés à virgules fixes (documentation Mathworks)



Format virgule fixe (signé)



 La représentation d'un nombre binaire signé avec virgule fixe est :

$$x_{(10)} = \sum_{i=-m}^{n-1} b_i 2^i - b_{n-1} 2^n$$

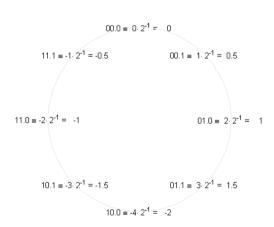
- n nombre de bit de la partie entière
 m nombre de bit de la partie fractionnaire
- avec $-2^{n-1} \le x \le 2^{n-1} 2^{-m}$
- Exemple :

$$n=3, m=2$$
 101. $01_{(2)} = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} - 1 \cdot 2^3 = -2.75$

VTF - MSR

3

Cercle des nombres signés à virgules fixes (documentation Mathworks)



Conversion



Conversion binaire vers décimal

- 1. noter le format : n , le nombre de bits de la partie entière m, le nombre de bits de la partie fractionnaire
- 2. convertir le nombre binaire en décimal signé sans tenir compte de la virgule (entier de n+m bits)
- 3. diviser par 2 à la puissance m(nombre de bit de la partie fractionnaire)

Exemples

$$n=3, m=4$$
 $010.1100_{(2)} = 0101100_{(2)} \cdot 2^{-4} = \frac{44_{(10)}}{16} = 2.75_{(10)}$

$$\mathbf{101.0100}_{(2)} = \mathbf{1010100}_{(2)} \cdot 2^{-4} = \frac{-44_{(10)}}{16} = -2.75_{(10)}$$

27/01/2011

VTF - MSR

1

Exercices

Convertir en décimal les nombres binaires signés (le premier bit est le bit de signe) :

- a) 01001.101
- b) 1011.10
- c) 110.1001
- d) 01011.111

Solutions

- a) 9.625
- b) -4.5
- c) -1.4375
- d) 11.875

Conversion



Conversion décimal vers binaire

- 1. choisir le format : n , le nombre de bits de la partie entière m, le nombre de bits de la partie fractionnaire
- 2. multiplier le nombre décimal par 2 à la puissance m
- 3. arrondir le résultat à un nombre entier
- 4. convertir en binaire signé
- Exemples (n=3, m=4)

$$\begin{array}{l} 3.\,29626_{(10)} \cdot 2^4 = 52.\,74_{(10)} \approx 53 = 0110101_{(2)} \\ \Rightarrow \ \ 3.\,29626_{(10)} \approx 011.\,0101_{(2)} = 3.\,3125_{(10)} \end{array}$$

$$\begin{array}{l} -3.\,29626_{(10)} \cdot 2^4 = -52.\,74_{(10)} \approx -53 = 1001011_{(2)} \\ \Rightarrow \ -3.\,29626_{(10)} \approx 100.\,1011_{(2)} = -3.\,3125_{(10)} \end{array}$$

27/01/2011

VTF - MSR

5

Exercices

Convertir en binaire signé les nombres décimaux suivants :

- a) 47.356 => n=7, m=3
- b) $-9.4789 \Rightarrow n=5, m=5$

Solutions

- a) 0101111.011
- b) 10110.10001

Arithmétique des nombres à virgules fixes : notation



- On utilise pour les nombres à virgules fixes la notation Q (format de Texas Instrument) ou la notation Matlab
- Qn.m désigne un nombre signé de n bits de partie entière (bit de signe inclus) et m bits de partie fractionnaire. Le nombre total de bits est n+m

Exemple: Q1.31 représente un nombre signé de 32 bits

 La notation Matlab Simulink est Fix_n+m_m (signé) ou UFIx n+m m (non-signé)

27/01/2011

Exemple: Fix_16_10 => nombre signé 16 bits (bit de signe inclus) avec 10 bits de partie fractionnaire

Nombres normalisés



- Il est courant et pratique de travailler avec des signaux de valeurs proches de |1| dans les intervalles [0,1[pour les nombres non-signés et]-1,+1[pour les nombres signés
- Ces nombres sont aux formats UFIx_m_m et Fix_m+1_m
- Exemples:

27/01/2011

VTF - MSR

 Un signal vidéo 8 bits peut être exprimé sous la forme UFIx_8_8:

Multiplication virgule fixe



- Soit deux nombres A et B dont les parties entières sont n_A , n_B et les parties fractionnaires m_A , m_B
- Le produit $C = A \cdot B$ a une partie entière $n_C = n_A + n_B$ et une partie fractionnaire $m_C = m_A + m_B$
 - Exemple: A (Fix_12_8) x B (Fix_16_9) = C (Fix_28_17)
- <u>Cas particulier</u>: multiplication de nombres normalisés signés. On peut diminuer de 1 le nombre de bit de la partie entière si les deux nombres sont strictement compris entre -1 et +1.
 - Exemple: A (Fix_12_11) x B (Fix_16_15) = C (Fix_28_26)

27/01/2011

• Le résultat peut être exprimé avec le format Fix_27_26 sans overflow

Addition virgule fixe



- Soit deux nombres A et B dont les parties entières sont n_A , n_B et les parties fractionnaires m_A , m_B .
- Pour additionner ces nombres les positions des virgules doivent être alignées
- La somme C = A + B a une partie entière $n_C = max(n_A, n_B) + 1$ et une partie fractionnaire $m_C = max(m_A, m_B)$
 - Exemple : A (Fix_12_8) + B (Fix_16_9) = C (Fix_17_9)

27/01/2011 VTF - MSR 9

Rapport signal/bruit



- Le bruit dans un circuit de traitement numérique provient de la conversion AD et de la quantification
- La dynamique est le rapport entre la puissance d'un signal d'amplitude maximale (avant écrêtage ou dépassement) et la puissance d'un signal de niveau minimum (niveau du bruit)
- Le rapport signal sur bruit (SNR) est le rapport entre la puissance d'un signal et celle du bruit $SNR = \frac{P_{signal}}{P_{bruit}}$ ou en dB :

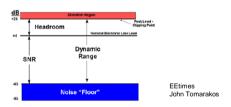
$$SNR = 20 \log_{10} \frac{A_{signal}}{A_{bruit}}$$

27/01/2011

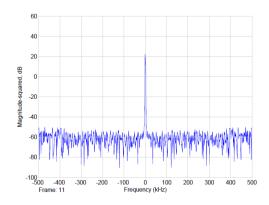
VTF - MSR

10

Attention: On doit normalement toujours donner le niveau du signal qui est pris en compte pour calculer ou mesurer un rapport signal sur bruit.



Exemple: mesure d'une sinusoïde de puissance 1 W en dBm : (1 dBm = 1mW sur 1 ohm)



Rapport signal/bruit de 90 dB signal +30dBm (amplitude = 1) bruit -60dBm





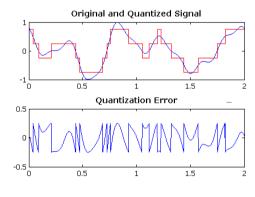
- L'erreur de quantification d'un signal par un ADC est la différence à l'instant d'échantillonnage entre la valeur du signal analogique en entrée et la valeur du signal échantillonné.
- Si le pas du convertisseur (Isb) est Δ , cette erreur est uniformément distribuée dans l'intervalle $\left[-\frac{\Delta}{2},+\frac{\Delta}{2}\right]$
- L'erreur de quantification est considérée comme <u>une source de bruit ajoutée au signal converti</u> de puissance $P_e = \frac{\Delta^2}{12}$

27/01/2011

VTF - MSR

11

Exemple d'erreur de quantification d'un signal quantifié en 4 bits signé



Erreur de quantification(2)



- Dans le cas un convertisseur AD de b bits, le SNR minimum théorique pour une sinusoïde est $1.761+6.02 \cdot b \ dB$ par rapport au niveau maximum (Full Scale) du signal échantillonné.
- Plus généralement le SNR d'un signal échantillonné sur b bits est environ $6.02 \cdot b \ dB$
- Réduire le nombre de bits d'un signal revient à augmenter le niveau de bruit – diminuer le SNR.
- Exemple: Un signal est échantillonné sur 16 bits, son SNR est 96 dB. Si nous supprimons les 4 bits de poids faible, le SNR devient 72 dB.

VTF - MSR

27/01/2011

Exemple de calcul



Rapport signal sur bruit maximum pour un signal sinusoïdal:

Soit un signal de type Fix_16_15

$$\Delta = 2^{-15} \text{ donc } P_e = \frac{2^{-30}}{12} = 7.76 \cdot 10^{-12} = -101.1 \ dB$$

L'amplitude maximale d'un signal sinusoïdal est 1. La puissance du signal sinusoïdal est $P_s = \frac{1}{2}$

Nous pouvons calculer le rapport signal sur bruit :
$$SNR = \frac{P_s}{P_e} = \frac{12}{2 \cdot 2^{-30}} \approx 98 \ dB$$

27/01/2011

La puissance d'une sinusoïde d'amplitude A est $P_s = \frac{A^2}{2}$	

Arrondi et troncature



- Faire un arrondi revient à ajouter un bruit de quantification de puissance $P_e = \frac{\Delta^2}{12} = \frac{(2^{-m})^2}{12}$ m étant le nombre de bit de la partie fractionnaire
- Exemples (avec 1 comme référence du 0dB):

$$P_e = \frac{1}{12} = -10.79 \ dB$$
 $P_s = \frac{127^2}{2} = 39.06 \ dB$ $SNR = 49.85 \ dB$

27/01/2011

• arrondi à la partie entière d'un 8 bits non-signé (UFIx_8_0)
$$P_e = \frac{1}{12} = -10.79 \ dB \qquad P_s = \frac{127^2}{2} = 39.06 \ dB \quad SNR = 49.85 \ dB$$
• arrondi à un nombre signé (Fix_12_8)
$$P_e = \frac{2^{-16}}{12} = -58.95 \ dB \qquad P_s = \frac{8^2}{2} = 15.05 \ dB \qquad SNR = 74.00 \ dB$$

 Une troncature (suppression de bits de poids faible) ajoute un bruit de même niveau que l'arrondi et un bias DC de $-\frac{\Delta}{2}$

Exemple : réglage de gain (1)

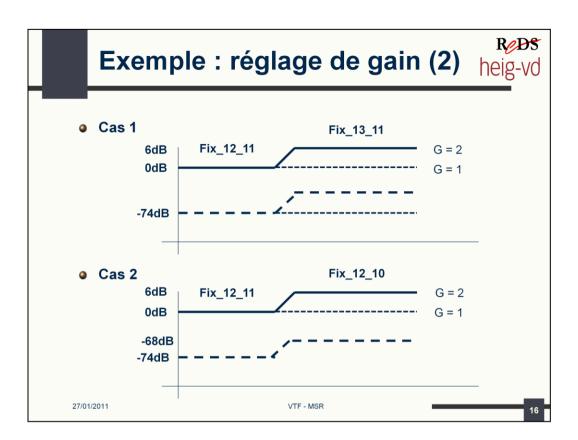






- G Gain non-signé 0 < G < 2, UFix_16_15
- Sout Résultat sans arrondi, Fix_28_26
- Arrondi de S_{out} à Fix 12_10 => Le SNR est dégradé de 6 dB pour G=1 et conservé pour G=2
- Le pas minimum du gain est $20 \log(1 + 2^{-15}) = 2.6 \cdot 10^{-4} dB$

27/01/2011 VTF - MSR 15



Signal / bruit : addition et multiplication



- ullet P_{As} , P_{Bs} puissance signaux A et B
- ullet P_{Ae} , P_{Be} puiss. bruit de quantification des signaux A et B
- \bullet P_e puiss. bruit de quantification (additionneur ou multiplicateur)
- Addition (calcul approximatif) => prendre le niveau max du bruit

$$P_{A+B} = P_{As} + P_{Ae} + P_{Bs} + P_{Be} + P_e$$

Multiplication (calcul approximatif)

$$P_{A \cdot B} = P_{As} \cdot P_{Bs} + P_{As} \cdot P_{Be} + P_{Bs} \cdot P_{Ae} + P_{Ae} \cdot P_{Be} + P_{e}$$

27/01/2011 VTF - MSR

17

Quantification d'un gain (ou atténuation)



- Calcul du nombre de bits optimum d'un réglage de gain : un pas d'atténuation doit correspondre à (au minimum) une variation d'un Isb du signal
- Soit une variable de réglage G de format UFix_n_n permettant une atténuation de $1-2^{-n}$ à 0

Soit k un entier dans l'intervalle $[2^n,1]$, alors $G=1-k\cdot 2^{-n}$ Le pas de variation du signal est :

$$\Delta_v = S[(1 - k \cdot 2^{-n}) - (1 - (k+1) \cdot 2^{-n})] = S \cdot 2^{-n}$$

- Pour un signal de m bits, le nombre de bits optimum du réglage de gain est celui du signal
 - Si le format de S est UFix_n_0, le pas maximum est 1
 - Si le format de S est Fix_n+1_n, le pas maximum est 2⁻ⁿ

27/01/2011 VTF - MSR

18