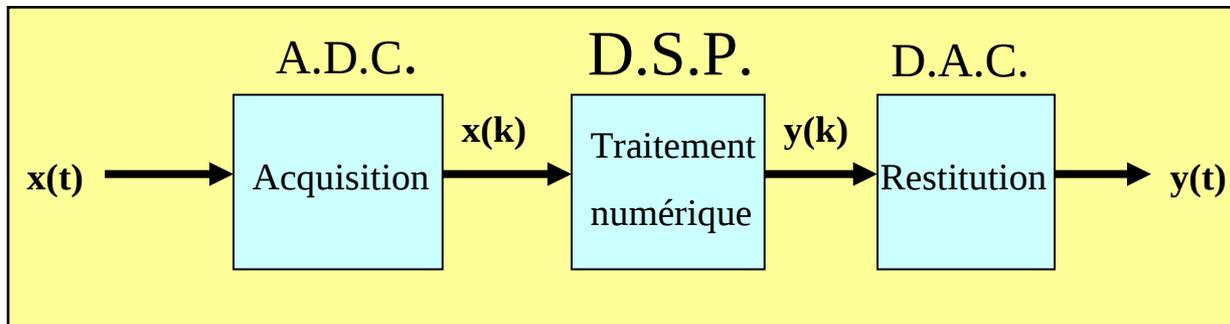


D.S.P.

(Digital Signal Processor)

- Processeur destiné au traitement du signal
- Les grandeurs physiques sont transformées en nombres
- Architecture des DSP optimisée pour le traitement numérique du signal



PROGRESSION

- Principaux traitements et domaines d'utilisation
- Comment choisir un D.S.P.
- Chaîne d'acquisition traitement restitution
- Les convertisseurs analogiques numériques
- Les convertisseurs numériques analogiques
- Etude du DSPIC 30F6014
- Calcul, Simulation et Programmation de filtre R.I.F. et R.I.I.
- Application au traitement du son.

PRINCIPAUX TRAITEMENTS

- Analyser un signal (FFT)
- Détecter, identifier des signaux (1750, DTMF, parole, image...)
- Filtrer
- Générer des signaux
- Moduler démoduler (modem, ADSL, Ethernet, AX25....)
- Crypter, décrypter

DOMAINES D 'APPLICATION

- Télécommunications
- Sons, Images, Vidéo
- Instrumentation
- Biomédical
- Commande de process
- Equipement automobile, avionique, radars...

COMPARAISON AUX AUTRES DISPOSITIFS DE TRAITEMENTS

Traitements microprogrammés

- microordinateurs (Traitements complexes mais rarement en temps réel)
- microprocesseurs classiques (<100 Kéch./s)
- DSP (< 1Méch/s)

Traitements câblés (architectures parallèles)

- FPGA,... <50 Méch/s...
- ASIC (réservé aux grosses productions)

COMMENT CHOISIR UN D.S.P.

Caractéristiques communes

- Architecture optimisée pour le traitement du signal (nombreux bus, Harvard, pipe-line...)
- Instruction multiplication accumulation en un cycle
- Modes d'adressages particuliers

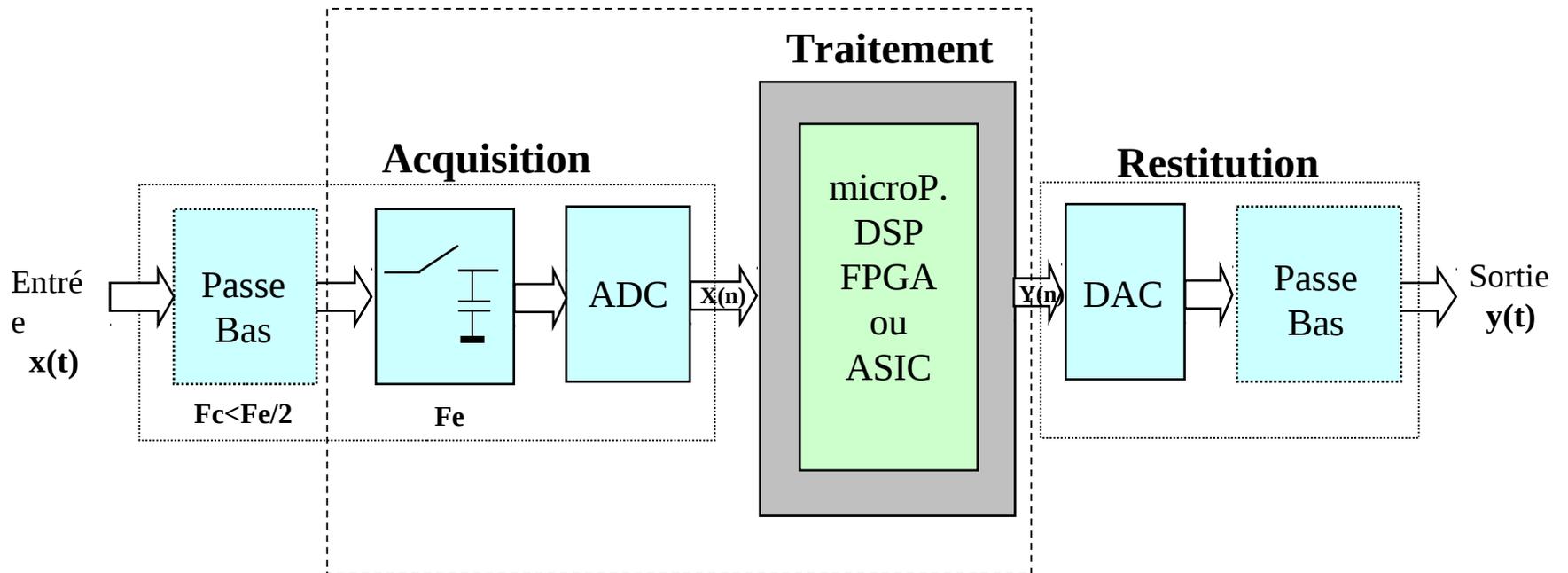
Les différentes familles de DSP

- DSP avec unité de calculs en virgule fixe
- DSP avec unité de calculs en virgule flottante câblé
- DSP avec des périphériques intégrés (contrôle moteur..)

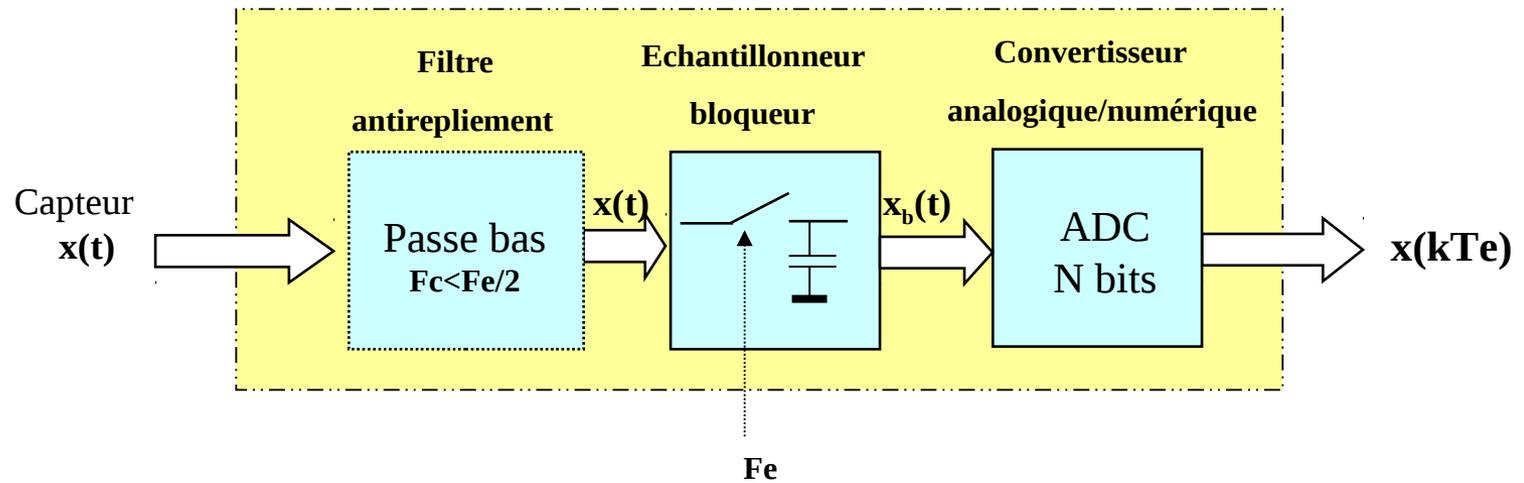
Les langages de programmation: L.machine, C, G. de C(ViSim..)

Constructeurs : AT&T, Texas Inst., Motorola, Analoge Devices...

ACQUISITION TRAITEMENT RESTITUTION



ACQUISITION



ACQUISITION

Produire une suite de nombre entiers $x(k)$ à partir du signal analogique $x(t)$

Filtre antirepliement du spectre (anti-aliasing)

- Choix de la fréquence d'échantillonnage (Shannon)
- Choix de la fréquence de coupure du filtre passe-bas

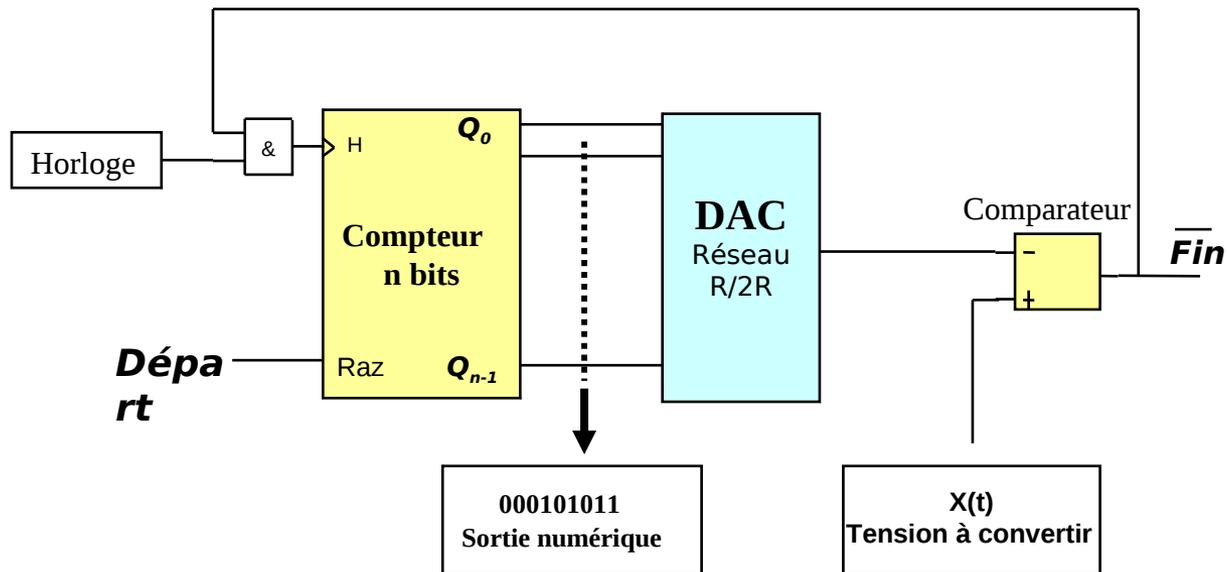
Echantillonneur-bloqueur

- Rôle
- Problèmes liés à l'impédance

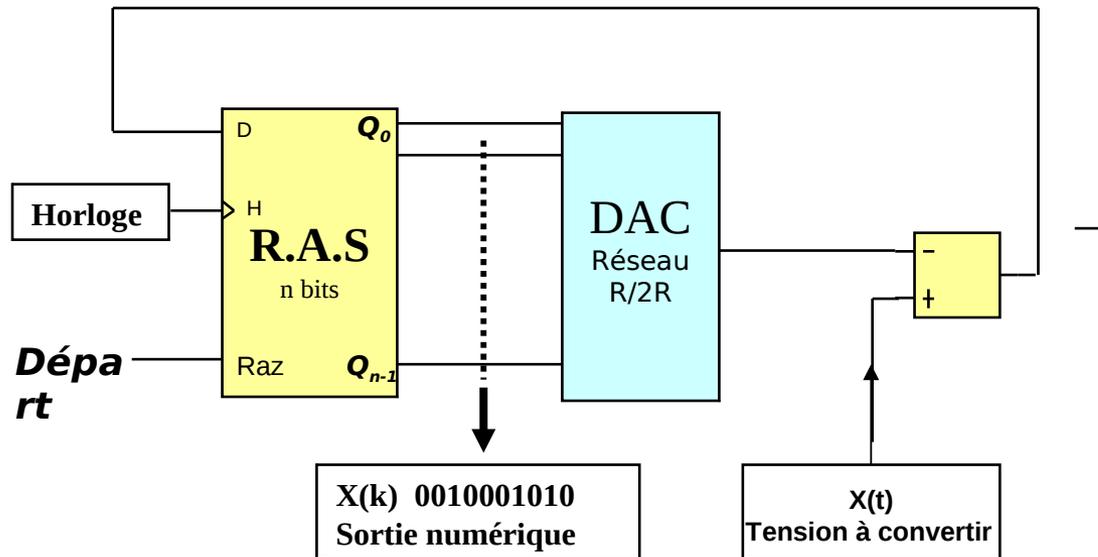
Quantification

- Résolution du convertisseur (A.D.C.)
- Temps de réponse
- Linéarité
- Précision et dynamique

Convertisseur Analogique - Numérique à rampe numérique



Convertisseur Analogique – à registre à approximations successives



Autres convertisseurs Analogique - Numérique

Convertisseur « Flash »

Convertisseur « sigma-delta »

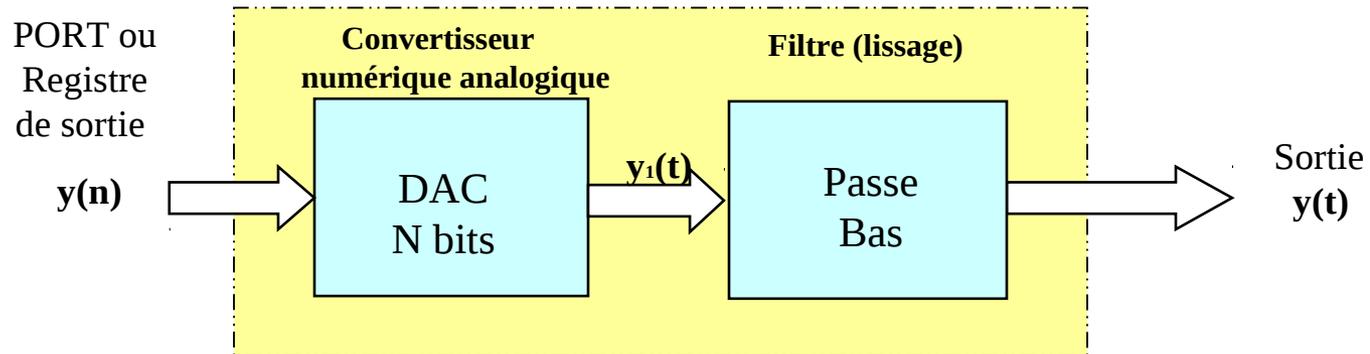
(voir TD 1)

RESTITUTION

- Bloqueur d'ordre 0 (registre de sortie)
- DAC ou MLI (PWM)
- Filtre passe-bas analogique

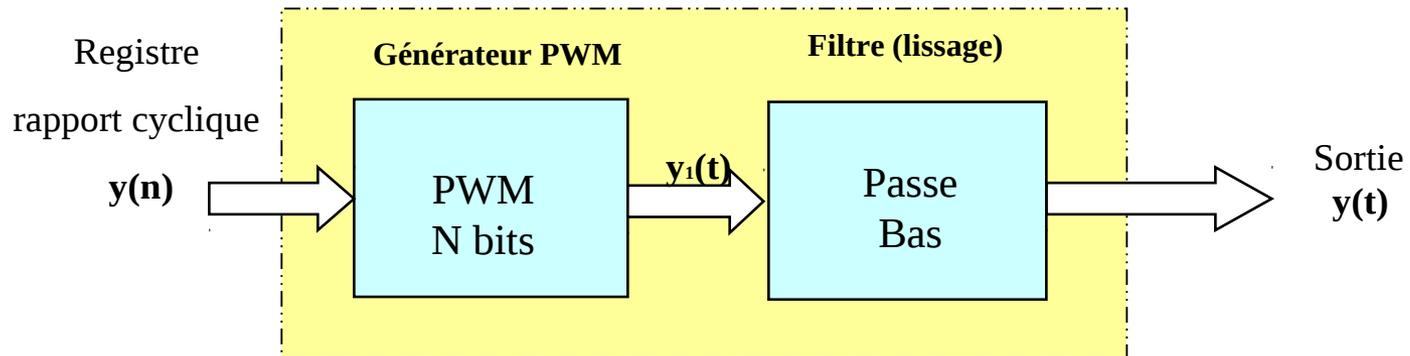
RESTITUTION

convertisseur analogique/numérique

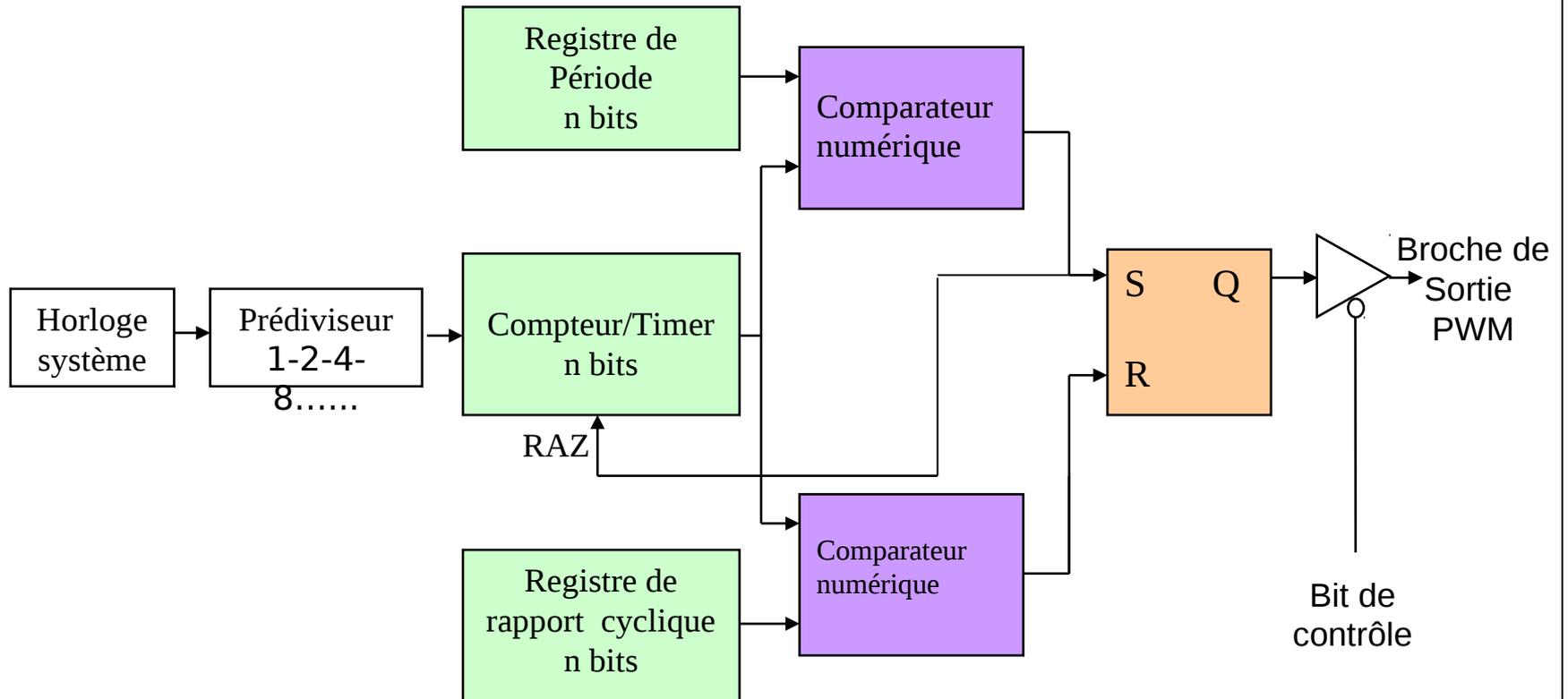


RESTITUTION

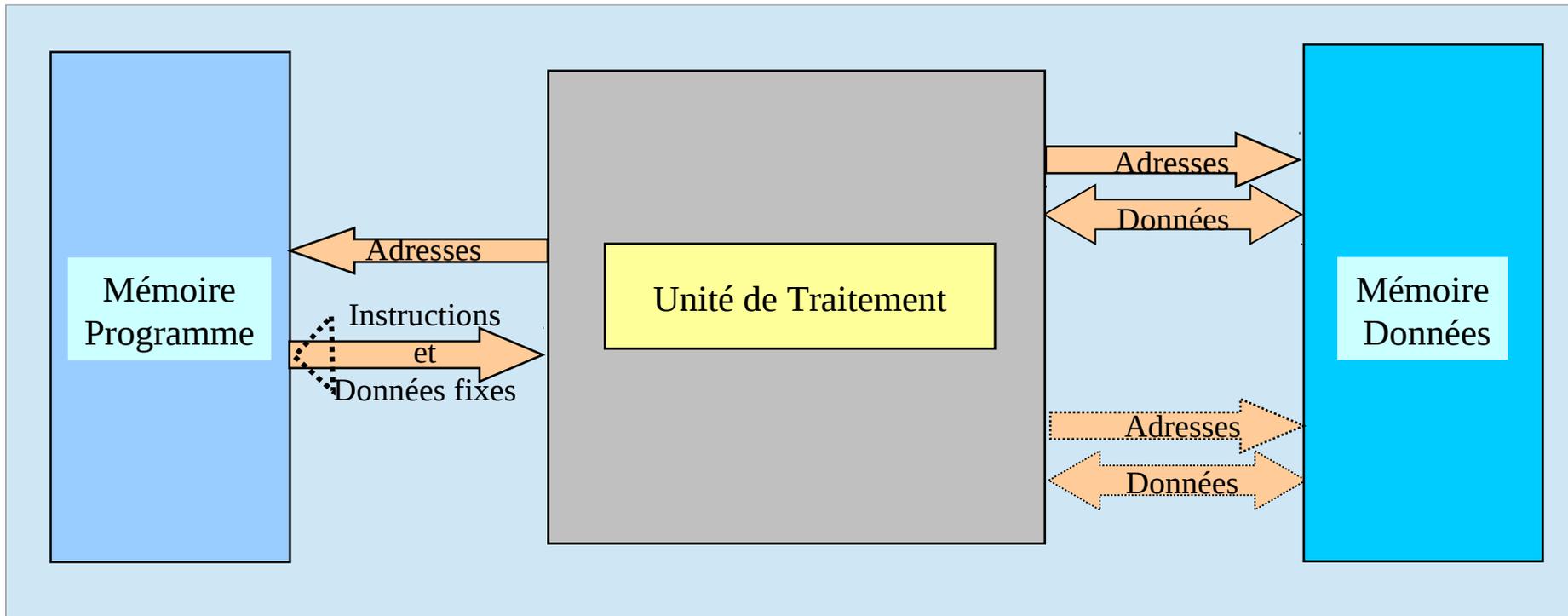
Par modulation de largeur d'impulsions



Génération PWM (Principe)



Architecture Harvard des DSPic (simplifiée)



PRESENTATION du DSPIC 33FJ16MCxxx (Microchip)

Documentation technique

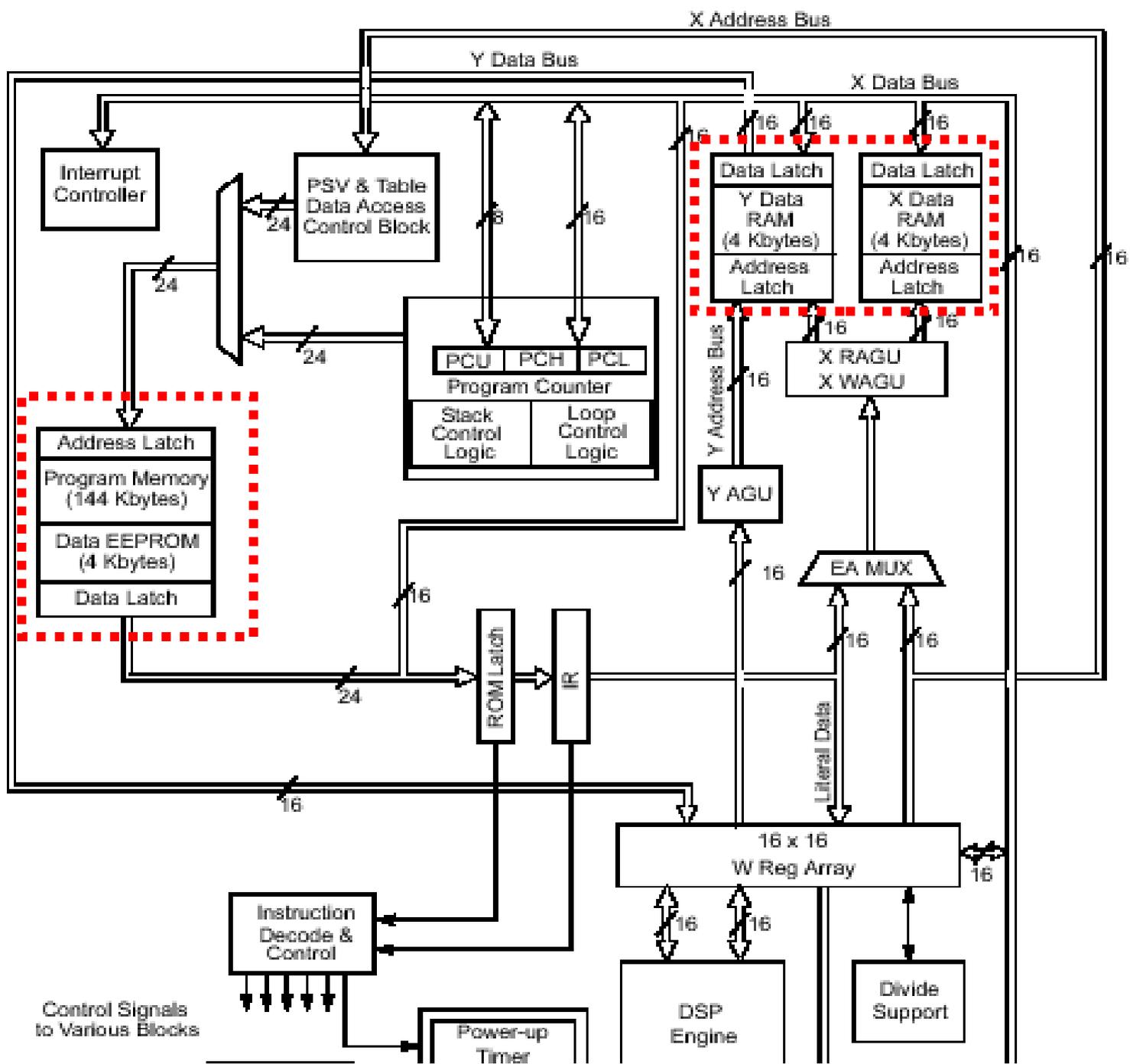
- Datasheet principal 33FJ16MC :	70283k.pdf
- CPU:	70204c.pdf
- Instructions :	70157f.pdf
- Ports (Entrées/Sorties)	70193d.pdf
- ADC :	70183d.pdf
- Timers	70205d.pdf
- uart	70188e.pdf
- Interruptions	70214c.pdf
- Horloge	70186e.pdf
- Watchdog	70196d.pdf
- Librairies	51456g.pdf
- Assembleur Linker :	51317h.pdf
- Compilateur XC16 :	50002071C.pdf

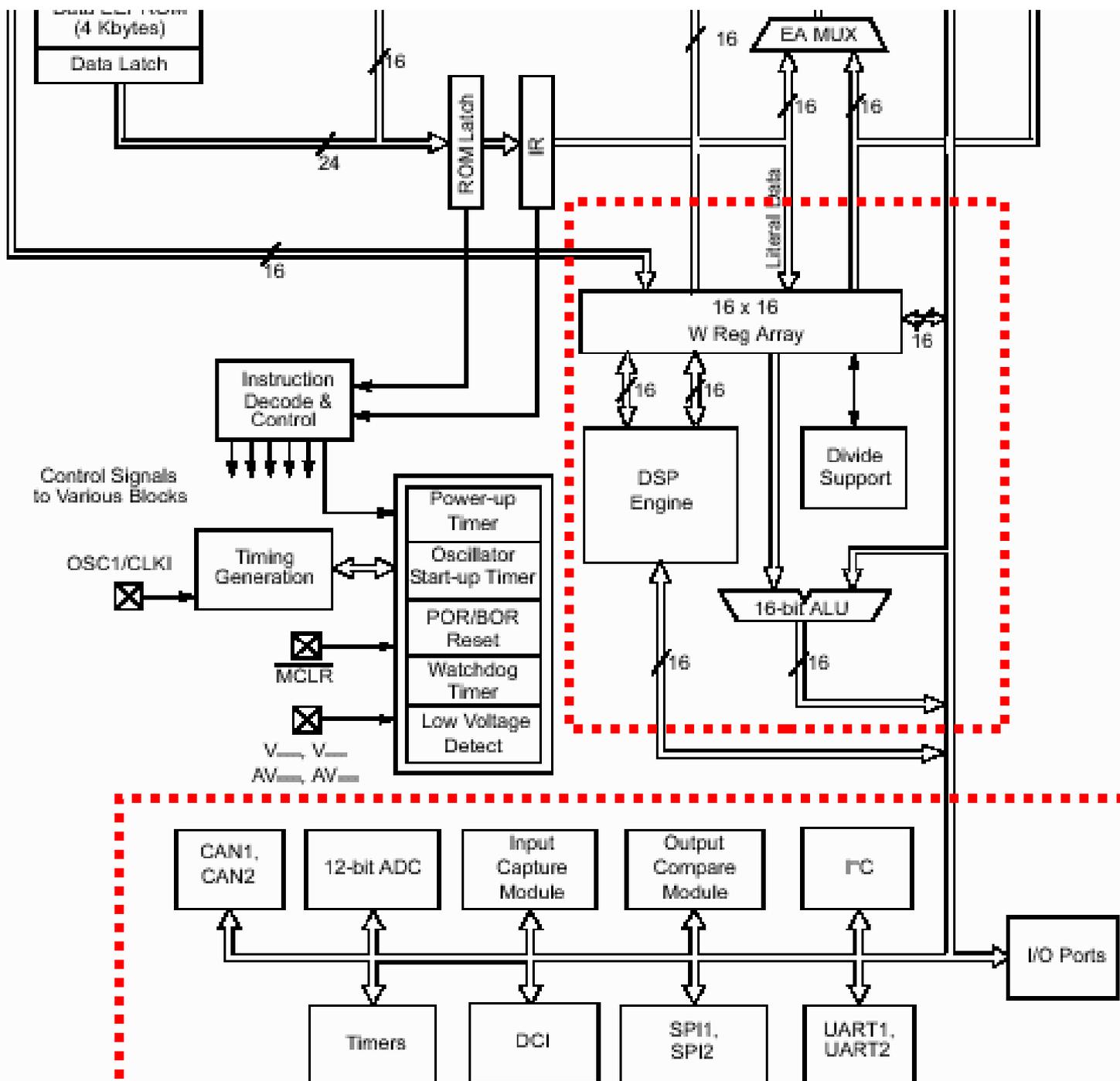
PRESENTATION du DSPIC 33f (Microchip)

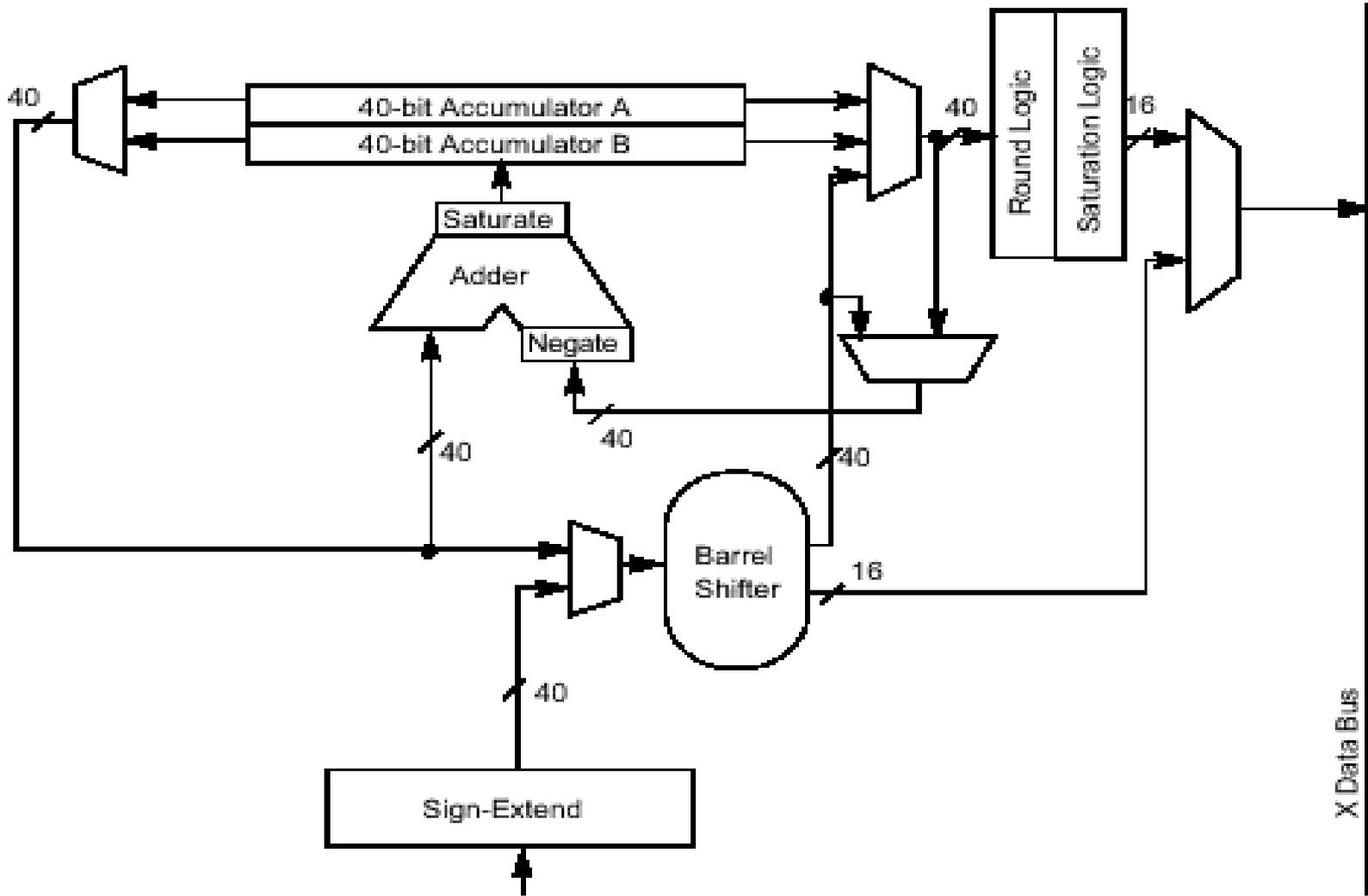
- Architecture
- Registres
- Mémoire
- Instructions et Modes d 'adressages
- Les principales fonctionnalités
- Les Timers
- Le convertisseur AD
- Gestion des interruptions

Caractéristiques du 33FJ16MC304

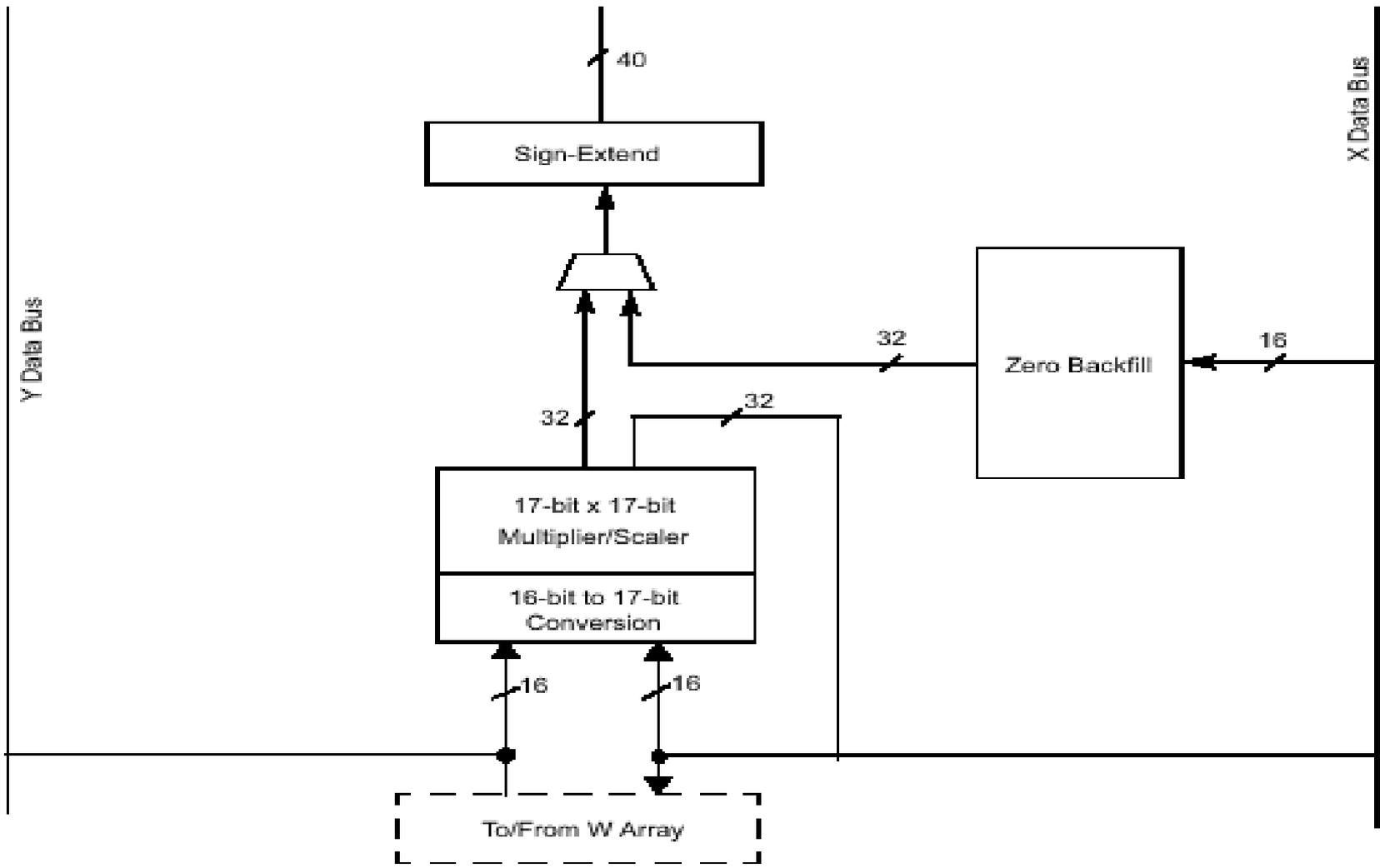
- Coeur 16 bits – 40Mips - 2ko de RAM- 16ko de Flash
- Alimentation : 3v à 3.6 V
- Oscillateurs internes **FRC=7.37MHz** ou LPRC=32.768kHz
- Oscillateur externe jusqu'à 40 MHz
- PLL et diviseur d'horloge intégrés configurable
- Fréquence cycle (instruction, Timers,etc..) **Fcy = Fosc/2**
- 9 ports d'entrées/sorties 8bits
- 9 entrées analogiques 12bits- 500kech/s ou 10bits-1100kech/s
- 3 Timers 16 bits , 1 Timer 32bits
- Sorties comparaison et PWM (10)
- Entrées de capture, QEP
- 1 liaisons série asynchrone SCI (Uart)
- 1 liaisons série synchrone SPI
- 1 bus I2C







X Data Bus



TROIS TYPES DE MEMOIRES

Mémoire programme:

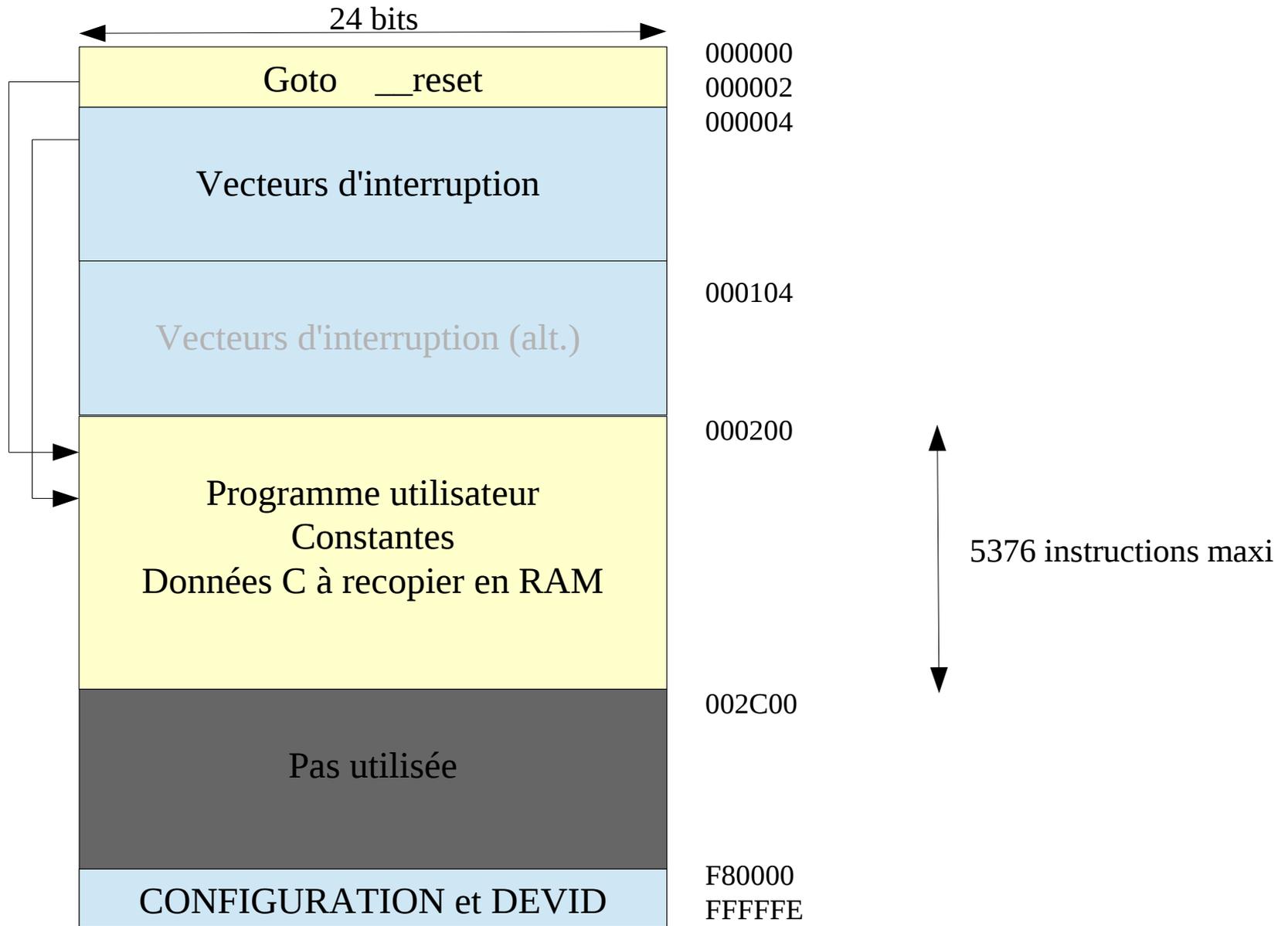
- flash 48 Kmots de 24 bits (144 k octets)

Mémoire données en EEPROM : 8k octets

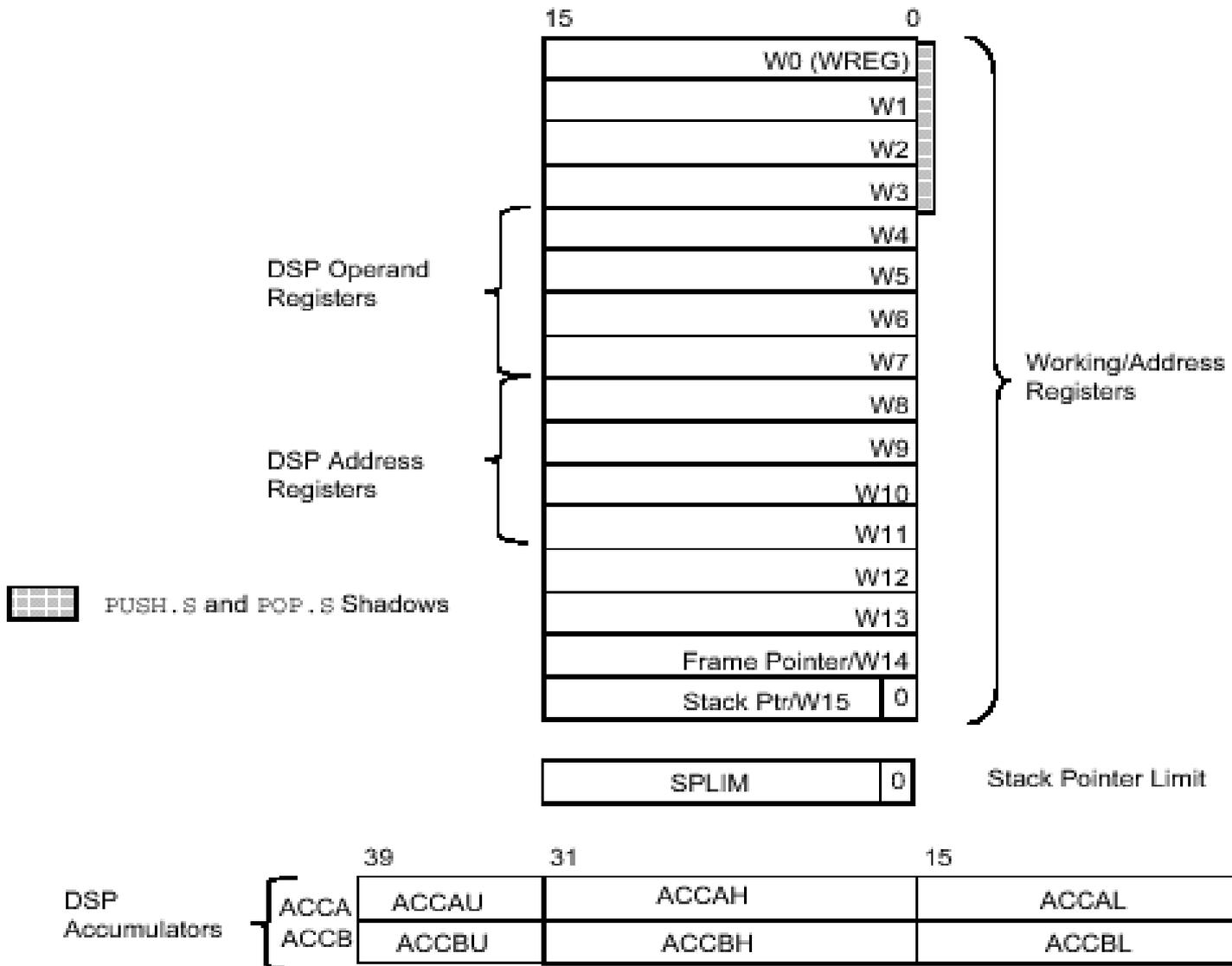
Mémoire données RAM: 32 Kmots de 16 bits (64 k octets)

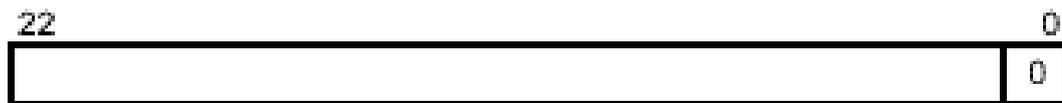
- Les registres système et périphériques internes (SFR) 1k mots (2k octets)
- RAM X 2k mots (4k octets)
- RAM Y 2k mots (4k octets)
- les derniers 16 k mots peuvent être utilisés pour lire la ROM

MEMOIRE Programme



LES PRINCIPAUX REGISTRES





Program Counter



Data Table Page Address



Program Space Visibility Page Address



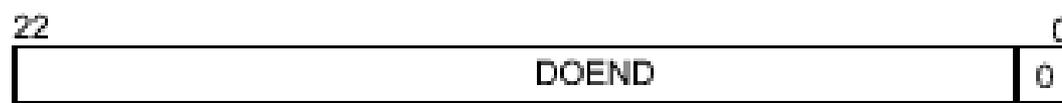
REPEAT Loop Counter



DO Loop Counter



DO Loop Start Address



DO Loop End Address



Status Register



Core Control Register

LES PRINCIPAUX MODES D 'ADRESSAGE

Une Instruction occupe 24 bits dans la mémoire programme

- Adressage IMMEDIAT:

La donnée se trouve à la suite de l 'instruction dans la mémoire FLASH

- Adressage DIRECT

L 'instruction contient l 'adresse (sur 13bits) de la donnée qui se trouve en RAM (accès limité à 8k adresses)

- Adressage indirect

Un des registres de travail W_n contient l 'adresse de la donnée qui se trouve en RAM

- post-incrémentation et post-décrémentation
- pre-incrémentation et pre-décrémentation
- offset

- Adressage « bit »

- permet de modifier ou de tester un bit d 'un registre en mémoire donnée

ADRESSAGE IMMEDIAT

Symbole #

Mnémonique #nombre, destination

- Exemples :

```
mov #0x01C2, W3
```

```
mov.b #0x05, W1
```

```
add #0x05, W0
```

- Autres instructions : sub, and, ior, xor...

- Taille du #nombre :

#lit10 signifie « literal » de 10 bits maxi....

ADRESSAGE DIRECT

*Mnémonique adresse ,
destination*
Mnémonique source , adresse

- Exemples :

```
mov 0x01C2, W3
```

```
mov.b W1, 0x0005
```

```
add 0x0105 , W0
```

```
rlc 0x01F0 , W0
```

```
rlc 0x01F0
```

- Autres instructions : rlncl, asl, ls, rrc, sub, and , ior, xor...

ADRESSAGE INDIRECT

Mnémonique [source] , destination
Mnémonique source , [destination]
Mnémonique [source] , [destination]

[Wx] Utiliser le contenu de l 'adresse Wx
[Wx ++] et [Wx - -] Utiliser le contenu de l 'adresse Wx puis ajouter ou retirer 1 à Wx
[++Wx], [- -Wx] Ajouter ou retirer 1 à Wx puis utiliser le contenu de l 'adresse Wx
[Wx+Wy] Utiliser le contenu de l 'adresse calculée

- Exemples :

```
mov [W2] , W3
mov [--W2] , W3
mov W5 , [W1+W0]
mov [W5--] , [++W2]

add Wx , [Wy++] , [Wz]
```

ADRESSAGE « bit »

Mnémonique destination, #numéro du bit

Exemples:

bset W6, #0

bclr 0x105A, #7

btsc [W5], #2

btss 0x105A, #3

Sauts et sauts conditionnels

Mnémonique (condition ,) adresse-programme

Label ou étiquette = adresse de la mémoire programme

- Saut simple

```
goto label1  
bra label2
```

- Saut conditionnel

Z, C, N, GE, GT, LE,LT,OA,....

Ex: bra z, label3

Appels de sous-programmes

APPEL :

- *CALL* *adresse-programme (ou label)*
 - L 'adresse de retour est sauvegardée dans la pile
 - W15 contient l 'adresse du sommet de la pile

RETOUR :

- *RETURN* ou *RETLW #nombre,Wx*
 - L 'adresse de retour est « dépilée »

Les instructions dédiées au traitement du signal

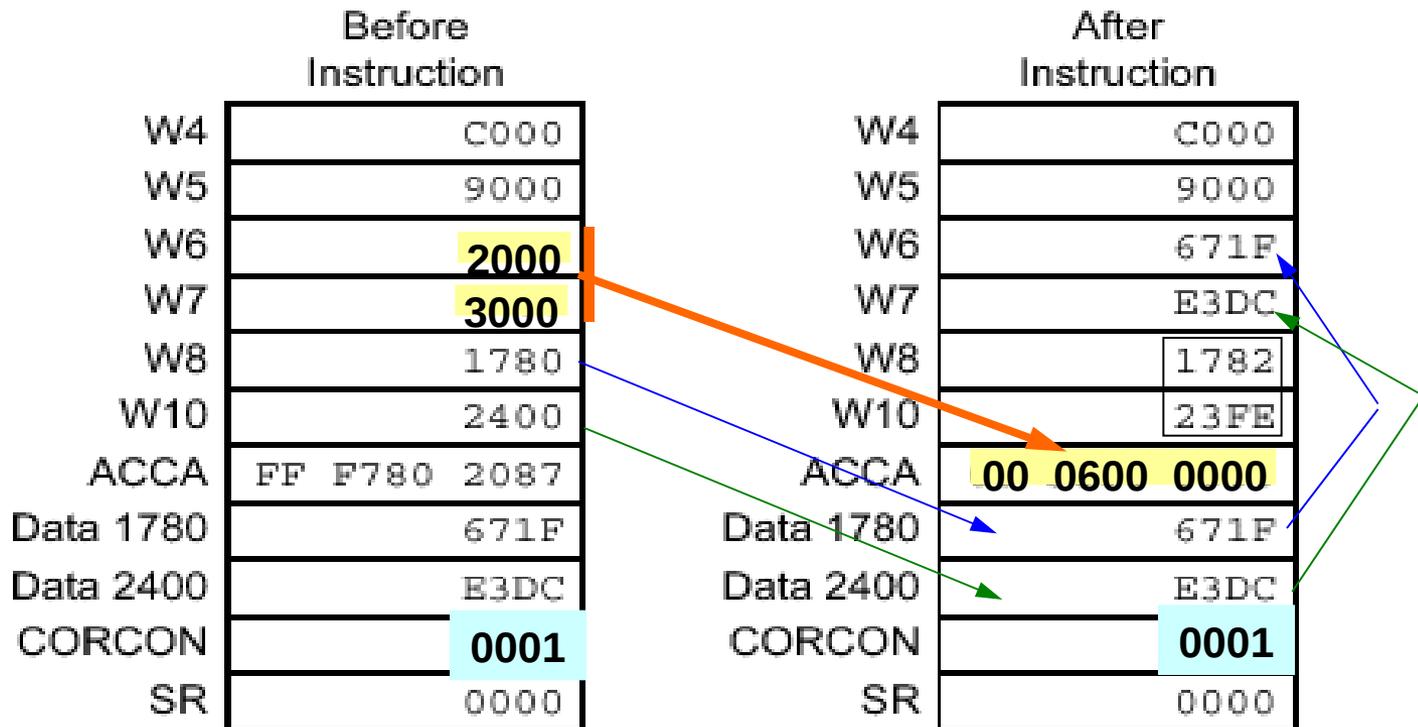
- Les modes de fonctionnement du multiplicateur (registre **CORCON**)
 - Signed/Unsigned
 - Saturation/Supersaturation (32 ou 40 bits)
 - Saturation enable/disable
 - Integer/Fractionnal (décalage du résultat pour format Q31)
- Les instructions:
CLR, MPY, MAC, MSC, MOVSAC, LAC, SAC, SFTAC
- Fonctionnement du multiplicateur (MPY , MAC, MSC)
 - Les grandeurs à multiplier doivent être dans **2 registres (W4...W7)**
 - Le résultat est stocké, ajouté ou soustrait à un accumulateurs (A ou B).
 - Le contenu des 2 registres peut être modifié, pour la multiplication suivante, grâce à l'adressage indirect
 - . de la RAM X par : [W8] ou [W9] +/-2, 4 6
 - . de la RAM Y par : [W10] ou [W11] +/-2, 4 6
 - Les bits 16 à 31 du deuxième accumulateur peuvent-être transférés dans [W13]

Fonctionnement du multiplicateur (MPY , MAC, MSC)

- Les grandeurs à multiplier doivent être dans **2 registres (W4...W7)**
- Le résultat est stocké, ajouté ou soustrait à un accumulateurs (A ou B).
- Le contenu des 2 registres peut être modifié, pour la multiplication suivante, grâce à l'adressage indirect
 - . de la RAM X par : [W8] ou [W9] +/-2, 4 6
 - . de la RAM Y par : [W10] ou [W11] +/-2, 4 6
- Les bits 16 à 31 du deuxième accumulateur peuvent être transférés dans [W13]

Exemple :

MPY W6*W7, A, [W8]+=2, W6, [W10] -=2, W7



Exemple :

MPY W6*W7, A, [W8]+=2, W6, [W10] -=2, W7

$A = W6 * W7$

$W6 = [W8] ; W8 = W8 + 2$

$W7 = [W10] ; W10 = W10 - 2$

	Before Instruction		After Instruction
W4	C000	W4	C000
W5	9000	W5	9000
W6	2000	W6	671F
W7	3000	W7	E3DC
W8	1780	W8	1782
W10	2400	W10	23FE
ACCA	FF F780 2087	ACCA	00 0C00 0000
Data 1780	671F	Data 1780	671F
Data 2400	E3DC	Data 2400	E3DC
CORCON	0000	CORCON	0000
SR	0000	SR	0000

$$Acc1 = Acc1 + Wm * Wn$$

$$Wm = [Wx]$$

$$Wn = [Wy]$$

$$W13 \text{ ou } [W13] = Acc2$$

MAC

multiply and Accumulate

Syntax: {label:} MAC Wm*Wn, Acc {,[Wx], Wxd} {,[Wy], Wyd} {,AWB}
{,[Wx]+=kx, Wxd} {,[Wy]+=ky, Wyd}
{,[Wx]-=kx, Wxd} {,[Wy]-=ky, Wyd}
{,[W9+W12], Wxd} {,[W11+W12], Wyd}

Operands: Wm*Wn ∈ [W4*W5, W4*W6, W4*W7, W5*W6, W5*W7, W6*W7]
Acc ∈ [A,B]
Wx ∈ [W8, W9]; kx ∈ [-6, -4, -2, 2, 4, 6]; Wxd ∈ [W4 ... W7]
Wy ∈ [W10, W11]; ky ∈ [-6, -4, -2, 2, 4, 6]; Wyd ∈ [W4 ... W7]
AWB ∈ [W13, [W13]+=2]

Operation: (Acc(A or B)) + (Wm)*(Wn) → Acc(A or B)
([Wx]) → Wxd; (Wx)+kx → Wx
([Wy]) → Wyd; (Wy)+ky → Wy
(Acc(B or A)) rounded → AWB

Status Affected: OA, OB, OAB, SA, SB, SAB

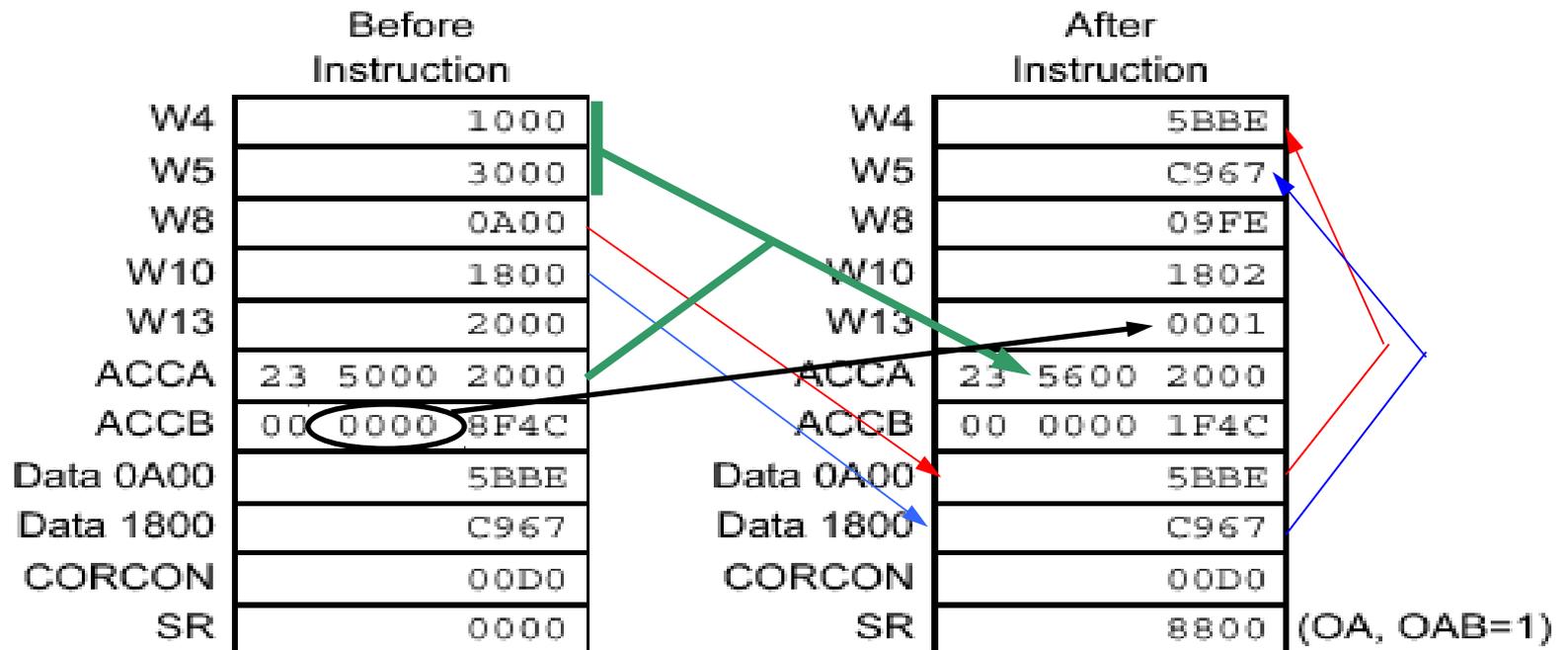
Exemple :

MAC W4*W5, A, [W8]-=2, W4, [W10] +=2, W5, W13

$$A = A + W4 * W5$$

$$W4 = [W8] ; W8 = W8 - 2$$

$$W5 = [W10] ; W10 = W10 + 2$$



Les buffers circulaires

Principe :

- On initialise les registres de contrôle
- On met l 'adresse de départ dans W_x
- W_x++ incrémente W_x
- lorsque W_x atteint la fin du buffer il revient automatiquement à l 'adresse de départ

Registres de contrôle :

MODCON :

- Autorisations des buffers X et Y
- Choix des registres pointeurs W_x et W_y ($W_0 \dots W_{14}$)

XMODSRT et YMODSRT : adresses de départ

XMODEND et YMODEND : adresses de fin

XBREV et YBREV : pour l 'adressage ' bit reverse '

Utilisations des Timers_

5 Timers indépendantes:

Génération de période d'échantillonnage

Bases de temps pour mesure de durée (Capture des Timers)

Compteur d'impulsions extérieures

Gestion de codeurs incrémentaux (QEP)

Bases de temps pour la génération automatique des signaux PWM

Gestion d'interruptions sur événements Timers, capture, comparaison...

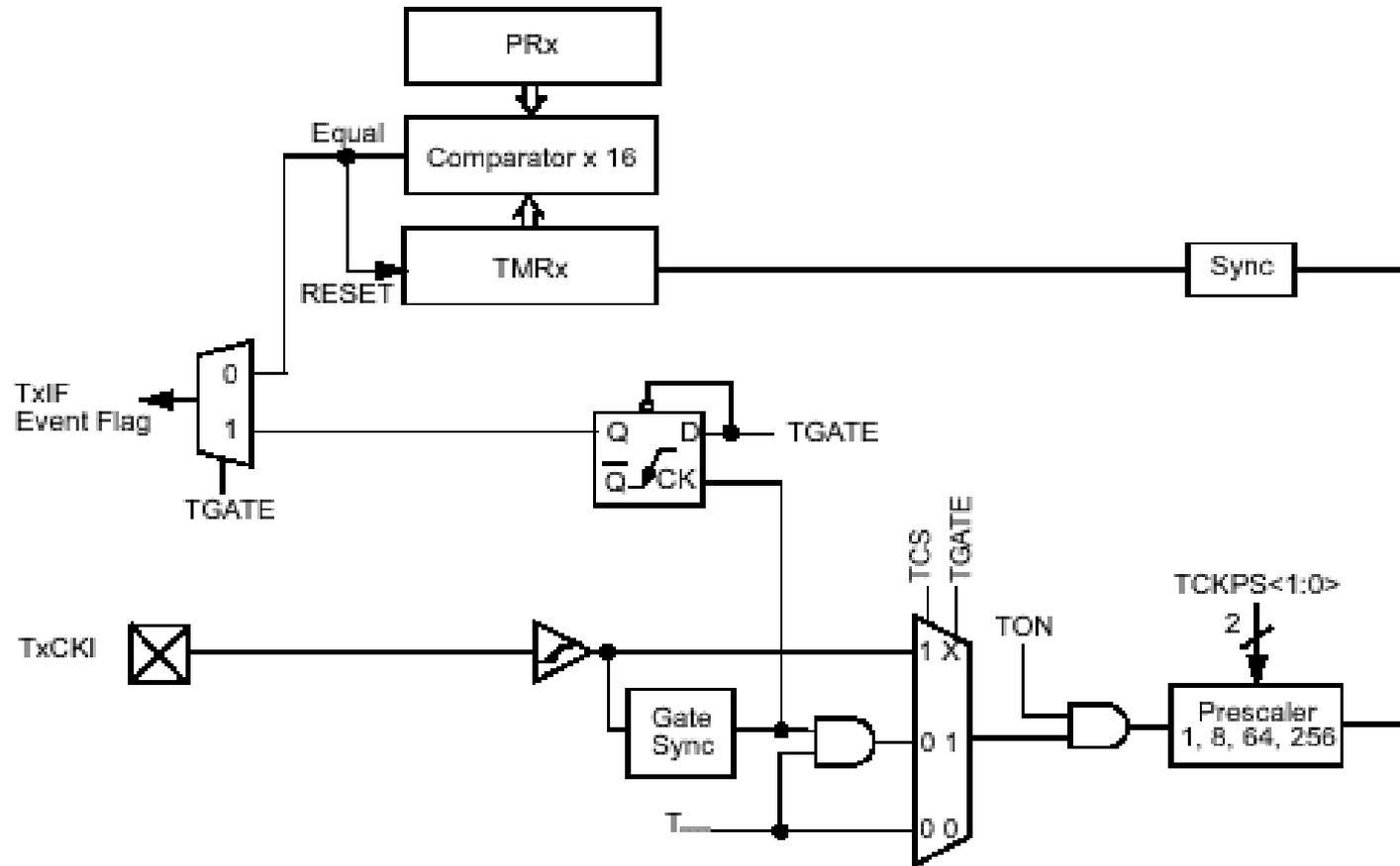
Registres associés à un timer

un registre compteur TMRx

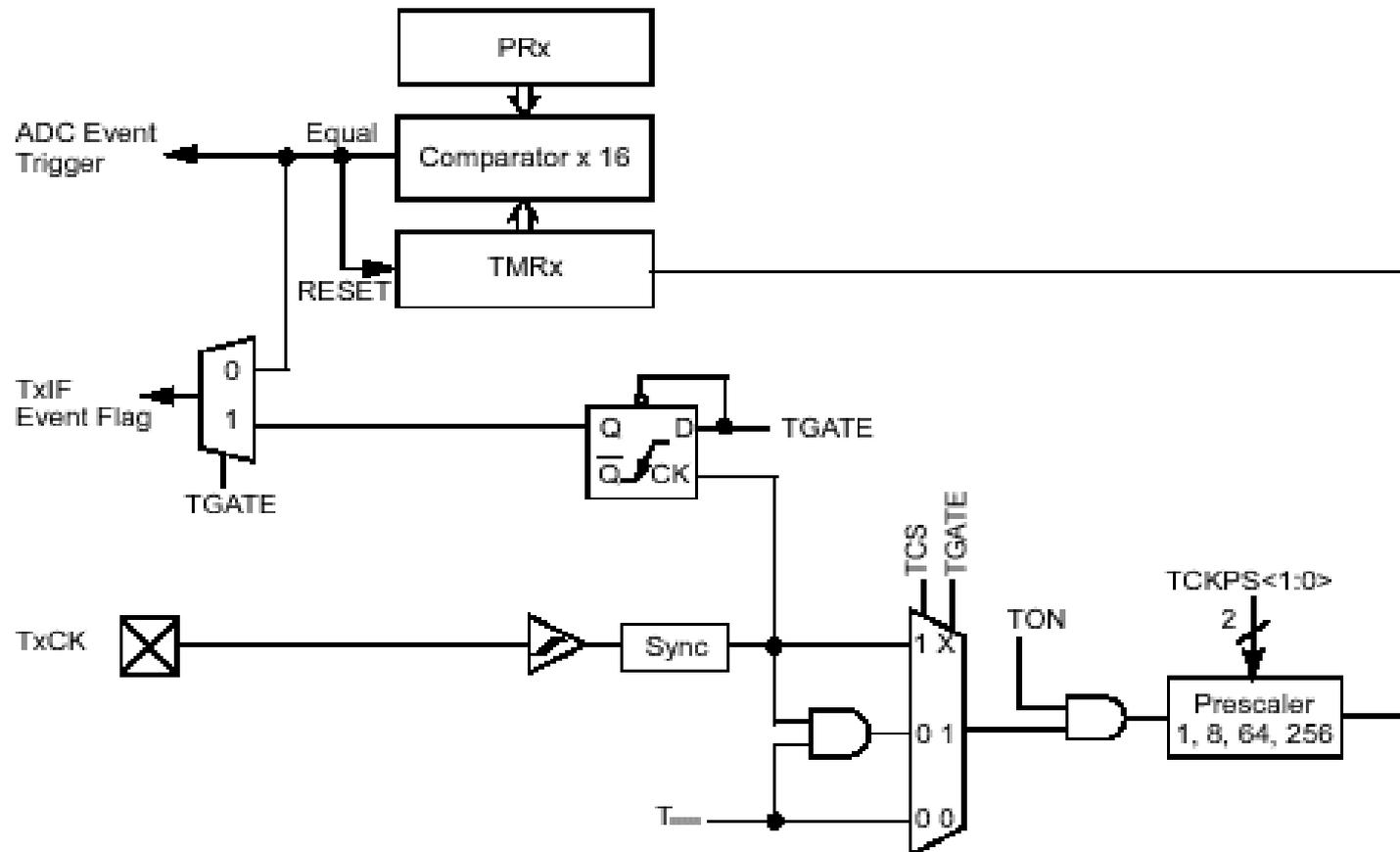
un registre de contrôle TxCON

un registre période PRx

TIMER 2

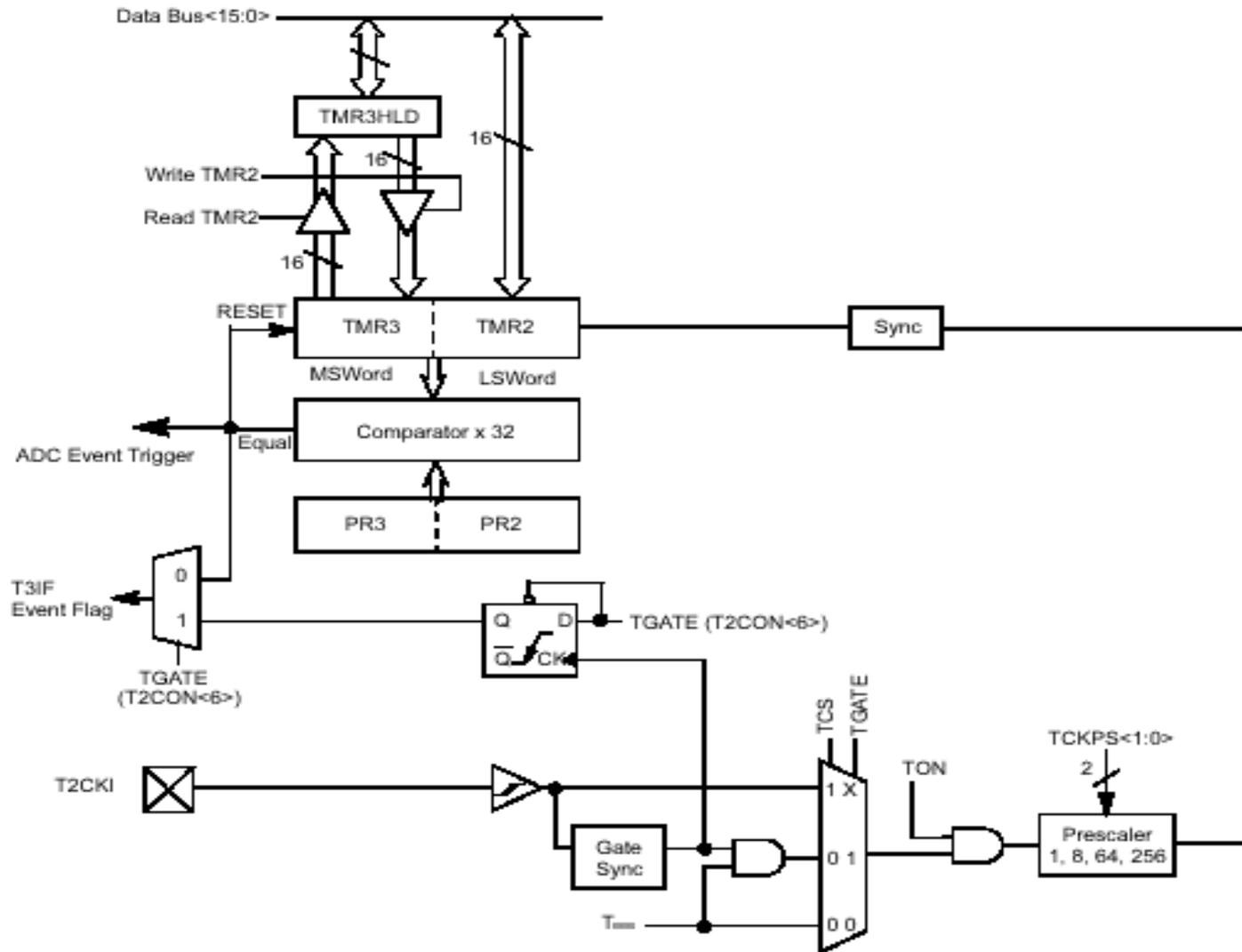


TIMER 3

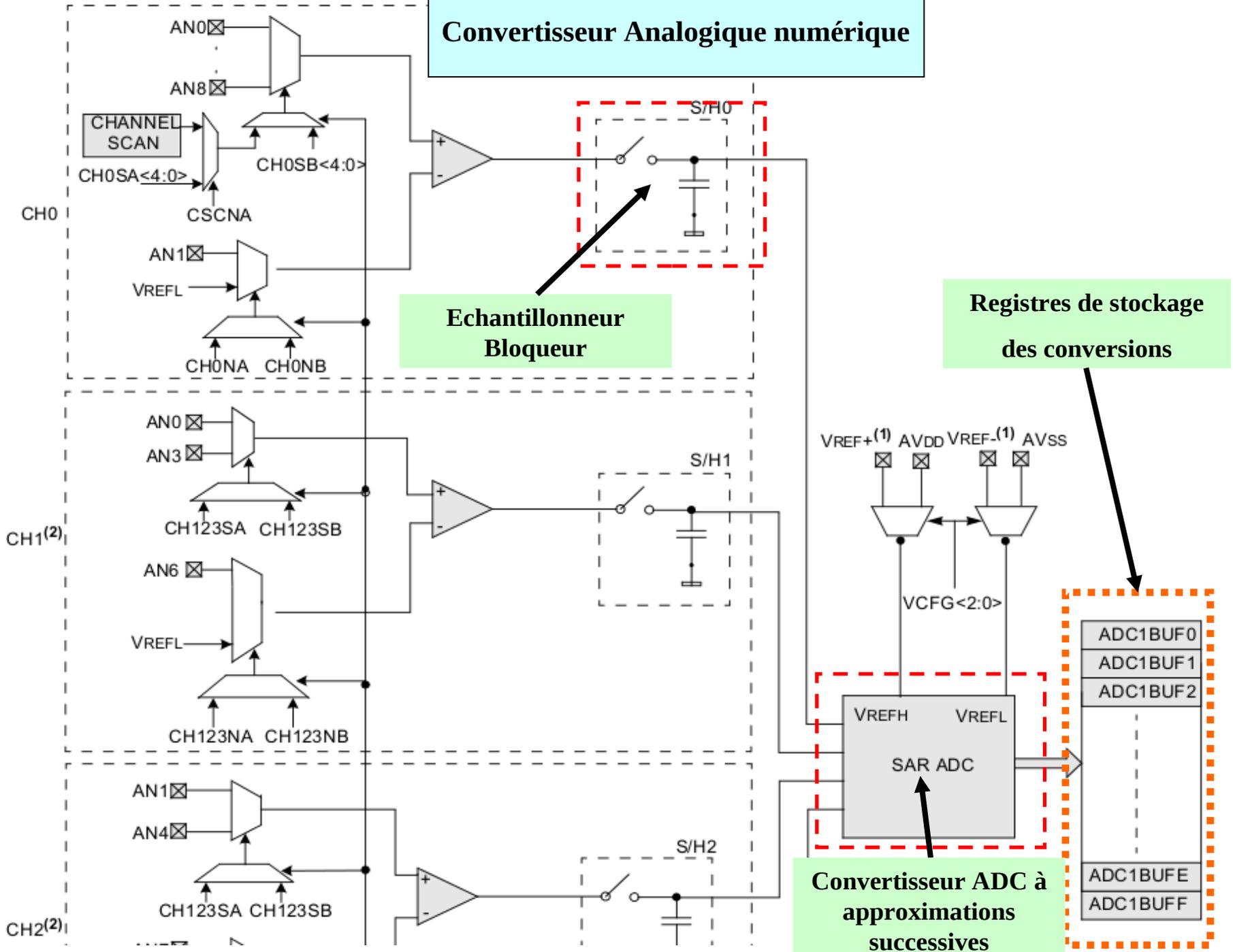


Note: In certain variants of the dsPIC30F family, the TxCK pin may not be available. Refer to the device data sheet for the I/O pin details. In such cases, the timer must use the system clock ($F_{\text{system}}/4$) as its input clock, unless it is configured for 32-bit operation.

TIMER 2+3 (32bits)



Convertisseur Analogique numérique



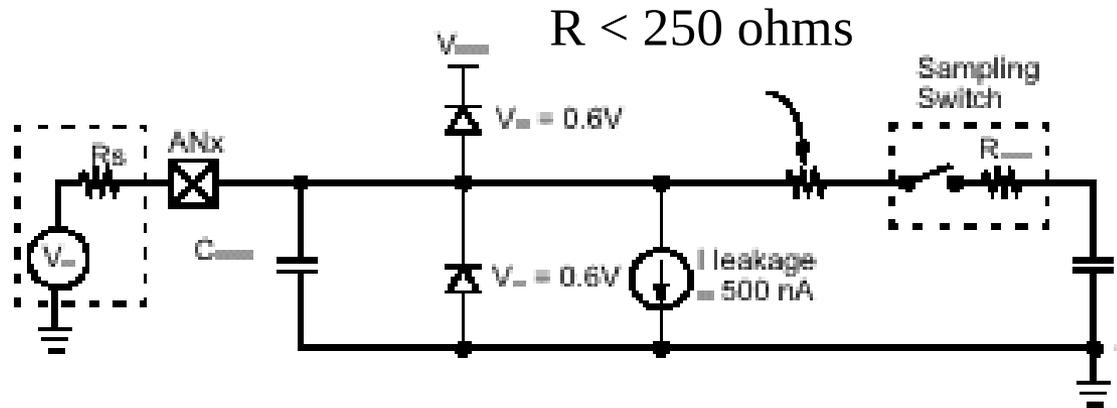
Echantillonneur Bloqueur

Registres de stockage des conversions

Convertisseur ADC à approximations successives

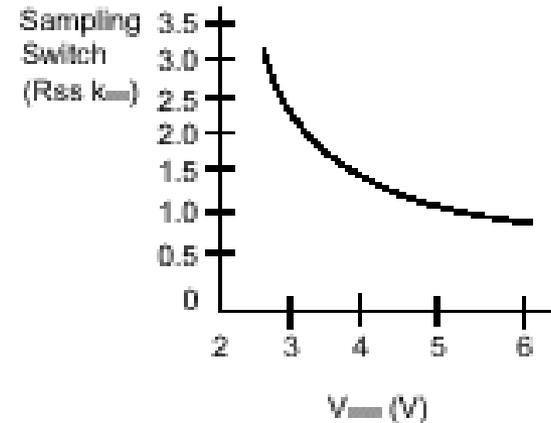
- ADC1BUF0
- ADC1BUF1
- ADC1BUF2
- ...
- ADC1BUFE
- ADC1BUFF

Caractéristiques électriques d'une entrée analogique



$C = 4.4 \text{ pF}$ (10bits)
 $C = 18 \text{ pF}$ (12bits)

- Legend:
- C_{in} = input capacitance (approximately 5 pF)
 - V_D = threshold voltage
 - $I_{leakage}$ = leakage current at the pin due to various junctions
 - R_{int} = interconnect resistance
 - SS = sampling switch
 - C_{SH} = sample/hold capacitance (from DAC)



Note: C_{in} value depends on device package and is not tested. Effect of C_{in} negligible if $R_s \leq 2.5 \text{ k}\Omega$.

Temps d'échantillonnage et de conversion

- Echantillonnage:

$$T_{\text{SMP}} = 0.5 \mu\text{s} + C_{\text{HOLD}} \cdot (R_{\text{IC}} + R_{\text{SS}} + R_{\text{S}}) \cdot \ln(2 \cdot 4095)$$

- Période d'horloge du convertisseur:

$$T_{\text{AD mini}} = 667\text{ns}$$

$$T_{\text{AD}} = T_{\text{CY}} * (0.5 * (\text{ADCS} \langle 5:0 \rangle + 1))$$

- Temps de conversion :

$$10 \text{ bits} = 12 * T_{\text{AD}}$$

$$12 \text{ bits} = 14 * T_{\text{AD}}$$

Calcul du résultats d'une conversion

- Tension d'entrée V_e $V_{ref-} < V_e < V_{ref+}$

- $V_{ref+} \text{ maxi} = 5V$

- $V_{ref-} \text{ mini} = 0$

- Résultat n sur 12 bits $0 < n < 2^{12} - 1$

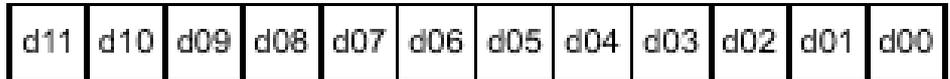
$$n = \text{round}(4095 * (V_e - V_{ref-}) / (V_{ref+} - V_{ref-}))$$

- Résolution : $1 \text{ lsb} = (V_{ref+} - V_{ref-}) / 4095$

- La Précision dépend de V_e

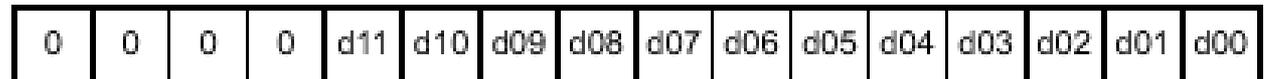
Formats du résultat de la conversion

RAM Contents:



Read to Bus:

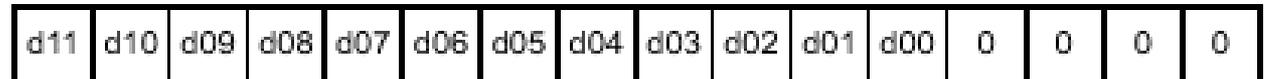
Integer



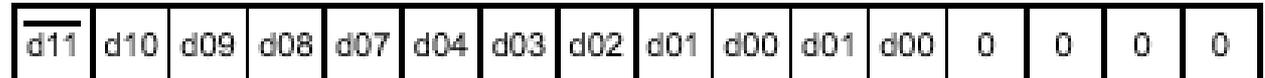
Signed Integer



Fractional



Signed Fractional (1.15)



Les registres du convertisseur Analogique Digital

- **AD1CON1** **Format du résultat, déclenchement,..**
- **AD1CON2** **Choix Vref+/-, mode de fonctionnement..**
- **AD1CON3** **Paramétrage de l'horloge de conversion**
- **AD1CHS0 ou 123** **Choix des entrées, références...**
- **AD1PCFGL** **Configuration des broches du circuit**
- **AD1CSSL** **Choix des entrées à balayer (scan)**
- **ADCBUF0ADCBUFF** **Registres résultats**

Utilisation du convertisseur Analogique Digital

INITIALISATION

- Configurer les entrées (analogiques ou numériques)
- Choisir les tensions de référence V_{ref+} et V_{ref-}
les entrées à convertir,
l'horloge de conversion,
le mode de déclenchement
- Activer le convertisseur
- . Configurer les interruptions si nécessaire

Traitement par POLLING:

- . Lancer la conversion (si pas automatique)
- . Attendre la fin de conversion
- . Lire le résultat

Traitement par routine d'interruption ADC

- . La routine est appelée automatiquement en fin de conversion
- . Lire le résultat et réinitialiser l'indicateur d'interruption
(La conversion peut être lancée automatiquement par le timer3 ou par le programme principal)

Les sources d'interruptions

- 8 non masquables de priorité fixe 8 à 15 (hard et soft)
- 118 masquables de priorité configurable (périphériques)
- Registres INTCON1 , INTCON2
- Registres pour autoriser/interdire IEC0, IEC1 et IEC2 (1 pour autoriser)
- Registres indicateurs IFS0, IFS1 et IFS2 (mis à 1 lors d'une demande d'int.)
(à remettre à 0 par soft)
- Registre pour gérer les priorités IPC0....IPC10 (0 faible ...7 élevée)
- Registres également concernés (INTREG, SR et CORCON)

Table 6-1: Trap Vector Details

Vector Number	IVT Address	AIVT Address	Trap Source
0	0x000004	0x000084	Reserved
1	0x000006	0x000086	Oscillator Failure
2	0x000008	0x000088	Address Error
3	0x00000A	0x00008A	Stack Error
4	0x00000C	0x00008C	Arithmetic Error
5	0x00000E	0x00008E	Reserved
6	0x000010	0x000090	Reserved
7	0x000012	0x000092	Reserved

Decreasing Natural Order Priority

Reset – GOTO Instruction	0x000000
Reset – GOTO Address	0x000002
Reserved	0x000004
Oscillator Fail Trap Vector	:
Address Error Trap Vector	:
Stack Error Trap Vector	:
Math Error Trap Vector	:
DMA Error Trap Vector	:
Reserved	:
Reserved	:
Interrupt Vector 0	0x000014
Interrupt Vector 1	:
~	:
~	:
~	:
Interrupt Vector 52	0x00007C
Interrupt Vector 53	0x00007E
Interrupt Vector 54	0x000080
~	:
~	:
~	:
Interrupt Vector 116	0x0000FC
Interrupt Vector 117	0x0000FE
Reserved	0x000100
Reserved	0x000102
Reserved	:
Oscillator Fail Trap Vector	:
Address Error Trap Vector	:
Stack Error Trap Vector	:
Math Error Trap Vector	:
DMA Error Trap Vector	:
Reserved	:
Reserved	:
Interrupt Vector 0	0x000114
Interrupt Vector 1	:

IVT

AVT

See [Table 32-1](#) for Interrupt Vector details.

Table 32-1: Interrupt Vectors

Vector Number	IVT Address	AIVT Address	Interrupt Source
0	0x000004	0x000104	Reserved
1	0x000006	0x000106	Oscillator Failure
2	0x000008	0x000108	Address Error
3	0x00000A	0x00010A	Stack Error
4	0x00000C	0x00010C	Math Error
5	0x00000E	0x00010E	Direct Memory Access (DMA) Error
6-7	0x000010-0x000012	0x000110-0x000112	Reserved
8	0x000014	0x000114	External Interrupt 0 (INT0)
9	0x000016	0x000116	Input Capture 1 (IC1)
10	0x000018	0x000118	Output Compare 1 (OC1)
11	0x00001A	0x00011A	Timer1 (T1)
12	0x00001C	0x00011C	DMA Channel 0 (DMA0)
13	0x00001E	0x00011E	Input Capture 2 (IC2)
14	0x000020	0x000120	Output Compare 2 (OC2)
15	0x000022	0x000122	Timer2 (T2)
16	0x000024	0x000124	Timer3 (T3)
17	0x000026	0x000126	SPI1 Error (SPI1E)
18	0x000028	0x000128	SPI1 Transfer Done (SPI1)
19	0x00002A	0x00012A	UART1 Receiver (U1RX)
20	0x00002C	0x00012C	UART1 Transmitter (U1TX)
21	0x00002E	0x00012E	ADC1 Convert Done (AD1)
22	0x000030	0x000130	DMA Channel 1 (DMA1)
23	0x000032	0x000132	Reserved
24	0x000034	0x000134	I2C1 Slave Events (SI2C1)
25	0x000036	0x000136	I2C1 Master Events (MI2C1)
26	0x000038	0x000138	Comparator Interrupt (CMP)

Exécution d'une routine de traitement d'interruption

- *L'adresse de retour et les indicateurs sont sauvegardés automatiquement sur la pile*
- Exécution du programme de traitement de l'interruption
 - Sauvegarde des registres utilisées sur la pile
 - Le traitement.....
 - Ré-initialisation de l'indicateur d'interruption concerné
 - Restauration des registres précédemment sauvegardés
 - RETFIE pour retourner au programme interrompu
- *Les indicateurs et l'adresse de retour sont dépilées automatiquement*

Gestion des priorités

- L 'instruction ***DISI #n*** interdit les interruptions pendant n cycles
- Une interruption peut-être interrompue par un autre interruption de priorité plus élevé (imbrication)
- On peut interdire l 'imbrication (*bit 31 de INTCON1*)
- On peut modifier les priorités (*0 à 7*) des int.
- Priorité 0 = int. interdite

Codage des nombres entiers en complément à 2

Soit X un nombre codé sur N bits en complément à

2



bit de signe :

$$b_{N-1} = 0 \text{ si } X > 0$$

$$b_{N-1} = 1 \text{ si } X < 0$$

$$X = -2^{N-1} \cdot b_{N-1} + \sum_{i=0}^{i=N-2} b_i \cdot 2^i$$

$$X_{\max} = 2^{N-1} - 1$$

$$X_{\min} = -2^{N-1}$$

Opération sur des nombres codés en complément à 2

- Addition de 2 nombres de N bits \Rightarrow Résultat sur N+1 bits au plus (overflow-carry)
- Produit de 2 nombres de N bits \Rightarrow Résultat sur 2N bits

Les 2 bits de poids forts du produit sont identiques (2 fois le bit de signe)

Le MSB est donc inutile.

Le résultat peut être décalé automatiquement d'un bit à gauche

Représentation des 'réels' en virgule fixe

Le terme « Q_k » indique k bits après la virgule

Si X est codé sur N bits en format Q_k .

Partie entière					Partie fractionnaire			
b_{N-1-K}	b_{N-2-K}	b_1	b_0	b_{-1}	b_{-2}	b_{-K}

$$X = -b_{N-1-K} \cdot 2^{N-1-K} + b_{N-2-K} \cdot 2^{N-2-K} + \dots + b_0 + b_{-1} \cdot 2^{-1} + \dots + b_{-K} \cdot 2^{-K}$$

- Erreur : $< 2^{-K}$
- Précision : $2^{-K} / X$ dépend de X (mauvaise si X est petit...)
- Valeurs limites : $X_{\max} = 2^{N-1-K} - 2^{-K}$
 $X_{\min} = -2^{N-1-K}$
 $|\Delta X_{\min}| = 2^{-K}$

Calculs en virgule fixe

Nombres codés dans le même format Qk

- On multiplie X par 2^k .
- Cela correspond à la représentation de l'entier $Y = \text{round}(X \cdot 2^k)$
- Les calculs se font alors exactement comme sur les entiers

Exemple de calcul en Q15

Soit la multiplication suivante: $a * b$

Avec: $a = 0.2$ Format Q15

$b = 0.5$ Format Q15

- On multiplie a et b par 2^{15}

$$a = 0.2 \times 2^{15} = 6553,6d \equiv 6554d = 199Ah$$

$$b = 0.5 \times 2^{15} = 16384d = 4000h$$

- On multiplie les entiers obtenus

$$199Ah \times 4000h = 666\ 8000h \text{ au format Q30}$$

- Résultat mis en Q15 par :

- Décalage à gauche de 1 (en Q31)

- et copie des 16 bits de poids forts

$$666\ 8000h = 0000\ 0110\ 0110\ 0110\ 1000\ 0000\ 0000\ 0000\ b$$

$$\text{Après décalage : } 0000\ 1100\ 1100\ 1101b = 0CCDh = 3277$$

- On divise par 2^{15} pour vérifier : $3277/2^{15} = 0.100006$

- On constate une erreur d'arrondi de $6 \cdot 10^{-6}$

Calculs en virgule fixe (suite)

Nombres codés dans des formats différents Q_{k1} et Q_{k2}

Addition: Il faut aligner les virgules en évitant les débordement
Donc convertir deux nombres dans le même format

Multiplication:

Le produit de 2 nombres sur N bits donne un résultat sur $2N$ bits
Les 2 bits de poids forts sont identiques (2 bits de signe..)
Seuls $2N-1$ bits sont nécessaires pour le résultat

Produit de 2 nombres en formats différents:

$$Q_{K1} \times Q_{K2} : \text{Résultat en } Q_{K1+K2}$$

Si on souhaite un résultat en format Q_{K3} , on doit effectuer:

- un décalage vers la droite de $K1+K2-K3$
- ou un décalage vers la gauche de $N - (K1+K2-K3)$