Etude comparative de diverses structures de filtres numériques : application aux signaux à très large bande et au corrélateur ALMA

Pascal Camino

▶ **To cite this version:**

Pascal Camino. Etude comparative de diverses structures de filtres numériques : application aux signaux à très large bande et au corrélateur ALMA. Micro et nanotechnologies/Microélectronique. Université de Bordeaux 1, 2008. Français. NNT : . tel-01237642

HAL Id: tel-01237642
https://hal.science/tel-01237642

Submitted on 3 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre : 3570

THÈSE

présentée à

L'UNIVERSITÉ DE BORDEAUX I

ÉCOLE DOCTORALE DE SCIENCES PHYSIQUES ET DE L'INGENIEUR

par Pascal CAMINO

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : ÉLECTRONIQUE

ETUDE COMPARATIVE DE DIVERSES STRUCTURES DE FILTRES NUMÉRIQUES. APPLICATION AUX SIGNAUX À TRÈS LARGE BANDE ET AU CORRÉLATEUR ALMA.

Soutenue le : 28 Mars 2008

Après avis des rapporteurs :

M.	Philippe BENABES	Professeur (Supélec)	Rapporteur
M.	Patrick GARDA	Professeur (Université P. et M. Curie, Paris VI)	Rapporteur

Devant la commission d'examen composée de :

M.	Jean-Marc HURE	Professeur (Université Bordeaux 1,LAB)	Président
M.	Alain BAUDRY	Astronome (LAB)	Directeur de thèse
M.	Dominique DALLET	Professeur (IMS, Université Bordeaux I)	Co-directeur de thèse
M.	Giovanni COMORETTO	Astronome associé (Osservatorio di Arcetri, Firenze)	Examinateur
M.	Philippe BENABES	Professeur (Supélec)	Examinateur
M.	Patrick GARDA	Professeur (Université P. M. Curie, Paris VI)	Examinateur

– 2008 –

Table des matières

Table des figures

Introduction Générale

1 Le projet ALMA

Le projet ALMA vise à construire un grand réseau interférométrique composé de 66 antennes mobiles sur un vaste plateau désertique afin d'observer l'Univers froid qui rayonne dans le domaine des ondes millimétriques et sub-millimétriques. Ce réseau d'antennes est situé dans le désert d'Atacama au Chili, à une altitude de 5000 m présentant les meilleures conditions de transparence atmosphérique pour le domaine d'observation. Les études ont été initiées au début des années 90 et le prototypage a démarré dès 1999. La détection du signal astronomique est réalisée par des systèmes d'électronique numerique qui intègrent divers algorithmes de traitement du signal. Ce système de détection est composé de 2 parties majeures : le banc de filtre et le corrélateur.
Actuellement, le site du Chili est prêt à accueillir le système de filtrage et le système corrélateur dans le bâtiment prévu à cet effet. Le système corrélateur a été finalisé par l'équipe du NRAO et est testé sur un site, dit d'intégration, à Charlottesville (Virginie, USA). Le système de filtrage (nommé Tunable Filter Bank ou TFB), quant à lui, a été développé par l'équipe européenne. Il est la pierre angulaire de la flexibilté du systéme de détection ALMA : associé au corrélateur, il offre de hautes résolutions spectrales dans divers modes d'observations et permet d'observer dans diverses fenêtres spectrales simultanément. Lui aussi est testé sur le site de Charlottesville, en association avec le corrélateur. Notre système de filtrage, du fait des signaux large bande à analyser, requiert des architectures massivement parallèles [1].
A la suite des tests réalisés avec le système de corrélation, il s'est avéré que l'architecture de filtrage développée en [2] engendrait une dissipation thermique relativement importante : 75 W par carte de filtrage comportant chacune 16 FPGAs ; 512 cartes sont utilisées dans le système Corrélateur[1]. Au vu de la configuration du site (la salle renfermant le système Corrélateur est situé à 5000 m) où sont stockés les systèmes de filtrage - corrélation , la dissipation thermique de l'ensemble est un élément primordial. Le but principal de cette étude a donc été l'optimisation en consommation du dit système de filtrage. En parallèle, nous avons recensé et envisagé d'implanter diverses structures de filtres numériques dans le cadre de signaux d'entrée à large bande. Il faut aussi souligner le fait que si aujourd'hui le filtrage numérique est largement utilisé dans de nombreux domaines (télécommunications, applications multimédia...), les projets actuels de radioastronomie – utilisant des systèmes numériques – constituent une nouvelle génération d'instrument. Pour cette raison, des études concernant la sensibilité des systèmes numériques et l'efficacité de calcul ont été menées.

[1] le terme Corrélateur avec un 'C' est utilisé pour désigner l'ensemble du système de détection filtrage numérique + corrélation

2 Organisation du document

Le Chapitre 1 présente le projet ALMA dans sa globalité. La radioastronomie y est dans un premier temps introduite brièvement avec un aperçu des principes élémentaires et avantages de l'interférométrie. La collaboration internationale du projet est ensuite présentée et les caractéristiques principales du systèmes sont données. Chaque « sous-ensemble » de l'instrument est abordé avec un développement plus important du système de détection. Le système de filtrage est alors décrit en détail, aboutissant aux problèmes de dissipation thermique rencontrés. Ce chapitre introduit le travail effectué durant la thèse : la finalisation du système de filtrage ainsi que la mise en place des tests permettant la validation de la fonctionnalité qui ont soulignés les problèmes de consommation de la carte de filtrage.
Dans l'optique de la diminution de la consommation du Corrélateur, différentes options de filtrage numérique ont été investiguées. Les 3 chapitres qui suivent présentent les solutions envisagées, cœur du travail de la thèse.

L'objectif du Chapitre 2 est de proposer une optimisation du premier étage du filtre retenu initialement pour ALMA, filtre à décimation de facteur 32. Il aborde un type de filtre potentiellement intéressant pour notre application, en vue du remplacement de cet étage pour diminuer sa consommation : un filtre pouvant fournir un facteur de décimation important et possédant une architecture électronique simple. Le probleme majeur est la très large bande caractérisant le signal à traiter. L'architecture de ce filtre est adaptée au format de ce signal et différentes réalisations sont comparées entre elles.

Le Chapitre 3 s'attache à l'optimisation du second étage du filtre TFB. Il fait une synthèse des structures de filtre RII à faible encombrement et présentant la caractéristique de linéarité de phase (caractéristique imposée par le projet pour le système de filtrage). Des méthodes de linéarisation de la phase basées sur une modification de l'architecture électronique ainsi que sur des algorithmes de réduction de modéles sont exposées. Ces méthodes sont alors employées et comparées afin de déterminer leur possible utilisation dans le système TFB.

Le Chapitre 4 présente une architecture de filtrage alternative à l'architecture TFB complète actuellement utilisée : un banc de filtre polyphase. Ce système est présenté comme étant moins volumineux mais moins flexible que le TFB. Une étude de 2 structures basées sur ce principe est réalisée afin de conclure sur le rapport complexité-flexibilité des structures polyphases et TFB.

Enfin, le Chapitre 5 propose un récapitulatif des structures les plus interessantes dans l'objectif du remplacement des différents étages du système de filtrage TFB ou du système lui-même. La solution retenue y est exposée en détails : la validation du traitement apporté par cette nouvelle structure ainsi que les performances obtenues y apparaissent.

Une introduction des notions de traitement numérique du signal essentielles à la compréhension du document est reportée en annexe.

Chapitre 1

L'Instrument ALMA

1 Introduction

Le projet ALMA est un projet de radioastronomie d'envergure internationale. Le système développé par le laboratoire d'électronique de l'observatoire de Bordeaux est une partie du Corrélateur ALMA, à savoir le système de filtrage. Dans ce chapitre va donc être exposé tout particulièrement le système Corrélateur composé du système de filtrage et de la partie corrélation des données.
La radioastronomie ainsi que les principes de l'interférométrie sont tout d'abord introduits afin de comprendre les spécifications nécessaires au système Corrélateur. L'historique du projet est ensuite brièvement retracé. Le traitement du signal tout au long des différents blocs constituant l'instrument est alors présenté, la structure électronique du corrélateur et du système de filtrage étant bien sûr plus développée. La dernière section est dédiée à l'étude de la dissipation thermique engendrée par la carte de filtrage.

2 La radioastronomie

Le but de ce chapitre est de se familiariser avec certaines notions, méthodes employées en radioastronomie et d'en comprendre les objectifs. Il n'aborde pas en détails cette science qu'est la radioastronomie [3].

2.1 Aspects scientifiques

Notre connaissance de l'univers provient en majeur partie de l'étude des ondes électromagnétiques (l'électromagnétisme est l'un des fondements de la radioastronomie, ces principes sont établis à la fin du XIX^{eme} siècle) émises ou absorbées dans le cosmos et captés à la surface de la Terre ou par satellite dans son environnement immédiat. L'Astronomie a d'abord consisté en une observation des astres à l'œil nu, puis les inventions de la lunette et du télescope ont participé à l'essor de l'astronomie optique moderne. Plus récemment les progrès de la technique ont permis l'étude des ondes électromagnétiques dans d'autres gammes de fréquences que celle du visible. Ainsi la seconde moitié du XX^{eme} siècle a vu le formidable développement de la Radioastronomie, science et technique associées à l'étude des ondes du domaine radio.
On peut considérer que l'atmosphère terrestre présente deux principales fenêtres de transparence aux ondes électromagnétiques (Figure 1.1) : la fenêtre du visible (domaine optique) et celle de la

radio (quelques fenêtres secondaires sont tout de même accessibles dans le domaine infra-rouge).

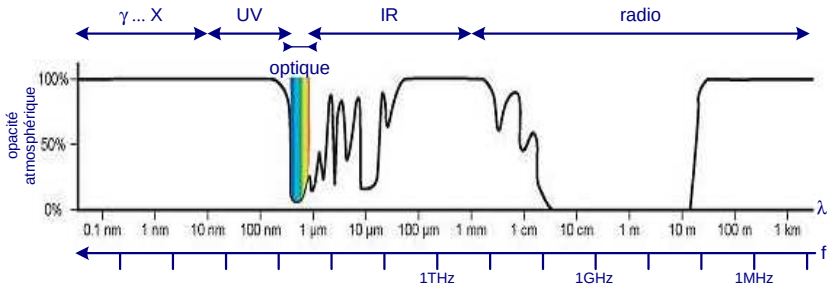


FIG. 1.1 – *Spectre électromagnétique et opacité atmosphérique*

De part et d'autre de la fenêtre radio, les phénomènes physiques expliquant l'opacité de l'atmosphère et donc l'incapacité à collecter des ondes à la surface de la Terre sont distincts. Pour des longueurs d'ondes supérieures à quelques dizaines de mètres, l'ionosphère (couche ionisée de la haute altitude de l'atmosphère) devient conductrice et réfléchit les ondes électromagnétiques. A des longueurs d'ondes inférieures au centimètre, les ondes électromagnétiques sont absorbées dans la basse atmosphère par les molécules d'eau. Ceci explique le fait que les radiotelescopes soient construits sur des sites en altitude où l'atmosphère est relativement dépourvue de vapeur d'eau, permettant ainsi de repousser les limites d'observations jusqu'aux longueurs d'ondes millimétriques ou sub-millimétriques. Pour explorer de manière continue le domaine sub-millimétrique et les autres domaines du spectre électromagnétique, l'observation ne peut se faire qu'au delà de l'atmosphère terrestre, à partir de satellites.
Les diverses radiosources, tel que les rayonnements des planètes, des étoiles, des nébuleuses gazeuses, des nuages de gaz atomique ou moléculaire, des quazars ou encore des galaxies, peuvent émettrent un spectre continu (continuum) ou un spectre de raies. Les phonomènes physiques relatifs à ces deux types d'émission – ainsi que les méthodes d'étude associées – sont fort différents et sont présentés par la suite. Le point commun de tous ces rayonnements réside dans leur nature aléatoire. En effet, pour des intervalles de temps et de fréquence suffisamment restreints, le signal reçu est un bruit blanc gaussien. Les caractéristiques de ces signaux sont développées dans l'Annexe intitulée « Notions importantes de Traitement Numérique du Signal ». Le caractère aléatoire de ces sources limite la connaisance que l'on peut déduire de l'analyse d'un signal collecté, pour une bande de fréquence et une durée d'observation nécessairement limitées. La précision d'analyse est déterminée par la loi des radioastronomes :

$$\sigma \propto \frac{1}{\sqrt{B\tau}} \tag{1.1}$$

où B est la largeur de bande observée et τ le temps d'intégration.
La précision avec laquelle est connu le signal est donc inversement proportionnelle à la racine carré de la bande de fréquence analysée et du temps d'observation de ce signal. Les temps d'observation requis peuvent atteindre des heures voire plusieurs jours pour obtenir un rapport signal sur bruit suffisant, ce qui exige des méthodes d'observations appropriées pour combattre les fluctuations rapides de gain des systèmes électroniques et les variations rapides de l'atmosphère.

2.1.1 La radiométrie

La radiométrie pourrait être définie comme étant un pan de la radioastronomie s'attachant à mesurer la puissance et la direction du rayonnement d'une radiosource. Cette puissance est appelée densité de flux. C'est l'équivalent d'un éclairement en photométrie soit le produit de la brillance de la source observée par l'angle solide sous lequel est vue cette source. Pour des émissions continuum le flux est constant dans une gamme de fréquence au voisinage de la fréquence à laquelle la mesure de flux est effectuée.
Pour l'électronique d'un radiotélescope, l'antenne adaptée qui collecte le rayonnement est équivalente à une impédance (pour laquelle la partie résistive domine) portée à la température T. Cette impédance présente à ses bornes une tension de bruit dont la puissance P, délivrée dans une bande de fréquence B, est donnée par la formulation thermodynamique :

$$P = kTB \tag{1.2}$$

où k est la constante de Boltzmann.
La mesure du flux sur une large gamme de longueur d'onde permet de caractériser l'indice spectral et ainsi de déterminer si l'émission d'une radiosource est de nature thermique (rayonnement d'un corps noir ou rayonnement des électrons en équilibre thermique dans un gaz ionisé) ou non (rayonnement gyromagnétique ou rayonnement synchrotron).

2.1.2 La spectroscopie

La spectroscopie consiste à analyser le contenu spectral du rayonnement des radiosources atomiques et moléculaires. Atomes et molécules n'émettent ou n'absorbent le rayonnement qu'à des longueurs d'onde bien déterminées. Chaque substance chimique produit un spectre de raies qui lui est propre et dont la « signature » apparaît dans les domaines optique ou radio pour les substances portées à des températures basses. Ainsi l'analyse spectrale du rayonnement permet d'identifier les constituants du gaz interstellaire.
La physique quantique indique qu'il n'existe pour un atome ou une molécule qu'un ensemble discret de niveaux d'énergie possibles. Lorsque spontanément ou sous l'action d'un rayonnement extérieur, une particule passe d'un état quantique d'énergie E à un autre état quantique d'énergie $E + \Delta E$, un rayonnement de fréquence ν est émis ($\Delta E < 0$) ou absorbé ($\Delta E > 0$). Cette transition quantique qui correspond à un réarrangement sur différentes orbites des électrons gravitant autour du noyau, ou à une modification des mouvements de rotation et de vibration de la molécule, est décrite par $\Delta E = h\nu$.
L'effet Doppler permet d'accéder à des informations concernant la vitesse de déplacement des sources de raies. En effet, si une source de rayonnement de fréquence ν se déplace à la vitesse relative Δv (vitesse radiale) par rapport à un observateur, le spectre observé fait apparaître une raie à la fréquence $\nu + \Delta\nu$ où :

$$\Delta v = -c\frac{\Delta\nu}{\nu} \tag{1.3}$$

$\Delta\nu$ étant positif pour une vitesse d'éloignement, c étant la célérité.
L'étude spectroscopique du rayonnement produit par un nuage de gaz permet donc de déterminer la vitesse de déplacement, la constitution chimique, la température, la distribution de population et par suite, à l'aide de modèles, les conditions physiques de ce nuage.

2.2 Aspects techniques

Les techniques mises en œuvre en radioastronomie sont similaires à celles utilisées en radiocommunications, la principale différence étant que la radioastronomie est une technique passive où l'on capte et analyse le rayonnement de sources naturelles.
Un radiotélescope est principalement composée de 3 parties : l'antenne, le récepteur et le détecteur. Un exemple concret de récepteur et de détecteur est donné en section 4.
L'antenne est l'élément qui transforme les champs électriques et magnétiques induits par le rayonnement en grandeurs électriques pouvant être traitées par l'électronique des radiotélescopes. Il faut noter que la précision des surfaces (maillage constituant la surface de l'antenne) qui servent à capter le rayonnement des radiosources est proportionnelle à la gamme de longueurs d'ondes à laquelle travaille l'instrument. Soulignons également la nécessité que les champs électriques et magnétiques collectés en différents points de l'antenne ne présentent pas de différence de phase, c'est la notion de cohérence. L'objectif est d'estimer la puissance totale du rayonnement capté en intégrant l'ensemble des puissances ponctuelles d'un même plan d'onde. Les antennes collectent après réflexion les puissances ponctuelles du rayonnement au foyer de la parabole, ou au foyer secondaire d'un ensemble parabole-hyperbole, permettant de placer le récepteur à l'arrière du collecteur principal.
La notion de sensibilité est directement liée à la surface de captation et à la sensibilité propre du récepteur placé derrière l'antenne. Une antenne est d'autant plus sensible que sa surface (A) est importante. Deux cas d'observation doivent être distingués pour le calcul de la puissance collectée et de la température d'antenne équivalente : l'observation d'une source ponctuelle où la totalité du flux émis est collecté par l'antenne et l'observation d'une source étendue (Figure 1.2).

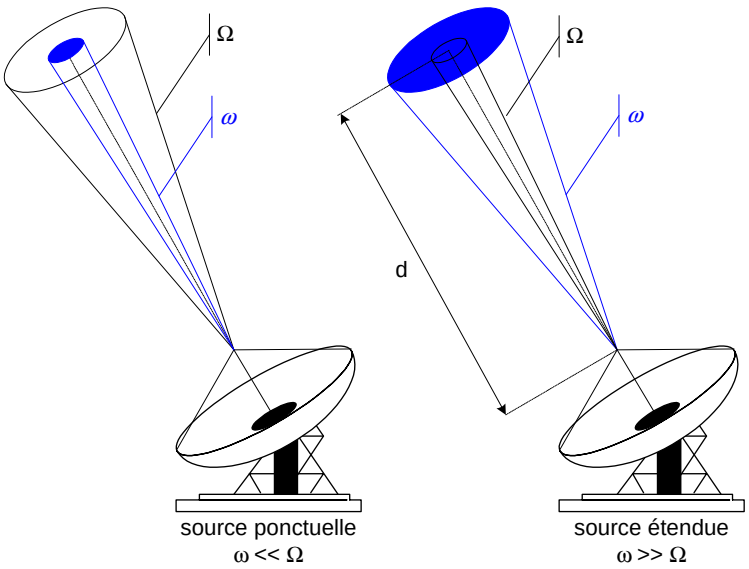


FIG. 1.2 – *Observation d'une source*

L'antenne est caractérisée par :
- une surface $A = \eta A_{physique}$, où $A_{physique}$ est la surface totale et η le rendement en surface de l'antenne
- un lobe principal d'angle solide $\Omega = \lambda^2/A$
- la source est vu sous l'angle solide ω

L'autre caractéristique essentielle d'une antenne est le pouvoir séparateur ϕ_{min}. Il s'agit du plus petit angle $\Delta\phi$ séparant deux sources pour lequel les sources peuvent être dissociées par l'antenne. Pour une antenne de surface A, le pouvoir séparateur vaut approximativement :

$$\phi_{min} = \frac{\lambda}{A} \tag{1.4}$$

avec λ la longueur d'onde du rayonnement.
L'optimisation des deux principales caractéristiques d'une antenne (sensibilité et pouvoir séparateur) nécessite donc une augmentation de la surface collectrice. Cette recherche de qualité pose des problèmes technologiques très complexes pour la fabrication des antennes du point de vue : de la tolérance sur la surface réflective (de l'ordre de $\lambda/10$), assurant la cohérence des rayonnements collectés, des montures associées (mobilité et précision du pointage impliquant un suivi des sources lors de l'intégration du signal à environ 1/15 du lobe principal).

2.3 Interférométrie, réseau d'antennes

Un radiotélescope à antenne unique peut, par définition, fournir l'image d'une région du ciel par des observations successives dans différentes directions de visée. Les limites d'un tel instrument, en terme de sensibilite et surtout de pouvoir séparateur, ont été exposés dans la section précédente. Le principal intérêt des techniques interférométriques [3] réside dans la possibilité de concevoir des instruments atteignant des pouvoirs séparateurs nettement supérieur à ceux d'un radiotélescope à antenne unique. Dans cette section sont exposés brièvement les principes de base de l'interférométrie.
En Figure 1.3 est schématisé un interféromètre à 2 antennes observant une source monochromatique de longueur d'onde λ dans un espace unidimensionnel.

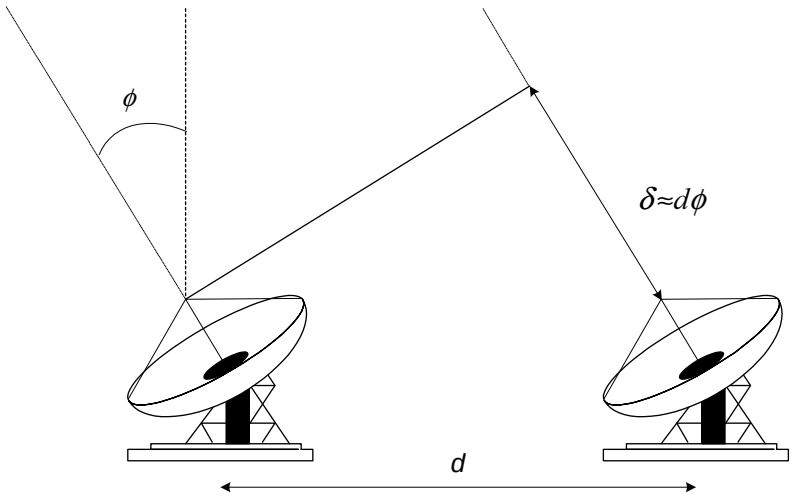


FIG. 1.3 – *Interféromètre à 2 antennes*

Du fait de la différence de marche $\delta = d\phi$, la détection en puissance totale sur la somme des signaux collectés par chacune des antennes, comme la détection par corrélation entre ces signaux, produit une fonction sinusoïdale en ϕ de période spatiale λ/d. Les motifs produits par suite du déplacement de la source sont appelés franges d'interférence, l'amplitude et la phase de ces franges déterminent complètement l'harmonique spatiale $u = d/\lambda$ de t_A (transformée de Fourier

de la température d'antenne T_A). Il faut noter que ces franges ne se forment que si les dimensions angulaires de l'objet observé sont petites devant λ/d.
Concernant les performances d'un tel instrument, le pouvoir séparateur est désormais lié à la distance entre les antennes et non plus aux dimensions physiques des antennes individuelles. Quant à la sensibilité, elle reste par contre limitée par les surfaces élémentaires d'antennes. Si le grand avantage d'un interféromètre est son pouvoir séparateur, sa limitation vient de la nécessité d'effectuer un grand nombre d'observations pour couvrir l'espace des fréquences spatiales du mieux possible. En effet, la caractérisation des différentes harmoniques spatiales permettant de construire la distribution de température résulte d'observation pour différents espacements d'antennes. Cette dernière caractéristique interdit l'étude de sources dont les variations sont rapides. Pour pallier ce défaut, on utilise des réseaux d'antennes mobiles, ce qui permet d'accéder simultanément à plusieurs harmoniques spatiales et de couvrir le plan des fréquences spatiales (u, v) de manière optimale avec un minimum de configurations géométriques du réseau. La rotation de la Terre qui induit des modifications apparentes de la configuration du réseau observant la source est également prise en compte dans cette recherche d'optimisation. Un réseau de n antennes est équivalent à $n(n-1)/2$ interféromètres indépendants.

3 Historique du projet, Présentation générale

ALMA – pour Atacama Large Militer Array – est un projet international visant à la mise en place, dans le désert d'Atacama (Chili), d'un réseau interférométrique aux performances révolutionnaires. Cet instrument de radioastronomie permettra de construire l'image de régions du ciel rayonnant dans les domaines millimétrique et sub-millimétrique avec une résolution spatiale inférieure à la seconde d'arc. Cette dernière pourra atteindre quelques millisecondes d'arc pour les observations réalisées aux plus hautes fréquences (approchant le THz) avec les espacements d'antennes maximaux (environ 10 km). Les 64 antennes offriront une très grande sensibilité permettant l'étude de galaxies lointaines, observées au cours de leur processus de formation, dans l'état qui était le leur au début de l'Univers. De plus, du fait de la grande capacité de reconfigation du réseau d'antennes, ALMA possèdera une grande capacité à imager avec une excellente fidélité et à différentes échelles spatiales pour de vastes régions du ciel.
Les principales caratéristiques techniques d'ALMA [2] sont présenté dans le tableau 1.1.
Pour atteindre les capacités scientifiques présentées, ALMA requiert un site d'observation aux conditions optimales dans les bandes d'observations, des développements techniques innovants et une organisation de projet sans précédent dans la communauté astronomique. En effet, la collaboration autour de ce projet est véritablement internationale : les partenaires d'Amérique du Nord sont les Etats-Unis (National Science Fondation) et le Canada (National Research Council) ; en Europe, aux membres actuels de l'ESO (European Southern Observatory : Allemagne, Belgique, Danemark, France, Ialie, Pays-Bas, Potugal, Royaume-Uni, Suède et Suisse) s'ajoute l'Espagne ; le Japon participe aussi au projet en amenant un réseau de 16 antennes qui se fondront dans le parc déjà présent ; et bien sur le Chili, en temps que pays hôte et à travers le conseil scientifique d'ALMA (Alma Scientific Advisory Comittee), participe au projet.
Afin de produire les images millimétriques et sub-millimétriques les plus précises, les radiotélescope requièrent un site sec. En effet, c'est la vapeur d'eau présente dans l'atmosphère (cf. section 2.1) qui absorbe les ondes millimétriques et sub-millimétriques dégradant ainsi la sensibilité de l'instrument et limitant les fenêtres de transparence à certains domaines de longueur d'onde. Les scientifiques ont choisi le site du désert d'Atacama au Chili (cf. Figure 1.4) car il a été démontré, par l'intermédiaire de mesures détaillées et prolongées, que le ciel au-dessus de

TAB. 1.1 – *Principales caractéristiques techniques d'ALMA*

Réseau	
Nombre d'antennes (N)	64[1]
Surface collectrice totale $\pi/4 \times ND^2$	7238 m^2
Longueur collectrice totale (ND)	768 m
Résolution angulaire	$0''2 \cdot \lambda$(mm)/ligne de base (km)
Configuration du réseau	
Compacte	150 m
Zoom continu	$200 - 5000$ m
Résolution maximale	14 km
Nombre de stations d'antennes	216
Antennes	
Diamètre	12 m
Précision de surface	25 mm RMS
Pointage	$0''6$ RSS (pour un vent de 0.9 ms^{-1})
Système de réception	
4 bandes prioritaires parmi 10 prévues à terme	
Bande 3	$84 - 116$ GHz
Bande 6	$211 - 275$ GHz
Bande 7	$275 - 370$ GHz
Bande 9	$602 - 720$ GHz
Radiomètre de vapeur d'eau	183 GHz
Fréquence intermédiaire (IF)	
Largeur de bande	8 GHz×2 polarisations
Numérisation du signal	4 GS.s^{-1} sur 3 bits - 8 niveaux
Transmission depuis les antennes	Numérique, via fibre optique
Corrélateur	
Nombre de lignes de base corrélée	2016 ($N(N-1)/2$)
Largeur de bande	16 GHz par antenne
Nombre de canaux spectraux	4096 par IF
Rythme des données	
Transmission des données	120 Gb.s^{-1}
Traitement des données par le corrélateur	1.6×10^{16} MAC.s^{-1}
	MAC : Multiplication - Accumulation

[1] 50 antennes dans un premier temps (et 14 antennes en option) auxquelles s'ajoutent les 16 antennes du réseau Japonnais

ce site présente des conditions uniques de transparence et de stabilité, propriétés essentielles au fonctionnement optimal du système de détection du projet.

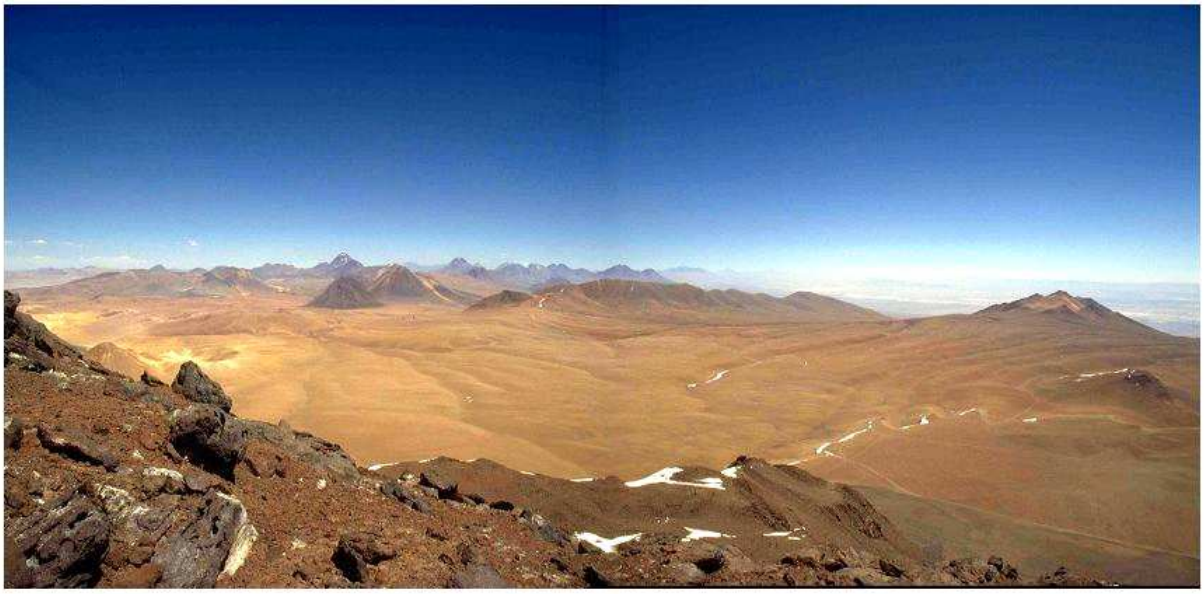


FIG. 1.4 – *Site ALMA dans le désert d'Atacama au Chili*

Le désert d'Atacama est situé à 5000 m d'altitude et sera seulement peuplé du réseau d'antennes (Array Operation Site - AOS). Les batiments techniques (Operations Support Facilities - OSF), où sont présents les ressources de calcul, les laboratoires techniques et le personnel concerné par les observations astronomiques, sont situés à 3000 m d'altitude pour des raisons de commodités (condition physique du personnel). L'OSF est distant de 30 km de l'AOS.

ALMA constituera donc pour les astronomes un instrument puissant et souple qu'ils pourront utiliser pour étudier, avec une haute précision angulaire, les émissions en onde millimétrique et sub-millimétrique d'une grande variété de sources astronomiques. En effet, la richesse du ciel dans les domaines millimétrique et sub-millimétrique est principalement liée à l'émission thermique de poussières célestes et de corps solides, ceux-la même rayonnant fortement en infrarouge, et à l'émission discrète du gaz moléculaire ou atomique froid situé dans les nuages interstellaires ou à la périphérie de certaines étoiles. Actuellement, dans cette fenêtre, l'observation se fait par l'intermédiaire de télescopes orbitaux. ALMA offrira des résolutions spectrales équivalentes mais des résolutions spatiales très supérieures à la prochaine génération de télescopes spatiaux. L'interféromètre devra aussi se caractériser par une longue durée de vie (au moins 30 ans) et une capacité d'adaptation et d'évolution afin de répondre aux besoins des astronomes au cours des prochaines décennies. Certains critères ont été établis afin de définir les besoins et caractéristiques des différents sous-systèmes afin d'atteindre les caractéristiques scientifiques d'ALMA. Ce point est développé en [2] et peut être résumé sous la forme de 3 critères de choix :

- Le premier de ces critères concerne la sensibilité des radiotélescopes. Cette dernière est liée à trois caractéristiques principales : la transparence atmosphérique, la performance en bruit des détecteurs et la surface collectrice totale. Pour satisfaire ce critère ALMA doit être 20 à 30 fois plus sensible que les instruments actuels de même nature.

– Le deuxième concerne le gain en résolution spectrale et spatiale. L'amélioration des résolutions fréquentielles (de l'ordre de 5 à 10 kHz) et l'augmentation des dimensions physiques du réseau (jusqu'à 14 kilomètres) permettent d'atteindre les spécifications.

– Le troisième critère concerne la capacité à délivrer des images à haute fidélité. Ceci requiert un nombre élévé de lignes de base afin de couvrir suffisamment le plan des fréquences spatiales (cf. 1.1). Les antennes doivent posséder une haute qualité de surface avec une grande précision de pointage. La précision de la calibration des phases interférométriques est aussi très importante ; elle est améliorée, particulièrement aux très hautes fréquences, grâce au radiomètre à vapeur d'eau qui permet de corriger la phase des fluctuations liées à la vapeur d'eau.

4 Flot du traitement du signal

En Figure 1.5 est présenté un schéma simplifié du flot de données dans le système ALMA. Ce dernier est composé de différentes structures dont les fonctionnalités sont définies dans les sections suivantes.

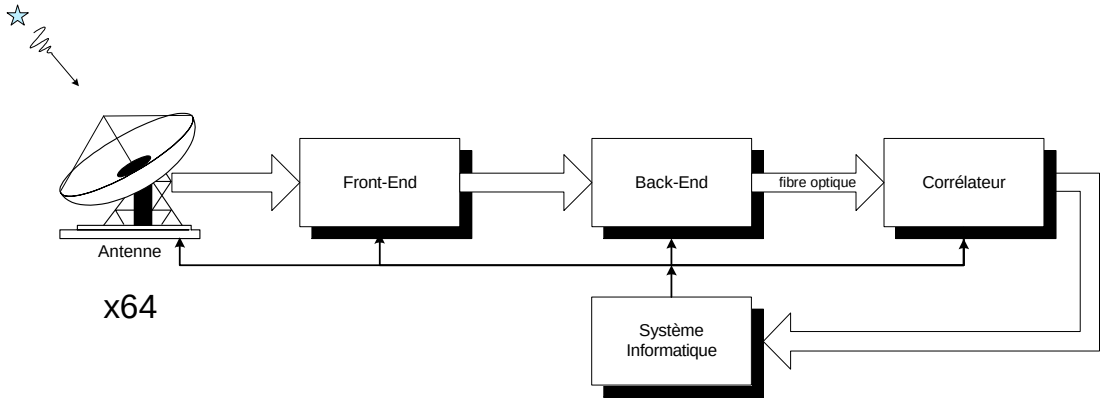


FIG. 1.5 – *Schéma simplifié de la chaîne de détection du système ALMA*

4.1 Le réseau d'antennes

Les 64 antennes du réseau (50 antennes sont prévus pour l'instant) possèdent toutes une surface collectrice paraboloïde de 12 m de diamètre. Les différents récepteurs du Front-End (cf. section 4.2) sont placés à des positions fixes dans le plan focal, la sélection de l'un de ces récepteurs s'effectuant par un ajustement du pointage du réflecteur principal. Chaque antenne est orientable ; 85 % de la sphère céleste est observable depuis le site d'Atacama. Les antennes sont déplaçables et peuvent être positionnées sur 216 stations. La possibilité de déplacer les antennes permet à ALMA d'atteindre les objectifs de résolution angulaire spécifiés par les scientifiques. Les configurations de réseau aussi compactes que quelques centaines de mètres permettent d'étudier des objets étendus ou peu brillants, tandis que les configurations de plusieurs kilomètres (pouvant atteindre 14 km) permettent l'étude d'objet compacts et très brillants.

4.2 Le Front-End ou récepteurs

Chaque antenne est équipée d'un système de réception hétérodyne, ou Front-End, qui effectue un abaissement de la fréquence des signaux reçus avant traitement de ces derniers. Les diverses bandes de fréquence sont données dans le tableau 1.1. Le Front-End forme un ensemble de récepteurs cohérents puisque chaque récepteur (10 par antenne dans la version finale d'ALMA) utilise un signal issu d'un même Oscillateur Local (LO) pour toutes les antennes de façon à translater le signal astronomique incident en une fréquence intermédiaire plus basse (rôle de l'hétérodyne) que l'on sait amplifier. L'oscillateur local de référence est différent pour chaque bande de réception et doit être adapté à l'intérieur d'une bande donnée à l'observation proposée.

Les composants du Front-End spécifiques à l'une des bandes de fréquence forment des assemblages dissociés (des cartridges), chacun d'entre eux nécessite un refroidissement cryogénique afin de diminuer le bruit de réception. Ces cartridges sont placés au coeur de l'antenne, au plus près du plan focal, dans une chambre sous vide. Chaque cartridge est équipé de 2 récepteurs travaillant dans des sens orthogonaux de polarisation linéaire. Le signal est alors séparé en deux polarisations linéaires (H et V, les techniques de séparation diffèrent selon les bandes considérés [2]) et propagé par guide d'onde jusqu'au système mélangeur suivi d'un amplificateur. Le signal qui en résulte est un signal de bande moyenne fréquence (Intermediate Frequency) $4-12$ GHz (8 GHz sont transmis pour chaque polarisation). A noter que chaque antenne est équipée d'un radiomètre à vapeur d'eau (raie d'émission à 183 GHz) permetant de mesurer la colonne de vapeur d'eau au-dessus de l'antenne. Ainsi, il est possible de corriger les effets parasites dus à la vapeur d'eau sur le signal astronomique et particulièrement sur la fluctuation de la phase en interférometrie.

4.3 Le Back-End ou système de transmission de données

Pour chaque antenne, les signaux provenant du Front-End sont transmis à la partie analogique du Back-End par des câbles coaxiaux situé sur l'un des cotés de la cabine de l'antenne. Le signal IF est alors mélangé avec un LO $8-14$ GHz pour être translaté en fréquence dans la bande $2-4$ GHz appelée Bande de Base (BB). Le système réalisant cette opération est appelé Down Converter (DC). Chaque DC possède un LO indépendant et réglable, controlé en fréquence et en phase. La bande de 8 GHz est alors découpée en 4 bandes de 2 GHz (toujours sous deux polarisations). De plus amples détails sur la conversion des bandes sont donnés en [2].

Les huit BBs associées à chaque antenne sont alors transmises par câble coaxial au système numérique du Back-End. Dès lors dans le flot de traitement du signal, les bandes sont organisées par paires de polarisation et sont traités identiquement. Chaque BB ($2-4$ GHz) est numérisée sur 3 bits - 8 niveaux à 4 GS.s^{-1}. Un ASIC (Application Specific Integrated Circuit) a dû être développé puisqu'au début du projet aucun circuit commercial ne pouvait réaliser la numérisation d'un tel signal. L'architecture retenue pour le CAN, dite Flash, est représentée en Figure 1.6 [4].

Le flot de données (3 bits à 4 GS.s^{-1}) généré par le CAN est immédiatement démultiplexé par 16, cela permet de former un flot de données équivalent de 48 voies parallèle à 250 Mb.s^{-1} (technonologie à logique LVDS différentielle retenue). Ce format peut être alors pris en charge par le système de transmission numérique des données. C'est un câble de 12 fibres optiques qui assure la transmission (à une vitesse de 10 Gb.s^{-1}) jusqu'au bâtiment technique central où les signaux issus de toutes les antennes du réseau sont corrélés.

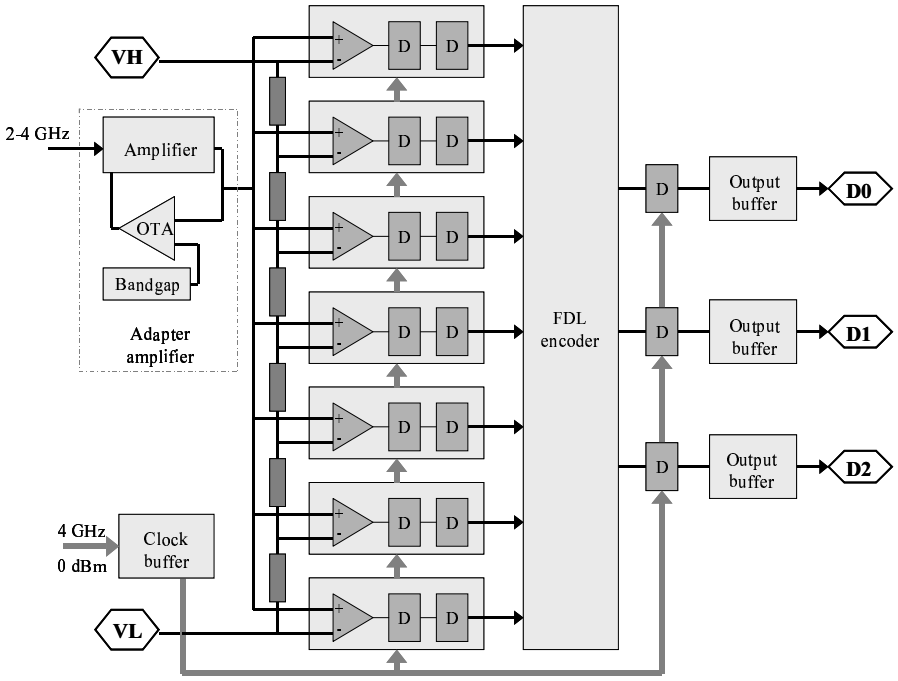


FIG. 1.6 – *Schéma simplifié du CAN dessiné pour le système ALMA*

4.4 Le Corrélateur

Le terme « corrélateur » désigne évidemment le système électronique calculant les coefficients de corrélation (leads et lags[2]) du signal observé, mais par extrapolation, il désigne également l'ensemble de l'électronique réalisant le traitement numérique du signal. Dans ce cas là, le terme corrélateur sera écrit avec un « C » majuscule.

Le Corrélateur est un système où l'on combine pour toutes les paires d'antennes du réseau les signaux transmis par le Back-End. Il calcule aussi tous les produits d'auto-corrélation afin de mesurer la puissance détectée par chacune des antennes du réseau ainsi que les produits de cross-corrélation. Le réseau ALMA comporte $\frac{64 \times 63}{2} = 2016$ paires d'antennes et à chaque antenne correspond 4 paires de BBs de 2×2 GHz (2 polarisations différentes), soit un flot de données de 16 GS.s^{-1} par antenne. Le corrélateur doit alors, pour traiter toutes ces données, effectuer plus de 10^{16} multiplications-accumulations par seconde. L'architecture du corrélateur est du type XF où les coefficients sont calculés à partir d'échantillons décalés dans le temps. Les images des sources sont alors obtenues par transformée de Fourier inverse des coefficients de corrélation.

4.5 Système informatique

Le système informatique général d'ALMA a pour tâches la programmation des observations, le contrôle de tous les intruments incluant le pointage des antennes, la surveillance des performances des antennes et des paramètres environnementaux, la gestion des flots de données à travers l'électronique du Back-End et l'envoi des données au corrélateur. Le traitement des données produites par le corrélateur est assuré par un réseau de 16 PCs, plus un PC pour le contrôle du réseau. Cet ensemble de PCs est appelé Correlator Data Processor (CDP). Quatre de ces PCs

[2]les lags sont les coefficients de corrélation pour un retard $\tau \geq 0$ et les leads pour un retard $\tau < 0$

sont connectés chacun au Long Term Acummulator (LTA) de l'un des quadrants (terme défini en section 5.1). Le CDP effectue une succession de traitements sur les données comme la correction de quatification, le fenêtrage, la Transformée de Fourier ou encore diverses corrections dues au traitement subi par le signal observé.

4.6 Intéret du filtrage numérique

Dans les systèmes électroniques de TNS d'un radiotélescope, le Corrélateur permet, entre autre, de calculer la Densité Spectrale de Puissance (DSP) des signaux. Un paramètre important caractérisant une DSP est sa résolution spectrale, ou plutôt la résolution spectrale fournie par le système. En pratique la résolution spectrale d'un corrélateur est définie comme suit :

$$R_s = \frac{B}{N_c} \tag{1.5}$$

où B est la bande traitée par le corrélateur et N_c le nombre de canaux du corrélateur (où nombre de coefficients de corrélation calculés). Ce nombre de canaux est fixé par l'architecture électronique du corrélateur, c'est à dire le nombre de « lags » disponibles dans les puces de corrélation (voir section 5.1). Pour améliorer la résolution spectrale (améliorer la précision avec laquelle les motifs spectraux apparaissent sur la DSP) il faut donc accepter de diminuer d'autant la bande de signal à analyser. Le rôle des filtres numériques est d'extraire de la bande de fréquences fournie par le Back-End les sous-bandes (SBs) que l'on souhaite analyser avec plus de précision. L'opération réalisée par les filtres consiste en une opération de filtrage et de ré-échantillonnage, ou opération de décimation dans notre cas (la décimation de facteur D est une opération qui consiste à conserver un échantillon sur D du signal traité). C'est cette opération de ré-échantillonnage qui permet de réduire la bande de signal à analyser (effet de « zoom » spectral). L'opération préalable de filtrage, dont le rapport de bande doit coïncider avec le facteur de décimation (filtre de Nyquist), permet d'atténuer les repliements spectraux (phénomène d'aliasing) qui ont lieu lors de la décimation du signal. L'effet de zoom spectral est donc assuré par l'opération de décimation. En Figure 1.7 est illustré cet effet de zoom spectral avec un filtre quart de bande suivi d'une décimation de facteur $D = 4$.

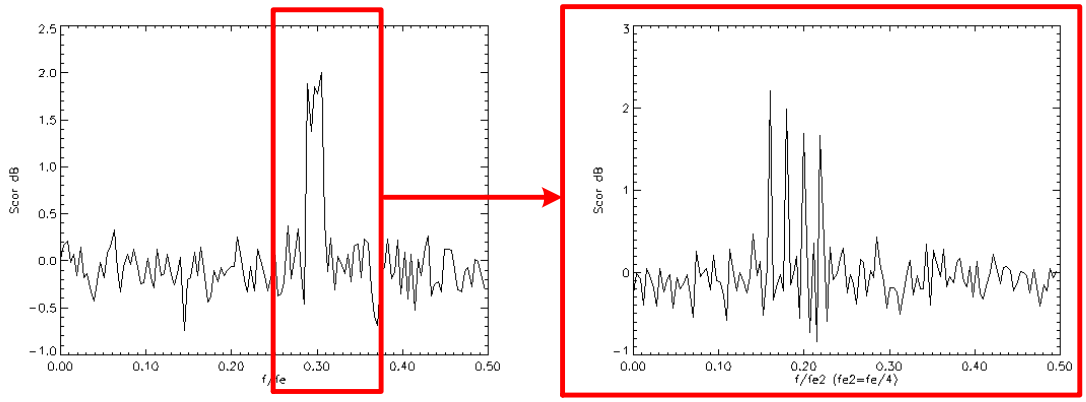


FIG. 1.7 – *Analyse spectrale d'un signal Radioastronomique, avec N_c fixe*

L'amélioration de la résolution spectrale (d'un facteur 4) y est flagrante. Il est à souligner que pour obtenir un même SNR, le temps d'intégration doit être augmenté du même facteur 4 puisque le SNR est proportionnel à $\sqrt{B\tau}$, où τ est le temps d'intégration.

Notons que le signal d'entrée utilisé pour les différentes modélisations mathématiques rencontrées dans ce document est composé de sinusoïdes noyées dans un bruit blanc gaussien représentant un signal de type radioastronomique (4 sinusoïdes sont ici utilisées).

5 L'architecture du Corrélateur ALMA

5.1 Structure générale

L'électronique du Corrélateur est organisée en quadrants, chaque quadrant traitant une paire de BBs (sous deux polarisations) pour toutes les antennes [5] (un quadrant représente 1/4 des ressources Corrélateur). Si les quatres paires de BBs fournies par le Back-End ne sont pas utiles, il est possible de transmettre le flot de données d'une seule BB, par exemple, à plusieurs quadrants. La résolution spectrale s'en trouve améliorée puisque les ressources de plusieurs quadrants calculent alors des lags différents pour le même signal. Il n'existe en revanche aucune connexion entre les quadrants du Corrélateur.
Le traitement effectué par un quadrant se divise en 4 étapes principales pour lesquelles des cartes électroniques distinctes (travaillant à 125 MHz) sont développées. Ces dernières sont représentées en Figure 1.8 : le système de filtrage (filtres RIF), les « station cards », le corrélateur et le système *LTA*.

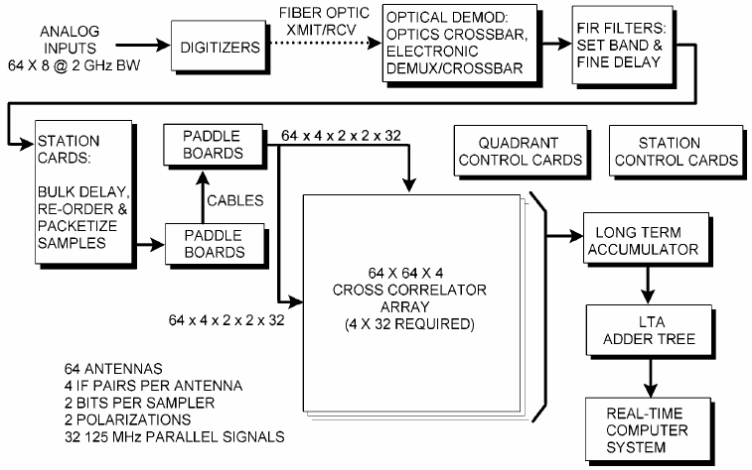


FIG. 1.8 – *Les sous-sytèmes composant le Corrélateur*

Les station cards sont composées de mémoires permettant de stocker 4 MS (soit 1 ms d'intégration). Le flot d'échantillons provenant des cartes de filtrage est en fait réorganisé dans ces mémoires afin d'être redistribué vers les « Planes » composant le Corrélateur. 32 Planes sont disponibles par paire de BBs (2 polarisations, H et V) et chaque Plane est constitué de 4 cartes de corrélation (Figure 1.9).
Une carte de corrélation est composé d'une matrice 8×8 de puces de corrélation (Figure 1.10) permettant de calculer les lags et les leads pour 32 antennes (soit les leads et lags en cross-corrélation de $\frac{32 \times 31}{2}$ combinaisons d'antennes et les lags en auto-corrélation pour les 32 antennes considérées).
Chaque puce de corrélation intègre une matrice 4×4 de blocs corrélateur de 256 lags (ou leads) pour un total de 4096 lags permettant de traiter les 4 modes de base de corrélation qui sont les

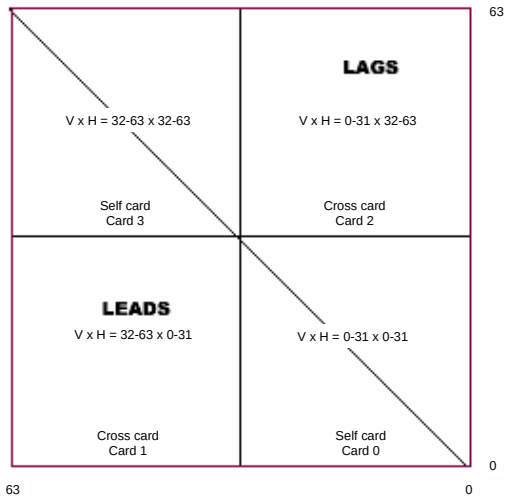


FIG. 1.9 – *Schéma d'un Plane du Corrélateur*

suivants :

- corrélateur 256 lags pour la polarisation H seulement (produit H.H).
- corrélateur 256 lags pour la polarisation V seulement (produit V.V).
- 2 corrélateurs 128 lags pour H et V (produits H.H et V.V).
- 4 corrélateurs 64 lags pour H et V (produits H.H, V.V, H.V et V.H).

Il est à noter que chaque puce de corrélation traite les signaux d'une bande de base provenant de 4 antennes sous les polarisations (R et L sur la Figure 1.10 ou H et V dans le texte). Cette puce a, tout comme le CAN, fait l'objet d'un développement ASIC.
Le nombre total de cartes nécessaires pour corréler (cross-corrélations et auto-corrélations) les 4 paires de BBs des 64 antennes est donc :
4 quadrants (ou BBs) x 32 planes x 4 cartes de corrélation = 512 cartes de corrélation.
Concernant le LTA, son rôle est d'accumuler les résultats fournis par les puces de corrélation si des temps d'intégration supérieur à 1 ms sont nécessaires.

5.2 Architecture DHXF

L'architecture du Corrélateur ALMA actuel [6], retenue pour le projet, appelée Digital Hybrid XF ($DHXF$), est intermédiaire entre les architectures traditionnelles XF et FX (qui correpondent aux méthodes dites du corrélogramme et du périodogramme, respectivement). Le principe de base de cette méthode consiste à diminuer les ressources de corrélation requises pour obtenir une résolution spectrale donnée, en divisant la bande à corréler en plusieurs sous-bandes (SBs) de fréquence contigües corrélées individuellement. Cette architecture permet de diminuer la complexité du corrélateur mais au détriment du système de filtrage numérique. Il est à noter que la complexité d'un corrélateur est proportionnelle au carré du nombre d'antennes alors que la complexité du système de filtrage numérique est proportionnelle au nombre d'antennes. L'architecture DHXF présente donc des avantages par rapport à une architecture XF surtout lorsque le nombre d'antennes composant le système devient important. Dans cette nouvelle architecture, le système de filtrage numérique joue un rôle majeur. Il est décrit en détail en section 6.

Les architectures XF et DHXF sont présentées en Figure 1.11 pour une bande à analyser de 2 GHz.

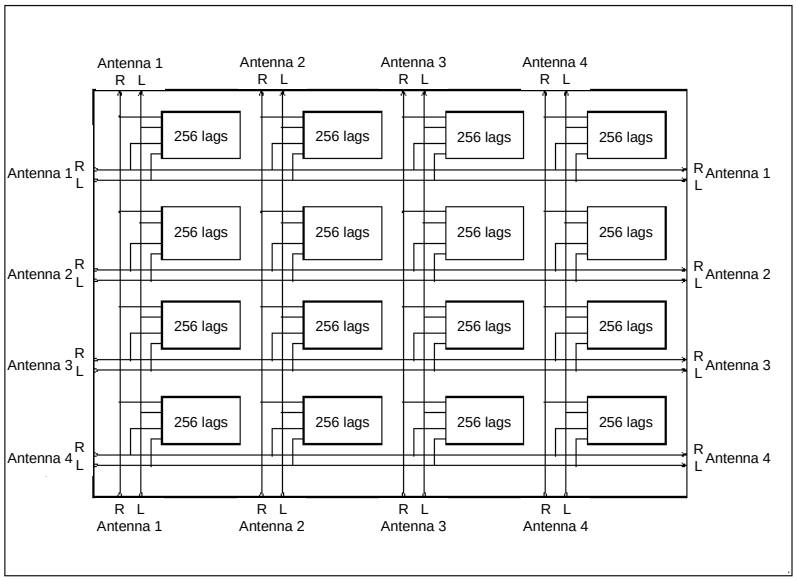


FIG. 1.10 – *Schéma d'une puce de corrélation*

La résolution spectrale offerte par l'architecture DHXF est supérieure à celle offerte par l'architecture XF d'un facteur égal au nombre de SBs utilisées pour représenter la BB, et ceci pour une complexité du corrélateur équivalente. L'utilisation optimale de la structure DHXF est obtenue lorsque la largeur de la SB et la fréquence de corrélation sont liées par le critère de Nyquist, ce qui correspond, pour une fréquence de fonctionnement du corrélateur de 125 MHz, à une largeur de SB de 62.5 MHz. Ceci revient à diviser la BB de largeur 2 GHz en 32 SBs adjacentes. Ainsi un seul Plane du corrélateur suffit pour corréler chaque SB. L'analyse du signal est alors effectuée par démultiplexage fréquentiel qui se substitue à la technique de démultiplexage temporel de l'architecture XF (chaque Plane du corrélateur traite un intervalle de temps égal à $\frac{1}{32}$ ms ce qui correspond à 125 MS.s^{-1}, $\frac{1}{32}$ du rythme initial de 4 GS.s^{-1}).
Le spectre total de la BB n'est reconstruit qu'après TF, en juxtaposant les 32 spectres des différentes SBs (la reconstruction des spectres et les corrections associées sont présentées en [7]).

6 Le système de filtrage numérique ALMA

Le système de filtrage numérique, développé par l'équipe européenne du Corrélateur ALMA, se nomme Tunable Filter Bank (TFB). Il apporte plus de puissance et de souplesse que le système initial prévu pour l'analyse spectrale sans demander aucune modification des autres sous-systèmes (en particulier le sous-système corrélateur). Dans ce chapitre, l'architecture de l'un des filtres composant le système TFB ainsi que différents points importants concernant le traitement numérique du signal à l'intérieur de ce même sous-système sont présentés.

6.1 Le filtre TFB

Le schéma de principe d'un filtre du TFB [2] est présenté en Figure 1.12.
Le signal d'entrée est composé de 32 voies parallèles de 3 bits à 125 MHz, représentant un flot de données équivalent à 4 GS.s^{-1}. Les 32 voies sont donc dépendantes les unes des autres. Ce signal

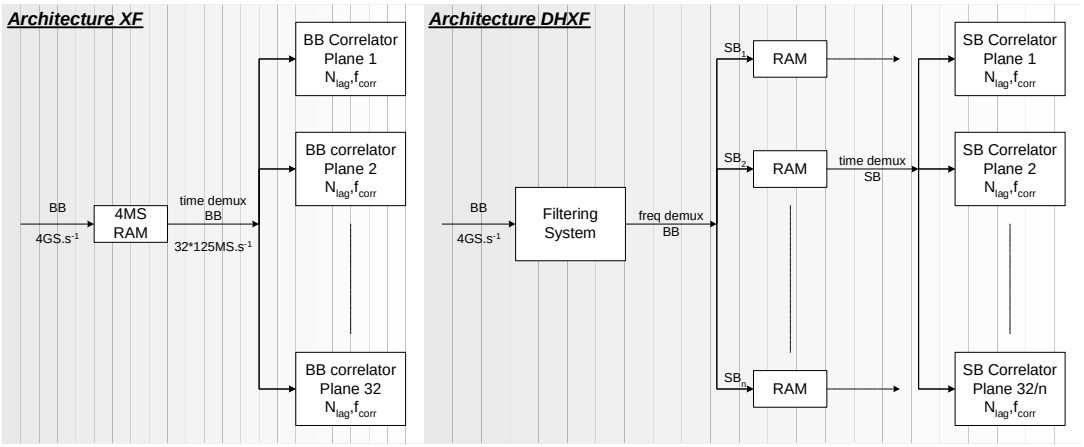


FIG. 1.11 – *Comparaison entre une architecture XF et DHXF*

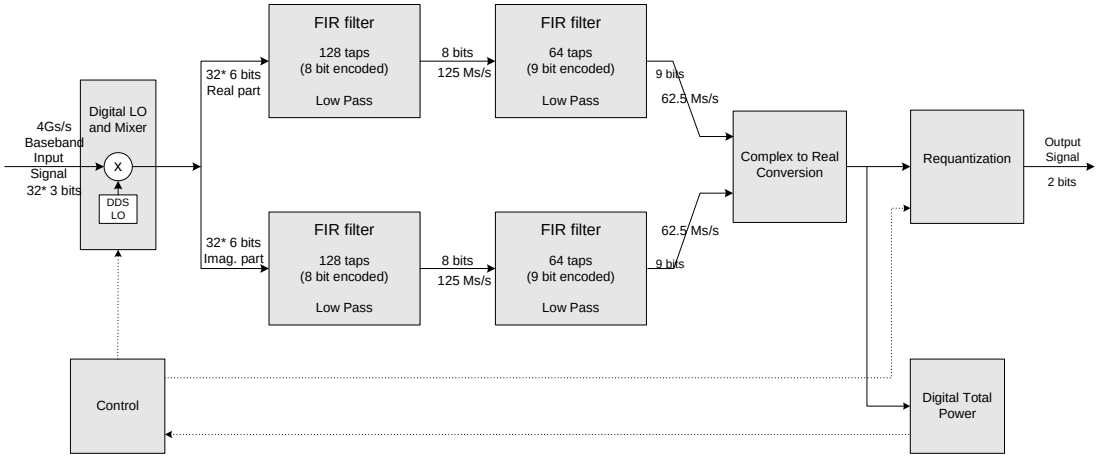


FIG. 1.12 – *Architecture d'un filtre composant le TFB*

est traité par un mélangeur numérique piloté par un oscillateur local numérique (Direct Digital Synthesizer - DDS) permettant de translater le signal en fréquence. Ceci permet de translater n'importe quelle fréquence de la BB à traiter (avec une certaine résolution) en fréquence centrale, le signal translaté est alors complexe. Intervient alors l'opération de filtrage (et de décimation) qui permet d'obtenir l'effet de zoom spectral sur la fenêtre d'analyse souhaitée. Le traitement s'effectue sur deux voies parallèles (parties réelle et imaginaire) comportant toutes les deux la même architecture de filtrage découpée en deux étages. Cette découpe permet de diminuer la complexité totale du filtre. Les performances sont équivalentes à un étage unique composé d'un nombre de poids égal au produit du facteur de décimation du premier étage par le nombre de poids du second étage.
Le premier filtre, un passe-bas avec une région de transition grossière, procure juste assez de sélectivité pour effectuer la décimation du signal par un facteur 32, sans aliasing significatif. Le deuxième étage, un passe bas avec une région de transition étroite, a pour but de donner la forme de bande finale de la SB. Le signal est ensuite décimé par un facteur 2 puis converti en un signal réel en re-combinant les parties réelle et imaginaire. La requantification du signal s'effectue sur 2 ou 4 bits (suivant l'efficacité de corrélation souhaitée) ; elle est fonction de la puissance du signal

en sortie du filtre ce qui assure une efficacité de corrélation optimale. Les différents paramètres indiqués en Figure 1.12 résultent de réflexions menées entre efficacité du traitement numérique et ressources électroniques disponibles dans les composants cibles.
Les architectures retenues pour les deux étages de filtrage sont développées dans les sections qui suivent.

6.1.1 Le premier étage TFB

Comme cela a été evoqué précédemment, la position de la SB sélectionnée est déterminée par le système DDS - mélangeur. Le premier étage permet de sélectionner $\frac{1}{32}^{ieme}$ de la BB translatée. Les spécifications du projet pour ce filtre sont les suivantes : 47 dB d'atténuation, bande passante égale à $\frac{1}{64}^{ieme}$ de la bande originale (i.e. de largeur $f_{ech}/128$), ce qui conduit à une largeur de la bande de transition importante (cette dernière ne doit pas empiéter sur la SB finale). Le tableau 1.2 présente les spécifications en fréquence du filtre.

TAB. 1.2 – *Gabarit en fréquence du* 1^{er} *étage TFB*

	Début	fin
Bande passante	0	$f_{ech}/128$
Bande d'arrêt	$3 \cdot f_{ech}/128$	$f_{ech}/2$

Le filtre a été synthétisé par l'intermédiaire de l'algorithme de Remez [8] et est composé de 128 poids symétriques assurant la linéarité de la phase nécessaire en radio-astronomie. Après quantification des coefficients sur 8 bits, la fonction de transfert obtenue est celle tracée en Figure 1.13 (axe des fréquences gradué en MHz afin de faire apparaitre la largeur de la SB).

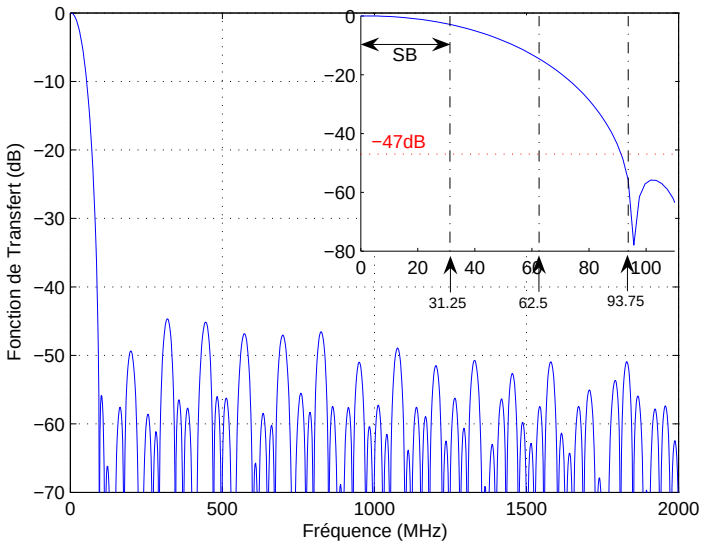


FIG. 1.13 – *Fonction de transfert du* 1^{er} *étage TFB (quantifiée sur* 8 *bits)*

Le zoom permet de constater que les spécifications sont remplies. La large région de transition y apparait clairement (de 31.25 MHz à 93.75 MHz). Aprés décimation par 32 ($\frac{f_{ech}}{2}$ est alors égal à 62.5 MHz), la région de transition est alors repliée sur elle même, dans des régions extérieures à la SB. La SB n'est donc pas polluée par le repliement, les autres régions repliées se trouvant

dans la bande atténuée. Le signal, complexe, obtenu après décimation contient de l'information utile dans la bande $[-31.25\text{MHz}, 31.25\text{MHz}]$.
L'architecture utilisée pour l'implémentation du filtre est représentée en Figure 1.14.

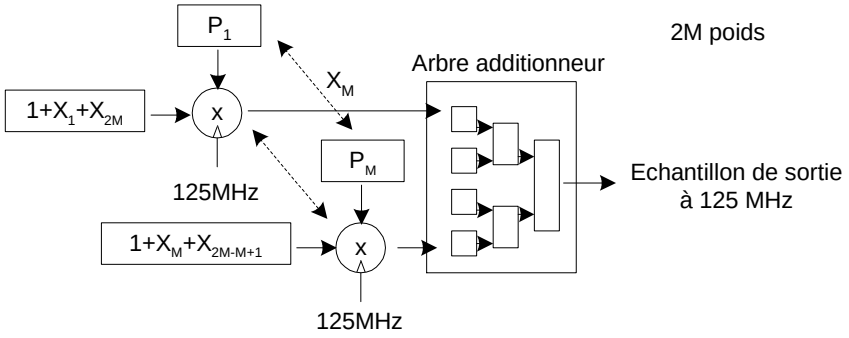


FIG. 1.14 – *Architecture du premier étage de filtrage TFB*

Le signal fourni par le DDS est codé sur 6 bits en complément à deux [3]. Les 32 voies sont dirigées vers un registre à décalage de profondeur 4, permettant de disposer de 128 échantillons successifs nécessaires au calcul de la fonction de convolution. Les échantillons correspondant aux poids symétriques sont additionnés avant d'être multipliés par leur poids commun (qui sont stockés en ROM). L'adjonction de 1 au résultat de la somme permet d'obtenir en sortie un codage du signal symétrique (contrairement à un signal codé en complément à deux) et de ce fait de moyenne nulle (le signal reçu par le Corrélateur étant de nature gaussienne à moyenne nulle). L'échantillon en sortie du filtre RIF est obtenu en sommant les résultats des diverses multiplications au moyen d'un arbre d'addition réalisé en pleine échelle (signal codé sur 21 bits en sortie). Le signal est alors requantifié afin de limiter la complexité du second étage de filtrage. Une troncature à 8 bits du signal est effectué (la méthode pour tronquer un signal sans détérioration de l'information est exposée en section 6.2.2). A noter que la décimation par un facteur 32 est intrinsèque puisque pour 32 échantillons en entrée, l'arbre d'addition n'en délivre qu'un.

6.1.2 Le second étage TFB

Le second étage de filtrage fixe la forme finale de la SB de largeur 62.5 MHz dans sa version nominale ; c'est un filtre décimateur de facteur 2. Les spécifications de ce dernier sont les suivantes : 47 dB d'atténuation, ondulation maximale dans la bande passante 0.2 dB et largeur de la bande de transition égale à $\frac{1}{32}^{ieme}$ de f_{ech}. Le tableau 1.3 retrace les spécifications en fréquence du filtre.

TAB. 1.3 – *Gabarit en fréquence du 2^{nd} étage TFB*

	Début	fin
Bande passante	0	$15 \cdot f_{ech}/64$
Bande d'arrêt	$17 \cdot f_{ech}/64$	$f_{ech}/2$

Le filtre a été synthétisé avec l'algorithme de Remez couplé à un algorithme de minimization (*Amoeba simplex minization*) permettant de compenser la bande passante du filtre utilisé en amont. Cet algorithme va permettre de compenser la chute dans la bande passante engendré

[3] Cette arithmétique est destinée à rendre cohérent le résultat d'opérations entre des nombres binaires signés

par le 1^{er} étage TFB (Figure 1.13, intervalle $[0, 31.25\ \text{MHz}]$). La réponse impulsionnelle obtenue est composée de 64 échantillons symétriques. La fonction de transfert de ce filtre est tracée en Figure 1.15 pour des coefficients codés sur 9 bits.

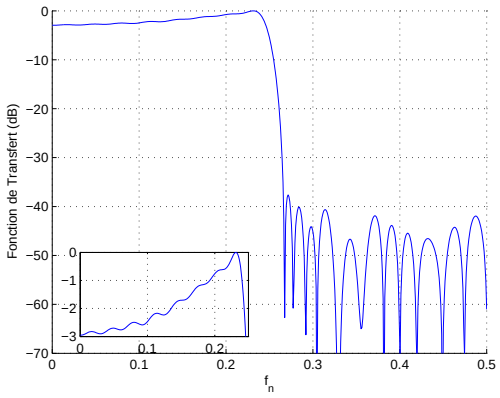


FIG. 1.15 – *Fonction de transfert du 2^{eme} étage TFB*

Dans le but d'améliorer encore la résolution spectrale, il a été décidé que le 2^{nd} étage permettrait d'implémenter un filtre appelé « demi-bande »($\frac{1}{2}$SB = 31.25 MHz, filtre décimateur de facteur 4). La principale contrainte est le nombre de poids que doit comporter ce demi-bande : 64, afin d'utiliser la même architecture pour les deux gabarits de filtre. Dans ce cas, l'amélioration de la résolution spectrale se fait au détriment de la sélectivité du filtre, de la raideur de sa bande de transition. En effet la compléxité d'un filtre est proportionnelle au rapport de bande. A compléxité donnée (64 poids), la bande de transition du demi-bande est moins performante que celle du filtre générant la SB. En Figure 1.16 est tracée la fonction de transfert de ce filtre générant une demi-SB.

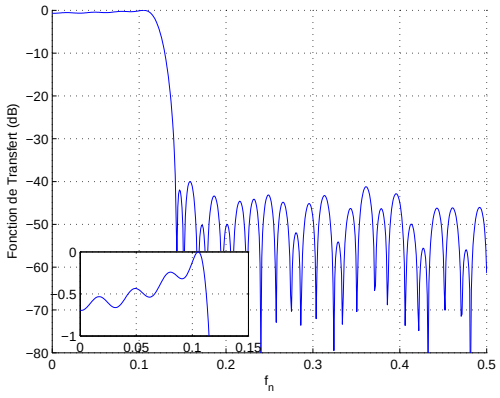


FIG. 1.16 – *Fonction de transfert du 2^{eme} étage - demi-bande - TFB*

L'architecture utilisée permettant l'implémentation de ce filtre est celle présentée en Figure 1.17.

Pour ce second étage de filtrage, la fréquence à laquelle doivent être délivrés les échantillons filtrés est 62.5 MHz (pour le mode nominal de filtrage). Il est donc possible de diviser par deux le nombre de multiplieurs à implémenter en fonctionnant à la fréquence d'horloge de 125 MHz,

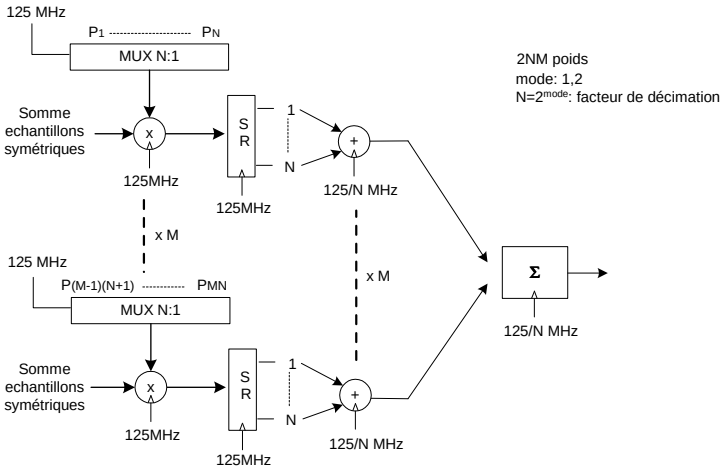


FIG. 1.17 – *Architecture du second étage de filtrage TFB*

en utilisant la technique dite de recirculation. Cette technique illustrée en Figure 1.17 consiste à effectuer la multiplication de 2 couples échantillons - poids différents à l'aide d'un même multiplieur, mais sur deux cycles d'horloge différents. A la suite de la multiplication, les produits $x_i \cdot p_i$ sont traités par un arbre d'addition après avoir été parallélisés au moyen d'un registre à décalage de profondeur N (le facteur de recirculation, 2 en version nominale du filtre ou 4 en version demi-bande). La logique en aval des multiplieurs fonctionne alors à la fréquence d'horloge de $125/N$ MHz et débouche sur une sortie unique au débit de $125/N$ MS.s^{-1} (décimation d'un facteur N intrinsinque). Comme dans le premier étage, le nombre de cellules multiplicatrices peut être diminué encore par deux puisque le filtre possède une réponse impulsionnelle symétrique. Ainsi le nombre de multiplications nécessaires à l'implémentation du filtre est $n/4$ (n étant l'ordre du filtre, 64).
Concernant l'implémentation des multiplications dans les composants sélectionnés, ce sont les blocs DSP qui s'adapte le mieux au besoin et au format des données (DSP 9x9 ; poids codés sur 9 bits et somme des échantillons correspondant au même poids sur : 8bits + 8bits = 9 bits). Ceci explique la nécessité de coder le signal de sortie du 1^{er} étage sur 8 bits, permettant une utilisation optimale des blocks DSP. Les poids sont reconfigurables dynamiquement et stockés en RAM.

6.1.3 Etage de conversion et re-quantification dynamique

Bien qu'il soit possible de concevoir des corrélateurs traitant des signaux complexes, le choix s'est porté pour le projet ALMA sur un format de corrélation réel. Le signal complexe résultant de la translation en fréquence induite par le DDS doit être reconverti en un signal réel transportant le même contenu spectral avant d'être traité par le corrélateur.
Dans de nombreux ouvrages de traitement du signal la notion de signal analytique est introduite. Ce signal analytique x_A est un signal complexe comportant la même information spectrale que le signal réel x_R dont il est dérivé, mais présente la particularité de ne véhiculer aucune information spectrale dans la bande des fréquences négatives (Figure 1.18).
Le signal complexe est alors obtenu par simple translation en fréquence ($\frac{1}{4}$ du spectre). Dans le cas de l'étage de conversion du TFB, nous devons effectuer en quelque sorte la transformation inverse : partant d'un signal complexe, nous voulons obtenir le signal réel correspondant.

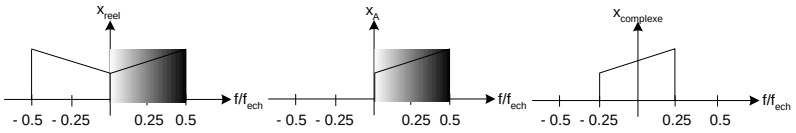


FIG. 1.18 – *Conversion complexe-réel*

Par définition le signal analytique x_A est lié au signal réel dont il provient par la formule :

$$x_A(k) = x_{reel}(k) + j\mathrm{TH}\left(x_{reel}(k)\right) \tag{1.6}$$

où TH désigne la transformée de Hilbert réalisée à partir d'un filtre caractérisé par une réponse impulsionnelle anti-symétrique et une réponse en amplitude égale à l'unité à travers le domaine fréquentiel [9]. Dans le cas général ou le signal complexe s'étend sur l'intervalle de fréquences normalisées $[-0.5, 0.5]$, la formule suivante doit être appliquée afin d'obtenir le signal réel équivalent :

$$x_{reel}(k) = \Re\left(x_{complexe}(k)\right) - \mathrm{TH}\left(\Im(x_{complexe}(k))\right) \tag{1.7}$$

Dans notre cas, le signal complexe reconstitué en sortie du 2^{nd} étage de filtrage couvre le domaine spectral $[-0.25, 0.25]$. Par une simple translation en fréquence de $\frac{1}{4}$ de spectre, on se ramène à la situation du signal analytique de la Figure 1.18. Il ne reste plus qu'à récupérer la partie réelle du signal analytique pour obtenir le signal désiré :

$$x_{reel}(k) = \Re(x_A(k)) = \Re\left(x_{reel}(k) + j\mathrm{TH}\left(x_{reel}(k)\right)\right) \tag{1.8}$$

Il suffit donc d'appliquer au signal complexe résultant des 2 étages de filtrages le traitement suivant pour obtenir le signal réel à contenu spectral équivalent :

$$x_{reel}(k) = \Re\left(x_{complexe}(k) \cdot exp\left(\frac{2j\pi k}{4}\right)\right) \tag{1.9}$$

L'implémentation d'une telle fonction de conversion est basée sur le schéma de principe illustré en Figure 1.19.

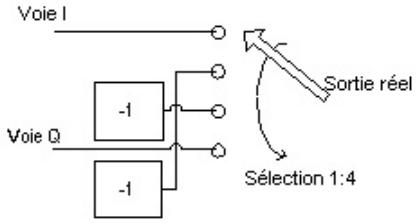


FIG. 1.19 – *Schéma de principe de la conversion*

Le signal réel est obtenu en sélectionnant alternativement les parties réelle et imaginaire du signal complexe (multiplié ou non par -1). Cependant les parties réelle et imaginaire issues des 2 voies de filtrage (Figure 1.12) doivent être décalées l'une part rapport à l'autre d'une période d'horloge afin de bien respecter l'enchevètrement des échantillons pairs – voie réelle – et impairs

– voie imaginaire.

Suite à cette conversion, la « re-quantification » du signal avant corrélation est effectuée (sur 2 bits ou 4 bits). Elle ne peut être ici qu'une simple troncature de bits. Il s'agit en effet de véritablement comparer le signal de sortie du convertisseur à des seuils références (fonction de l'écart type du signal lui-même) et d'encoder ce dernier en conséquence. Pour un signal à puissance variable en entrée du fitre (SBs), la puissance du signal en sortie (avant corrélation) doit être stable. L'adaptation des seuils de re-quantification est donc requise pour s'adapter à la puissance du signal de chaque SB et ainsi conserver une efficacité[4] de corrélation optimale pour le format de donnée choisi [10, 11].
Deux solutions sont envisageables afin de re-quantifier dynamiquement le signal en sortie de l'étage de conversion :

- Faire varier les seuils de re-quantification en fonction de la puissance du signal à traiter.
- Appliquer un facteur d'echelle au signal qui est fonction de la puissance de ce dernier avec des seuils de re-quantification fixes.

Il a été choisi, dans le cadre ALMA, d'utiliser la seconde solution qui nécessite moins de ressources logiques. En Figure 1.20 est présenté le schéma de principe de cette étage de re-quantification.

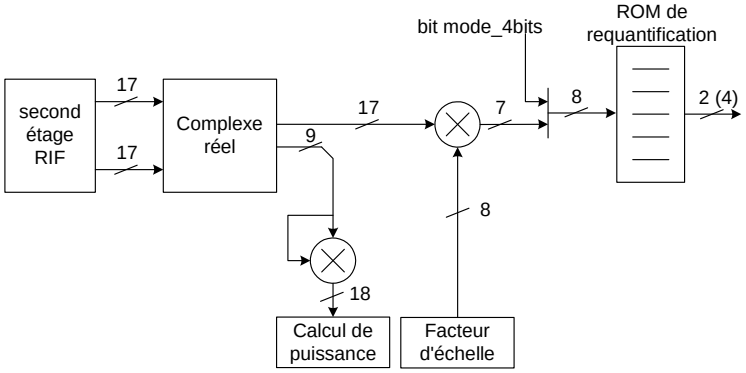


FIG. 1.20 – *Chaine de traitement de la sortie du second étage à la re-quantification*

Sur ce schéma apparaît aussi l'étage de converion complexe-réel. Il est à noter que cette conversion est effectuée avant toute opération de troncature de la sortie du 2^{nd} étage de filtrage. Ainsi la composante continue induite par cette troncature (8 LSBs sont supprimés, Least Significant Bit) effectuée après conversion apparaît dans le canal central du spectre 6.2.2, canal non utilisé lors de l'interprétation des données.
Après multiplication par le facteur d'échelle, le signal est codé sur 25 bits (17 bits signés provenant de la sortie du convertisseur multiplié par le facteur d'échelle codé sur 8 bits non-signé). Ce signal est alors tronqué sur $7bits$ auquel est concaténé le signal $\text{mode}_{4\text{bits}}$ qui va permettre d'atteindre suivant sa valeur (0 en mode 2 bits et 1 en mode 4 bits) différentes régions de la RAM de requantification. Le calcul de puissance, effectué par le système informatique d'ALMA, permet le calcul d'un facteur d'échelle qui à son tour permet de ré-ajuster la puissance du signal de facon à correpondre au seuils de quantification optimaux préalablement établis. La valeur RMS optimale (σ_{optS}) du signal après décalage est donné par l'équation :

$$\sigma_{optS} = \text{facteur d'échelle} \times \sigma_{optE} \tag{1.10}$$

[4] l'efficacité η d'un TNS est définie comme suit : $\eta = \frac{SNR_{\text{avec quantification}}}{SNR_{\text{sans quantification}}}$

La valeur optimale σ_{optS} choisie doit prendre en compte le nombre de bits sur lequel le facteur d'échelle est codé (dans notre cas, 8 bits sont utilisés). Ce nombre de bits caractérise la précision nécessaire à la requantification (permettant d'obtenir une valeur σ_{optS} aussi proche que possible de la valeur optimale) ; il est aussi fonction de la valeur de σ_{optE}. Enfin la valeur optimale σ_{optSs} doit être une puissance de 2 car la requantification obtenue par ce processus est l'équivalent d'un prélèvement de bits. Dans le tableau 1.4 sont données les valeurs de σ_{optS} ainsi que des seuils assurant une efficacité optimale [12]. Ces seuils sont répartis uniformément sur la fonction de distribution qui, ici, est de nature gaussienne.

TAB. 1.4 – *Valeurs optimales des seuils de quantification*

	2 bits	4 bits
Valeur optimale des seuils (σ_{optE})	0.996σ	0.335σ
σ_{optS}	2048	512
Efficacité obtenue	0.881154	0.988457

Le calcul du facteur d'échelle permettant de ré-ajuster la puissance du signal au contenu de la ROM réalisant la fonction de re-quantification s'effectue alors de la manière suivante :

$$\text{Pour le mode 2 bits : } f_{scale} = 2048/\left(0.996 \times \sigma\right)$$
$$\text{Pour le mode 4 bits : } f_{scale} = 512/\left(0.335 \times \sigma\right)$$

où σ est la valeur RMS, ou écart type, du signal à traiter déduite du calcul de puissance.
En Figure 1.21 est présenté l'encodage du signal sur 2 ou 4 bits avec les valeurs de pondération associées à chaque état de codage utilisé pour la corrélation des données.

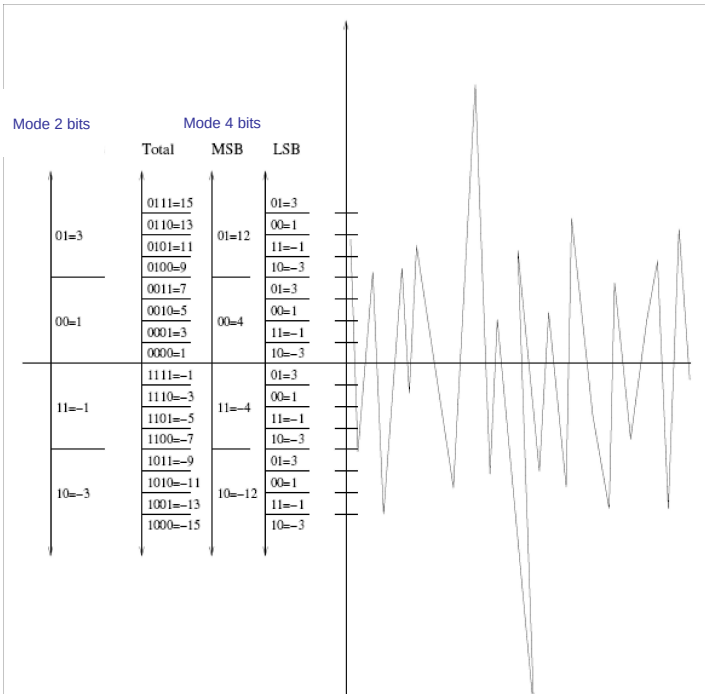


FIG. 1.21 – *Quantification du signal pour les modes 2 bits et 4 bits*

Le mode corrélation 2 ou 4 bits est choisi en fonction de l'efficacité nécessaire pour la corrélation et de la largeur de bande à analyser (expliqué en section suivante).

6.1.4 Modes de fonctionnement secondaires du TFB

Il a été prévu dans le cadre d'observation en continuum, où de larges bandes de fréquence sont utilisées, de conserver un mode permettant de corréler directement la BB sans découpage fréquentiel. Ce mode est similaire à celui présenté en section 5.2 appelé « time demultiplexed mode ». Il a donc été implémenté une fonction permettant de « court-circuiter » le filtre TFB. Cette fonction est appelé mode Bypass. Pour réaliser cette fonction, chaque puce de la carte (cf. section 7) traite (requantifie) 2 lignes du signal d'entrée (32 lignes de 3 bits) et fournie un signal sur 2 bits. Un mode Bypass 3 bits a aussi été implémenté. Ceci n'est possible qu'avec une seule BB ; de plus la corrélation à 3 bits implique qu'il y ait 4 fois moins de points spectraux par BB comparé au cas de celle à 2 bits.

Un mode Half-Delay est aussi implémenté. Il permet de réaliser la fonction « twice Nyquist » destinée à améliorer la sensibilité du corrélateur. Afin de réaliser cette fonction, deux filtres sont nécessaires pour la synthèse d'une SB, le deuxième filtre traitant les mêmes échantillons que le premier mais décallés dans le temps (un delai de 16 échantillons dans notre cas).

La corrélation des données sur 4 bits est aussi possible visant à améliorer l'efficacité de corrélation des données. Pour réaliser cette fonction, 2 filtres du TFB synthétisent la même SB, l'un en mode de corrélation 4 bits qui fournit ainsi les 2 LSB et l'autre en mode 2 bits qui fournit les 2 MSBs (Most Significant Bit). Ces modes de fonctionnement permis par la re-quantification dynamique.

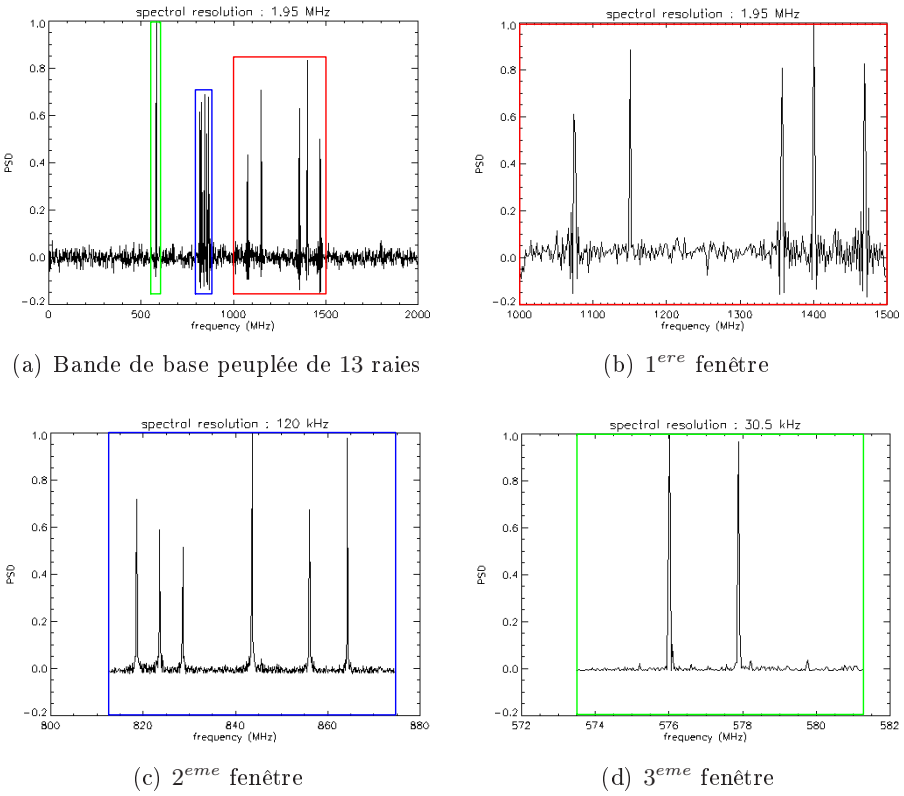
Un compromis entre la souplesse d'analyse et la quantité de données (ou largeur de bande) pouvant être traitées est donc nécessaire.

6.1.5 Flexibilité du système TFB

Le mode nominal du système de filtrage ALMA consiste à extraire 32 SBs de 62.5 MHz de la BB de 2 GHz, permettant ainsi de couvrir intégralement cette BB. De plus, les SBs peuvent être placées à n'importe quel endroit dans la BB (à la résolution du DDS près, 31.5 kHz) permettant à l'utilisateur de couvrir intégralement cette BB ou seulement les parties les plus intéressantes. L'ensemble de ces propriétés offre au système de filtrage TFB une grande souplesse d'analyse.
Un autre critère de souplesse concerne la possibilité de mélanger les modes de filtrage. En effets dans le système TFB, les différents filtres de SBs sont indépendants. Comme cela a été décrit précédemment (cf. section 5.1), les ressources du corrélateur peuvent être redistribuées afin de fournir plus de puissance de calcul à une SB en particulier (ou à une BB). Les différentes résolutions spectrales pouvant être obtenues grâce au système Corrélateur sont décrite de manière exhaustive en [6].
Ainsi, en utilisant les spécificités de l'architecture du corrélateur et du système de filtrage, il est possible d'analyser simultanément, par exemple, une large bande spectrale en continuum et une SB étroite du spectre qui présente un motif particulier faisant l'objet d'une étude spectroscopique (Figure 1.22).

6.1.6 Configuration des paramètres du TFB

Le système TFB comprend plusieurs paramètres pouvant être modifiés dynamiquement tel que l'incrément du DDS qui détermine la fréquence de translation et donc la position de la SB dans la BB ou encore les poids du 2^{nd} étage de filtrage. Le système TFB doit aussi fournir



(a) Bande de base peuplée de 13 raies

(b) 1^{ere} fenêtre

(c) 2^{eme} fenêtre

(d) 3^{eme} fenêtre

FIG. 1.22 – *Spectre illustrant la souplesse d'analyse du système de filtrage*

au système ALMA certains paramètres comme la puissance mesurée par SB. Tous ces aspects de communication sont gérés « à la manière » d'un microcontrôleur. Les puces TFB, en plus de la fonctionnalité de filtrage, intègrent donc des registres de contrôle, des registres de mode de fonctionnement et des registres de données qui peuvent être sélectionnés ou modifiés par l'intermédiaire de 4 signaux de contrôle et d'un port de données 8 bits [13].

6.2 Particularités du flot de traitement du TFB

Dans ce paragraphe, certains points clefs du TNS (telle la représentation des signaux ou encore l'opération de troncature) intervenant à différents endroits de la chaîne de traitement vont être développés.

6.2.1 La représentation du signal

Pour un échantillon v codé sur n bits en complément à deux, l'intervalle couvert est le suivant : $v \in \left[-2^{n-1}, 2^{n-1}-1\right]$. Une valeur codée ν du signal correspond alors à l'intervalle de valeurs $[\nu, \nu+1]$. Une composante continue (DC) apparait alors au centre du spectre représentant le signal due au codage non-symétrique. Cette dernière peut être une source d'erreur d'interprétation pour l'observateur puisqu'elle ne présente aucun intérêt scientifique, c'est un parasite. Une convention a été établie dans le cadre du projet ALMA concernant le codage du signal ; une valeur ν numérisée doit correspondre à l'intervalle $[\nu - 1/2, \nu + 1/2]$ (représentation décalée). Pourtant cette représentation ne peut pas être utilisée lors de processus arithmétiques « complexes » comme des multiplications ou des additions qui sont des opérateurs fonctionnant

en complément à deux standard. Le signal doit être converti pour être compatible. Les valeurs utilisées sont donc décalées de $0,5$. La Figure 1.23 présente une méthode permettant de compenser ce décalage lors de l'addition de 2 échantillons (addition des échantillons symétriques pré-multiplication dans le filtre du 2^{nd} étage par exemple).

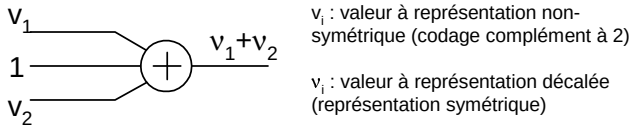


FIG. 1.23 – *Représentation décalée du signal*

Une fois ceci réalisé, le signal résultant de l'opération peut être réutilisé tel quel pour une autre opération arithmétique « complexe » à condition que le format du signal n'ait pas été modifié entre temps. Dans le cas d'une opération de troncature en sortie d'un étage de filtrage par exemple (équivalente à une re-quantification du signal), on retrouve la représentation initiale du signal, la représentation non-symétrique. Si une nouvelle opération doit être effectuée, la technique explicitée ci-dessus doit être ré-employée pour l'obtention d'une représentation symétrique (décalée).

6.2.2 Effets des multiples re-quantifications

Dans le cas du système de filtrage ALMA TFB FIR composé de 2 étages, une étude concernant les pertes de sensibilité liées à l'adoption d'une structure multi-étages, et les troncatures qui y sont associées, a été menée [12]. Ces troncatures (suppression de LSBs) ont deux principales influences : d'une part elles diminuent la dynamique du signal de par l'augmentation du niveau de bruit, d'autre part elles entraînent l'apparition d'un biais dans le signal (apparition d'une composante continue). En effet une troncation à n bits est équivalente à une « re-quantification » du signal avec un pas de quantification égal à n bits, i.e. 2^n unités. Une valeur de codage ν couvre alors l'intervalle $[2^n \times \nu, 2^n \times (\nu + 1) - 1]$ avec une valeur assumé de $\nu + 1/2$ (cf. section 6.2.1). Ceci entraine un léger déplacement de l'intervalle de $1/2$ bit (ou de $1/2^{n+1}$ dans les nouvelles unités de ν) par rapport au vrai centre de ce dernier.
Pour évaluer le biais induit par les troncatures intervenant à chaque sortie d'étage (de filtre) deux paramètres sont nécessaires : le gain du filtre et le nombre de bits supprimés en sortie de l'étage précédent. Le biais total est la somme des biais induits par les différentes troncatures.
Dans le système ALMA TFB, ce biais est compensé dans l'étage de conversion complexe-réel [12].Le calcul de ce biais est détaillé ici. Le gain dans la bande passante d'un filtre RIF d'ordre N est calculé à l'aide de l'équation suivante :

$$G_{filtre} = \sum_{k=0}^{N-1} h_k \tag{1.11}$$

Comme décrit précédemment, le biais induit par une troncature de n bits est $\frac{1}{2^{n+1}}$. Plus le nombre de bits tronqués est faible, plus le biais est important. Ce dernier est alors multiplié par le gain du filtre suivant, s'il y a, et additionné au biais induit par la troncature à la sortie de ce nouvelle étage, et ainsi de suite.

La troncature sur 8 bits (9 LSBs supprimés) appliquée en sortie du premier étage TFB introduit un biais égal à $\frac{1}{1024}$. Le gain du filtre suivant, le second étage TFB, est lui égal à 544.

Ceci résulte en un biais parasite de valeur avoisinant $0,5$. Dans l'optique de diminuer la valeur de ce biais, aucune troncature du signal en sortie du 2^{nd} étage de filtrage n'est réalisée (sortie codée sur 17 bits), seuls les MSBs non significatifs sont éliminés en respectant la loi exposée dans la remarque qui suit. La troncature du signal à 9 bits nécessaire au calcul de la puissance (utilisation d'un block DSP 9×9) est réalisée après conversion du signal, le biais en résultant apparaît alors dans le canal central du spectre réel, canal qui est dans la suite du traitement éliminé.
L'amplitude parasite correspondant au seul biais engendré par la structure de filtrage est de l'ordre de -82dB ce qui représente une valeur comparable au bruit du système dans un canal après plusieurs heures d'intégration et peut donc être négligée. Ce bruit parasite est cependant indissociable du signal. Il a été décidé d'utiliser l'étage de conversion complexe-réel pour déplacer ce biais au centre du spectre après conversion, en lieu et place de la composante continue dont le canal n'est pas utilisé. Voici l'explication de la méthode utilisée qui permet de dissocier ce bruit du signal utile :
Soit y la valeur non biaisée et $x = y + b$ la valeur biaisée. On sait que $b = 0,5$. Dans le cas d'une représentation classique en complément à 2, les valeurs sont représentées ainsi :

$$\begin{aligned} \text{valeurs positives} &: y + b \\ \text{valeurs négatives} &: -y - b \end{aligned}$$

Cette représenation a comme conséquence l'apparition d'une composante continue. En soustrayant la valeur 1 à chaque valeur et en utilisant un complément à 1, on obtient la représentation suivante :

$$\begin{aligned} \text{valeurs positives} &: y + b - 1 \\ \text{valeurs négatives} &: -y - b + 1 - 1 \end{aligned}$$

Dans le cas présent, avec $b = 0,5$, le biais a été déplacé dans le canal central du spectre. Le codage devient pour les valeurs positives $y - 0,5$ et pour les valeurs négatives $-y - 0,5$. La Figure 1.24 est une schématisation de ce qui est réalisé electroniquement. Une astuce de conception est ici

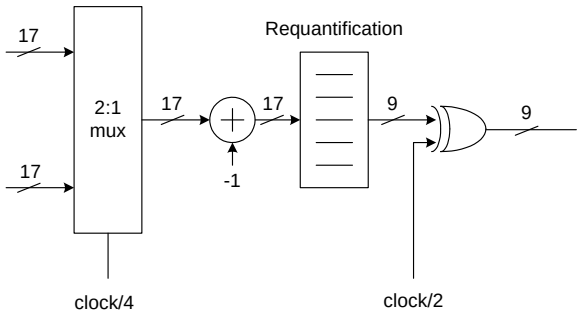


FIG. 1.24 – *Structure du module de conversion complexe-réel modifié*

utilisée : signed(not$(x - 1)$) qui est la représentation signée en complément à 1 du signal x est égale à $-x$. Cette astuce est appliqué au signal lors de la reconversion réelle de ce dernier.

Remarque : troncation des MSBs et dynamique du signal Dans le document [14] est exposé le traitement des différentes re-quantifications opérant tout au long du traitement analytique du filtre TFB. La troncature d'un signal, caractérisée par l'élimination de LSBs, a comme conséquence l'apparition d'un biais, point déja abordé. Cependant, lors de la re-quantification d'un signal, un certain nombre de MSBs non significatifs sont aussi abandonnés, cette opération

se nomme le « clipping ». Ces opérations déterminent la dynamique du signal.
Le bruit ajouté, dû au clipping, est de l'ordre de $\frac{1}{\sqrt{12}} \times q$ (q étant le pas de quantification) tant que la valeur RMS du signal est inférieur à 1/8 de l'échelle totale de quantification [14], et que les valeurs dépassant l'échelle de quantification sont proprement limitées (saturation du codage, pas de « wrapped around »). Dans le cas contraire, le bruit de clipping devient le bruit dominant. Il en est de même concernant le bruit de compression (lié à la dynamique du signal). Si la règle (1.12) permettant la détermination du MSB est respectée, il est inutile d'appliquer de correction par l'intermédiaire de l'algorithme de Van Vleck [15]. Ce dernier point ne sera pas développé ici puisqu'utilisé après corrélation.

$$MSB = \log_2 (8 \cdot \sigma_{signal}) \tag{1.12}$$

Par exemple, pour une quantification sur 8 bits (256 unités de codage), l'amplitude RMS ne doit pas dépasser 5 bits (32 unités de codage) ce qui correspond bien à un rapport 8. Le bruit ajouté dans ce cas là est de l'ordre de -41 dB $\left(20 \log \left(\frac{1}{32 \cdot \sqrt{12}}\right)\right)$.
La règle (1.12 est celle appliquée, tout au long du traitement propre au TFB, au signal lors d'opérations de requantification afin de déterminer le MSB. La valeur de $8 \cdot \sigma$ a été choisie puisqu'elle assure, pour un signal à distribution gaussienne, une probabilité d'apparition extrêmement faible au delà de ce seuil.

6.2.3 Reconstruction du spectre de la BB (overlap)

En [1] et [14] est abordé le problème de la reconstruction du spectre d'une BB à partir de 32 SBs. Après division et filtrage de la BB, les SBs résultantes sont ré-ageancées informatiquement afin de recréer la BB analysée avec une résolution spectrale accrue. Cependant des problèmes peuvent survenir durant l'alignement des SBs. La re-quantification dynamique des SBs en sortie des filtres, calculée en fonction de leur puissance respective, permet ici de compenser d'eventuels problèmes dits de « platforming » (ou « décrochage » de niveau de SB à SB). Ainsi les SBs sont ré-alignées à un même niveau de puissance.
La question des effets de bords est tout aussi essentielle puisqu'ils déterminent la précision avec laquelle le spectre complet est reconstruit dans les régions de transition. Ce problème est directement lié aux performances du filtre. Plus la bande de transition est étroite, moins le spectre de la BB est altéré. Il a été choisi de ne conserver que les canaux spectraux correspondant aux bandes passantes des différentes SBs, et réaliser ainsi un chevauchement des SBs en les juxtaposant. Une étude a été menée à ce sujet [7], démontrant les effets de la juxtaposition des SBs sur la phase et l'amplitude du signal. Il a été envisagé d'abandonner 1 ou 2 canaux de part et d'autre de la bande passante d'un filtre (excepté dans le cas des SBs aux extrémités de la BB). Cependant, cette technique a comme inconvénient de ne pas pouvoir couvrir la BB entièrement. Dans le cas d'abandon de 2 canaux par SBs, la bande utile représente 97% de celle d'origine (Figure 1.25) et dans le cas de l'abandon de 4 canaux par SBs, elle représente 93.74%.

Il a été décidé de supprimer 4 canaux par SB (résultant en une SB finale de largeur 58.59 MHz) permettant de conserver, en accord aves les spécifications scientifiques d'ALMA, une erreur en phase inférieure à 0.5 ° et une erreur en amplitude de l'ordre de 2 % au point de recoupement des SBs.

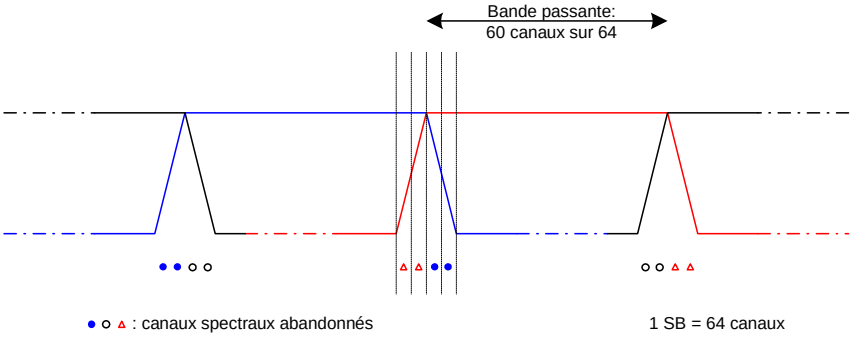


FIG. 1.25 – *Juxtaposition des SBs*

7 La carte TFB

Une fois l'architecture du TFB établie, la question de l'implémentation de la fonctionnalité s'est posée. Il a été imposé de regrouper sur une même carte de filtrage l'ensemble des ressources relatives à l'ensemble d'une BB. La fonctionnnalité (filtrage et système de configuration des filtres) ne pouvant pas être intégrée dans un unique circuit numérique, une étude comparitive des diverses solutions envisageables [2] a été menée. Il a été décidé de répartir la fonctionnalité sur 16 puces de type FPGA (Field Programmable Gate array) Stratix I (des puces Stratix I, EP1S40 ont été choisies initialement). La carte de filtrage représentée en Figure 1.26 est donc composée de 16 puces FPGA (numérotées de puce0 à puce15 sur la figure) contenant chacune deux filtres synthétisant chacun une SB.

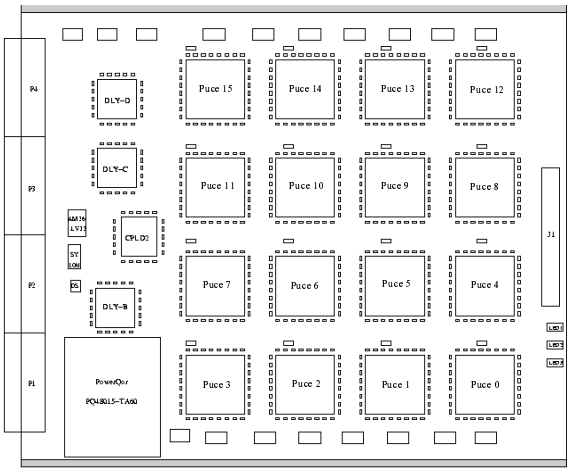


FIG. 1.26 – *Schéma de la carte de filtrage TFB*

La présence de deux autres types de puce ayant des fonctionnalités différentes est aussi à noter. Les puces nommées DLY-X sont des puces delay servant de « buffers » du signal d'entrée. La puce CPLD2 permet la programmation des puces de filtrage, de delay ainsi que l'interfaçage de la carte avec le système de contrôle du corrélateur. La carte est de dimension $6U$ (format A4) et apparait en Figure 1.27.
Tout ce qui concerne la distribution des signaux parmi les puces de filtrage est expliqué en [2].
Pour filtrer toutes les données provenant des 64 antennes, 512 cartes de filtrage sont nécessaires :

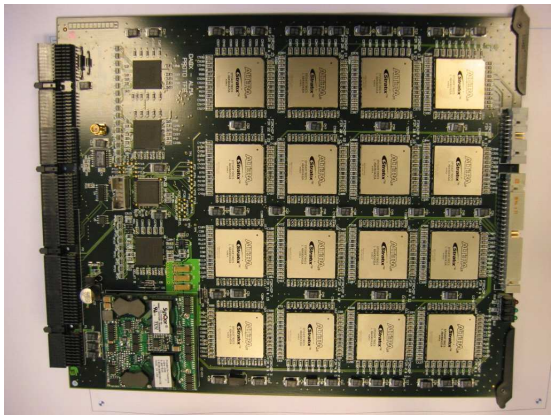


FIG. 1.27 – *La carte TFB peuplée de puces Stratix*

4 bandes de base ×2 polarisations ×64 antennes.

7.1 Tests permettant la vérification de la fonctionnalité

7.1.1 La carte pré-prototype

Avant de lancer la fabrication des cartes TFB, il a été décidé de produire une carte pré-prototype comportant seulement 4 composant Stratix I : 3 puces « RIF »et 1 puce « Test »afin de tester la distribution des données. Nous avons établi un rapport de test à la suite de la validation de la distribution qui apparaît en [16].
Cette carte a permis de tester la distribution des signaux entre les 3 puces FIR. Chaque puce est composée d'un module distribution, d'un module LFSR (Linear Feedback Shift Register) et d'un module test. La puce test, quant à elle, comporte seulement un module LFSR. Le LFSR permet de générer une séquence pseudo-aléatoire d'échantillons [17]. La Figure 1.28 illustre l'architecture électronique d'un LFSR.

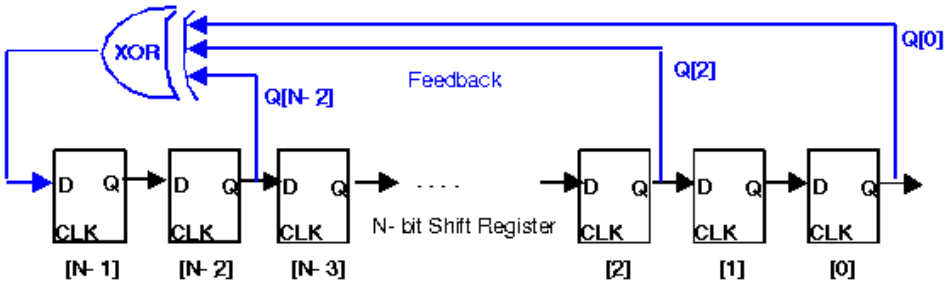


FIG. 1.28 – *Architecture électronique d'un LFSR*

C'est une structure constituée d'un registre à décalage de N étages dont certaines sorties sont assemblées pour ensuite être redirigées à l'entrée du registre. Si ces contre-réactions sont judicieusement choisies, on obtient en sortie du registre une séquence de longueur maximale $T = (2^N - 1)$. Les contre-réactions doivent être choisies de façon à former un polynôme irréductible (polynôme dont les seuls diviseurs sont les éléments inversibles, cf. Annexe). Ainsi, la structure de la Figure 1.28 correspond au polynôme $P(x) = 1 + x^2 + x^{N-2}$. Deux portes logiques peuvent être

utilisées pour le traitement des sorties du registre : XOR et $XNOR$. La première entraine le blocage du registre lorsque l'état $'0'$ est présent sur toute la chaîne alors que c'est l'état $'1'$ qui bloque le registre lors de l'utilisation de la seconde.
La séquence générée est dite pseudo-aléatoire parce que reproductible. En effet, une séquence peut être obtenue à volonté en initialisant les entrées des différentes bascules avec une « graine » donnée. De plus en nourissant, pendant un cours instant, un LFSR avec la séquence générée par un autre, il développera la même séquence que ce dernier.

Cette carte à permis de réaliser différents tests (listés en section 7.1.2) permettant de valider la distribution des signaux et donc le routage de la carte électronique. Le principe général en est le suivant : la puce test envoie une séquence de synchronisation à chacun des LFSR contenu dans les puces RIF. Chaque puce RIF redistribue alors le bit généré aux puces voisines (Figure 1.29).

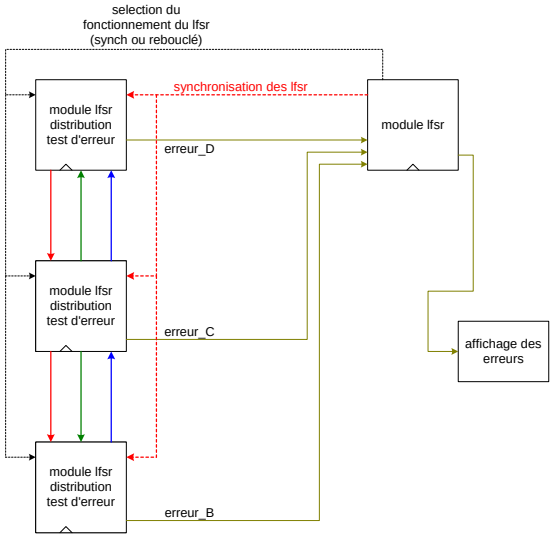


FIG. 1.29 – *Schéma de la vérification de la distribution des signaux*

Les séquence sont alors comparées en interne à chaque puce et les erreurs sont transmises à la puce de test qui gère un afficheur.

Ces tests ont permis de définir la tension d'alimentation des entrées - sorties (IOs) des puces FPGA. Lors de l'étape de placement routage, le standard choisi ainsi que le courant assigné à chacune des IOs sont spécifiés. Une évaluation de la fréquence maximale de fonctionnement du circuit est effectuée après programmation des puces. Le tableau 1.5 retrace les résultats obtenus pour différentes tensions et courants disponibles pour le composant.
L'alimentation 1.8 V a été retenue pour l'alimentation des IOs « signal » avec des « drivers » de courants supportant 4 mA.

7.1.2 Les tests

Pour réaliser les test sur les cartes de filtrage, un système spécifique développé par l'équipe du NRAO a été utilisé (Figure 1.30).
Ce dispositif, nommé Test Fixture, reproduit le sytème de contrôle ALMA (reproducion de la « Station Card » installée dans les « Station Racks » du Corrélateur), permet de générer des signaux d'entrées de type « astronomique » et calcule les coefficients de corrélation du signal filtré. Ce système est conçu pour recevoir deux cartes, une carte de test comportant, entre autre, un microcontroleur et une puce de corrélation ALMA et la carte TFB à tester. Il est relié à un

TAB. 1.5 – *Impact de l'alimentation des IOs sur la distribution des signaux*

Tension (V)	Courant (mA)	Frequence max. (MHz)
1.5	2	115
	4	165
	8	205
1.8	2	145
	8	185
	12	205
3.3	2	165
	4	165
	8	165
	12	165

PC muni d'une interface graphique permettant le lancement de tests destinés à la validation soit du programme implémenté dans les puces, soit des cartes elles-mêmes.
4 types de tests sont alors effectués par l'intermédiaire de ce Test Fixture [18] :

Tests de communication Ces tests permettent de vérifier la communication entre chaque puce de filtrage et le système de contrôle. Pour cela, une séquence pseudo alétoire est inscrite dans les registres des puces puis lue en retour et comparée.

Tests de distribution Ces tests permettent de déterminer précisément si une piste de la carte ou une bille BGA (Ball Grid Array) des puces FPGA assemblées sur la carte TFB est détériorée. Des données de nature pseudo-alétoire sont envoyées à la carte par le Test Fixture puis récupérées par ce dernier. Les erreurs peuvent être identifiées pour chaque ligne éléctrique (en entrée 32 lignes sur 3 bits, en sortie 32 lignes sur 2 bits). A partir des résultats de ces tests, un diagramme de l'oeil de la carte est établi en réglant les phases d'horloge des FPGAs pour la génération des données et la capture de ces dernières en les comparant à l'horloge de la carte. Un diagramme de l'oeil pour les données d'entrée et de sortie est donné en Figure 1.31.

Validation du traitement du signal Ces tests permettent de verifier la bonne fonctionnalité du filtre TFB. Les échantillons de sorties des filtres sont comparés à ceux prédits mathématiquement (par l'intermédiaire d'une modélisation mathématique du système de filtrage) pour un jeu de 80 « patterns » de test obtenus pour différents réglages des paramètres du filtre (entrée statique, fréquence du DDS, phase du DDS, facteur d'échelle pour la requatification finale). A la suite de cela une analyse spectrale est réalisée (à l'aide d'un programme réalisant une Transformée de Fourier des coefficients de corrélation calculés par le Test Fixture). Elle permet de vérifier les principales caracéristiques du système de filtrage comme le déplacement en fréquence des SBs, l'ondulation dans la bande passante et l'atténuation des repliements spectraux.

Tests de consommation Ces test sont réalisés par l'intermédiaire d'une routine du software du Test Fixture. Elle permet d'afficher la consommation de la carte de filtrage pour les différentes tensions utilisées : 3.3 V pour les IOs communication, configuration et sortie de filtre, 1.8 V pour les IOs restantes (signaux d'entrées), 1.5 V ou 1.2 V pour le coeur du filtre suivant la technologie de la puce. La consommation totale de la carte peut alors en être déduite.

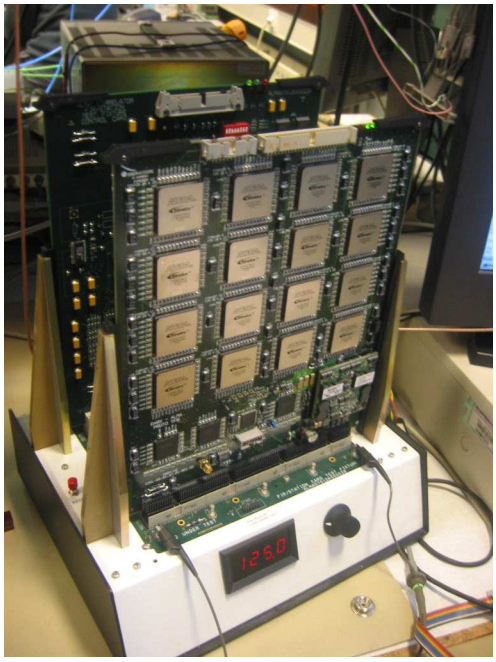
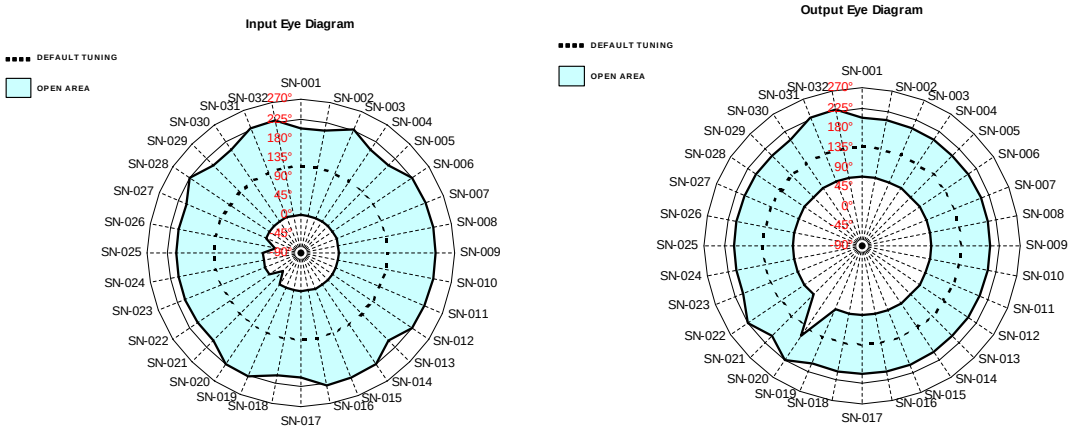


FIG. 1.30 – *Système de test des cartes*

7.2 Problème thermique et évolution du design

Une première carte prototype peuplée de puce Stratix I $EP1S40$ (16 puces FPGA) a été développée (technologie 0.13 μm, coeur alimenté en 1.5 V, 40000 LEs disponibles - LE : Logic Element). La consommation totale de cette carte prototype s'élève à environ 150 W. Cette consommaton a été jugé trop importante pour le projet ALMA compte tenu de la difficulté d'évacuer des calories dans les racks situés à 5000 m d'altitude où l'air est raréfié. Dans le but de diminuer cette consommation, une fonction nommé « low power » a été développée. Elle permet « d'éteindre » un filtre lorsqu'il n'est pas sollicité par le système de contrôle. En fait, cette fonction permet de bloquer le DDS et injecte des échantillons nuls en entrée ce qui permet de diminuer la consommation de la carte lorsque toutes ses ressources de filtrage ne sont pas utilisées.
Il a aussi été envisagé de changer de technologie de puce afin de diminuer de manière plus drastique la consommation. Une solution HardCopy ALtera (technologie 0.13 μm, 1.5 V de coeur) et une solution Stratix II Altera (technologie 90 nm, 1.2 V de coeur) ont été retenues. La solution HardCopy permet de réduire la consommation de coeur de la puce d'environ 40 % par rapport à une solution Stratix I (donnée constructeur). Cette solution permet de diminuer la complexité de la puce en implémentant de manière irréversible (équivalent d'un ASIC) un design préalablement validé sur un FPGA par exemple. Ainsi, le nombre de couches nécessaires à l'implémentation de la fonctionnalité et la surface de silicium nécessaire sont réduits, un grand nombre de couches dans le FPGA étant alloué à la reconfigurabilité de ce dernier et certaines ressources étant abandonnées.
Afin d'évaluer la solution optimale, une simulation de la consommation des puces Stratix I, HardCopy et Stratix II contenant la fonctionnalité TFB a été effectuée à l'aide de l'outil de conception Quartus II (cf. section 2 en p. 127 du Chapitre 5). Une simulation post synthèse (outil Modelsim) doit être effectuée afin d'obtenir un fichier permettant de réaliser la simulation fournissant l'estimation de la consommation de la fonctionnalité implémentée (outil Quartus II).



(a) Diagramme pour les signaux d'entrée (b) Diagramme pour les signaux de sortie

FIG. 1.31 – *Digramme de l'oeil pour les cartes* $SN-01$ *à* $SN-32$

Dans un second temps, le même design a été implémenté sur une carte de dévelopement Stratix II afin d'estimer la consommation de ce design ; la consommation pour la technologie Stratix I a été mesurée sur la carte TFB. Les résultats concernant la solution HardCopy sont obtenues simplement en soustrayant 40 % de la consommation obtenue pour le Stratix I (cf. tableau 1.6), diminution annoncée par Altera.

TAB. 1.6 – *Comparaison des solutions envisageables (Simulation)*

	Stratix I	HardCopy (−40% SI)	Stratix II
Simulation (Quartus II)			
Puissance estimée pour une puce TFB	7.25 W	4.58 W	4.72 W
Extrapolation pour une carte TFB	147.5 W	97.3 W	100 W
Mesure			
Puissance mesurée pour une puce TFB	7.48 W	/	5.5 W
Extrapolation pour une carte TFB	151 W	/	114 W

Malgré les résultats de moins bonne qualité de la solution Stratix II (en effet le programme ayant été implémenté sur une carte de développement, la consommation donnée prend en compte les différents éléments présents sur la carte, il est difficile alors d'en déduire la consommation exacte de la fonctionnnalité implémentée) par rapport à la version HardCopy estimée, elle est la solution retenue. La diminution de consommation affichée n'est pas négligeable, elle est de l'ordre de 25 %. De plus, contrairement à la solution HardCopy, la solution Stratix II permet la reconfiguration de la fonctionnalité et donc l'évolution du design.
Les puces Stratix II, tout comme les puces Stratix I, sont intégrées dans des boitiers BGA. La densité choisie pour la version Stratix II correpond à la puce $EP2S30$ qui comporte 30000 LEs, quantité suffisante à l'implémentation de la fonctionnalité TFB. La cellule élémentaire de la famille Stratix II est appelée ALM (Adaptative Logic Module). Chaque ALM contient entre autres, une variété de ressources dont la base est la LUT (Look Up Tabe), deux additionneurs et deux bascules flip-flop. La Figure 1.32 montre les différentes configurations que peut supporter une simple cellule ALM.
Des fonctions logiques possédant jusqu'à 7 entrées ainsi que des fonctions complexes arithmétiques peuvent être implémentées dans une ALM. Alors que la cellule de base de la famille Stratix

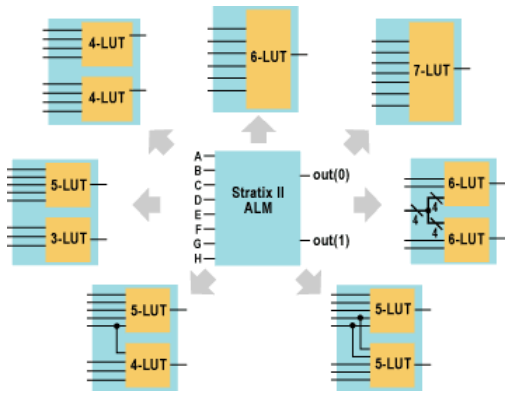
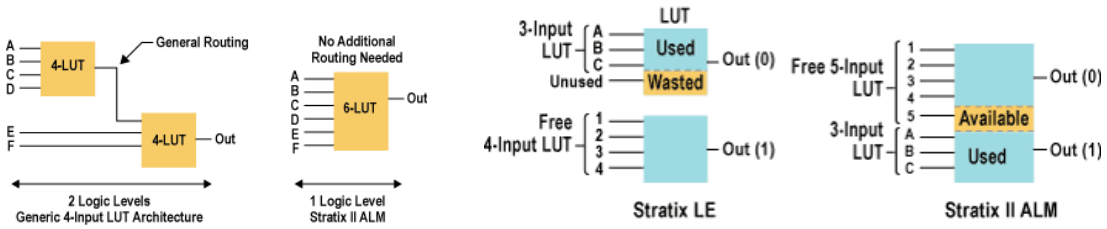


FIG. 1.32 – *Configurations possibles pour une ALM*

I était constituée d'une LUT à 4 entrées et d'une flip-flop, l'utilisation d'ALM permet d'améliorer certaines performances comme illustré sur les Figures 1.33 : routage amélioré d'une cellule à une autre et partage des cellules amélioré.



(a) Performance accrue de la cellule ALM (b) Efficacité accrue de la cellule ALM

FIG. 1.33 – *Comparaison des cellules LE (Stratix I) et ALM (Stratix II)*

Pour les raisons résumées précédemment, l'implémentation de la fonctionnalité développée dans la puce Stratix II entraîne une diminution du nombre de ressources nécessaires. En Figure 1.34 sont présentées les différents types de ressources disponibles dans ce composant.
Une comparaison avec la version précédente des différents types de ressource utilisés est donnée en tableau 1.7 avec le pourcentage d'occupation de la puce dans chacun des cas.
Une carte prototype Stratix II a alors été développée puis testée à l'aide du système décrit précédemment. Après validation du design de la carte, cette dernière a été sélectionnée pour la production des 512 cartes TFB requises pour le Corrélateur ALMA. La consommation mesurée de cette nouvelle carte s'élève à 75 W en moyenne, ce qui correspond à une amélioration de l'ordre de 47 % par rapport à celle peuplée de puces Stratix I. Une mesure de la température de jonction de chaque puce a aussi été réalisée afin d'évaluer la durée de vie de chacun des composants. Un seuil limite de température au-dela duquel la tenue en fréquence de la fonctionnalité implémentée n'est plus garantie a été fixé à 130 °C par Altera. En Figure 1.35 apparaît l'évolution de cette température de jonction en fonction du flot d'air délivré par un ventilateur).
Le gradient de température de la carte est notable du fait que les puces soient disposées en une matrice de $4 \cdot 4$ (la température de l'une amenant un accroissement de la température de la suivante) et du fait de la position du convertisseur DC/DC (Figure 1.26) qui possède une assez forte dissipation thermique. Pour limiter le réchauffement des puces (et ainsi augmenter

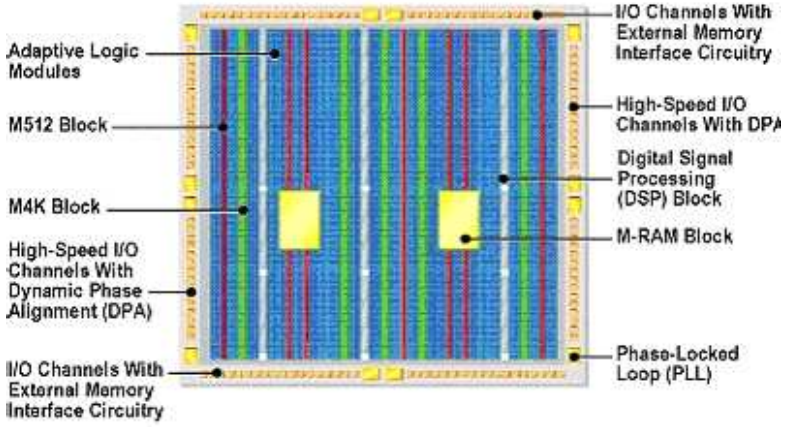


FIG. 1.34 – *Architecture d'une puce Stratix II*

TAB. 1.7 – *Utilisation des ressources des puces Stratix I et Stratix II*

	ALMs	LEs	M512	M4k	M-RAM	Mult. $9*9$ bit	PLLs
Stratix I (*EP1S40*)							
Ressources disponibles	-	41250	384	183	4	112	12
Ressources utilisées	-	19979	97	104	0	68	1
Pourcentage	/	*48%*	*25%*	*57%*	*0%*	*61%*	*17%*
Stratix II (*EP2S30*)							
Ressources disponibles	13552	-	202	144	1	128	6
Ressources utilisées	10970	-	96	103	0	70	1
Pourcentage	*81%*	-	*48%*	*72%*	*0%*	*55%*	*17%*

leur durée de vie) un système de ventilation a été mis en place dans les armoires qui contiennent les cartes de corrélation et de filtrage (Figure 1.36, la vitesse initiale de l'air prévue dans le Corrélateur est de l'ordre de 200 feet/min). Ces cartes de filtrage sont insérées horizontalement, avec d'autres cartes, dans les deux racks de gauche. Les deux racks de droite contiennent les cartes du corrélateur.

8 Conclusion

Ce chapitre a permis de présenter l'instrument ALMA et tout particulièrement le système de filtrage TFB.
Au début de la thèse, il existait du système TFB seulement le filtre. La première étape de mon travail a donc été de participer à la finalisation de ce système de filtrage. Dans un premier temps, l'achèvement de différents modules tels que la requantification dynamique ou encore les modes de fonctionnement secondaires a été réalisée. Une fois le système finalisé, il a fallu mettre en place et effectuer les divers tests visant à valider le routage de la carte et le fonctionnement du système de filtrage. Après validation de la fonctionnalité, des mesures de consommations ont été réalisée et ont révélés une consommation importante de la carte TFB (150 W pour la solution StratixI et 75 W pour la StratixII).
Au jour d'aujourd'hui, un modèle de dissipation thermique a été développé par le NRAO afin d'étudier le comportement théorique du système. A posteriori, une mesure de la distribution de

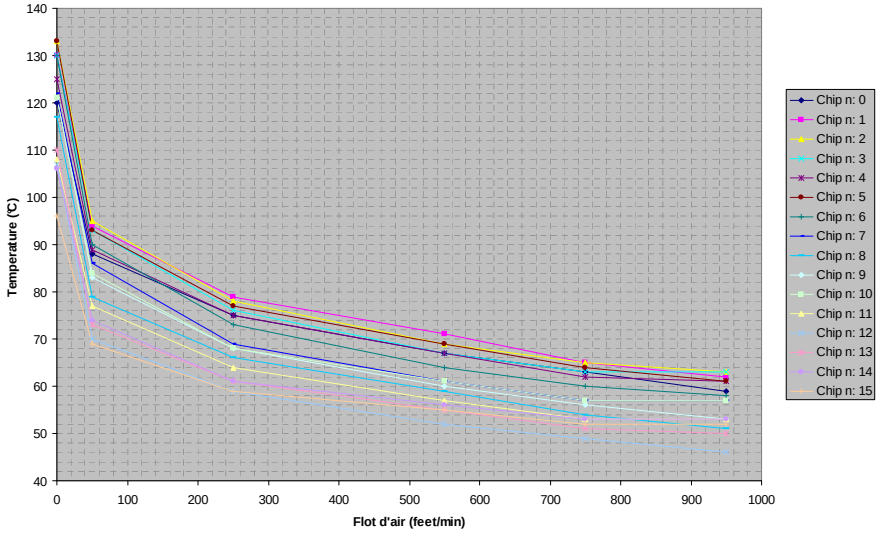


FIG. 1.35 – *Température de jonction en fonction du flot d'air délivré par les ventilateurs*

température a été effectuée sur le site de test à Charlottesville. Suite à ces études, la valeur optimale de la température de jonction à travers la carte TFB a été jugé inférieure ou égale à 70 °C. Trois ventilateurs par rack ont été prévus.
Nous avons cependant envisagé, en parallèle aux travaux concernant l'optimisation du flot d'air à travers les racks de l'armoire, un redesign du système de filtrage (cf. Chapitre 2, 3 et 4). Cette étude constitue la majeur partie du travail effectué durant cette thèse. Différentes architectures du système de filtrage ont été alors investiguées afin de proposer une solution optimale en terme de consommation d'energie et de dissipation thermique. Dans un premier temps a été envisagé le remplacement du premier étage de filtrage TFB. Ensuite, l'idée de l'utilisation de filtre de type RII pour le second étage TFB a été étudiée. Enfin un redesign complet du système basé sur une architecture polyphasée a été mené.

FIG. 1.36 – *Armoire Corrélateur qui contiendra* 1/8 *des cartes de l'ensemble du système*

Chapitre 2

Filtre à Haut Taux de Décimation Appliqué aux Signaux Large Bande

1 Introduction

Dans le chapitre présentant, entre autres, le système de filtrage ALMA, nous avons exposé l'architecture et le rôle des deux étages de filtrage. Dans un souci d'optimisation de la consommation de la carte TFB et suite au problème thermique rencontré sur la carte TFB (température de jonction des puces), une étude a été lancée afin d'établir les diverses possibilités pouvant s'appliquer à notre type de signal (entrée démultiplexée) et au traitement souhaité. Le fait que le premier étage de filtrage TFB soit un filtre agissant comme un étage de décimation – avec une large bande de transition – amène à considérer la solution du filtre *Cascaded Integrator Comb* (CIC). Le filtre RIF de l'actuel premier étage utilise 1730 ALMs.
Une définition du filtre CIC ouvre ce chapitre, donnant l'utilisation et l'architecture électronique classique de ce type de filtre. Différentes structures, dérivées de cette définition, sont ensuite présentées dans le but de trouver celle qui s'applique le mieux à notre utilisation : architecture récursive, non-récursive, à entrée démultiplexée ou encore une architecture multi-étages. Suit une étude d'adaptation de la structure au cas ALMA ainsi qu'une d'optimisation de cette dernière. Enfin les différentes architectures retenues sont comparées entre elles en terme de ressources logiques dans l'optique de la diminution de la dissipation thermique du système de filtrage. Aucun autre type de ressources disponibles dans la puce que l'ALM n'est utilisé afin de rendre plus aisée la comparaison des architectures.

2 Le filtre CIC

2.1 Présentation

Il a été montré que le CIC est un élément efficace dans les systèmes à haut taux de décimation [19]. Ce type de filtre est habituellement utilisé comme premier étage de décimation d'un convertisseur analogique-numérique de type $\Sigma\Delta$. Il est à noter que l'implémentation de ce filtre ne nécessite aucune multiplication, contrairement à un filtre RIF, ainsi qu'une utilisation limitée des ressources de stockage. Cette simplicité architecturale conduit pourtant à une sélectivité hors bande d'intérêt très intéressante, caractéristique essentielle dans la synthèse des filtres. De ce fait, une telle architecture peut s'avérer être un atout en terme d'amélioration de la consommation comparée à une structure de filtre RIF remplissant des spécifications identiques.

La transformée en z de la fonction de transfert du filtre CIC est la suivante :

$$H(z) = \left(\sum_{k=0}^{D-1} z^{-k}\right)^N \tag{2.1}$$

$$= \left(\frac{1-z^{-D}}{1-z^{-1}}\right)^N \tag{2.2}$$

où D est le facteur de décimation et N l'ordre du filtre. La réponse impulsionnelle de type RIF (cf. équation 2.1) est constituée de D coefficients tous égaux à 1. Cette suite géométrique peut s'écrire sous la forme d'un filtre RII (cf. équation 2.2) qui peut être décomposé en une partie Intégrateur (le dénominateur) et une partie Comb ou « peigne » (le numérateur). Le couple de paramètres (D, N) permet d'obtenir différents gabarits de filtre illustrés en Figure 2.1.

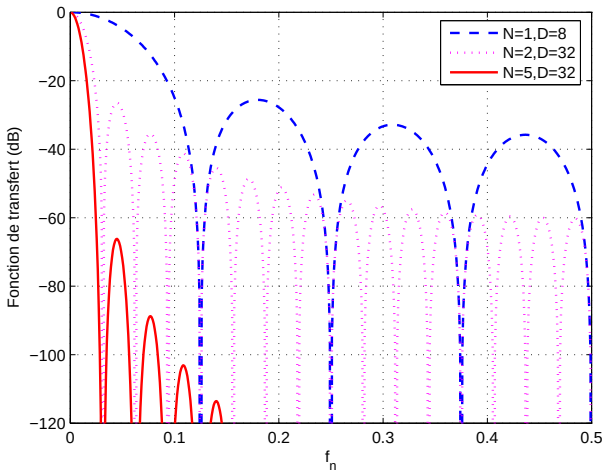


FIG. 2.1 – *Réponse en amplitude du CIC pour différents jeux de paramètres*

L'augmentation de l'ordre N du filtre entraîne certes une amélioration de l'atténuation, mais au détriment d'une chute plus rapide de l'amplitude dans la Bande Passante (BP). Cette figure met aussi en évidence la répartition des zéros sur l'axe des fréquences normalisées apparaissant à chaque $\frac{k}{D}$, $\forall k \in [1, \frac{D}{2}]$. La linéarité de la phase à l'intérieur de chaque lobe mérite d'être soulignée.
Le fait que la réponse en fréquence du CIC soit seulement tributaire des paramètres D et N peut être considéré comme un inconvénient, limitant les gabarits possibles de filtre. Un autre phénomène très important à prendre en compte lors d'un processus de décimation est celui du repliement, ou *aliasing*, afin d'éviter toute pollution de la bande d'intérêt. La Figure 2.2 illustre le phénomène de repliement.

Les bandes repliées dans la bande d'intérêt après décimation sont délimitées par les lignes hachurées situées dans les régions $[\frac{k}{D} - BI, \frac{k}{D} + BI]$ de largeur $2 \times BI$ (BI étant la bande d'intérêt). Afin de déterminer si l'atténuation du filtre est suffisante, il suffit de calculer cette atténuation à la fréquence critique normalisée $f_c = \frac{1}{D} - f_{BI}$ (Figure 2.2, position du cercle). Elle correspond au pire cas d'aliasing pouvant se produire. La formule permettant de calculer ce point atténuation

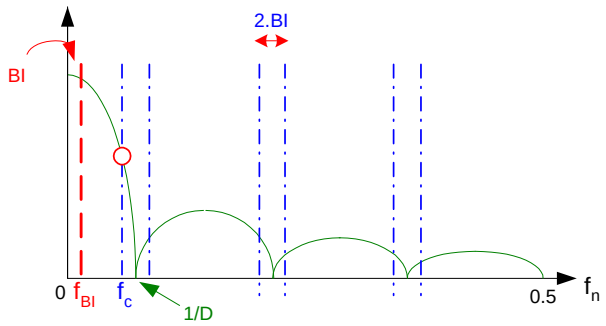


FIG. 2.2 – *Zoom sur le phénomène de repliement , f_n : fréquence normalisée*

critique est la suivante et est tirée du module de $H(\omega)$:

$$H(\omega) = \left(\frac{\sin(\omega\pi D/2)}{\sin(\omega\pi/2)}\right)^N \tag{2.3}$$

$$\text{Soit} \quad Att(\omega_c) = 20 \cdot N \log\left(\frac{\sin(\omega_c\pi D/2)}{\sin(\omega_c\pi/2)}\right) \tag{2.4}$$

avec $\omega_c = 2\pi f_c$.
Le tableau 2.1 contient les différentes atténuations obtenues pour différents ordres N et différentes valeurs de f_{BID}, $f_{BID} = f_{BI} \cdot D$ (Figure 2.2).

TAB. 2.1 – Atténuation à f_c pour différentes valeurs de paramètres

	N=1	N=2	N=3	N=4	N=5
f_{BID}=1/4	10dB	20dB	31dB	42dB	52dB
f_{BID}=1/8	17dB	34dB	51dB	68dB	84dB
f_{BID}=1/16	23dB	47dB	70dB	93dB	116dB
f_{BID}=1/32	28dB	58dB	86dB	115dB	144dB
f_{BID}=1/64	35dB	71dB	105dB	140dB	175dB

Le tableau 2.2 quant à lui présente la chute dans la bande passante pour les mêmes jeux de paramètres.

TAB. 2.2 – Atténuation à f_{BI} pour différentes valeurs de paramètres (chute dans la bande passante)

	N=1	N=2	N=3	N=4	N=5
f_{BID}=1/4	.91dB	1.82dB	2.74dB	3.65dB	4.56dB
f_{BID}=1/8	.22dB	.45dB	.67dB	.9dB	1.12dB
f_{BID}=1/16	.06dB	.11dB	.17dB	.22dB	.28dB
f_{BID}=1/32	.01dB	.03dB	.04dB	.06dB	.07dB
f_{BID}=1/64	0dB	.01dB	.01dB	.01dB	.02dB

Ces tableaux illustrent le lien entre l'atténuation et la chute dans la BP en fonction des paramètres du CIC sélectionnés. Il est à noter que le produit de la bande d'intérêt f_{BI} par le facteur de

décimation caractérise la complexité d'implémentation du filtre. Pour une atténuation donnée, plus ce produit est faible, plus l'ordre du filtre est bas.
Par exemple, dans le cas d'une décimation par 32 et pour une bande d'intérêt correspondant à 1/128 de la bande originale (cas ALMA, aboutissant à $f_{BID} = 1/4$), pour atteindre une atténuation de $47dB$ il est nécessaire d'employer un filtre d'ordre $N = 5$ (cf. tableau 2.1). Dans ce cas la chute dans la bande passante atteint $4.56dB$ (cf. tableau 2.2), ce qui est un résultat non négligeable nécessitant une compensation obtenue par l'application d'un deuxième étage de filtrage.

2.2 Réalisation d'un CIC

L'architecture classique d'un filtre CIC décimateur d'ordre 2 est représentée en Figure 2.3 (un filtre d'ordre N est composé d'une cascade de N blocs Intégrateur et Comb).

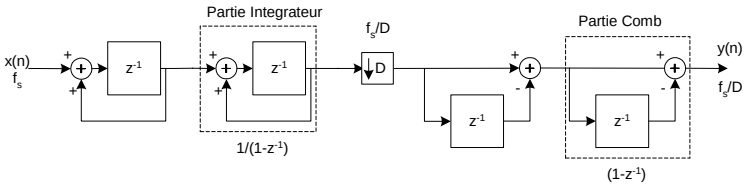


FIG. 2.3 – *Architecture classique du filtre CIC*

Les deux premiers étages sont les parties « Intégrateur » (I) et les deux dernières les parties « Comb » (C). Toutes les deux sont séparées par l'opération de décimation. Les premières parties opèrent à un taux d'échantillonnage élevé f_s, alors que les parties Comb opèrent à un taux réduit f_s/D. L'intégrateur est en fait un accumulateur correspondant à un filtre à pôle unique (situé en $z = 1$) et le Comb un différenciateur (ou dérivateur). Le fait d'effectuer la décimation entre les deux étages permet d'utiliser seulement une cellule retard au lieu de D (2.2) dans la partie différentiateur, ceci en prenant compte des identités remarquables des sytèmes multicadences [20]. Ces dernières sont illustrées en Figure 2.4 où $G(z)$ est une fonction de transfert rationnelle, le bloc M un décimateur de facteur M et le bloc L un interpolateur de facteur L.

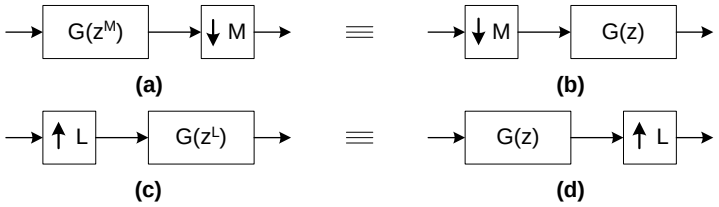


FIG. 2.4 – *Identités remarquables des systèmes multicadences*

En (a) le décimateur est précédé par la fonction de transfert. Cette cascade est équivalente à celle représentée en (b), où $G(z)$ opère a un taux M fois inférieur à celui de $G(z^M)$. L'équivalence se vérifie aussi pour (c) et (d) dans le cas de l'interpolation.

Il est à souligner que chaque partie accumulateur possède une rétroaction positive à coefficient unitaire. Cette rétroaction est une source de problèmes car elle entraine un accroissement de la taille des registres au fur et à mesure des étages et donc une augmentation de la complexité d'implémentation. Hogenauer a présenté une méthode permettant la limitation de la taille des différents registres du filtre CIC (des différents étages I et C) sans dérérioration du signal traité,

et donc de l'encombrement de ce dernier [19]. La limite supérieure des registres (*MSB* - Most Significant Bit) doit supporter l'amplitude attendue à la sortie du filtre CIC. Pour la déterminer, le gain du filtre ainsi que le nombre de bits utilisés au codage du signal d'entrée sont nécessaire. De (2.1) se déduit le gain du filtre CIC de paramètres D et N :

$$G = \left(\sum_{k=0}^{D-1} h_{cic}(k) \right)^N = D^N \tag{2.5}$$

On en déduit le nombre de bits suffisant pour coder le signal de sortie, le MSB :

$$B_{max} = \lceil N log_2 D + inBit \rceil \tag{2.6}$$

où $inBit$ est le nombre de bits permettant de coder le signal d'entrée et $\lceil x \rceil$ correspond à l'entier le plus proche supérieur à la valeur x.
B_{max} est la limite supérieure de chaque étage, que ce soit dans la partie intégrateur (bien que la sortie de cette dernière diverge) ou Comb (dérivateur) ; l'ensemble étant perçu comme une « boite noire » (filtre RIF) ayant une énergie finie.
La divergence de la sortie des étages I provoque cependant un dépassement du registre de chaque accumulateur. Ce phénomène peut cependant être contourné en utilisant une arithmétique basée sur le complément à 2 pour l'implémentation qui permet le *wrap-around* (lien) entre la valeur la plus positive et la plus négative lors de l'apparition d'un dépassement du registre (opposé à la notion de saturation où, lors du dépassement, la donnée est définitivement perdue, Figure 2.5).

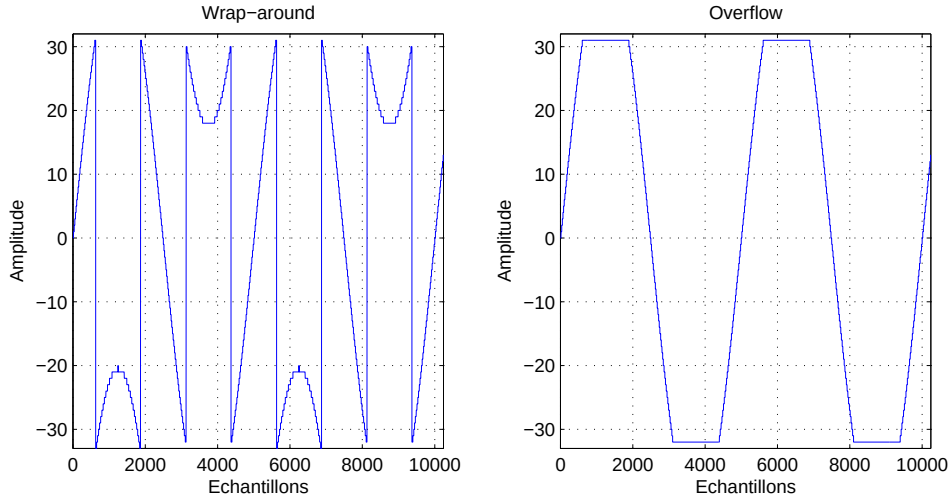


FIG. 2.5 – *Wrap-around et Saturation (signal signé codé sur 6 bits)*

Dans la suite du chapitre, la notion de *wrapped adder* est utilisée, elle renvoie à la notion de *wrap-around* appliquée à un additionneur.
L'utilisation de ce résultat associé à la limite haute des registres calculée précédemment permet d'assurer l'intégrité du signal à la sortie du filtre.
Une étude concernant les LSBs des registres a aussi été menée en [19]. Cette étude, basée sur une étude statistique des caractéristiques du filtre CIC, fournit une méthode permettant de limiter le nombre de bits utilisés dans chaque registre ou étage I-C à l'aide de troncatures ou d'arrondis, tout en minimisant l'erreur produite en sortie. Le nombre de LSBs supprimés doit évoluer de

manière croissante d'étage en étage.
Voici les équations permettant le calcul du nombre de LSBs à éliminer pour chaque étage j (B_j avec N étages I et N étages C) d'après [19] :

$$B_j = \lfloor -log_2 F_j + log_2 \sigma_{T_{2N+1}} + \frac{1}{2} log_2 \frac{6}{N} \rfloor \tag{2.7}$$

où N est l'ordre du filtre avec

$$\begin{cases} F_j^2 = \sum_k h_j^2(k) \text{ pour } j \in [1, 2N] \\ F_j^2 = 1 \text{ pour } j = 2N+1 \end{cases} \tag{2.8}$$

h étant la réponse impulsionnelle et j caractérisant l'étage étudié
et

$$\sigma^2_{T_{2N+1}} = \sigma^2_{2N+1} F^2_{2N+1} \tag{2.9}$$

$$\text{avec } \sigma^2_{2N+1} = \frac{1}{12} \cdot 2^{B_{2N+1}} \tag{2.10}$$

$$\text{et } B_{2N+1} = B_{max} - B_{out} + 1 \tag{2.11}$$

B_{max} étant la limite supérieure des registres et B_{out} le nombre de bits souhaités en sortie, fixé par l'utilisateur. $\sigma^2_{T_{2N+1}}$ représente la variance totale du système constitué de $2N+1$ étages.

2.3 Architectures alternatives au filtre CIC classique

Dans l'optique d'améliorer les performances de filtrage ou encore de minimiser l'utilisation des ressources logiques, une étude exhaustive des différentes architectures électroniques envisageables pour l'implémentation du filtre CIC a été menée. Chacune d'entre elles posséde son propre champs d'application aboutissant à l'amélioration de telle ou telle caractéristique de filtrage. Dans un premier temps, une architecture permettant d'améliorer les performances de filtrage est présentée. Elle est suivie par la description de deux architectures dont le but est la diminuation des ressources logiques.

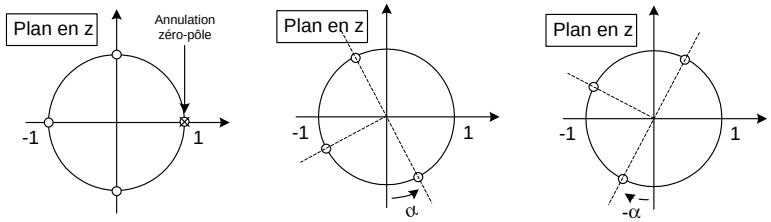
2.3.1 Le filtre CIC modifié '*rotated-sinc*'

Cette structure vise à redistribuer, sur le cercle unité, les zéros de la fonction de transfert du filtre CIC (2.1). De cette façon, une nouvelle fonction de transfert est définie, permettant d'améliorer l'atténuation de la bande repliée du filtre pour un ordre donné [21]. La Figure 2.6(a) fait apparaître le placement du pôle et des zéros d'un filtre CIC d'ordre 1 sur le cercle unité, ces derniers répondant à : $z_k = e^{j\frac{2\pi}{D}k}$ $k \in [0, D-1]$.
En appliquant une rotation de α radians à chaque zéro, la répartition devient celle présentée sur la Figure 2.6(b) qui correspond à l'équation :

$$H_1 = \frac{1 - z^{-D} e^{j\alpha D}}{1 - z^{-1} e^{j\alpha}} \tag{2.12}$$

De la même manière, en appliquant la rotation opposée (Figure 2.6(c)), on obtient :

$$H_2 = \frac{1 - z^{-D} e^{-j\alpha D}}{1 - z^{-1} e^{-j\alpha}} \tag{2.13}$$



(a) positionnement des zéros et du pôle (b) rotation α (c) rotation $-\alpha$

FIG. 2.6 – *Position des zéros d'un filtre CIC dans le plan en z, $D = 4$*

Le filtre CIC modifié est la cascade des deux filtres H_1 et H_2 dont les coefficients sont complexes conjugués. Le filtre qui en résulte comporte donc des coefficients réels.

$$H_T = H_1 \cdot H_2 = \frac{1 - 2cos(\alpha D)z^{-1} + z^{-2D}}{1 - 2cos(\alpha)z^{-1} + z^{-2}} \tag{2.14}$$

Le paramètre α doit être choisi de manière à améliorer l'atténuation des bandes repliées définies par $[\frac{k}{D} - BI, \frac{k}{D} + BI]$ où BI représente la bande d'intérêt. Un bon choix quant à la détermination de ce paramètre est fonction de la bande d'intérêt et de la fréquence d'échantillonnage :

$$\alpha = q2\pi\frac{BI}{f_s} = q2\pi f_{BI} \tag{2.15}$$

q prenant des valeurs proches de 1.
Les zéros sont désormais placés aux positions $\frac{k}{D} \pm \frac{\alpha}{2\pi}$ (Figure 2.7).

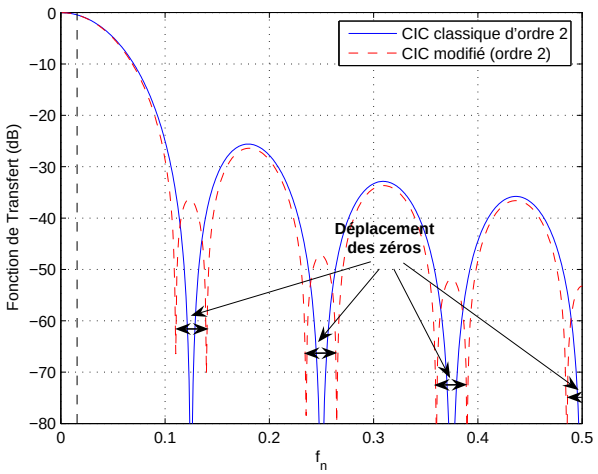


FIG. 2.7 – *Position des zéros de la fonction de transfert du filtre CIC modifié $D = 8$, $q = 0.95$*

L'augmentation de l'atténuation ne devient intéressante que lors d'une cascade du CIC modifié (d'ordre 2 (2.14)) dit *rotated-sinc* avec un étage CIC classique d'ordre N afin d'obtenir un filtre CIC d'ordre $N + 2$ amélioré. En Figure 2.8 apparaît l'amélioration en terme d'atténuation apportée par l'introduction de la structure modifié dans la cascade d'étages CIC. Deux fonctions de transfert d'un filtre CIC d'ordre 3 classique et à structure modifiée ($N = 1 + 2$) sont tracées.

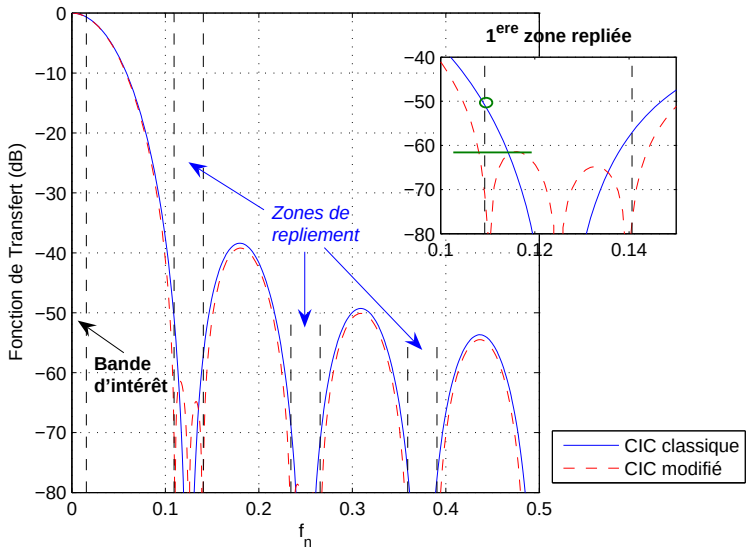


FIG. 2.8 – *Comparaison CIC et cascade rotated-sinc + CIC*

L'amélioration obtenue par cette cascade avoisine les 10 dB sans pour autant augmenter la chute dans la bande d'intérêt (cf. zoom sur la région de repliement).
La réalisation électronique de cette structure est donnée en Figure 2.9, la base reste celle du filtre CIC classique (Figure 2.3). L'apparition des deux coefficients $b = 2\cos(\alpha)$ et $c = 2\cos(\alpha D)$ complexifie légèrement l'architecture et donc son implémentation (utilisation de multiplieurs). De plus le format des données en sortie s'accroit très rapidement en fonction de l'ordre du filtre.

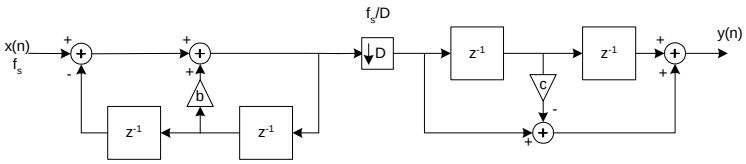


FIG. 2.9 – *Architecture récursive d'un filtre CIC « rotated-angle »de deuxième ordre*

Toujours en [21], d'autres modifications sont présentées afin d'améliorer par exemple la chute dans la BP par l'intermédiaire de cellule retard ou de structure dite *sharpened*. Le filtre total, amélioré, est obtenu par cascade de toute ces architectures modifiées respectant la formule suivante :

$$H_n(z) = H_b^{n+1}(z) \sum_{k=0}^{n} \frac{(n+k)!}{n! \cdot k!} \left[1 - H_b(z)\right]^k \tag{2.16}$$

où H_b est un filtre de type CIC rotated-sinc. La cascade engendrée par cette équation complexifie considérablement la structure de filtrage, mais en améliorant les caractéristiques du filtre décimateur.

2.3.2 Architecture polyphasée

La structure polyphasée adaptée au filtre CIC décimateur permet d'opérer à des taux d'échantillonnage beaucoup moins élevés que celui du signal d'entrée et de ce fait de diminuer la consommation [22]. La décomposition polyphasée est basée sur un traitement parallèle des données assurant la décimation [23]. Il ne s'agit ici que de modifier l'architecture électronique, la fonction de transfert restant la même.
Soit une fonction de transfert $H(z) = \sum_{n=0}^{+\infty} h(n)z^{-n}$ caractéristique d'un filtre numérique qui peut s'écrire de la manière suivante :

$$\begin{aligned} H(z) &= \left[h(0) + h(2)z^{-2} + h(4)z^{-4} + h(6)z^{-6} \ldots\right] \\ &\quad + z^{-1}\left[h(1) + h(3)z^{-2} + h(5)z^{-4} \ldots\right] \end{aligned} \tag{2.17}$$

En utilisant les abréviations suivantes :

$$\begin{aligned} F_0(z) &= \sum_{n=0}^{+\infty} h(2n)z^{-n} \\ F_1(z) &= \sum_{n=0}^{+\infty} h(2n+1)z^{-n} \end{aligned} \tag{2.18}$$

l'équation 2.17 se transforme en :

$$H(z) = F_0(z^2) + z^{-1}F_1(z^2) \tag{2.19}$$

Ce résultat correspond à une décomposition polyphasée à deux éléments, la fréquence d'échantillonnage est alors divisée par 2. Les éléments F_0 et F_1 sont appelés les composants polyphases. Une décomposition à D éléments de la réponse impulsionnelle du filtre CIC peut être obtenue de manière identique après développement de cette dernière :

$$H(z) = \left(\sum_{k=0}^{D-1} z^{-k}\right)^N = \sum_{i=0}^{D-1} z^{-i}F_i(z^D) \tag{2.20}$$

où $F_i(z)$ sont les composants de cette décomposition opérant à f_s/D. En utilisant les propriétés énoncées en Figure 2.4, la schématique représentée en Figure 2.10 est établie.

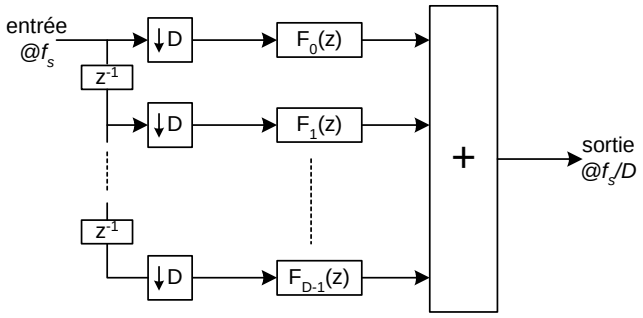


FIG. 2.10 – *Décomposition polyphasée du CIC décimateur (D,N)*

Une étude plus approfondie de ce type de structure est détaillée en section 2.1, p. 109 du Chapitre 4, elle n'est que brièvement décrite dans cette section.

Dans la littérature, cette structure polyphasée est utilisée dans le but de réduire la vitesse de calcul nécessaire au filtre et ainsi diminuer sa concommation. Elle est souvent chainée avec un autre filtre décimateur (CIC, RIF). La décimation attendue $D = D1 \cdot D2$ est alors obtenue par cascade de la décimation $D1$ de la structure polyphase avec celle du deuxième filtre décimateur $D2$, ceci dans le but de diminuer la complexité du filtre équivalent à cette cascade.
Cette structure permet donc d'opérer à un taux d'échantillonnage réduit dès l'entrée. En contrepartie, elle nécessite l'utilisation de coefficients non unitaires (coefficients obtenus après développement de la réponse impulsionnelle (2.20) dans le cas d'un ordre N supérieur à un) autrement dit de ressources mémoire et de multiplieurs, au contraire d'un filtre CIC classique. Cette structure est donc un compromis entre l'efficacité d'un filtre RIF et l'optimisation amenée par un filtre CIC.

2.3.3 Architecture non-récursive

La structure proposée dans cette partie est dérivée de l'architecture classique répondant à l'équation 2.1. Par simple factorisation, en s'assurant que $D = 2^M$, $M \in \mathbb{N}$, on peut écrire l'équation suivante [24] :

$$H(z) = \left(\sum_{k=0}^{D-1} z^{-k} \right)^N = \prod_{i=0}^{(log_2 D)-1} \left(1 + z^{-2i}\right)^N \tag{2.21}$$

Il en résulte une cascade de filtres RIF de faible ordre, tous de structures identiques mais opérant à des taux d'échantillonnages différents (Figure 2.11).

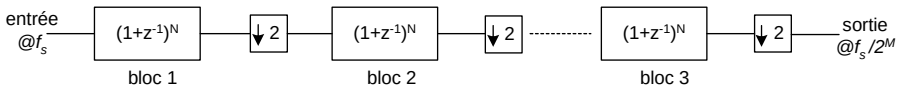


FIG. 2.11 – *Architecture non-récursive d'un filtre CIC*

L'accroissement de la taille des registres tout au long du traitement du signal est contenu en appliquant la formule délimitant la taille de sortie B_{out} d'un bloc. Cette formule est fonction de l'ordre N du filtre et du nombre de bits d'entrée du bloc B_{in} :

$$B_{out} = B_{in} + N \tag{2.22}$$

L'avantage d'une telle structure comparée à l'implémentation récursive classique est le caratère non récursif de sa structure interne conduisant à l'utilisation de registres de taille moins importante (spécialement dans les premiers étages).

3 CIC à entrée démultiplexée - Adaptation de l'architecture au projet ALMA

Comme énoncé dans la section 6 en p. 25 du Chapitre 1, le signal fourni par le digitizer ALMA au TFB est démultiplexée sur 32 voies à 125 Ms/s, entraînant l'utilisation d'une architecture démultiplexée. Dans cette section, certaines des architectures présentées en section 2.3 sont adaptées à ce flot de données. Dans l'optique d'une diminution de la consommation du système de filtrage TFB ALMA, la première idée fut de remplacer le premier étage du filtre ALMA TFB par un filtre CIC remplissant les mêmes spécifications : une atténuation de $47dB$,

une ondulation dans la bande passante inférieure à $0.2dB$ et un facteur de décimation égal à 32. L'inconvéniant d'un tel filtre est la chute induite dans la bande passante (ou « passband drop ») qui doit être compensée en fonction de son importance. Le fait que le signal en entrée du second étage doit être codé sur 8 bits (cf. section 6.1.2 en p. 6.1.2 du Chapitre 1) constitue une autre caratéristique de sortie du filtre.
En se basant sur le tableau 2.1, on détermine qu'un filtre CIC d'ordre 5 est nécessaire pour atteindre $47dB$ d'atténuation avec $f_{BID} = 1/4$ pour un facteur de décimation $D = 32$. La chute dans la bande passante obtenue pour de telles spécifications est égale à $4.56dB$ (cf. tableau 2.2). Cette dernière doit être compensée afin de répondre à la spécification d'ondulation dans la BP (le deuxième étage TFB est entre autre utilisé dans ce but).

3.1 Adaptation de l'architecture classique

Dans un premier temps, nous avons proposé le remplacement du 1er étage TFB par un filtre CIC d'ordre $N = 5$. Le schéma présenté en Figure 2.3 doit subir quelques modifications pour être compatible à une entrée démultiplexée. Seule la partie Intégrateur doit être adaptée ; la partie Comb étant précédée par l'opération de décimation, elle ne traite qu'une seule voie. Cette nouvelle structure est donc composée de 5 étages intégrateur cascadés possédant chacun une entrée sur 32 voies. La sortie de ces 5 étages Intégrateur s'effectue, après décimation, sur 1 voie qui nourrit les 5 étages comb cascadés similaires à ceux présentés en Figure 2.3.

3.1.1 Modification de l'étage intégrateur

L'entrée démultiplexée est composée de $V = 32$ voies nommées X_i pour $i \in [0, 31]$ (Figure 2.12).

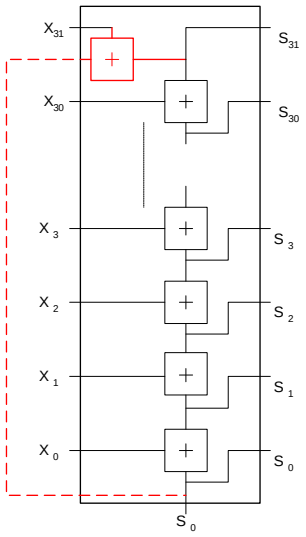


FIG. 2.12 – *Structure d'un intégrateur (accumulateur) modifié*

Chaque sortie d'étage intégrateur de S_0 à S_{31} est redirigée vers l'entrée du prochain, c'est ainsi que la cascade est effectuée. La décimation de facteur D est réalisée en redirigeant la sortie S_0 du 5^{eme} étage vers l'entrée des étages comb.

3.1.2 Taille des registres

En appliquant la méthode présentée en section 2.2 [19], l'encombrement du filtre peut être amoindri. En utilisant (2.6), la limite haute des registres ($2N+1$ registres, N I, N C plus le registre de sortie) est fixée à $B_{max} = 31$. Pour chaque étage (Intégrateur ou Comb), un certain nombre de LSBs peuvent être supprimés sans perte significative d'information en utilisant (2.7), (2.8) et (2.8) avec une sortie fixée à 8 bits. L'évolution de la taille des registres est représentée à travers la Figure 2.13. Au bas de chaque étage de la Figure 2.13, le nombre de LSBs supprimés est indiqué.

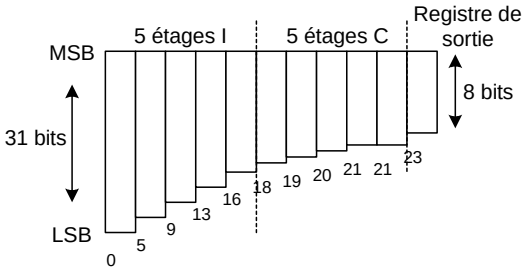
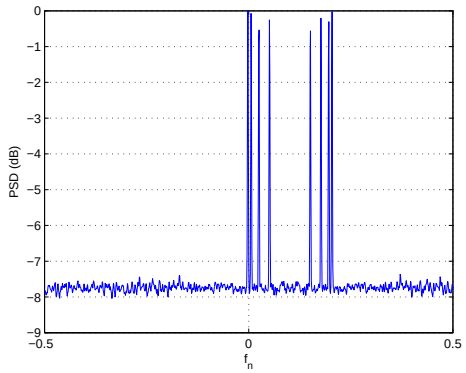


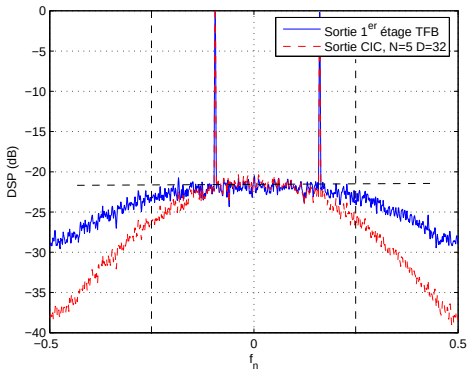
FIG. 2.13 – *Taille des registres, suppression des LSBs*

3.1.3 Modélisation

Le signal d'entrée est un signal complexe codé sur 6 bits provenant du DDS. Ce signal est composé d'un bruit blanc gaussien et de 4 sinusoïdes modélisant un signal de type radio-astronomique (la puissance des raies a été volontairement très accentuée par rapport à un signal réel). Sa représentation spectrale est donnée en Figure 2.14(a) en fonction de la fréquence normalisée.



(a) Signal issu du DDS et appliqué à l'entrée du filtre



(b) Sortie sdu filtre CIC et du 1^{er} étage TFB

FIG. 2.14 – *Spectres de sortie du filtre CIC et du* 1_{er} *étage TFB, après décimation*

Sur la Figure 2.14(b) sont tracées les réponses spectrales, après décimation, de la sortie du filtre CIC ($D = 32$ $N = 5$) ainsi que celle du 1^{er} étage de filtrage TFB pour comparaison. Ces spectres résultent de la combinaison des parties réelle et imaginaire du signal complexe ayant subies individuellement le même traitement. Le spectre de sortie ne présente aucun repliement spectral provenant des bandes adjacentes (Figure 2.2) dans la bande d'intérêt ($[-0.25, 0.25]$ sur l'échelle

des fréquences normalisées). Ils permettent de mettre en évidence la différence de chute dans la bande passante pour l'une et l'autre version. Dans le cas du filtre TFB, cette dernière avoisine les $2dB$ à la fréquence $f_n = 0.25$ correspondant à la limite de la bande d'intérêt finale – aprés application du 2^{nd} étage TFB. Cette chute est justement compensée par la bande passante du 2^{eme} étage de filtrage TFB. Le filtre CIC provoque, quant à lui, une chute de l'ordre de $5dB$ dans la bande passante à cette même fréquence (ce qui correspond à la valeur prédite théoriquement). Pour compenser celle-ci, la modification des poids de l'actuel second étage est nécessaire.

3.1.4 Implémentation

Cette structure a été traduite en langage VHDL de manière à estimer le nombre de ressources logiques nécessaires. Une vue RTL produite par le synthétiseur est présentée en Figure 2.15.

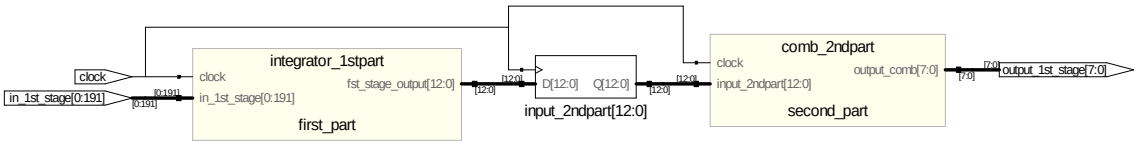


FIG. 2.15 – *Schématique RTL*

Cette structure utilise 6434 ALMs atteignant une fréquence de fonctionnement de 145 MHz. L'implémentation en hardware ne sera pas effectuée puisque cette structure n'apporte aucune amélioration notable à la solution utilisée actuellement. Au vu de ce résultat, un ordre 5 est inacceptable pour une architecture à entrée démultiplexée. Il faut choisir un produit bande utile - facteur de décimation ($f_{BID} = f_{BI} \cdot D$) plus faible. En effet dans le cas ALMA $f_{BI} = \frac{1}{128}$ (demi bande passante : 31.25 MHz, fréquence d'échantillonnage : 4 GHz), ce qui donne pour $D = 32$, $f_{BID} = \frac{1}{4}$. Cependant l'utilisation d'un f_{BID} plus faible conduit à repenser l'architecture du filtre TFB : un découpage en plusieurs étages (supérieur à deux) doit alors être envisagé.

3.2 Architecture polyphasée

Nous avons alors pensé utiliser une structure polyphasée (Figure 2.10) qui est naturellement adaptée à une entrée démultiplexée. En utilisant (2.20) avec les paramètres $D = 32$ et $N = 5$, les composants F_i sont calculés. Cela aboutit à une réponse impulsionnelle d'ordre $(D-1)N = 155$ avec a priori une trop importante répartition des coefficients. De plus, cette solution résulte en une architecture similaire à celle utilisée pour l'implémentation du 1^{er} étage TFB. De ce fait, elle n'est pas retenue pour l'implémentation du filtre CIC.

3.3 Architecture non-récursive

L'utilisation de la transformation présentée en section 2.3.3, pour les valeurs $D = 32$ et $N = 5$, conduit à une structure composée de 5 blocs d'ordre 5 cascadés. Chaque bloc est constitué d'une entrée démultiplexée pour correspondre au format d'entrée du système de filtrage ALMA. Le premier bloc possède une entrée sur 32 voies, le suivant, une entrée sur 16 voies et ainsi de suite jusqu'au 5^{eme} étage qui fournit une sortie composée d'une seule voie. Certes le filtre n'est plus récursif, mais il est désormais composé d'une grande quantité de bloc de base $\left(1 + z^{-1}\right)$.

Ceci entraîne une utilisation de ressources importante, principalement due aux premiers blocs massivement parallèles. Cette méthode n'est donc pas optimale dans le cas présent.

3.4 Architecture démultiplexée non-récursive

Cette architecture a été spécialement établie à partir de (2.1) afin d'être directement adaptable à une entrée démultiplexée. L'équation qui découle de cette étude [25] est la suivante :

$$C_D^N = S_D^N(1-z^{-1})^{N-1} + z^{-1}\sum_{i=1}^{N-1} A_i(D)C_D^{N-i}(1-z^{-1})^{i-1} \tag{2.23}$$

Un schéma de la structure électronique est présenté en Figure 2.16.

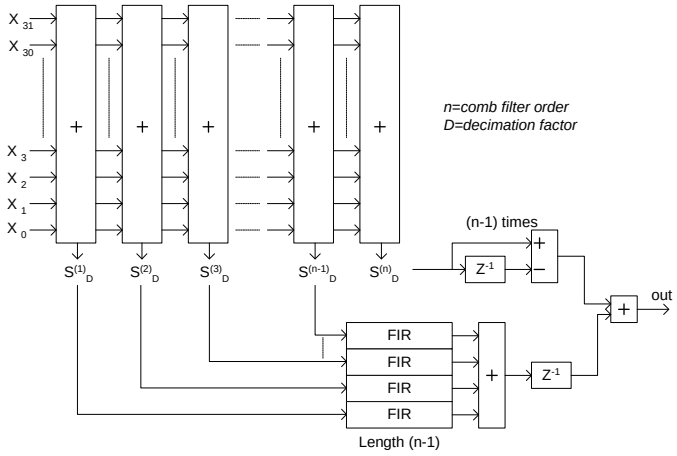


FIG. 2.16 – *Architecture non-récursive d'un filtre CIC (N, D)*

Le terme S_D^N correspond à la n^{ieme} sortie des blocs additionneurs de la figure. Le terme C_D^N correspond à la fonction de transfert du filtre CIC d'ordre N qui est une combinaison linéaire des filtres CIC d'ordres inférieurs dont les coefficients sont ceux des filtres RIF notés $A_i(D)$ ainsi que de la sortie du N^{ieme} bloc d'addition (2.23. $A_i(D)$ est une notation permettant d'exprimer le coefficient binomial suivant : $\begin{pmatrix} D+i-1 \\ i \end{pmatrix} = \frac{(D+i-1)!}{i!\,(D-1)!}$.
Cette architecture est l'équivalent d'un filtre CIC d'ordre 5 et de facteur de décimation 32. Comme cela a été décrit précédemment (Figure 2.16), l'architecture de cette solution est directement liée au nombre de voies parallèles présentes en entrée. Ici, la structure possédant 32 voies en parallèle résulte forcément en un facteur de décimation de 32. Contrairement à la structure classique (Figure 2.12), les blocs additionneurs sont une cascade de 31 additions successives, sans rétroaction entre l'entrée X_{31} et la sortie S_0. Les sorties de ces blocs sont ensuite redirigées vers le bloc de filtres RIF de faible ordre (égal à 4). Chaque sortie n ($n \in [1,5]$) correspond à celle d'un filtre CIC d'ordre n.

3.4.1 Taille des registres

Le fait de chaîner 5 blocs additionneur entraîne un accroissement de la taille des registres mais de moindre importance, du fait de leur caractère non-récursif, comparé à celui de la cascade de 5 blocs Intégrateur. Cet accroissement peut encore être limité sans perte notoire d'efficacité

en appliquant (2.22) :

$$B_{out} = B_{in} + log_2(N_{add}) \tag{2.24}$$

où N_{add} est le nombre d'additions effectuées successivement. Ce qui conduit à l'utilisation de 32 bits pour coder notre signal en sortie du 5^{eme} bloc. L'architecture peut encore être optimisée en considérant que les deux opérandes des additionneurs n'ont pas le même format. Ceci a été menée sans détérioration du processus de traitement du signal amenant une légère diminution du nombre de bits utilisés dans ces blocs et donc du nombre de ressources nécessaires. En effet, cette équation s'applique plutôt aux arbres d'addition, ce qui n'est pas le cas ici puisque à chaque entrée d'additionneur, les deux données présentes ne possèdent pas le même format. Il est à noter que l'accroissement de la taille des registres dû à la partie RIF est négligeable devant celui dû à la partie additionneur.

3.4.2 Modélisation

Nous utilisons le signal provenant de la sortie du DDS en tant que signal d'entrée (Figure 2.14(a)). En Figure 2.17 est tracée l'évolution des réponses complexes du filtre d'ordre 1 à 5.

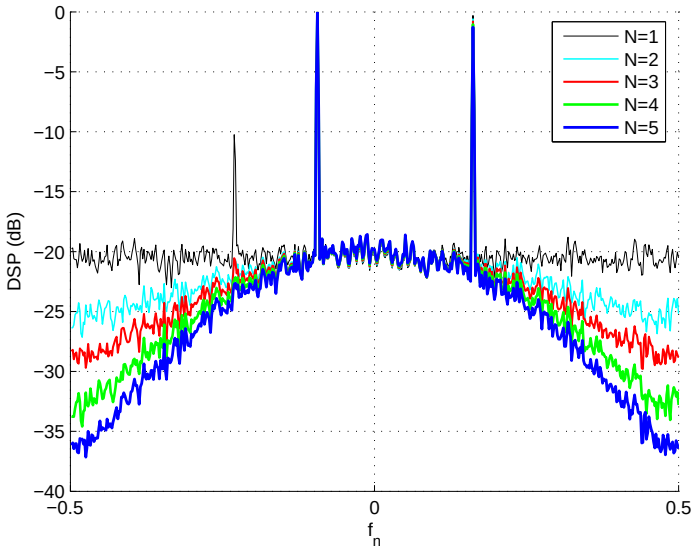


FIG. 2.17 – *Spectre en sortie du filtre démultiplexée non-récursif pour différents ordres*

La chute dans la bande passante obtenue en sortie n°5 est identique à celle obtenue avec un filtre CIC classique, à savoir $5dB$ à la fréquence $f_n = 0.25$.

Remarque : Sur le spectre correspondant à $N = 1$ une raie parasite fait son apparition. L'atténuation obtenue avec un tel filtre n'étant pas assez importante, lors du recouvrement après décimation, la bande passante se retrouve polluée. La spécification d'atténuation doit être assez sévère afin que de tels phénomènes ne se produisent pas.

3.4.3 Implémentation

Sur la Figure 2.18 est représentée la vue RTL de la structure non-récursive développée.

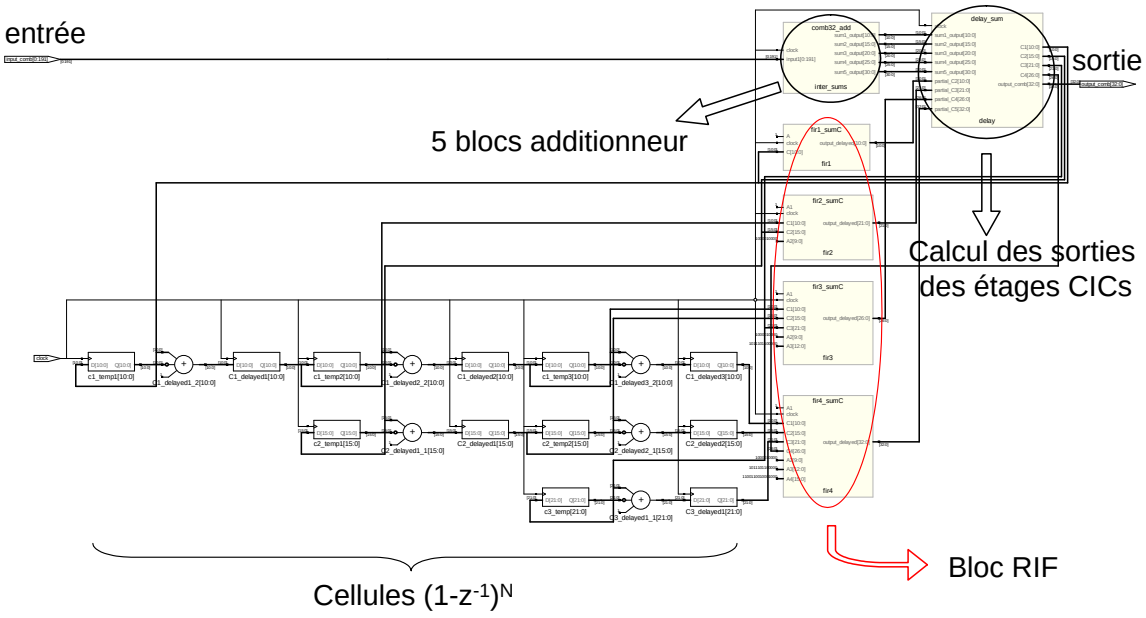


FIG. 2.18 – *Schématique RTL (correpondance avec la Figure 2.16)*

L'entrée comporte 32 voies sur 6 bits. Apparaissent sur cette figure, les 5 blocs additionneur ainsi que les 4 filtres RIF respectivement d'ordre 4, 3, 2 et 1. Différentes cellules retard sont aussi présentes sur cette vue RTL. Tous ces éléments retranscrivent l'équation 2.23 permettant d'obtenir le comportement d'un filtre CIC décimateur avec une structure démultiplexée non-récursive. Ici encore la sortie a été tronquée à 8 bits.
Cette solution requiert 2000 ALMs et atteint une fréquence de fonctionnnement de l'ordre de 110 MHz. La fréquence de fonctionnement du TFB requise (i.e. $125MHz$) n'est donc pas atteinte. Pour cela l'architecture doit être légèrement modifiée en utilisant la technique du *pipelining*. Cependant, ces modifications amènent inévitablement une augmentation du nombre de ressources qui est déjà supérieur à celui de la structure RIF TFB. Cette structure n'aboutit donc pas à l'optimisation escomptée.

4 Décimation multi-étages

Les résultats présentés dans la section précédente ne satisfont pas les attentes liées à l'utilisation d'un filtre CIC. La limite provient du fait que l'ordre du filtre est trop élevé pour une structure démultiplexée. Le produit $f_{BI} \cdot D$ est trop grand, il vaut $\frac{1}{4}$ (cf. tableau 2.1). Une solution permettant de diminuer la valeur de ce produit est le découpage du 1^{er} étage de filtrage en plusieurs étages afin de partitionner la décimation de facteur D. Remarquons au passage que le système TFB est déja un système de filtrage multi-étages (2 étages), nous proposons, ici, de découper cette opération de filtrage en plus de deux étages. Dans cette section sont présentés différents systèmes multi-étages utilisant les architectures de la section précédente. Dans le cas spécifique du projet ALMA, les structures polyphases et démultiplexées-non-recursives ne sont pas adaptables à un filtrage multi-étages puisque leur décimation est intrinsèque et liée au nombre de voies démultiplexées en entrée.

4.1 Multiples étages CIC

Une première solution consiste à cascader 2 filtres CIC [26] dont le produit des facteurs de décimation D_1 et D_2 est égal à 32 ($D = D_1 \cdot D_2$). Le choix du taux de décimation doit être fait dans l'optique de la minimisation de l'ordre de chaque filtre. Voici la méthode à suivre pour déterminer les différents paramètres en utilisant le tableau 2.1 :

- Choisir N_1 le plus faible possible afin de minimiser la complexité du premier CIC.
- Choisir D_1 (ou f_{BID1}, f_{BI} étant fixée) le plus grand possible afin de minimiser la complexité du second CIC.
- Le couple (f_{BID1}, N_1) doit satisfaire le critère d'atténuation.

Une fois le premier étage caractérisé, les paramètres du second doivent être choisis en conséquence. D_2 étant déjà fixé puisque $D = D_1 \cdot D_2$, il reste à déterminer N_2 qui est l'ordre le plus faible satisfaisant l'atténuation recherchée couplé avec la valeur de la bande d'intérêt $f_{BID2} = \frac{1}{4}$ ($= f_{BI2} \cdot D_2 = f_{BID1} \cdot D_2 = f_{BI1} \cdot D_1 \cdot D_2$).
L'application de ce raisonnement conduit à une structure présentée en Figure 2.19 pour les couples suivant de paramètres : ($N_1 = 2$, $D_1 = 8$) et ($N_2 = 5$, $D_2 = 4$).

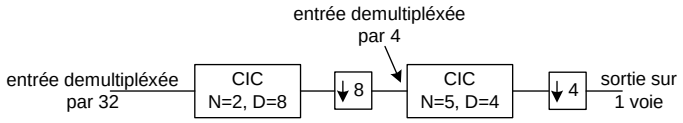


FIG. 2.19 – *Cascade de 2 filtres CIC*

En Figure 2.20 est tracé la fonction de transfert obtenue par la cascade de deux filtres CIC.

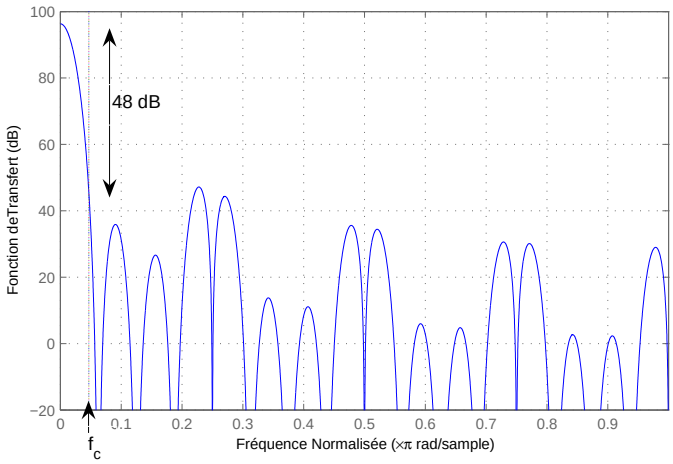


FIG. 2.20 – *Fonction de transfert de la cascade des 2 filtres CIC*

La chute dans la bande passante est, ici encore, non négligeable et identique à celle obtenue avec la structure CIC classique.
Le grand nombre d'étages nécessaires à l'obtention de l'atténuation spécifiée ($N_2 = 5$ dans le second étage) conduit cependant à remettre en question l'implémentation d'une telle architecture. La technique de réduction du nombre de ressources nécessaire à l'implémentation d'une structure cascadée, présentées en [26], ne permet pas d'abaisser ce nombre à une valeur inférieure, voire comparable, à celle de la structure TFB.

4.2 Cascade d'un CIC et de filtres RIF

Une autre possibilité considérée est la cascade du premier filtre CIC décrit dans la section précédente ($N = 2$, $D = 8$) cascadé avec une solution RIF composée d'un ou plusieurs étages de filtres. Le but étant d'assurer une décimation totale de 32 en conservant les mêmes spécifications d'atténuation. La structure (le filtre ou la cascade de filtres) qui suit doit donc induire un facteur de décimation égal à 4.
Dans une première partie, une étude des différentes solutions envisageables pour l'implémentation du filtre CIC ($N = 2$, $D = 8$) est présentée. La seconde partie concerne l'étude des structures possibles pour l'implémentation de l'étage RIF : un filtre quart de bande ou une cascade de deux filtres demi bande.

4.2.1 Le CIC

Le choix des paramètres du filtre CIC se porte donc sur les valeurs suivantes : $D_1 = 8$ et $N_1 = 2$ (cf. méthode décrite précédemment). Le filtre CIC est réalisé sous deux formes différentes afin de comparer les structures électroniques en terme d'ALMs : la forme récursive et la forme non-récursive, chacune d'elles possédant une entrée démultiplexée sur 32 voies et délivrant une sortie démultiplexée sur 4 voies (facteur de décimation $D = 8$).

4.2.1.1 Forme récursive En appliquant la méthode de Hogenauer qui permet de diminuer la taille des registres en fixant le MSB et en supprimant un certain nombre de LSBs, nous obtenons une architecture électronique dont le codage du signal d'entrée des différents étages est présenté en Figure 2.21.

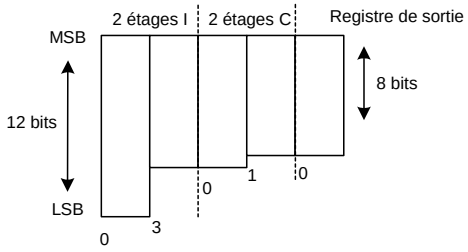
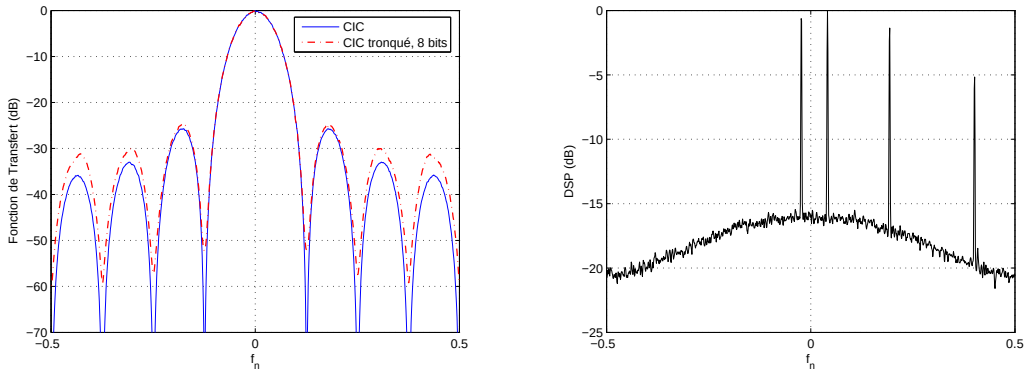


FIG. 2.21 – *Taille des registres, suppression des LSBs*

Au bas de chaque d'étage, le nombre de LSBs supprimés est indiqué, chaque étage I étant constitué de 32 voies parallèles (Figure 2.12) et chaque étage C de 4 voies parallèles. La sortie est fixée à 8 bits afin de limiter la complexité de la structure électronique. Ceci permet de conserver une dynamique du signal et une atténuation en sortie acceptable. Cette atténuation peut être analysée sur la Figure 2.22(a) où apparaît la sortie tronquée (méthode d'Hogenauer) et non-tronquée du CIC.
En Figure 2.22(b) est tracée la sortie après décimation de la modélisation du filtre CIC. Le signal d'entrée est toujours celui provenant de la sortie du DDS. La modélisation du filtre est réalisée de manière à correspondre à l'architecture électronique de la Figure 2.3 afin de pouvoir appliquer les modifications d'Hogenauer. Les signaux sont toujours de nature complexe.

Pour en revenir à la restriction du signal de sortie à un plus faible nombre de bits, une structure électronique aboutissant à une sortie codée sur 6 bits a été étudiée (échelle correspondant à celle du signal d'entrée). La fonction de transfert en résultant est tracée en Figure 2.23.



(a) Fonction de transfert du CIC (b) Sortie du CIC après décimation

FIG. 2.22 – *Caractéristique de sortie du filtre CIC, $D = 8$, $N = 2$*

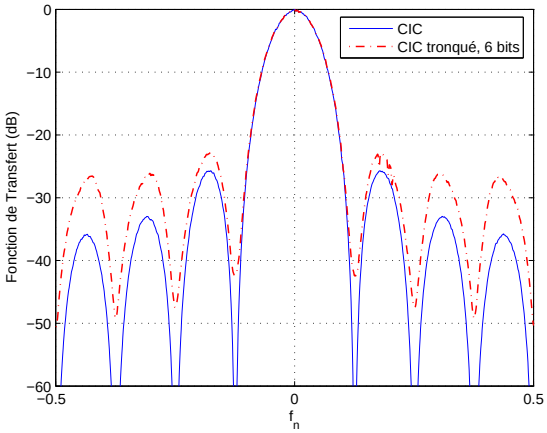


FIG. 2.23 – *Fonction de transfert du filtre CIC, $D = 8$, $N = 2$*

Le spectre en sortie possède, du fait de cette troncature, une dynamique moins importante mais aussi une atténuation dégradée avoisinant les 40 dB (puisque les troncatures se font tout au long de la structure électronique). La structure dont le traitement résulte en une sortie sur 8 bits possède un spectre plus conforme à la solution non tronquée, respectant l'attenuation attendue et conservant une meilleure dynamique en sortie.

Cette architecture, aprés avoir été traduite en VHDL, a été synthétisée. La vue RTL résultant de cette synthèse est donnée en Figure 2.24.
Les parties Intégrateur et Comb sont toutes deux constituées de deux blocs (CIC d'ordre 2). L'entrée comporte 32 voies sur 6 bits et la sortie 4 voies sur 8 bits. L'occupation des ressources correspond à 1200 ALMs.

4.2.1.2 Forme non récursive A partir de la formule 2.21, on établit la structure électronique présentée en Figure 2.25.
La structure est composée de 3 blocs, chacun d'ordre 2. Chaque bloc est suivi d'un bloc décimateur par deux, ce qui conduit bien à un facteur de décimation total de 8. Cette architecture est facilement adaptable à un format de données d'entrée démultiplexé. La sortie est tronquée sur

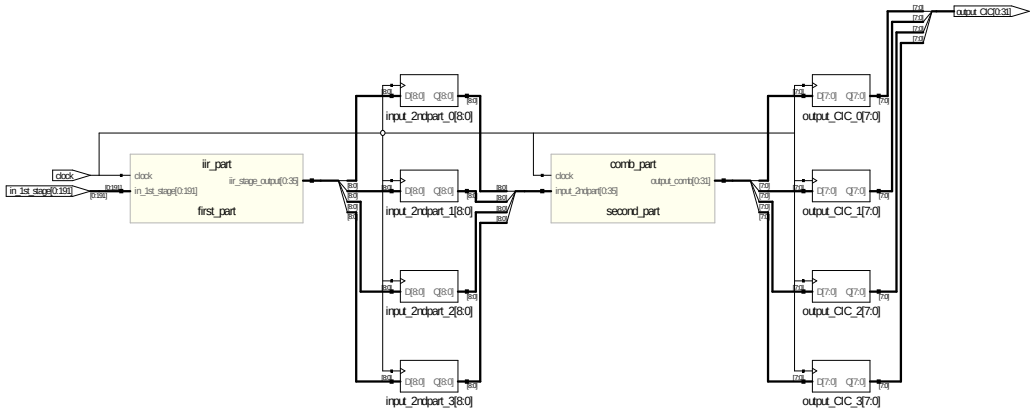


FIG. 2.24 – *Schématique RTL*

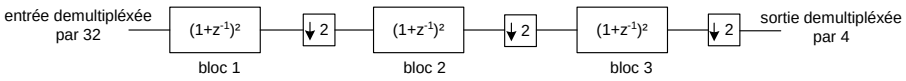


FIG. 2.25 – *Architecture du filtre CIC non-récursif,*$D = 8$ $N = 2$

8 bits toujours dans l'optique de limiter la complexité de la structure électronique. Le signal en sortie de cette architecture possède les mêmes caractéristiques électroniques (codé sur 8 bits) et spectrales que celui obtenu par la méthode récursive (Figure 2.26).

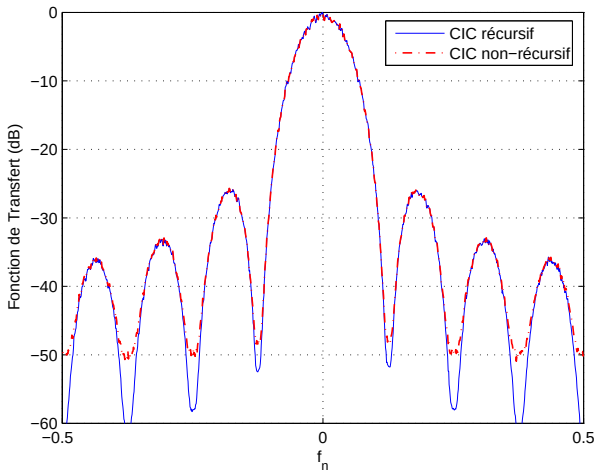


FIG. 2.26 – *Comparaison des fonction de transfert des structures électroniques (8 bits de sortie)*

L'architecture démultiplexée synthétisée n'occupe que 390 ALMs (Figure 2.27).

Du fait de la décimation par 2 effectuée entre chaque bloc, des simplifications de l'architecture sont appliquées et permettent de sauver un nombre de ressources important (Figure 2.28) : pour le second additionneur (dû à l'ordre 2) de chaque bloc, seulement la moitié des sommes est calculée.

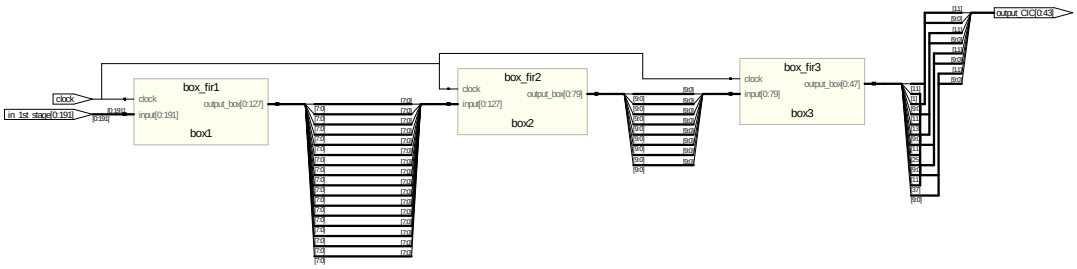


FIG. 2.27 – *Schématique RTL*

4.2.2 Les filtres RIF demi bande et quart de bande

Dans cette partie, nous évoquerons seulement l'utilisation de filtres type RIF à phase linéaire, les filtres RII étant écartés du fait de la nécessité d'une structure à entrée démultiplexée.Le filtre (ou la cascade de filtres) doit respecter la spécification d'atténuation, peut accepter une bande de transition médiocre et doit induire un facteur de décimation $D_2 = 4$. La synthèse de tels filtres est réalisée par l'intermédiaire de l'algorithme de Remez [8]. Cependant, un algorithme modifié, consistant à homogénéiser les coefficients en bout de réponse impulsionnelle, est présenté en [9]. Sur la Figure 2.29 sont tracées deux fonctions de transfert de filtre RIF de même ordre $N = 35$, résultant toutes les deux du même gabarit, chacune synthétisée par un algorithme différent.
Cette figure illustre clairement l'avantage de la méthode modifiée : l'atténuation dans la bande d'arrêt décroît en $\frac{1}{f}$, alors qu'elle était constante dans le cas de la première méthode.

Différentes structures électroniques peuvent être utilisées afin d'implémenter un filtre RIF, les principales sont exposées dans le paragraphe qui suit. Dans le deuxième et troisième paragraphe, l'utilisation d'un RIF quart de bande et d'une cascade de deux RIF demi bande sont étudiées. Le format de l'entrée de la structure utilisée doit correspondre à celui de la sortie de l'étage CIC précédent, à savoir un format de données démultiplexé sur 4 voies.

4.2.2.1 Structures de filtre envisageables

Forme directe C'est l'implémentation classique d'un filtre RIF. Elle a été présentée en Annexe « Notions importantes de Traitement Numérique du Signal ». En utilisant la propriété de symétrie des coefficients, une structure nécessitant moins de ressources est obtenue (Figure 2.30).

Structure cascade Le principal intérêt d'utiliser une structure cascade par rapport à une structure directe est de diminuer la sensibilité de la réponse du filtre à la quantification de ses coefficients. De cette manière ces derniers peuvent être codés sur un plus faible nombre de bits diminuant d'autant la taille des mots résultant du traitement.
Deux méthodes sont présentées en [27], l'une basée sur des cellules à coefficients réels et l'autre sur un mélange de cellules réelles et complexes. En Figure 2.31, on montre le regroupement des zéros pour ces deux méthodes aboutissant à la création de plusieurs cellules d'ordre 2.
Les coefficients réels sont obtenus en couplant chaque zéro avec son conjugué complexe (Figure 2.31(a)). Il est à noter que la sensibilité est d'autant plus grande que les zéros sont proches de l'axe réel. Les cellules complexes sont utilisées, dans la deuxième méthode, pour représenter les zéros qui sont situés hors du cercle unité (Figure 2.31(b)). La distance entre une paire de

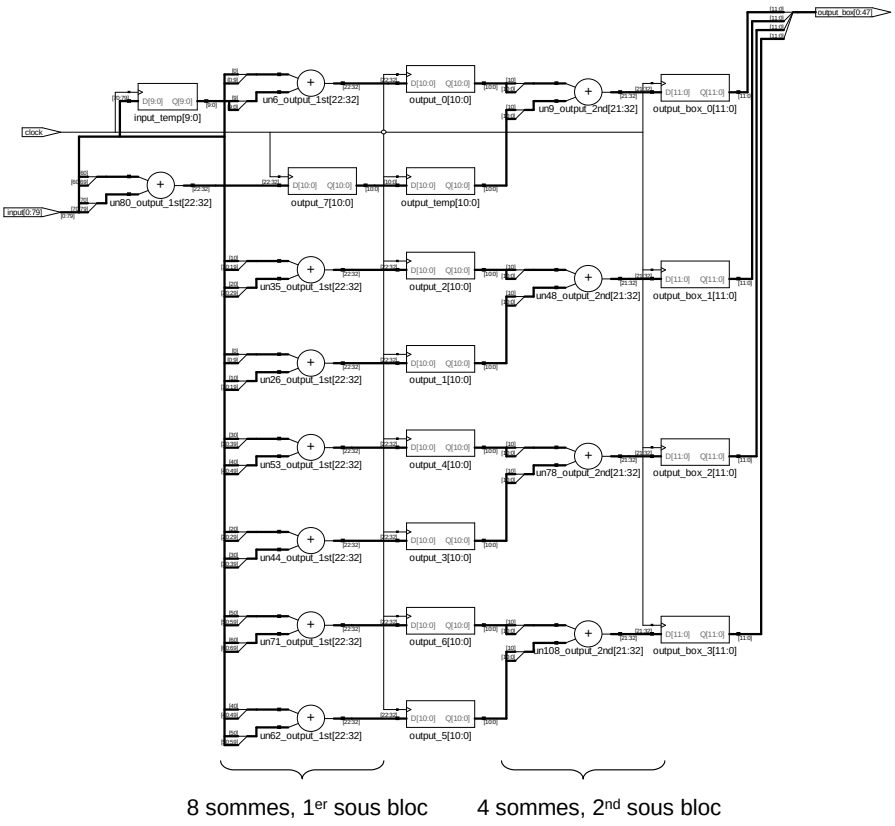
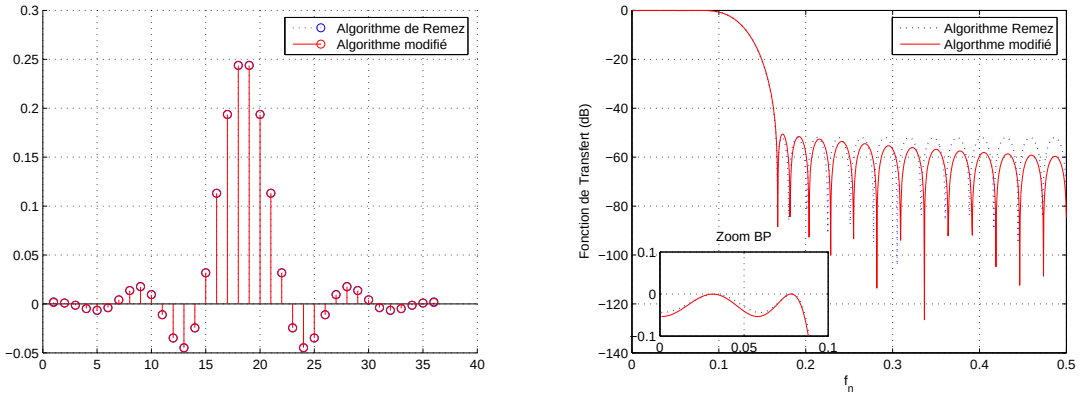


FIG. 2.28 – *Schématique RTL du bloc* 3 *:* 8 *vers* 4

zéros est alors plus grande que dans la méthode conventionnelle (cellules réelles). Ceci résulte en une sensibilité moindre même lorsque les zéros sont proches de l'axe des réels [27].
Le filtre est alors composé par une cascade des différents regroupement de zéros. Dans le cas du filtre quart de bande d'ordre 35 (zéros représentés en Figure 2.32), l'architecture obtenue est composée de 17 cellules d'ordre 2 et 1 cellule d'ordre 1.
En Figure 2.33(a) apparaît la structure électronique de la cellule d'ordre 2 des coefficients réels. La Figure 2.33(b) représente celle de la cellule d'ordre 2 des coefficients complexes ; cette structure s'effectuant sur deux voies (partie réelle et imaginaire), une recombinaison de celles-ci est ensuite effectuée.

Remarque sur la position des zéros : Le fait que les zéros positionnés sur le cercle unité occupent l'intervalle $\left[\frac{\pi}{2} : \pi\right]$ indique que le filtre est de type passe pas. Leur quantification, bonne ou mauvaise, joue sur la qualité de l'atténuation dans la bande d'arrêt. Le fait que ces zéros sur le cercle occupe les $\frac{3}{4}$ du cercle indique que c'est un filtre quart de bande.

Structure polyphasée Dans le cas d'une entrée démultiplexée, une structure polyphasée est une solution plus intéressante. La transformation de la structure du filtre est réalisée par l'intermédiaire de l'équation 2.20 avec le facteur de décimation $D = 4$ par exemple.
L'application de l'équation 2.20 au filtre d'ordre $n = 35$ utilisé précédemment conduit à une décomposition en 4 filtres RIF composés chacun de 9 coefficients (structure similaire à celle de la



(a) Réponse Impulsionnelle (b) Fonction de Transfert

FIG. 2.29 – *Méthode de Remez et Méthode de Remez modifiée*

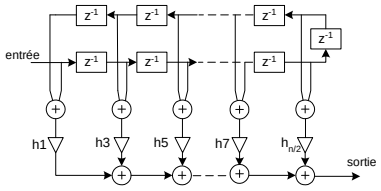
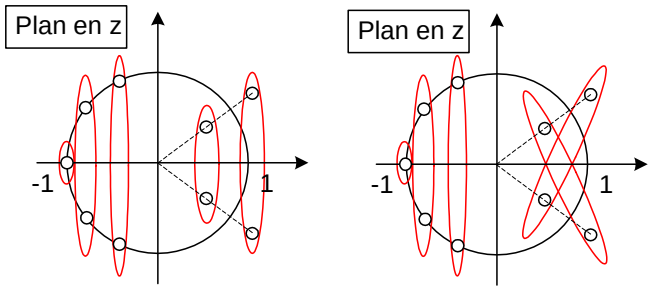


FIG. 2.30 – *Structure directe symétrique (nb de coeff. pair)*

Figure 2.10 avec $D = 4$, chaque filtre F_i possédant l'architecture de la Figure 2.30 avec $n = 9$). Aprés décomposition en 4 sous-éléments RIF, chacune des 4 réponses impulsionnelles a perdu son caractère symétrique. Cependant, les réponses sont symétriques deux à deux permettant de stocker une réponse impulsionnelles sur deux en mémoire (18 coefficients stockés en mémoire). Cette solution nécessite l'utilisation d'un registre à décalage de profondeur 8 pour chaque « sous-filtre » »(4 registres sont donc nécessaires). Chaque sous-filtre requiert 9 multiplications, soit un total de 36. Ces résultats peuvent être généralisés à tout ordre n.

Cette solution est comparable, en terme d'architecture, à celle utilisée dans le cas du 1^{er} étage TFB. En effet, à ordre donné et à facteur de décimation donné, toutes deux comportes des registres à décalage de même longueur, un bloc de multiplication, un processus de décimation



(a) Couplage des zéros - cellules réelles (b) Couplage des zéros - cellules réelles et complexes

FIG. 2.31 – *Couplage des zéros*

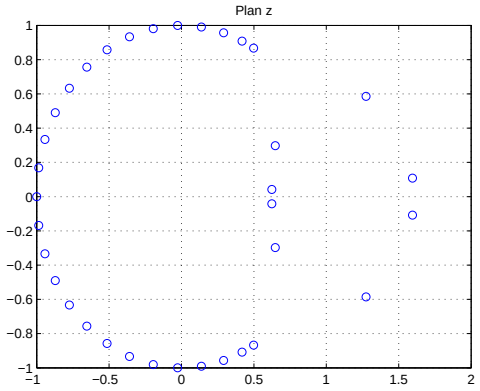
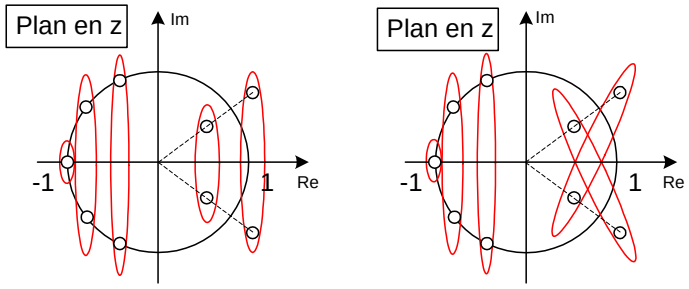


FIG. 2.32 – *Zéro du filtre quart de bande*



(a) Implémentation d'un coefficient réel

(b) Implémentation d'un coefficient complexe

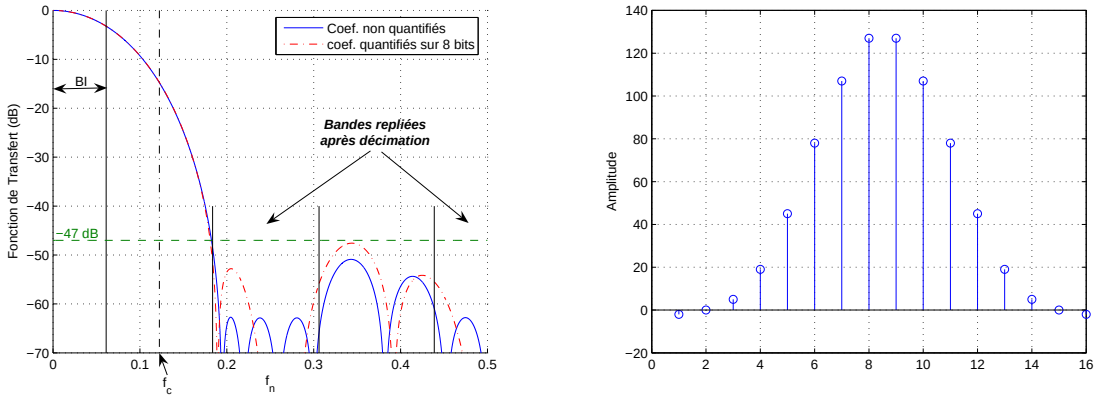
FIG. 2.33 – *Structure d'ordre deux limitant la sensibilité du filtre à la quantification*

(intrinsèque aux deux structures) et un arbre d'addition sur 4 voies. Cependant du fait de la décomposition polyphasée, les sous-filtres en résultant ne possède plus la caractéristique de symétrie. Dans ce cas, un nombre supérieur de multiplieurs est nécessaire. La solution type « 1^{er} étage TFB » est donc plus performante en terme d'occupation de ressources.

4.2.2.2 RIF quart de bande Le filtre quart de bande est la solution découlant directement de la nécessité d'obtenir une décimation par 4 aprés l'application du CIC ($N = 2$, $D = 8$). En appliquant l'algorithme de Remez aux spécifications suivantes :

- une bande d'intérêt de largeur $\frac{1}{16}$ (en fréquence normalisée) correspondant à la SB finale en sortie du TFB.
- deux bandes d'arrêt (la première : $\left[\frac{3}{16}, \frac{5}{16}\right]$, la seconde : $\left[\frac{7}{16}, \frac{1}{2}\right]$) correspondant aux bandes repliées aprés décimation
- une bande de transition sans contrainte particulière (chute dans la bande d'intérêt acceptée). Cela permet de limiter la complexité du filtre.

on obtient une fonction de transfert d'atténuation supérieure à 47 dB pour un ordre $n = 15$ (Figure 2.34(a)).
Une quantification des coefficients sur 8 bits permet de conserver les caratéristiques de départ. La réponse impulsionnelle est bien sûr symétrique (filtre à phase linéaire, Figure 2.34(b)). 8



(a) Fonction de transfert

(b) Réponse impulsionnelle

FIG. 2.34 – *Fonction de transfert du filtre quart de bande synthétisé*

multiplieurs et 15 additionneurs sont nécesaires pour implémenter un filtre d'ordre $n = 15$ à réponse impulsionnelle symétrique dans le cas d'une structure non-démultiplexée. L'implémentation « démultiplexée » a été réalisée en se basant sur la méthode utilisée pour l'implémentation du premier étage TFB, en appliquant un registre à décalage de profondeur 4 à l'entrée démultiplexée par 4.
Le spectre en sortie de la modélisation de la cascade CIC - Quart de bande est tracé en Figure 2.35.

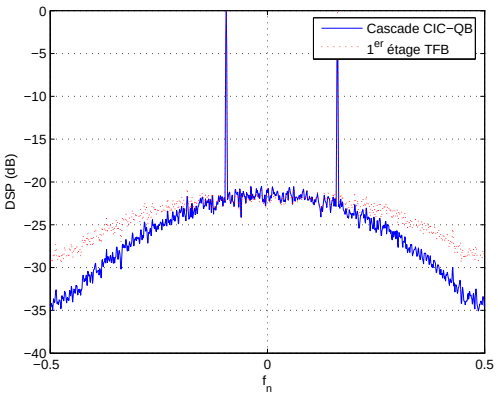


FIG. 2.35 – *Spectres en sortie de la cascade CIC-QB et du* 1^{er} *étage TFB*

Il est comparé à la sortie du 1^{er} étage TFB. Tous deux offrent une décimation par 32 du signal d'entrée avec une importante chute dans la bande passante qui est compensée par le second étage de filtrage TFB.

4.2.2.3 RIF demi bande Une autre solution consiste à cascader 2 filtres demi-bande (Half-band) suivis chacun d'une décimation par 2 afin d'obtenir la décimation totale de 32. L'utilisation de l'algorithme de Remez avec les spécifications de bandes suivantes :

- une bande passante de largeur $\frac{1}{8}$
- une bande d'arrêt appartenant à l'intervalle $\left[\frac{3}{8}, \frac{1}{2}\right]$ qui correspond à la bande repliée dans

la bande d'intérêt après décimation
- une bande de transition sans contrainte particuliaire

conduit à une atténuation de $-47dB$ pour un ordre égal à 10. Une quantification des coefficients sur 9 bits permet de conserver la spécification d'atténuation (Figure 2.36).

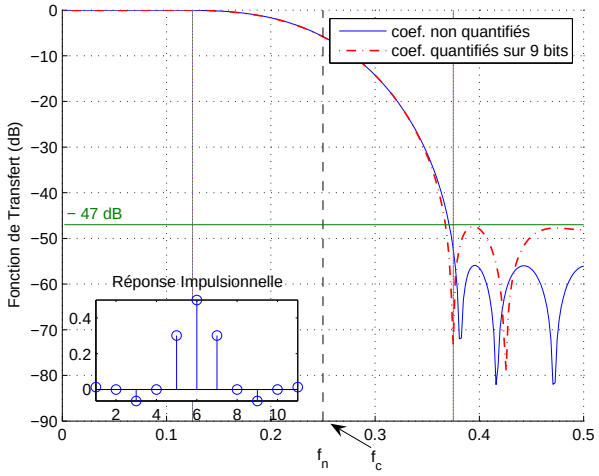


FIG. 2.36 – *Fonction de transfert du filtre demi-bande et réponse impulsionnelle*

Il ne faut cependant pas oublier le caractère démultiplexé du signal à traiter. Sur la Figure 2.37 est représentée le schéma bloc d'un filtre CIC avec $D = 8$ avec entrée démultiplexée cascadé avec deux filtres demi-bande.

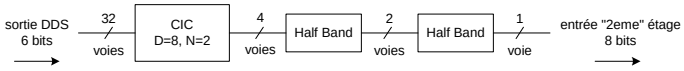


FIG. 2.37 – *Structure multi-étages*

Le format du signal d'entrée amène à repenser l'architecture classique du filtre RIF (Figure 2.30). Le filtre CIC délivre un signal sur 4 voies parallèles. Ceci implique une architecture démultiplexée 4 vers 2 pour le premier halfband et 2 vers 1 pour le deuxième.
En se référant à la Figure 2.36 où apparaît la réponse impulsionnelle du filtre demi bande, on observe que les coefficients d'indice pair de ce filtre sont nul, mis à part le coefficient central. Cette remarque a une répercussion instantanée sur l'architecture électronique du filtre (Figure 2.38).

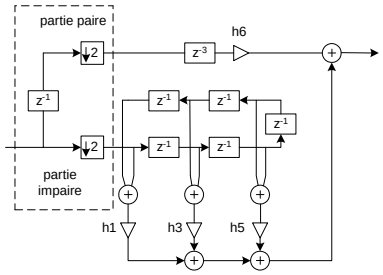


FIG. 2.38 – *Structure d'un filtre halfband*

Le signal d'entrée est séparé en échantillons pairs et impairs de façon à appliquer à chaque partie les coefficients correspondants. La décimation par 2 est intrinsèque à l'architecture. Le spectre en sortie de la cascade CIC ($N = 2, D = 8$) - deux filtres demi-bande est donné en Figure 2.39.

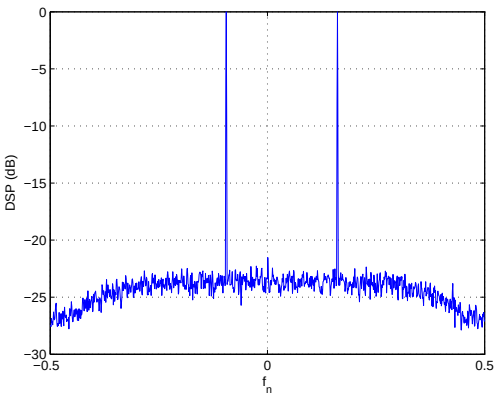


FIG. 2.39 – *Spectre en sortie de la cascade CIC ($D = 8, N = 2$) + 2 demi-bande*

Dans le cas présent, la bande passante est strictement plate et n'a donc pas besoin d'être compensée par l'étage de filtrage suivant.

Halfband 2 vers 1 En revenant à la Figure 2.38, on s'aperçoit que l'application de la structure au signal provenant de la sortie du 1^{er} halfband est instantanée. Cette architecture est retenue pour ce filtre demi-bande.

Halfband 4 vers 2 L'architecture précédente n'est pas directement applicable au cas présent. La solution nécessitant le moins de modifications de la structure du filtre RIF consiste à utiliser deux architectures identiques en parallèle semblables à celle présentée en Figure 2.30. Pour que la sortie soit composée de 2 voies démultiplexées (distribuant des échantillons alternés), il est nécessaire d'utiliser la structure présentée en Figure 2.40 avec $n = 10$.

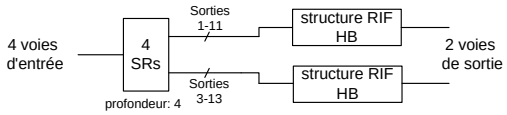


FIG. 2.40 – *Structure d'un filtre halfband 4 vers 2*

Les 4 registres à décalage (de profondeur 4) en entrée permettent d'avoir à chaque instant t, tous les échantillons nécessaires pour effectuer les 2 filtrages en parallèle. En effet, la réponse impulsionnelle comportant 11 coefficients, pour obtenir une sortie démultiplexée par 2, il faut décaler la fenêtre de convolution du deuxième filtre de deux échantillons par rapport celle du premier. Ceci explique la nécessité d'avoir 13 échantillons à chaque instant t en entrée.

4.3 Implémentation

Dans le tableau 2.3 sont reportés les résultats de l'implémentation des structures récursive et non-récursive du filtre CIC ($D = 8$, $N = 2$).

TAB. 2.3 – Comparaison de la solution CIC récursif et non récursif

	Ressources (Stratix2)
CIC récursif (Hogenauer)	1200 ALMs
CIC non-récursif	390 ALMs

Dans le tableau 2.4 sont reportés les résultats de l'implémentation des filtres quart de bande et demi bande.

TAB. 2.4 – Comparaison de la solution Quart de bande et 2 Demi bande

	Ressources (Stratix2)	nombre de multiplications
Quart de bande (TFB)	295 ALMs	8
2 demi bande (HB4-2 et HB2-1)	400 ALMs	12

De ces tableaux, est extraite la solution la plus intéressante, à savoir un filtre CIC non-récursif cascadé avec un filtre quart de bande.
Pour obtenir une sortie de ce filtre multi-étages (composé de 2 filtres) compatible avec le format imposé à l'entrée du deuxième étage du filtre TFB, certaines réductions du format des données sont nécessaires. A chaque sortie d'étage, une troncature (suppression des bits de poids faible) peut être réalisée afin de réduire ce format mais cela n'est pas sans conséquence (cf. section 6.2.2 en p. 36 du Chapitre 1).

5 Récapitulatifs des résultats

Suite aux différentes études menées dans l'optique de l'optimisation du premier étage TFB présentées dans ce chapitre, le tableau 2.5 peut être établi.

TAB. 2.5 – Résumé des solutions envisagées

	Ressources (Stratix2)	Fréquence max.
TFB 1^{er} étage	1775 ALMs	180 MHz
Solution récursive (D=32,N=5)	6500 ALMS	140 MHz
Solution demultiplexée [25]	1900 ALMs	114 MHz
CIC (D=8,N=2) non-rec 2HBs	769 ALMs	217 MHz
CIC (D=8,N=2) non-rec QB	653 ALMS	240 MHz

Ce tableau permet de mettre en valeur la solution CIC non-récursif associé au filtre quart de bande. Seule la solution CIC non-récursif associé aux filtres demi bande s'en approche. La particularité de cette dernière concerne la bande passante induite par cette structure : elle est strictement plate et ne nécessite donc aucune compensation par l'étage suivant.

6 Conclusion

L'enjeu de ce chapitre était de proposer une alternative au 1^{er} étage de filtrage du système TFB permettant de diminuer la consommation du dit système (75 W par carte). L'étude s'est tout de suite orientée sur le filtre CIC, filtre permettant d'important facteurs de décimation ($D = 32$ dans notre cas), dont le principal avantage est la simplicité architecturale.

La difficulté majeure a été d'adapter la structure classique du filtre CIC à notre entrée démultiplexée tout en minimisant le nombre de ressources utilisées. Après une étude approfondie des différentes architectures de filtre CIC, la solution qui s'est révélée la plus interessante est celle du filtre multi-étages composé d'un filtre CIC ($D = 8$,$N = 2$) à architecture non-récursive cascadé avec un filtre quart de bande assurant la décimation souhaitée. Cette structure entraine certes une complexification de l'architecture (deux étages au lieu d'un) mais aussi une diminution importante du nombre de ressources nécessaires à son implémentation.

Chapitre 3

Filtre RII - Linéarisation de Phase

1 Introduction

Le filtre RII est potentiellement intéressant en raison de son faible nombre de coefficients comparé à celui d'un filtre RIF possédant les mêmes spécifications en fréquence. De ce fait la structure électronique résultante utilise un plus faible nombre de ressources logiques. Cependant ce type de structures induit une phase non-linéaire ce qui ne corespond pas à la spécification imposée par le projet ALMA à ce sujet (le projet recommande une ondulation maximale de la phase de 0.3 °). L'objet de ce chapitre est donc d'étudier des techniques permettant l'obtention d'un filtre RII à phase linéaire ou approximativement linéaire. Le but serait alors, éventuellement, de remplacer l'actuel second étage de filtrage permettant de fixer la bande finale de fréquence à analyser. Certaines de ces techniques reposent sur une modification de l'architecture électronique d'un filtre RII alors que d'autres sont des algorithmes basés sur une réduction du modèle d'un filtre RIF à phase linéaire.
Différentes architectures permettant de limiter l'encombrement des filtres RII lors de l'implémentation sont dans un premier temps exposées : directe, parallèle, « lattice » et réseau de filtres « allpass ». Par la suite, certaines sont employées dans des structures aboutissant à la linéarisation de la phase. Le principal intérêt de ce travail de revue est de fournir un état de l'art des solutions de filtrage envisageables, par l'utilisation de filtres RII, mais aussi d'aboutir à une possible optimisation en consommation de la structure complète du TFB en remplaçant le second étage du filtre TFB.
Ces techniques sont donc appliquées à la conception d'un filtre demi-bande répondant aux spécifications de ce second étage de filtrage (47 dB d'atténuation, $\pm 0,2$ d'ondulation dans la bande passante et une région de transition étroite – $1/32^{eme}$).

2 Filtre RII faible encombrement

Il est connu que le filtre dit « elliptique » présente la bande de transition la plus étroite, pour un ordre donné, parmi les différents types de filtres analogiques à partir desquels un filtre RII peut être obtenu. Pour cette raison, seul ce filtre est pris en compte pour la réalisation des différentes structures qui vont être exposées.
Une classe particulière de filtres elliptiques nommée « Elliptic Minimal Q-factor » (EMQF) a attiré notre attention. Un facteur de première importance lors de l'implémentation d'un filtre dans un FPGA est la taille du mot codant les coefficients de ce dernier. En effet, ce paramètre délimite la taille des registres (ressources logiques) permettant d'effectuer la fonction de filtrage,

sans dégradation du gabarit d'origine. Cette classe de filtre permet justement de réduire la sensibilité du filtre à la quantification de ses coefficients. La méthode permettant la synthèse de ces filtres EMQF est brièvement présenté en première partie de cette section. Ils sont décrits en détails en [28].

Dans la seconde partie sont abordées les structures électroniques RII à faible encombrement pouvant être utilisées pour l'implémentation des filtres EMQF ou elliptic simple. 3 familles ont été étudiées : la structure cascade SOS utilisant un bloc multiplieur pour l'implémentation des coefficients (dérivée de la structure cascade SOS classique), la structure lattice et la structure allpass. Pour des raisons de clarté, seule la structure utilisée par la suite est présentée dans cette section : la stucture allpass. Les deux autres sont décrites dans les Annexes Bloc multiplieur et Structure Lattice.

2.1 Les filtres EMQF

Une caratéristique importante concernant un filtre elliptique est la possibilité de définir indépendemment l'ondulation dans la bande passante et dans la bande atténuée. Cependant cette caractéristique peut être abandonnée au profit d'une très faible sensiblilité aux variations des composants du filtre (i.e. les coefficients). Dans ce cas, le filtre est dit *minimal Q-factor*. Effectivement, la sensibilité aux valeurs des composants électroniques (dans le domaine analogique) est inversement proportionnelle au facteur Q (facteur de qualité, qui permet de déterminer la qualité du comportement d'un système) des pôles de la fonction de transfert du filtre en question. La formule permettant le calcul du facteur de qualité pour le pôle i est la suivante :

$$Q_i = -\frac{s_i}{2\Re(s_i)} = -\frac{1}{2\cos(\arg(s_i))} \tag{3.1}$$

Pour un ordre de filtre donné, il existe une relation entre le facteur d'ondulation ε et le facteur de sélectivité ξ qui permet de minimiser le facteur Q pour tous les pôles de la fonction de transfert :

$$\varepsilon_{Q_{min}} = \frac{1}{\sqrt{L_n(\xi)}} \tag{3.2}$$

où L_n le facteur de discrimination défini en [29]. Les relations générales d'un filtre elliptique liant ces derniers paramètres aux ondulations dans la bande passantee δ_p et dans la bande atténuée δ_a sont :

$$\begin{aligned} \delta_a &= \frac{1}{1+L} \\ \delta_p &= 1 - \frac{1}{1+\varepsilon^2} \end{aligned} \tag{3.3}$$

En appliquant (3.2) à (3.3), on obtient $\delta_a = \delta_p$, les deux ondulations sont alors dépendent l'une de l'autre.

L'utilisation de cette caractéristique associée à la nécessité d'une faible ondulation entraine une légère augmentation de l'ordre du filtre par rapport à un filtre elliptique d'ordre minimal. En contrepartie, la réponse dans la bande passante et dans la bande atténuée est moins sensible aux variations des valeurs des éléments du filtre. En [30] est présenté un algorithme permettant la synthèse de filtres EMQF analogique. Le filtre numérique est alors obtenu en appliquant la transformation bilinéaire.

Une remarque importante concernant les filtres EMQF est leur faible ondulation dans la bande passante, ce qui se traduit pour la réponse en phase par une faible déviation par rapport à une caratéristique linéaire.

Particularités concernant les pôles Un filtre RII dérivé d'un filtre EMQF analogique a la particularité d'avoir ses pôles (dans le plan z) sur un cercle dont le centre est situé sur l'axe des réels [31]. Il est montré que la position du centre de ce cercle est uniquement déterminée par la fréquence de coupure (fréquence à laquelle l'atténuation vaut -3 dB, Figure 3.1).

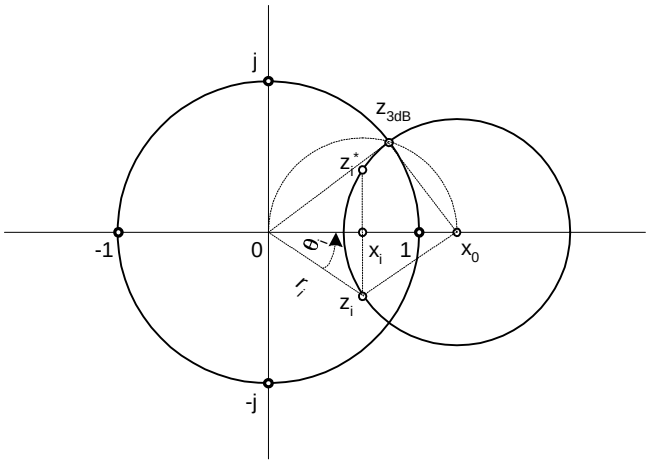


FIG. 3.1 – *Position des pôles d'un filtre EMQF dans le plan z*

2.2 Filtres récursifs *Allpass*

Dans ce paragraphe est developpée une architecture permettant la réalisation d'un filtre quelconque à partir de filtre allpass. Elle permet de minimiser l'utilisation des ressources lors de son implémentation. Elle est basée sur une structure parallèle à deux branches composées chacune d'une cascade de structures allpass.

2.2.1 Présentation

Un filtre allpass est caractérisé par une fonction de transfert dont le module est égal à 1 :

$$|H(z)|_{z=e^{j\theta}} = 1 \tag{3.4}$$

Un tel filtre est donc seulement caractérisé par sa réponse en phase. La fonction de transfert d'un filtre allpass du premier ordre est donnée ci-dessous :

$$H(z) = \frac{\alpha + z^{-1}}{1 + \alpha z^{-1}} \tag{3.5}$$

La réponse en phase de ce filtre est la différence de la phase résultant du numérateur et de celle résultant du dénominateur. Elle est tracée en fonction de la valeur de α en Figure 3.2.
La relation générale, pour un ordre M, est la suivante :

$$\begin{aligned} H(z) &= \frac{N_M(z)}{D_M(z)} = \frac{z^M D_M(z^{-1})}{D_M(z)} \\ &= \frac{a_M + a_{M-1}z^{-1} + a_{M-2}z^{-2} + \ldots + a_1 z^{-(M-1)} + z^{-M}}{1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_{M-1}z^{-(M-1)} + a_M z^{-M}} \end{aligned} \tag{3.6}$$

Le numérateur est le polynôme réciproque du dénominateur, ainsi est assurée l'égalité 3.4. Ce type de filtre est par exemple employé en tant qu'égaliseur de phase.

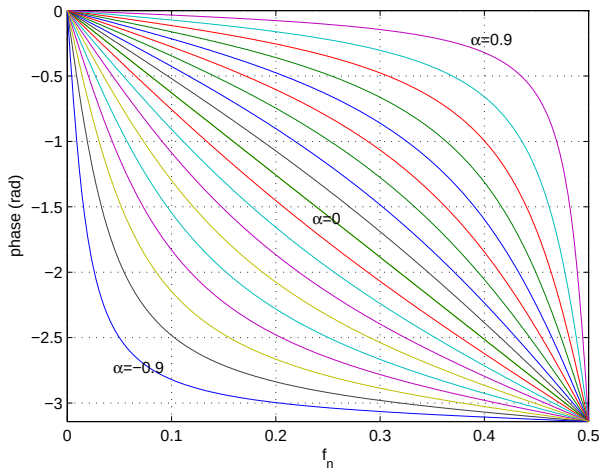
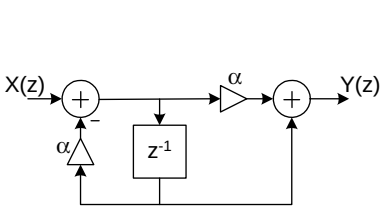


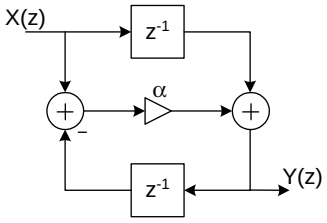
FIG. 3.2 – *Réponse en phase d'un filtre allpass d'ordre* 1

2.2.2 Implémentation de la structure Allpass

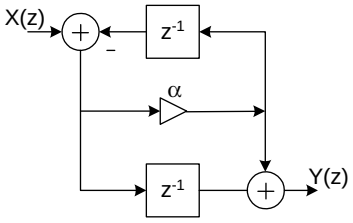
Il existe différentes manières d'implémenter un filtre allpass, chacune d'entre elles favorisant une réduction de l'utilisation d'un élément électronique de base. La première de ces architectures (cf. fig 3.3(a)) est composée de deux multiplieurs. Elle découle directement de la fonction de transfert.



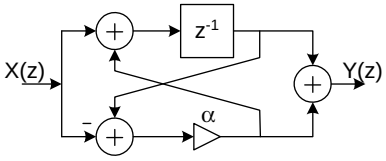
(a) Implémentation canonique



(b) Implémentation à multiplieur unique



(c) Implémentation à multiplieur unique duale



(d) Implémentation à multiplieur unique - alternative

FIG. 3.3 – *Implémentation d'un filtre allpass d'ordre* 1

Après quelques manipulations, en effectuant une pré-somme des deux entrées du multiplieur, l'architecture présentée en Figure 3.3(b) est obtenue. Elle n'utilise qu'un seul multiplieur mais double le nombre de delais. Une propriété intéressante de cette structure concerne le format des données dans sa branche de rétro-action (contenant la sortie de gain unité) et la branche

directe (*feed-forward*, contenant l'entrée retardée). Il peut être défini égal à celui du signal d'entrée sans pour autant amener une détérioration de l'information à traiter. Un autre avantage de cette structure apparaît lorsqu'on cascade plusieurs structures allpass d'ordre 1. Puisque la rétro-action d'un étage contient les mêmes données que la boucle directe de l'étage suivant, les deux délais peuvent alors être fusionnés en un.
La structure duale est présentée en Figure 3.3(c). Elle est obtenue en interchangeant les additionneurs et les noeuds simples de la structure précédente et en changeant le sens de circulation des données. Dans le cas présent, la dernière propriété citée dans le cas précédent n'est plus valable, ce qui rend cette structure moins intéressante.
La dernière structure (cf. fig 3.3(d)) nécessite seulement un delai mais en contre partie utilise 3 additionneurs. Cette structure est celle retenue pour la suite des travaux.

Remarque : En Figure 3.4 est représentée une structure allpass d'ordre 2 (cf. (3.11)) dérivée des Figures 3.3.

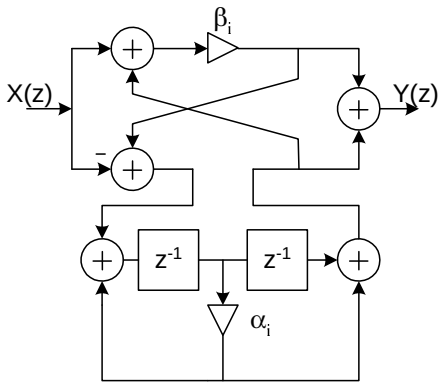


FIG. 3.4 – *Structure allpass d'ordre* 2

2.2.3 Filtres récursifs *two-path Allpass*

Un très large éventail de filtres multi-cadence peut être obtenu à partir d'une combinaison élémentaire de filtres allpass récursifs [9]. La structure de ces filtres est très différente des filtres récursifs traditionnels. Cette structure répond à l'équation suivante :

$$H(z) = \frac{H_1(z) \pm H_2(z)}{2} \tag{3.7}$$

Elle est basée sur la mise en parallèle de deux cascades de filtres allpass (réseau de filtres allpass, Figure 3.5).
En changeant le signe dans l'équation, on obtient dans un cas un filtre passe-bas (+) et dans l'autre un filtre passe-haut(−). Ce réseau de filtre allpass est caractérisé par les équations suivantes (cas du filtre passe bas) :

$$M(\omega) = \left|\cos\frac{1}{2}(\Phi_1 - \Phi_2)\right| \tag{3.8}$$

$$\Phi(\omega) = \frac{1}{2}(\Phi_1 + \Phi_2) \tag{3.9}$$

$$\tau(\omega) = \frac{1}{2}(\tau_1 + \tau_2) \tag{3.10}$$

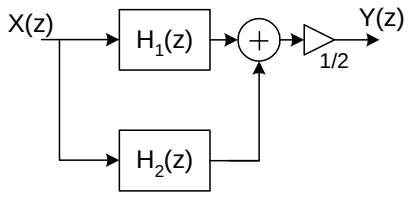


FIG. 3.5 – *Filtre two-path allpass passe-bas*

où $M(\omega)$, $\Phi(\omega)$ et $\tau(\omega)$ sont l'amplitude, la phase et le retard de groupe de la fonction de transfert $H(z)$.
Considérons maintenant les expressions des fonctions de transfert des deux structures parallèles [31] :

$$H_1(z) = z \prod_{i=1}^{\lceil (n+3)/4 \rceil} \frac{\beta_i + \alpha_i(1+\beta_i)z^{-1} + z^{-2}}{1 + \alpha_i + \alpha_i(1+\beta_i)z^{-1} + \beta_i z{-2}} \tag{3.11}$$

$$H_2(z) = \prod_{i=\lceil (n+7)/4 \rceil}^{(n+1)/2} \frac{\beta_i + \alpha_i(1+\beta_i)z^{-1} + z^{-2}}{1 + \alpha_i + \alpha_i(1+\beta_i)z^{-1} + \beta_i z{-2}} \tag{3.12}$$

Des filtres RII elliptique (d'ordre impair) peuvent alors être réalisés par l'intermédiaire du réseau de deux cascades de filtres allpass du 1^{er} et du 2^{nd} ordre. Après avoir fixé un gabarit de filtrage et obtenu les pôles correspondant du filtre elliptique souhaité, les composants du réseau de filtre allpass peuvent être établis.
Pour $\beta_1 = 0$, la cellule allpass de premier ordre du réseau correspond au pôle réel (dû au filtre d'ordre impair) de la fonction de transfert du filtre synthétisé. Elle est donnée par :

$$z\frac{\alpha_1 z^{-1} + z^{-2}}{1 + \alpha_1 z^{-1}} = \frac{\alpha_1 + z^{-1}}{1 + \alpha_1 z^{-1}} \tag{3.13}$$

En considérant que la position des pôles de la fonction de transfert est donnée par :

$$z_i = r_i e^{\pm j\theta_i} \tag{3.14}$$

les paramètres α_i et β_i de $H(z)$ sont déterminés de la manière suivante :

$$\begin{cases} \beta_1 = 0 \\ \alpha_1 = -r_1 \end{cases}$$
$$\begin{cases} \beta_i = r_i^2 \\ \alpha_i = -2\dfrac{r_i cos\theta_i}{1 + r_i^2} \; \forall \, i > 1 \end{cases}$$

le paramètre α_1 représentant le pôle réel.
Le nombre de multiplications nécessaires à l'implémentation de $H(z)$ correspond à l'ordre du filtre, chaque cellule allpass d'ordre 2 (ou 1) nécessitant seulement deux multiplications comme illustré en Figure 3.4(ou une multiplication). Ceci constitue une importante amélioration de l'implémentation comparé à une structure RII directe ou cascade (et parallèle). La faible sensibilité dans la bande passante est une caratéristiques très interessante de ce type de filtre. Nous en tirons partie dans la suite du chapitre. En contrepartie, la bande atténuée est plus sensible que dans le cas de l'utilisation d'une structure cascade SOS classique.

2.3 Implémentation d'un filtre EMQF avec une structure allpass

2.3.1 Présentation de la méthode

Une méthode permettant de diminuer le nombre de cellules de multiplications lors de l'implémentation d'un filtre EMQF sous forme d'un réseau de filtres allpass est exposée en [31]. Il a été expliqué en section 2.1 que la position des pôles dépend uniquement de la fréquence de coupure du filtre. A partir de la Figure 3.1, on détermine la valeur de la distance $(0, x_0)$ de la manière suivante, x_0 étant le centre du cercle sur lequel les pôles du filtre sont situés :

$$x_0 = \frac{1}{\cos(2\pi f_{3dB})} = \frac{1+\tan^2 \pi f_{3dB}}{1-\tan^2 \pi f_{3dB}} \tag{3.15}$$

avec $\tan^2 \pi f_{3dB} = \tan \pi f_p \tan \pi f_a$
où f_p est la fréquence à la limite de la bande-passante et f_a celle à la limite de la bande atténuée. Après quelques manipulations mathématiques, à partir de la Figure 3.1 et par l'intermédiaire de l'équation 3.15, on obtient les relations suivantes [31] reliant les paramètres α_i (composants des sections d'ordre 2 de la structure allpass (3.11)) à x_0. Le calcul des paramètres β_i, lui, reste identique :

$$\alpha_1 = -x_0 \left(1 - \sqrt{1 - \frac{1}{x_0^2}}\right) \text{ représente le pôle réel} \tag{3.16}$$

$$\alpha_i = \alpha = -\frac{1}{x_0} \; \forall \; i > 1 \tag{3.17}$$

$$\beta_i = r_i^2 \tag{3.18}$$

Il est à noter que les constantes $\alpha_i \; \forall \; i > 1$ sont toutes identiques à α. Cette propriété offre l'opportunité de choisir une valeur optimale pour α (la moins complexe à implémenter – somme de puissance de 2) permettant de répondre aux specifications du gabarit. Ainsi les coefficients α_i (pour $i \in]1, \frac{n+1}{2}]$, n étant l'ordre du filtre) peuvent être implémenté de façon exacte (sans quantification), sans aucun multiplieur, simplement à l'aide d'opérations logiques basiques (opération *shift-and-add*).
Un autre avantage de cette méthode concerne la distance des pôles avec le cercle unité qui est plus importante que dans le cas d'un filtre elliptique classique. Ceci assure une plus faible sensibilité des coefficients β_i à la quantification.
Pour des valeurs de f_a et f_p données, l'ordre de grandeur de α est déterminé ainsi :

$$-\frac{1-tan^2\pi f_p}{1+tan^2\pi f_p} < \alpha < -\frac{1-tan^2\pi f_a}{1+tan^2\pi f_a} \tag{3.19}$$

La valeur choisie doit être la plus proche de celle obtenue avec l'équation 3.20 :

$$\alpha \approx -\frac{1-tan\pi f_p tan\pi f_a}{1+tan\pi f_p tan\pi f_a} \tag{3.20}$$

La valeur de α fixée, on en déduit la nouvelle valeur de f_{3dB} et ensuite celle de $alpha_1$. En se fixant $F_a = f_a$, on en déduit la nouvelle valeur de la fréquence à la limite de la bande passante F_p ou vice et versa. Il ne reste alors plus qu'à calculer les valeurs des coefficients β_i à partir des pôles provenant du nouveau gabarit F_a, F_p (3.15).

Remarque : Le filtre RII demi-bande est un cas particulier des filtres RII dérivé d'un prototype analogique EMQF. Dans ce cas le cercle de centre x_0 (cf. fig 3.1) fusionne avec l'axe des imaginaires ($\alpha = 0$) et de ce fait, les pôles sont situés sur ce même axe entre les points $-j$ et j. Seulement le calcul des coefficients β_i est alors nécessaire.

Après avoir calculé les pôles aboutissant au gabarit fixé (méthode classique, programmes Matlab), il est nécessaire de répartir convenablement ces derniers parmi les deux groupes H_1 et H_2 (cf. équation 3.11). Les pôles de $H_1(z)$ et $H_2(z)$ sont distribués alternativement dans le plan z. Partant de cette remarque, les pôles sont ordonnés de la manière suivante dans l'ordre croissant des modules : $H_1(z)$ comprend le pôle réel et une paire de pôles conjugués sur deux, $H_2(z)$ comprend les pôles restant (cf. fig 3.6).

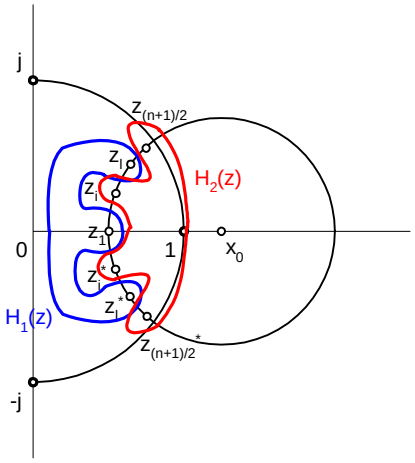


FIG. 3.6 – *Partage des pôles*

2.3.2 Alternative : implémentation sans multiplication

En [32], une méthode basée sur l'implémentation des coefficients sans l'utilisation d'un seul multiplieur est présentée. Elle est basée sur une analyse de la sensibilité de la phase du réseau de filtres allpass à la quantification des coefficients.
Comme exposé dans la section 2.3.1, $(n+1)/2$ constantes peuvent être implémentées sans quantification. Il est d'ailleurs montré que ces dernières sont les plus sensibles et peuvent être sources d'erreur. Les $(n-1)/2$ constantes restantes sont quantifiées à l'aide d'une technique de tolérence en phase.
Habituellement le gabarit d'un filtre est défini par des spécifications en amplitude caractérisant les ondulations dans la bande passante, l'atténuation de la bande d'arrêt (δ_p et δ_a respectivement) ainsi que les fréquences définissant la bande passante et la bande d'arrêt (f_p et f_a respectivement). Les tolérence en amplitude peuvent être définies ainsi :

$$T_A = \{\delta_a - A_a, \delta_p - A_p, f_a - F_a, f_p - F_p\} \tag{3.21}$$

A_a, A_p, F_a et F_p représentant les valeurs d'atténuation et de fréquence obtenues après synthèse du filtre.
Dans le cas présent, cette tolérence en amplitude doit être traduite en une tolérence en phase. La

tolérence en amplitude s'appliquait au gabarit $A(\omega) = \left|H(e^{j\omega})\right|$. Or A peut aussi être exprimé en fonction de la phase $\psi = \frac{\phi_1(\omega)-\phi_2(\omega)}{2}$ (ϕ_1 étant la phase de la branche supérieure de la structure composée de 2 cascades de filtres allpass) :

$$A(\omega) = cos(\psi(\omega)) \tag{3.22}$$

A l'aide des équations suivantes [32], le gabarit en fréquence est determiné :

$$D_p = cos^{-1}(1 - 10^{a_p/20}) \tag{3.23}$$

$$D_a = \frac{\pi}{2} - cos^{-1}(10^{a_a/20}) \tag{3.24}$$

où D_p et D_a représentent les variations permises de $\psi(\omega)$ dans la bande passante et dans la bande d'arrêt, respectivement. Les tolérences présentées en équation 3.21 peuvent être redéfinies ainsi :

$$T_P = \{D_a - \psi_a, D_p - \psi_p, f_a - F_a, f_p - F_p\} \tag{3.25}$$

avec ψ_a et ψ_p les ondulations obtenues après synthèse du filtre (Figure 3.7).

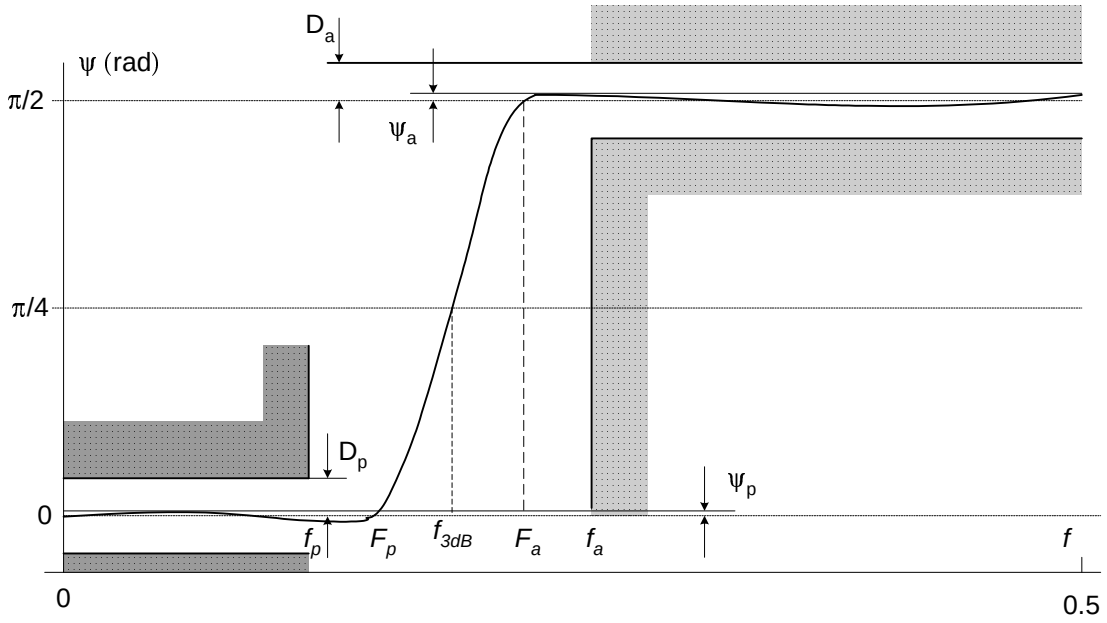


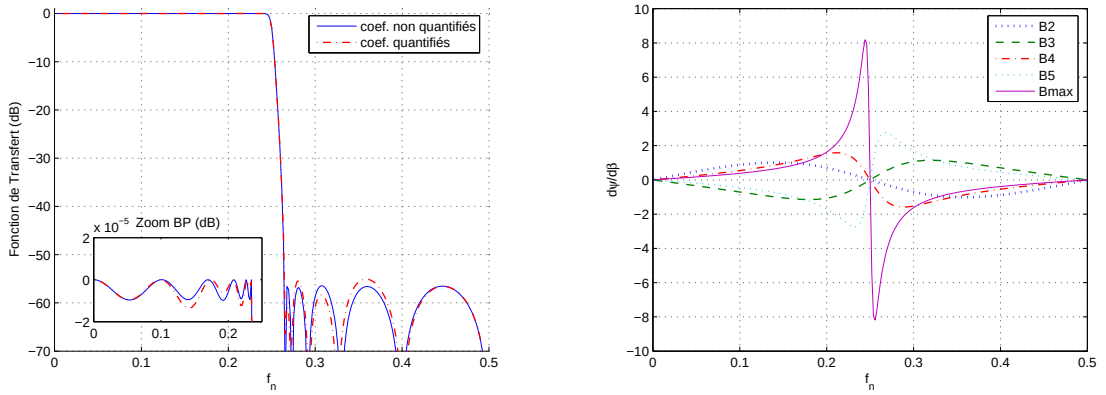
FIG. 3.7 – *Gabarit de tolérence en phase*

Du fait de la faible sensibilité de la structure *two-path allpass* dans la bande passante, les répercutions des différentes quantifications sur la phase sont seulement analysées dans la bande d'arrêt. La méthode à suivre afin de dimensionner correctement les constantes β_i quantifiées afin de les implémenter sans multiplieur est décrite en [32].

2.3.3 Exemple d'application : Filtre Half-band

Il a été énoncé dans l'introduction de ce chapitre que les différentes techniques exposées ont pour objectif l'optimisation en terme de ressources du 2^{nd} étage du filtre TFB. Pour obtenir un

filtre répondant aux spécifications (47 dB d'atténuation, une bande de transition de largeur 1/32 – de 15/64 à 17/64 en fréquence normalisée – et une ondulation dans la bande passante de .2 dB) du TFB, un filtre RII EMQF d'ordre 11 est nécessaire. Ceci se traduit en une structure *two path allpass* comportant 3 coefficients dans une branche et de 2 dans l'autre.
En Figure 3.8(a) est tracée la réponse en fréquence du filtre avant et après quantification des coefficients.



(a) Fonction de transfert du filtre

(b) Sensibilité des coefficients de la structure all pass

FIG. 3.8 – *Fonction de transfert d'un filtre d'ordre* 11 *(*5 *coefficients)*

On remarquera la marge considérable obtenue dans la bande atténuée par rapport à la spécification de départ. En effet, l'algorithme destiné au calcul des coefficients du filtre EMQF ne résulte, pour un ordre 9, qu'en une atténuation de l'ordre de 45 dB, inférieure à la valeur attendue (ce résultat est lié à l'étroitesse de la bande de transition requise qui limite l'algorithme). Pour cette raison, il a fallu augmenter la valeur de l'ordre du filtre, ce qui aboutit à une atténuation de l'ordre de 55 dB. Il est à noter que l'ondulation dans la bande passante est très faible(de l'ordre de 10^{-5} dB).
La sensibilité de la fonction de transfert en fonction des coefficients β_i est par ailleurs tracé en Figure 3.8(b). Le coefficient le plus sensible est B_{max}, ce qui conforte le raisonnement tenu lors de la quantification des différents coefficients [32].
Après application de cette méthode, basée sur la technique de tolérence en phase (cf. équation 3.25), les coefficients suivants sont obtenus :

$$\begin{cases} \beta_{max} = 1 - 2^{-4} \\ \beta_5 = 0,7982133110273367 \approx 1 - 2^{-2} + 2^{-4} - 2^{-6} + 2^{-9} \\ \beta_4 = 0,6076484943893845 \approx 2^{-1} + 2^{-3} - 2^{-6} \\ \beta_3 = 0,3569709841291784 \approx 2^{-1} + -2^{-3} - 2^{-6} \\ \beta_2 = 0,1093064542236781 \approx 2^{-3} - 2^{-6} + 2^{-10} \end{cases} \tag{3.26}$$

L'évolution de la phase dans la bande d'arrêt durant le processus de quantification des différents coefficients (du coefficients β_5 à β_2) est tracée en Figure 3.9.
La limite supérieure, à ne pas dépasser pour conserver les spécifications de départ, D_a répond à l'équation 3.24. Le résultat de la quantification est plus que satisfaisant (Figure 3.8(a)), permettant d'implémenter les coefficients sans aucun multiplieur et assure la stabilité du filtre (pôle compris à l'intérieur du cercle unité).
Le filtre, composé de 5 structures allpass, assure donc une atténutaion de l'ordre de 55 dB.

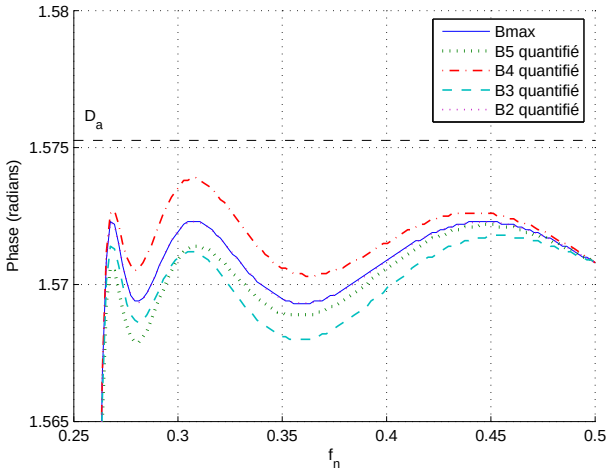


FIG. 3.9 – *Evolution de la caratéristique de phase durant le processus de quantification*

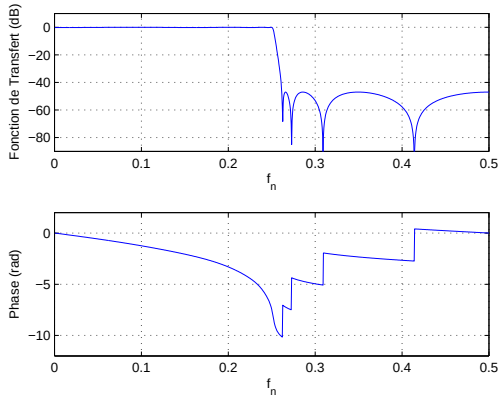
Cependant, en acceptant une légère dégradation de la spécification en atténuation (45 dB), la compléxité du filtre peut encore être diminuée par l'utilisation de seulement 4 structures allpass. Une remarque importante concernant l'implémentation d'un tel filtre sans aucune multiplication doit cependant être faite. Du fait de la réalisation des coefficients à l'aide d'additions et surtout de décalages de bits, le format du signal d'entrée doit être adapté. En effet, afin de ne pas dégrader le traitement du signal effectué par ce filtre, le signal d'entrée doit être codé sur un nombre de bits suffisament grand.

3 Linéarisation de la phase - modification de l'architecture électronique

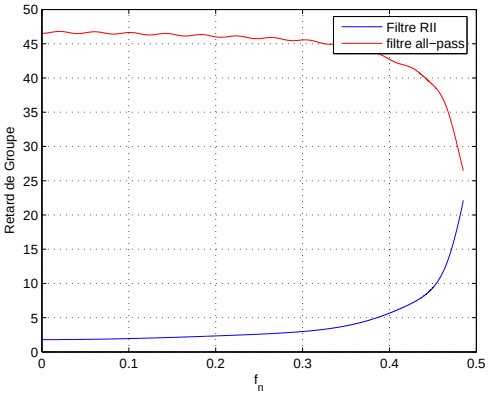
Après avoir présenté des méthodes aboutissant à des structures de filtre RII à faible encombrement, il est désormais nécessaire de s'intéresser à la linéarisation de la phase. La spécification du projet ALMA concernant la phase induite par les systèmes électroniques est stricte : phase linéaire avec une ondulation maximale de .3. Cependant la phase induite par les filtres RII est non-linéaire.
Les méthodes permettant la linéarisation de la phase présentées dans cette section sont toutes basées sur l'assemblage ou la modification de différentes structures électroniques. La solution la plus répandue est caractérisée par l'association d'un filtre récursif RII et d'un égaliseur de phase permettant de compenser la non-linéarité de la phase du premier.
D'autres méthodes basées sur des structures utilisant des filtres allpass y sont exposées. Différents auteurs [33, 34] ce sont intéressés à ces dernières qui utilisent une architecture *two path allpass* modifiée.
Une dernière méthode qui, théoriquement, fournit une phase strictement nulle caractérisée par la double application d'un filtre RII $H(z)$ est aussi présentée dans cette section.
Dans la suite, ces méthodes sont comparées en termes d'efficacité et d'encombrement résultant de l'implémentation. Dans chacune de ces méthodes, le filtre RII utilisé est un filtre dont l'architecture découle de la structure électronique la plus performante afin de fournir une solution finale optimale.

3.1 Filtre égaliseur de phase

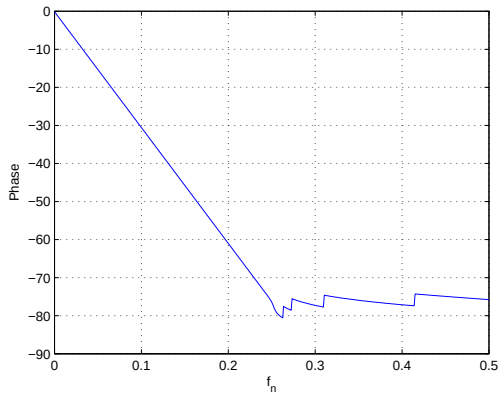
Une des premières méthodes employée dans l'optique de la linéarisation de la phase des filtres RII a été l'utilisation d'un filtre allpass qui, cascadé au filtre RII, permet de compenser la non-linéarité de la phase induite par ce dernier. Ce type de filtre est appelé filtre égaliseur de phase [35]. La méthode consiste à synthétiser un filtre allpass dont le retard de groupe permet de compenser celui induit par le filtre RII afin d'obtenir un retard de groupe total quasiment constant. Cette compensation peut seulement être effectuée dans la bande passante du filtre où la phase linéaire est désirée. En Figure 3.10 sont présentés les réponses en phase du filtre RII de type elliptique, du filtre allpass et de la cascade des 2.



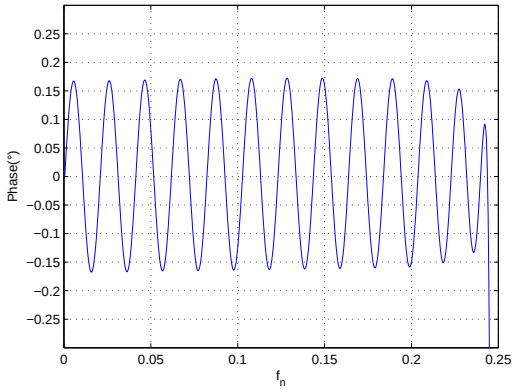
(a) Réponses du filtre RII (ordre 8)



(b) Retards de groupe des 2 étages



(c) réponse en phase de la cascade



(d) Ondulation de la phase résultante

FIG. 3.10 – *Retards de groupe et réponses en phase de l'ensemble filtre RII - égaliseur de phase*

La synthèse du filtre RII a été réalisée dans le but de respecter les spécifications du 2^{nd} étage TFB.
L'ondulation de la phase obtenue dans la bande passante avoisine les 0.3. Un filtre allpass d'ordre 24 a été utilisé afin d'aboutir à cette caractéristique à la limite de la spécification.
En comparant le nombre de coefficients nécessaires à l'implémentation de cette structure à celui d'une structure filtre RIF, nous pouvons évaluer l'intérêt de l'utilisation d'une telle structure. Le filtre RIF utilisé dans le filtre TFB est constitué de 64 coefficents symétriques, l'impléméntation de 32 coefficients est donc suffisante ce qui correspond à l'utilisation de 32 multiplications. En ce

qui concerne la cascade RII + allpass, 24 coefficients sont nécessaires à l'implémentation du filtre allpass sous forme lattice (Figure 3, les coefficients sont quantifiés sur 10 bits) et 4 coefficients pour le filtre RII demi-bande (d'ordre 8) en utilisant une structure *two-path allpass*. La cascade nécessite donc l'utilisation de 28 multiplications contre 32 pour la structure RIF (sans utiliser la méthode de recirculation des poids divisant encore le nombre de multiplieurs de la structure RIF par 2).
Le tableau 3.1 présente les résultats de la synthèse de cette structure.

TAB. 3.1 – Ressources en ALMs

	ALMs	Frequence max.
Structure IIR	239	98 MHz
Structure allpass - égaliseur	1268	82 MHz
Cascade	1507	97 MHz

Le filtre allpass permettant la linérisation de la phase occupe une partie non négligeable des ressources. Le fait que la fréquence de fonctionnnement n'atteigne pas celle escomptée (supérieure à 125 MHz) nous force à complexifier la structure par l'adjonction de bascule D (méthode de *pipelining*) permettant d'améliorer le routage (diminution des longueurs entre deux éléments du design) et donc la fréquence maximale de fonctionnement.

3.2 Modification de la structure *Two pass allpass*

Il est possible de modifier la structure présentée en section 2.2.3 afin de former un filtre comportant une phase approximativement linéaire (linéarité présente dans la majeure partie de la bande passante) [33].
Le filtre d'ordre N composé des deux cascades de filtre allpass H_1 et H_2 (de phase Φ_1 et Φ_2)doit répondre aux spécifications suivantes :

- Dans la bande passante, l'amplitude doit être approximativement égale à 1. En utilisant (3.8), on en déduit la condition suivante : $\Phi_1 \approx \Phi_2$. La condition de phase approximativement linéaire se traduit par $\Phi \approx -k\omega$ avec $k > 0$, entrainant $\Phi_{1,2} \approx -k\omega$.
- Dans la bande d'arrêt, l'amplitude doit avoisinner 0. On en déduit par l'intermédiaire de (3.8) que $\Phi_1 \approx \Phi_2 + \pi$.

La fonction de transfert de la branche supérieure d'ordre $n = 2\lfloor N/4 \rfloor + 1$ est de la forme :

$$H_1(z) = \frac{\sum_{i=0}^{n} \beta_{n-i} z^{-i}}{\sum_{i=0}^{n} \beta_i z^{-i}} \tag{3.27}$$

où $\beta_0 = 1$. La seconde branche possède une fonction de transfert identique à celle-ci mais d'ordre $n = 2\lfloor (N+2)/4 \rfloor$.
La phase induite par cette dernière s'écrit ainsi :

$$\Phi_{H_1}(\omega) = -n\omega - 2\arctan\left(\frac{-\sum_{m=1}^{n} \beta_m \sin m\omega}{\sum_{m=0}^{n} \cos m\omega}\right) \tag{3.28}$$

L'algorithme présenté en [33], permettant la conception d'un filtre *two pass allpass* à phase approximativement linéaire, est basé sur la résolution d'un système d'équations « sur-déterminé » dans le sens des moindres carrés. Le système d'équations, d'inconnues β_m, est obtenu en faisant varier $\omega = 2\pi f_n$ dans l'intervalle des fréquences normalisées :

$$\sum_{m=0}^{n-1} \beta_m \sin\left(\frac{1}{2}(\Phi - n\omega) + m\omega\right) = -\sin\left(\frac{1}{2}(\Phi + n\omega)\right) \tag{3.29}$$

l'équation étant une reformulation de (3.28). Ce système doit être appliqué dans un premier temps à la branche supérieure pour des valeurs de ω couvrant la bande passante. Dans cette bande de fréquence la phase Φ_1 est égale à $-k\omega$. Une fois les valeurs β_m trouvés, une factorisation de (3.30) est ensuite réalisée afin de déterminer les racines de cette dernière ou coefficients du filtre.

$$f(x) = \sum_{m=0}^{n} \beta_m x^m \tag{3.30}$$

La même méthode est ensuite appliquée à la branche inférieure mais cette fois-ci sur un intervalle de fréquence couvrant la bande passante (où $\Phi_2 = \Phi_1$) et la bande d'arrêt (où $\Phi_2 = \Phi_1 - \pi$).

Un cas particulier, permettant de simplifier l'implémentation matérielle, est obtenu en restreignant une des branches à un délai pur [9] (fixant la valeur de la constante k, Figure 3.11).

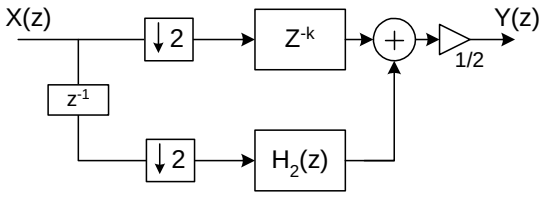
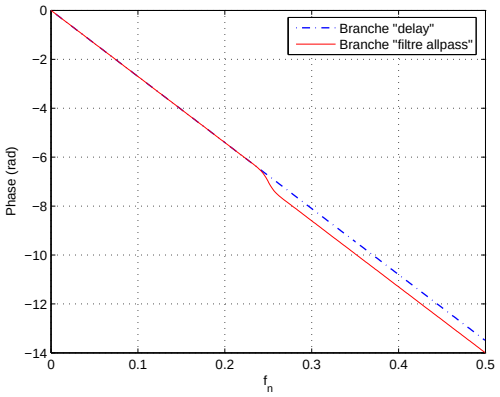
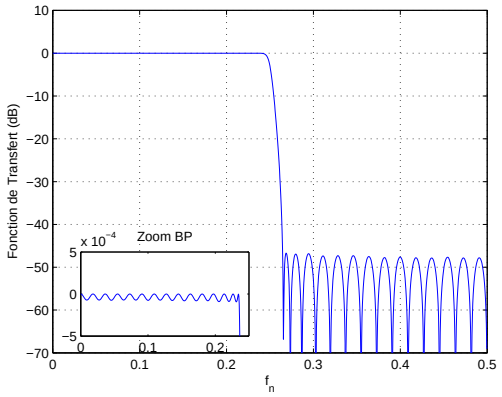


FIG. 3.11 – *Schéma du filtre allpass 2 branches à phase linéaire*

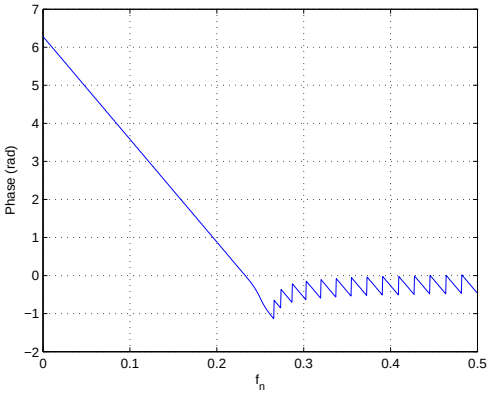
Du fait de cette manipulation, le processus de synthèse du réseau de filtres allpass perd des degrés de liberté et, à performances identiques, un filtre modifié nécessite des sections allpass additionnelles (section d'ordre 1 et d'ordre 2 en z^2). Pour un filtre allpass H_2 d'ordre n pair, entrainant le délai k à prendre la valeur de n, le filtre total est d'ordre $2n + 1$.
Une synthèse répondant aux spécifications du 2^{nd} étage de filtrage TFB (cf. tableau 1.3) a été réalisée à l'aide de l'algorithme utilisé en [9]. Afin de répondre à la spécification en phase (ondulation maximale de .3)L'ordre du filtre H_2 nécessaire est égal à 28. Ce filtre est composé de 20 coefficients découpés en 2 sections d'ordre 1 et 9 sections d'ordre 2. Pour l'obtention du même filtre sans la caractéristique de linéarité de la phase, seulement 4 coefficients sont nécessaires. En Figure 3.12 sont tracées les réponses en phase, en amplitude ainsi que le retard de groupe du filtre obtenu.
La phase résultant de cette architecture de filtrage est bien approximativement linéaire dans la bande passante et les spécifications en amplitude sont respectées. La quantification des coefficients amène une dégradation de l'ondulation de la phase et de l'atténuation dans la bande d'arrêt. Une quantification sur 12 bits permet de conserver des spécifications acceptables (ondulation de la phase $\pm 0.15°$, atténuation ≈ 47 dB).



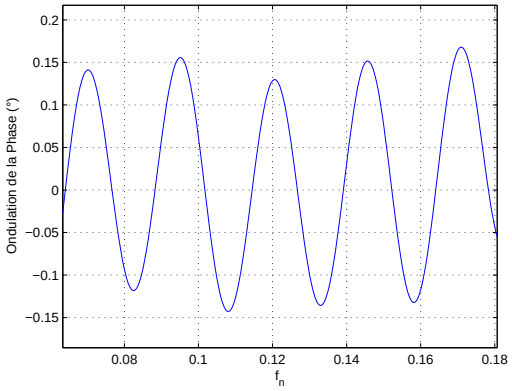
(a) Réponse en phase de chacune des branches



(b) Fonction de transfert de la structure assemblée



(c) Réponse en phase de la structure assemblée



(d) Ondulation de la phase dans la BP

FIG. 3.12 – *Caractéristiques du filtre allpass*

TAB. 3.2 – Ressources en ALMs

	ALMs	Frequence max.	DSP
Structure *two-path allpass* modifiée	984	74 MHz	non

Cette architecture a été synthétisée et les résultats de la compilation sont ceux présentés dans le tableau 3.2.
Il est à noter que les deux branches de la structure fonctionnent à une fréquence moitié de l'horloge principale (décimation par 2 en entrée de chaque branche, Figure 3.11). Ici encore, la synthèse a été réalisée dans le but de comparer l'encombrement des différentes structures, aucune optimisation de l'utilisation des différentes ressources n'a été effectuée.

3.3 Filtre « deux passages »

Cette solution est basée sur la double application d'un filtre de fonction de transfert $H(z)$ au signal à traiter.(Figure 3.13) :

$$H_{LP} = H(z) \cdot H(z^{-1}) = |H(z)|^2 \tag{3.31}$$

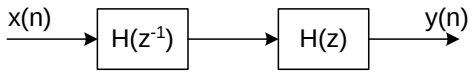


FIG. 3.13 – *Schéma du filtre à phase linéaire à deux passages*

La réponse impulsionnelle inverse, donc non causale, du filtre puis la réponse impulsionnelle causale sont appliquées à la suite. Cette manipulation permet l'obtention d'un filtre à phase linéaire et à caratéristiques doubles de celles du filtre de fonction de transfert $H(z)$ (puisque la nouvelle fonction de transfert H_{LP} est égale au carré du module de $H(z)$) : l'atténuation dans la bande d'arrêt et l'ondulation dans la bande passante sont doublées.
Pourtant, une telle structure n'est pas implémentable en hardware (dans une puce FPGA) à cause du filtre à réponse impulsionnelle non-causale (ou encore le traitement associé à ce filtre à effectuer en temps continu). En [36] est présentée une méthode permettant de contourner ce problème en implémentant le filtre $H(z^{-1})$ sur deux voies différentes, ces deux voies étant solicitées à tour de rôle et traitant les données d'entrée inversées. Ce traitement par bloc de données (*block processing*) consiste à diviser le flot d'entrée continu en section de longueur L. En Figure 3.14 est présentée la structure électronique permettant de réaliser cette fonctionnalité.

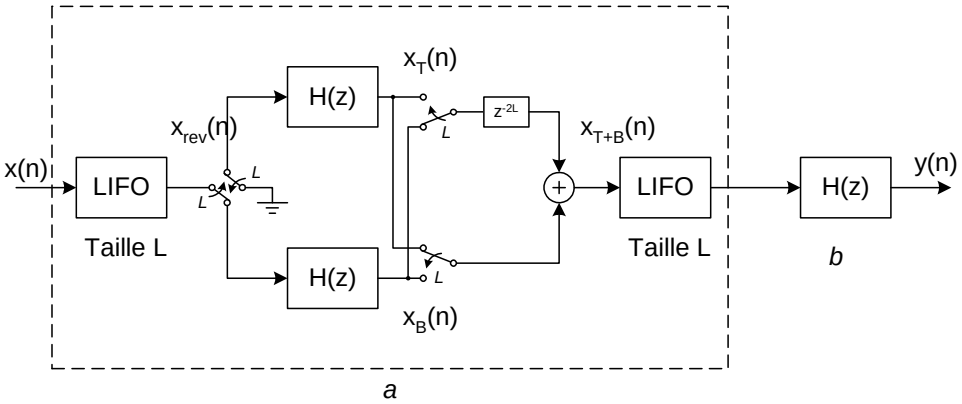


FIG. 3.14 – *Schéma électronique du filtre à phase linéaire, méthode overlap-add*

Le fonctionnement de cette structure électronique est expliquée en détails dans les sections qui suivent.

3.3.1 Block processing - présentation

Une condition nécessaire à la causalité d'un filtre est la durée finie de sa réponse impulsionnelle. Dans le cas d'un filtre RII, cette réponse impulsionnelle doit être tronquée afin de répondre à ce critère. Cependant, on peut considérer que la réponse impulsionnelle résultant de la réalisation physique d'un filtre RII est toujours de durée finie du fait de la quantification de ces coefficients, réalisée sur un nombre fini de bits.
Dans le cas d'une séquence d'échantillons d'entrée divisée en sections de longueur L, la réponse impulsionnelle du filtre implémenté doit être inférieure ou égale à ce même L. Ceci induit des blocs de sortie de longueur inférieure ou égale à $2L-1$ (longueur du résultat de l'opération de convolution). Pour réaliser le filtrage non-causal $H(z^{-1})$ avec un signal d'entrée de durée infinie, des sections de longueur L sont inversées par l'intermédiaire d'une LIFO (Last In First Out) et

traitées en parallèle par le filtre $H(z)$. Les données recombinées par l'intermédiaire d'une méthode dite *overlap-add* sont ensuite ré-inversées (LIFO de longueur L) afin de retrouver un flot de données continu et cohérent (Figure 3.14, *a*).

3.3.1.1 Overlap-add method La méthode *overlap-add* est illustrée en Figure 3.15, l'agencement des différentes trames y est exposé. Chaque sortie de filtre de longueur $2L-1$ ($2L$ en y ajoutant la phase de remise à zéro du filtre, correspondant à l'émission d'un zéro) est divisée en 2 parties : la partie *lead* (menante) et la partie *trail* (suivante). La recombinaison des signaux est réalisée en ajoutant deux à deux la partie lead du filtre bas avec la réponse du filtre haut présente $2L$ échantillons avant et la partie trail du filtre haut avec la réponse du filtre bas présente $2L$ échantillons avant.

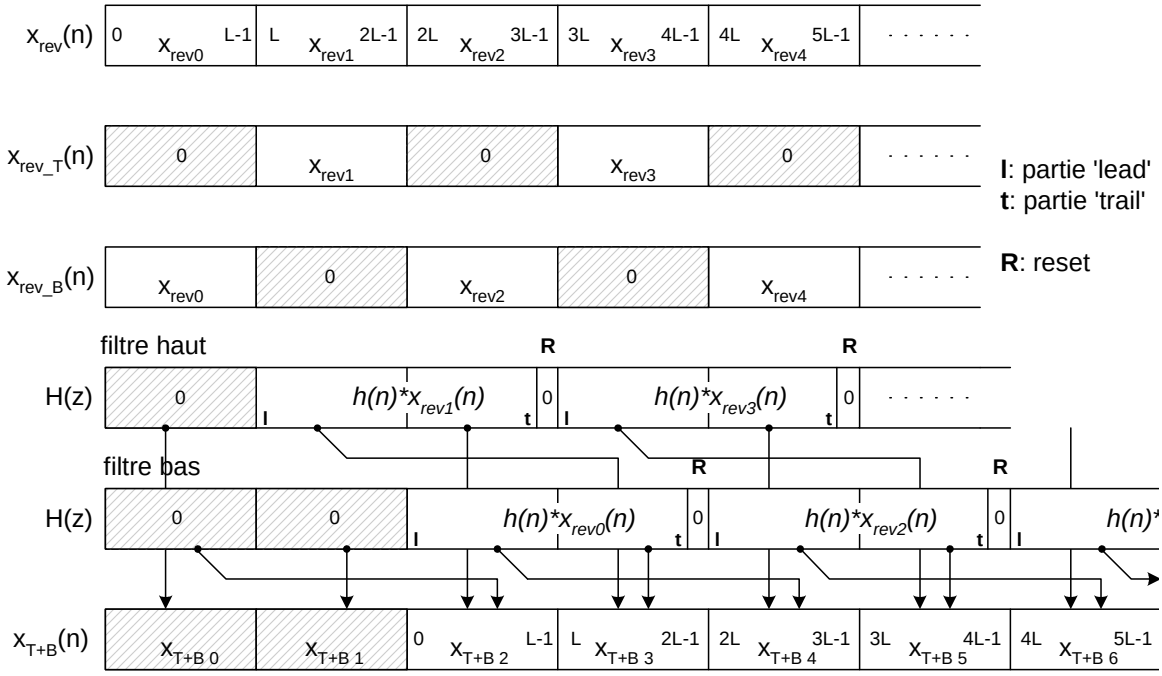


FIG. 3.15 – *traitement par blocs, méthode overlap-add*

3.3.1.2 LIFO La LIFO quant à elle permet d'inverser des sections d'échantillons de longueur L :

$$x_{rev}(n) = \{\underbrace{x(L-1),\ldots,x(0)}_{section\ 0},\underbrace{x(2L-1),\ldots,x(L)}_{section\ 1},\underbrace{x(3L-1),\ldots,x(2L)}_{section\ 2},x(4L-1),\ldots\}$$

Après ré-assemblage des données provenant des branches de filtre, le signal est une nouvelle fois inversé pour retrouver la cohérence temporelle de l'entrée $x(n)$ avant d'attaquer la deuxième phase de filtrage (2^{eme} passage dans le filtre $H(z)$).
Ces deux techniques permettent un filtrage en temps réel du signal d'entrée par le filtre non causal $H(z^{-1})$. Toutefois, la valeur de L a son importance dans le design de ce filtre RII à phase linéaire. Plus cette valeur est importante plus le temps de réponse du système est grand (et égal à $2L$). Dans notre cas, le temps de réponse n'est pas un paramètre d'importance puisque les

temps d'intégration du système ALMA peuvent atteindre plusieurs heures pour une fréquence d'échantillonage de 4 GHz).

3.3.2 Influence de la longueur L

Il est énoncé en [36] qu'une sortie provenant d'une convolution continue, effectuée avec un filtre à réponse impulsionnelle infinie, est équivalente à une sortie provenant d'une convolution sectionnée, effectuée avec un filtre à réponse impulsionnelle tronquée, pour une longueur de section L suffisamment importante. Le fait qu'une réponse impulsionnelle soit plus ou moins raccourcie a une répercution sur l'ondulation dans la bande passante (phase et amplitude).
En définissant la différence entre la réponse impulsionnelle originale et celle tronquée de longueur L comme suit ($h(m)$ étant la réponse impulsionnelle du filtre) :

$$\epsilon(L) = \sum_{m=L}^{\infty} |h(m)| \tag{3.32}$$

on peut établir un critère de choix pour la longueur minimale L' [36] :

$$\epsilon(L') = \sum_{n=L'}^{\infty} |h(n)| = min(2\delta_p, 2\delta_a) \tag{3.33}$$

La longueur minimale L' de la section permettant de ne pas dégrader l'efficacité de l'opération de filtrage est obtenue lorsque $\epsilon(L')$ est inférieur aux spécifications d'ondulation.

Remarque : Une structure basée sur le même principe que [36] est étudié en [37], cependant cette fois-ci le filtre non-causal et son homologue causal ne sont pas réalisés avec un même filtre $H(z)$ mais chacun par une partie d'un filtre two-pass allpass (H_a et H_b). Une représentation alternative de l'équation 3.7 d'une structure allpass est donnée ci-dessous :

$$H_{LP}(z) = (1 + H_a(z)H_b(z^{-1}))/2 \tag{3.34}$$

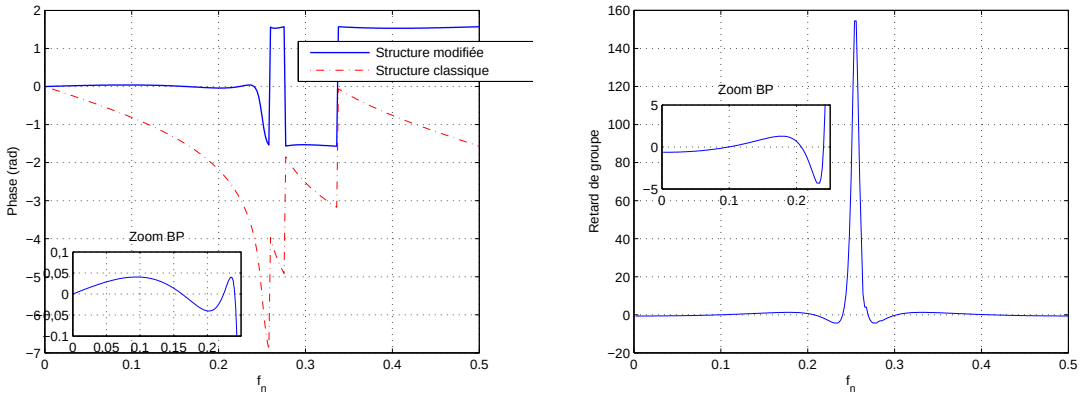
Cette réalisation conduit à l'obtention d'une phase approximativement linéaire (Figure 3.16).
Le retard de groupe résultant d'une telle structure, présenté en Figure 3.16(b), permet d'évaluer l'importance de la non-linéarité. Comparée à celle obtenue avec les structures présentées précédemment, la méthode est moins efficace puisque conduisant à une valeur plus importante de la non-linéarité de la phase. La réalisation électronique retranscrivant cette structure n'a, de ce fait, pas été effectuée.

3.3.3 Application au système ALMA

Dans l'optique du remplacement du 2^{nd} étage TFB FIR, la complexité d'implémentation d'une telle architecture a été étudiée. Comme annoncé précédemment, le filtre $H(z)$ à synthétiser doit présenter des caractéristiques deux fois moins importantes que celles du filtre H_{LP} résultant du traitement :

$$\begin{cases} A_{a_H} = \frac{A_{a_{H_{LP}}}}{2} = 24dB \\ A_{p_H} = \frac{A_{p_{H_{LP}}}}{2} = 0,1dB \end{cases} \tag{3.35}$$

En Figure 3.17, sont tracées les fonctions de transfert des filtres $H(z)$ et $H_{LP}(z)$.
Les caractéristiques du filtre final $H_{LP}(z)$ sont bien doublées par rapport au filtre initial conçu.



(a) Réponse en phase (b) Retard de groupe

FIG. 3.16 – *Caractéristique en phase et retard de groupe*

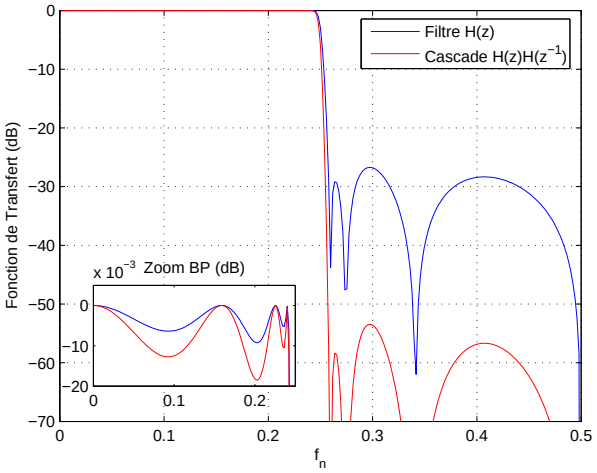


FIG. 3.17 – *Fonctions de transfert $H(z)$ et $H_{LP}(z)$*

3.3.3.1 Synthèse du filtre La méthode employée ici est celle décrite dans la section 2.3.1. Le filtre obtenu est d'ordre 7, 3 coeffcients sont nécessaires à son implémentation. En Figure 3.18 apparaisent les fonctions de transfert quantifiée (implémentation sans multiplieurs) et non-quantifiée.
La quantification des coefficients a été effectuée par l'intermédiaire de la méthode dite de « tolérence en phase ».

$$\begin{cases} \beta_{max} = 1 - 2^{-4} \\ \beta_3 \approx 1 - 2^{-2} - 2^{-6} \\ \beta_2 \approx 2^{-2} + 2^{-4} \end{cases} \tag{3.36}$$

La structure électronique retenue pour l'implémentation du filtre demi-bande est proposée en Figure 3.19.
L'implémentation des coefficients est donc réalisée à l'aide d'opérations simples tels que l'addition ou le décalage de bits (le nombre de bit supprimé est illustré par un chiffre au dessus de la ligne en question). Les opérations arithmétiques de chaque branche sont opérées à un rythme deux

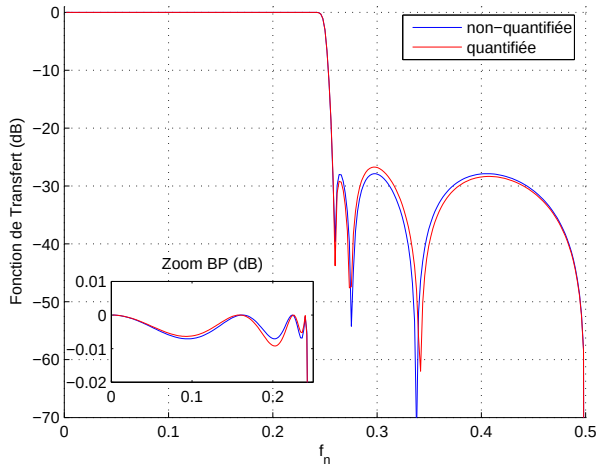


FIG. 3.18 – *Filtre $H(z)$ sans et avec coefficients quantifiés*

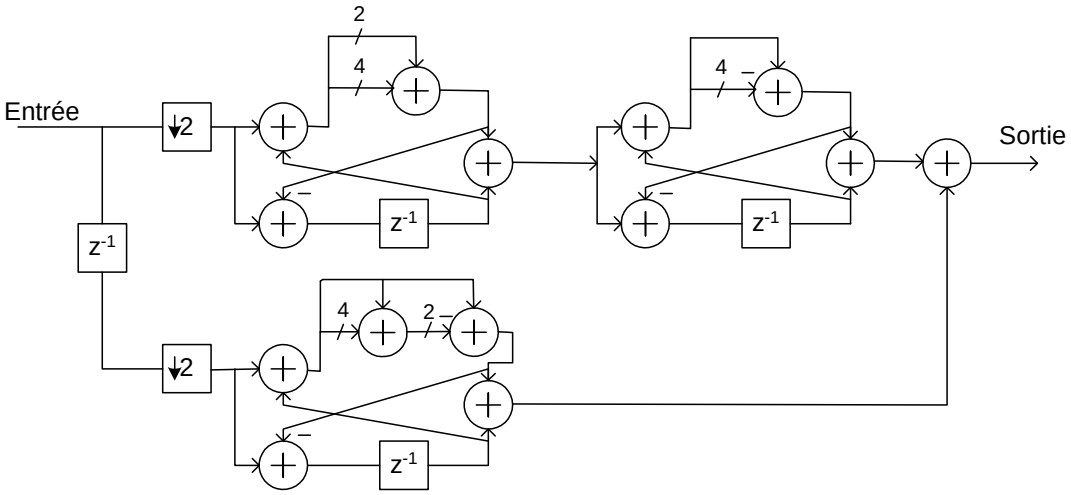


FIG. 3.19 – *Implémentation du filtre demi-bande $H(z)$ sous forme 2-path allpass*

fois moins élevé que celui de l'entrée, la décimation par 2 étant réalisée en amont des sections allpass.
Comme cela a été mentionnée en section 2.3.3, le format du signal d'entrée est un paramètre important lorsque une opération de décalage est effectuée. Cela doit être pris en considération lors de la réalisation électronique de la structure. En Figure 3.20 sont comparées différentes sorties du filtre pour différents formats d'entrée.
Les résultats d'implémentation présentés dans la suite du document tiennent compte de cette remarque.

3.3.3.2 Modélisation de la structure deux passages Le filtre $H(z)$ décrit précédemment, une fois ses poids quantifiés, possède une réponse impulsionnelle de longueur $l = 360$. Afin d'obtenir une erreur $\varepsilon < \min(2\delta_p, 2\delta_a) = 4.5 \cdot 10^{-3}$, On choisit L égal à 200, valeur légèrement supérieure à la valeur minimale acceptable pour L (cf. (3.33), L_{min} étant égal à 172).

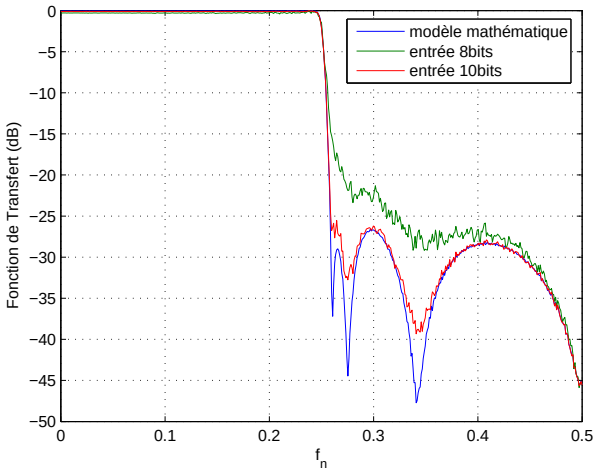


FIG. 3.20 – *Fonction de transfert de la structure électronique du filtre two-path allpass*

En appliquant le signal d'entrée de la Figure 3.21(a), qui correspond à la sortie du filtre CIC-2HB étudié précedemment, à notre structure de filtrage à deux passages, pour une LIFO de profondeur $L = 200$, on obtient les spectres de sortie des filtres $H(z^{-1})$ et $H(z)$ tracés en Figures 3.21(b) à 3.21(d).

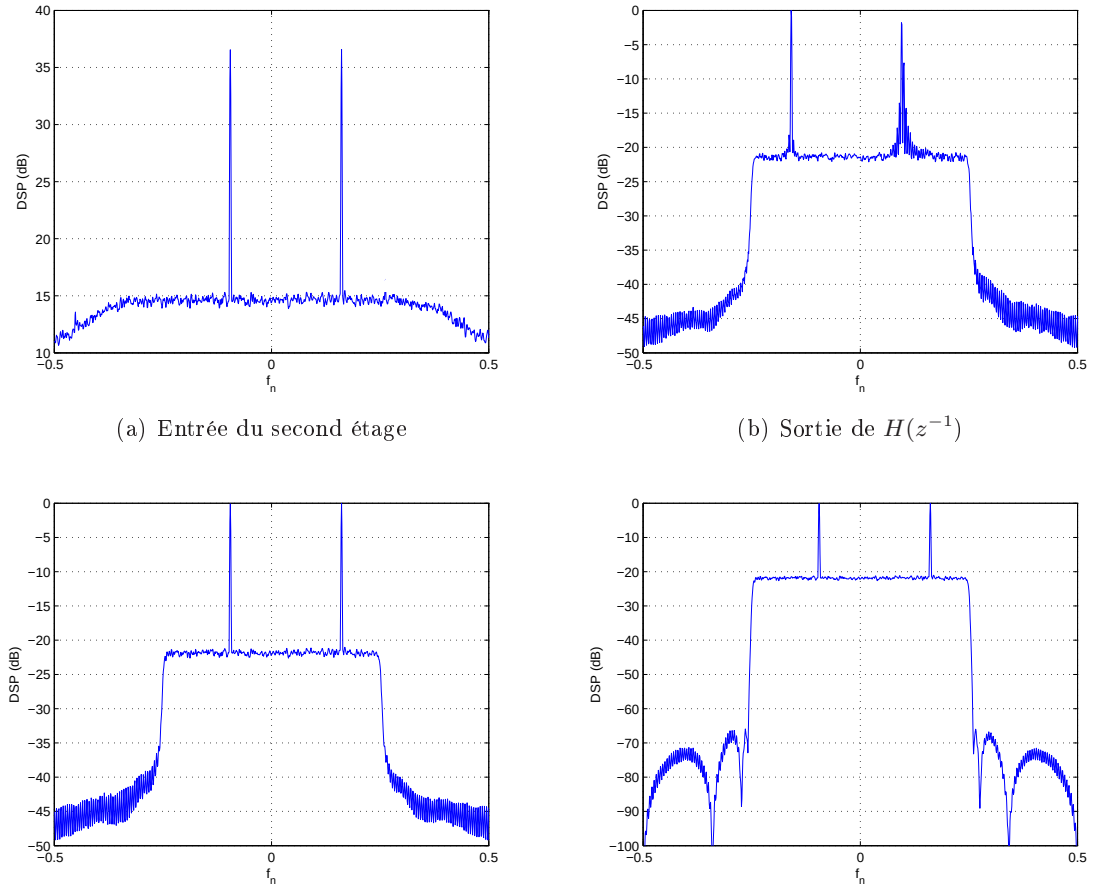
Les spectres sont de nature complexe, le traitement effectué ici est semblable à celui effectué par le filtre TFB originel sur le même signal d'entrée : le premier étage TFB a été remplacé par la cascade CIC-2HB et le second étage par le filtre présentement étudié (filtre RII à phase linéaire). Sur la Figure 3.22 sont tracées la phase induite par le filtre $H(z^{-1})$ ainsi que celle induite par la structure complète.
La caractéristique de phase linéaire est bien atteinte. Le zoom permet de se rendre compte de l'erreur induite : ondulation dans la bande passante de .23°.

3.3.3.3 Implémentation Deux implémentations de la méthode overlap-add ont été réalisées. L'une avec un filtre $H(z)$ a structure SOS et l'autre avec un filtre $H(z)$ basé sur une structure allpass. Le tableau 3.3 présente les résultats aprés synthèse du VHDL.

TAB. 3.3 – Ressources en ALMs

	ALMs	Frequence max.	L
LIFO	34 (+8 M512)	250 MHz	200
IIR allpass	107	147 MHz	
IIR SOS	371	70 MHz	
Structure LP-IIR SOS	1400	72 MHz	
Structure LP-IIR Allpass	415	147 MHz	

En Figure 3.23 sont présentées les vues RTL des sous-entités *LIFO* et *délai* de longueur $2L$.
Pour réaliser la LIFO, une mémoire RAM a été utilisée avec un multiplexage en entrée permettant de lire et d'écrire dans les deux sens. Le délai, quant à lui, a été implémenté dans une *altsynchram*, composant de base d'Altera.
Concernant, les deux structures implémentées, celle qui se révèle la plus intéressante est celle



(a) Entrée du second étage

(b) Sortie de $H(z^{-1})$

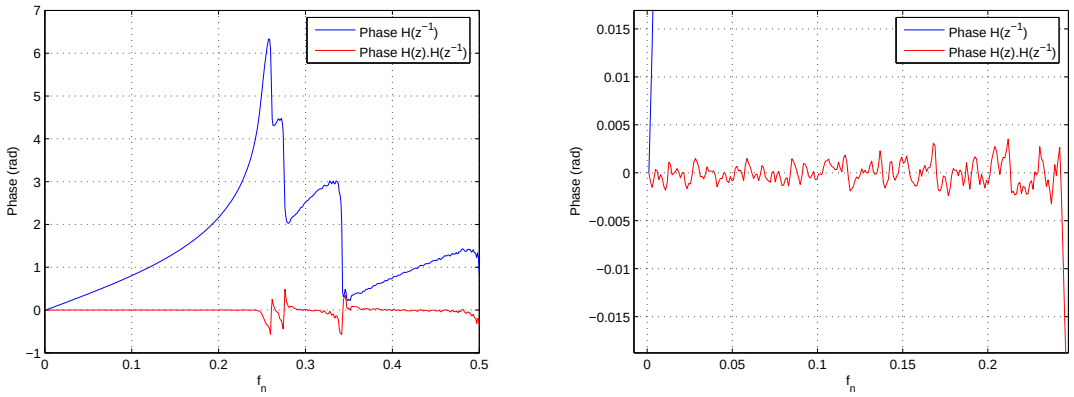
(c) Sortie de $H(z^{-1})$ ré-ordonnée

(d) Sortie de $H(z)H(z^{-1})$

FIG. 3.21 – *Entrée et sortie des différentes étapes lors de la linéarisation de la phase*

utilisant les filtres allpass puisqu'elle ne comporte que 3 coefficients contre 3 sections SOS dans le cas de l'autre structure.

Remarque : Les filtres RII obtenus par les méthodes de synthèse décrites précedemment ont cependant un inconvénient pour notre application. Le filtre du 2^{nd} étage doit en effet permettre de compenser la chute dans la bande passante induite par les étages de filtrage précédants. Les méthodes de synthèse de filtre utilisées ici, filtre à structure SOS et à structure allpass, ne sont pas assez « flexibles » (et tout particuliairement la structure two path allpass) pour délivrer une telle caractéristique. Un programme de minimisation, semblable à celui utilisé pour les filtre RIF doit être étudié.
Comme présenté en section 3.3.3.2, nous avons contourné ce problème en utilisant en amont du filtre RII, composé d'un réseau de filtre allpass, la cascade CIC-2HB qui n'introduit aucune chute dans la bande passante.



(a) Phase induite par les passages

(b) Zoom

FIG. 3.22 – *Phases introduites lors du premier passage (filtre $H(z^{-1})$) et du second $H(z)$*

4 Linéarisation de la phase - Algorithmes basés sur la réduction de modèles

Après avoir étudié et comparé différentes structures électroniques permettant la linéarisation de la phase d'un filtre RII, des algorithmes de synthèse aboutissant à cette même linéarisation sont exposés dans cette section.
Ces algorithmes, basés sur la méthode de réduction de modèles, permettent à partir d'un filtre RIF à phase linéaire d'ordre n l'obtention d'un filtre RII d'ordre m ($m << n$) à phase linéaire. Pour cela, le filtre RIF doit être présenté à l'algorithme sous sa forme « espace d'état ».
On peut en effet toujours représenter un filtre digital comme une équation différentielle d'ordre 1 :

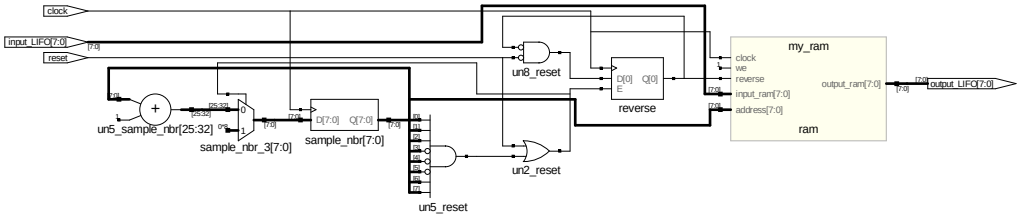
$$\begin{aligned} x(n+1) &= Ax(n) + bu(n) \\ y(n) &= cx(n) + du(n) \end{aligned} \tag{3.37}$$

où $u(n)$ est le vecteur d'entrée, $x(n)$ le vecteur d'état et $y(n)$ le vecteur de sortie. A est une matrice de taille $m * m$ où m est l'ordre du filtre, b est un vecteur colonne, c un vecteur ligne et d un scalaire. Ces élements sont les composants de la forme espace d'état (Figure 3.24).
En prenant la transformée en z des équations 3.37 et en les combinant, on montre l'équivalence de cette représentation avec la fonction de transfert du filtre :

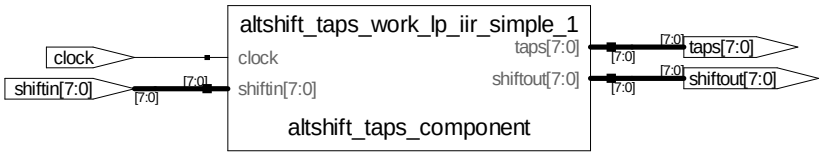
$$Y(z) = H(z)U(z) \text{ où } H(z) = c(zI - A)^{-1}b + d \tag{3.38}$$

avec I, la matrice identité.
La Figure 3.25 illustre le fait qu'un système peut toujours être séparé en deux parties distinctes.

Tout système est alors décomposable en une partie dominante et une partie « faible » ayant moins d'influence sur le résultat en sortie. Les composants A_r, b_r, c_r et d de la partie dominante constituent ce que l'on appelle le modèle réduit. Cette réduction est obtenue en séparant ces deux parties par l'intermédiare de différentes méthodes imposant certaines contraintes [38]. Elle



(a) LIFO



(b) Delai de longueur $2L$

FIG. 3.23 – *Vue RTL*

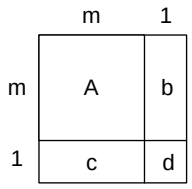


FIG. 3.24 – *Représentation state-space d'un filtre digital*

existe donc si et seulement si le modèle complet peut être organisé ainsi :

$$\begin{pmatrix} x_1(n+1) \\ x_2(n+1) \end{pmatrix} = \begin{pmatrix} A_R & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} + \begin{pmatrix} b_R \\ b_2 \end{pmatrix} u(n) \tag{3.39}$$

$$y(n) = \begin{pmatrix} c_R & c_2 \end{pmatrix} \begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} + d\ u(n)$$

avec A_R, b_R, c_R et d les composants du modèle réduit (Figure 3.25).
Dans notre cas, où le système étudié est un filtre numérique, cette réduction amène à un filtre possédant une réponse impulsionnelle très proche de celle du filtre original, permettant de conserver les caractéristiques spectrales de ce dernier (dont la caractéristique de phase linéaire par exemple).

Les algorithmes utilisés sont exposés en Annexe Présentation des algorithmes de réduction pour des raisons éidentes de clarté du document (algorithmes faisant appel à des notions rencontrées en théorie des sytèmes, notions étrangères au « traiteur » de signal habituel). Dans cette section ne sont présentés que les résultats obtenus par l'application de ces derniers. L'application de ces méthodes à un filtre demi-bande répondant aux spécifications du 2^{nd} étage TFB est dans un premier temps présentée. Une implémentation de la solution optimale est ensuite réalisée.

4.1 Application des méthodes à un filtre demi-bande (2^{nd} étage TFB)

Le filtre à réduire est le filtre RIF utilisé pour le 2^{nd} étage de filtrage du TFB. En Figure 3.26 sont tracés les réponses en phase et en amplitude de ce dernier.

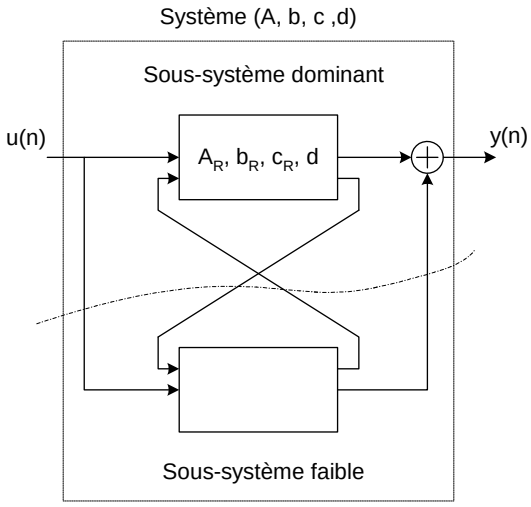


FIG. 3.25 – *Réduction du modèle*

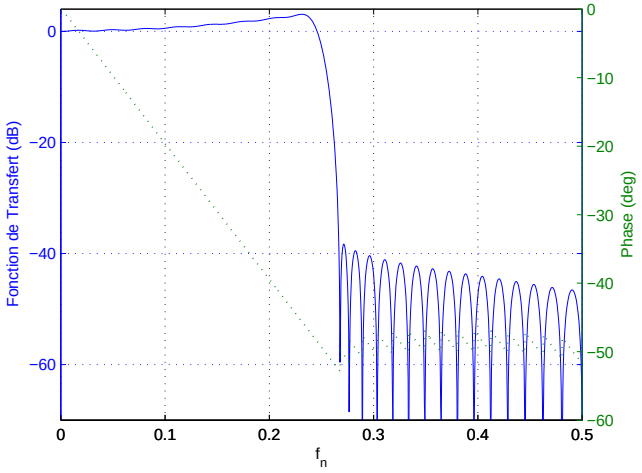
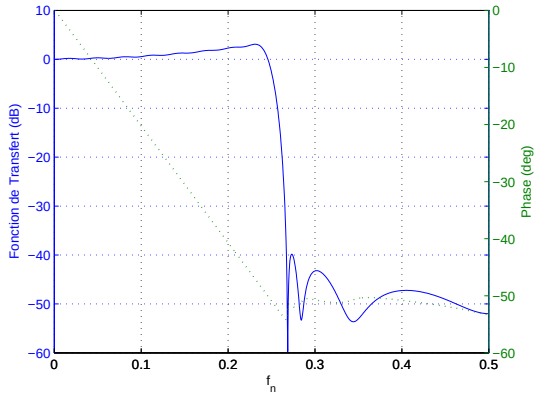


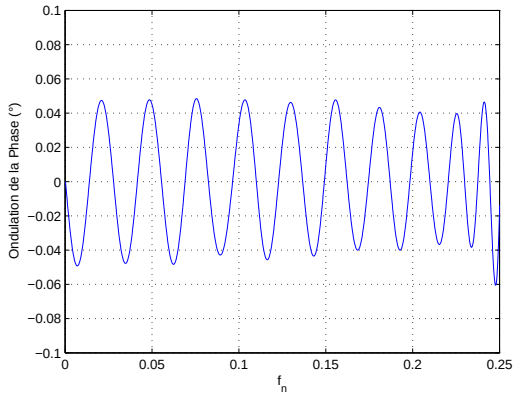
FIG. 3.26 – *Réponses du filtre demi-bande à réduire*

Les 2 premières méthodes utilisées (norme de Hankel et gramian de la réponse impulsionnelle) convergent toutes vers des spécifications en phase et en amplitude satisfaisantes (pas de dégradation de la réponse du filtre) pour un ordre $m = 23$. La 3^{ieme}, la réduction du modèle par décomposition en éléments singulier, affublée de ce même ordre m n'y succède pas. Le fait de choisir un ordre fixe pour les méthodes utilisées permet ainsi de comparer l'efficacité de ces dernières. En Figure 3.27, 3.28, 3.29 sont tracées les réponses obtenues après application de ces méthodes.

On peut remarquer que les 2 premières méthodes assurent une ondulation de la phase bien inférieure à la spécification du projet.La méthode nommée « Frequency weighted model reduction » n'a pas été modélisée car les résultats sont semblables à ceux déjà présentés.
La solution Weighted least square approximation devrait donner des résultats plus intéressants (filtre d'ordre 12 pour un filtre RIF de départ d'ordre 64). Cependant, certains problemes ont été

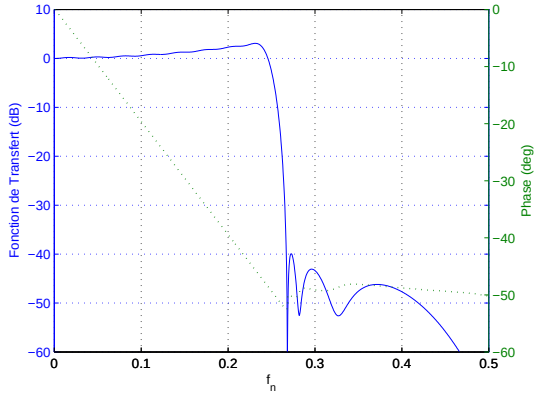


(a) Fonction de transfert et réponse en phase

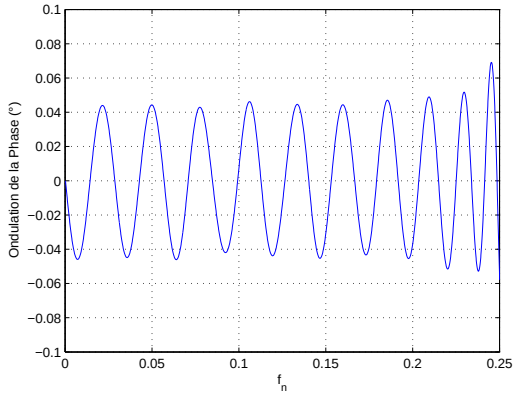


(b) Ondulation de la phase

FIG. 3.27 – *Méthode Hankel Norm, $m = 23$*



(a) Fonction de transfert et réponse en phase



(b) Ondulation de la phase

FIG. 3.28 – *Méthode Impulse Grammian, $m = 23$*

rencontrés lors du développement de la routine d'optimisation BFGS et la modélisation n'a pas été achevée. L'ordre 12 prévu est celui donné en [39] pour la réduction d'un filtre RIF d'ordre 64 répondant à un gabarit similaire. Il a tout de même été décidé d'implémenter ce dernier, puisqu'étant la solution optimale résultant des méthodes de réduction de modèles.

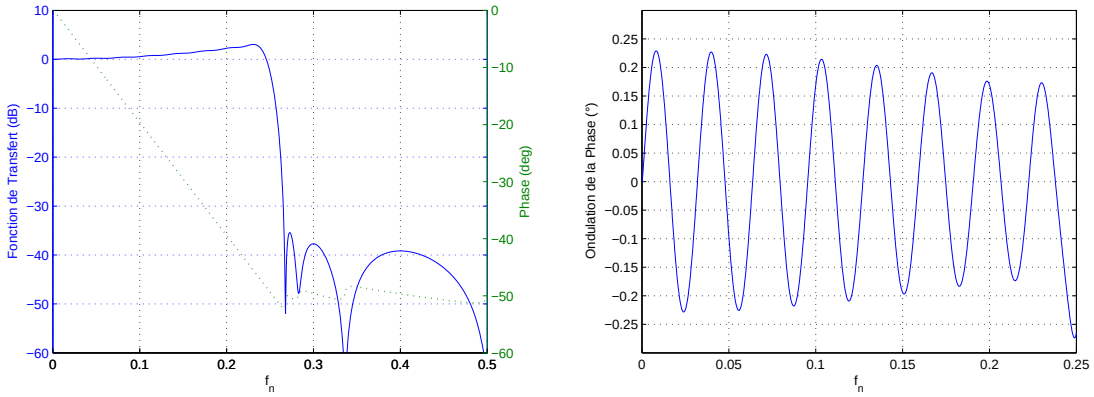
4.2 Implémentation

Dans le tableau 3.4 sont donnés les résultats de l'implémentation des filtres IIR à phase linéaire d'ordre 12.

TAB. 3.4 – Ressources en ALMs

	ALMs	Frequence max.	DSP blocks (9 bits)
IIR SOS	594	148 MHz	-
IIR SOS	27	216 MHz	5(36)

La structure utilisée est une structure de type SOS illustrée en Figure 3.30. 6 cellules SOS ont été



(a) Fonction de transfert et réponse en phase

(b) Ondulation de la phase

FIG. 3.29 – *Méthode Minimum Sensitivity,* $m = 23$

nécessaires à l'implémentation de ce filtre.

5 Comparaisons des différentes implémentations possibles

Dans un premier temps, la solution du 2^{nd} étage TFB actuel (c'est à dire un filtre RIF d'ordre 64 à coefficients symétriques) est comparée en terme de ressources électroniques (ALMs associés à des blocs DSP ou non) à un filtre RII répondant au même gabarit en fréquence, mis à part la phase qui est non linéaire dans la bande passante. Le but d'une telle comparaison est de mettre en évidence la faible complexité du filtre RII ainsi que l'architecture optimale permettant son implémentation.
En Figure 3.30 est présentée la structure de la cellule SOS qui a été utilisée pour l'implémentation du filtre RII (dans le cas de l'utilisation des blocs DSP, la structure utilisée est celle de la Figure 5(c)).

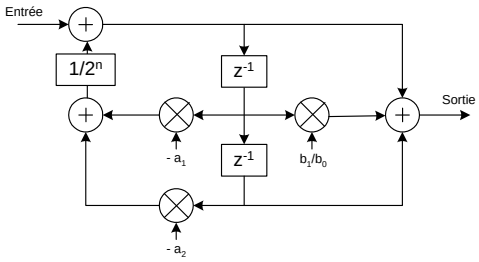


FIG. 3.30 – *Structure électronique d'une cellule* SOS

La variable n correspond au nombre de bits utilisés pour la quantification des coefficients.
Une cascade de 4 structures SOS est nécessaire à l'obtention d'un filtre RII répondant aux mêmes spécifications que le filtre « 2^{nd} étage TFB ».
L'utilisation d'une structure Lattice/Ladder a aussi été étudiée pour l'implémentation du filtre RII. Elle résulte en une structure composée de 8 cellules de base *lattice/ladder* (basée sur le même principe que celui de la Figure 2) amenant à l'utilisation de 8 coefficients lattice et 9 coefficients ladder.

Enfin, une structure *Allpass two path* a été utilisée dans le même but. Un filtre à 4 coefficients (et donc 4 structures Allpass, cf. section 2.2.3) permet d'atteindre une spécification d'atténuation de 45 dB. Une vue RTL de la structure est disponible en Figure 3.31.

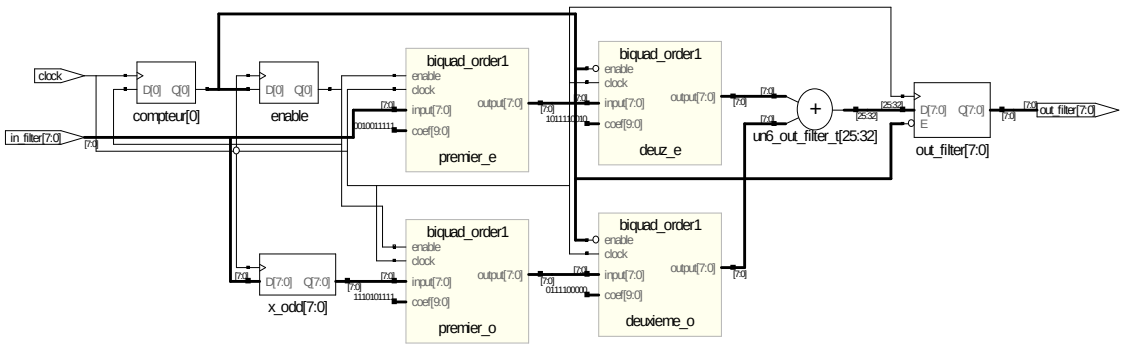


FIG. 3.31 – *Vue RTL du filtre two-path Allpass*

Les structures Allpass d'ordre 1 sont utilisées après l'opération de décimation par 2 et donc peuvent fonctionnner à la fréquence du système divisée par 2 (i.e. 62.5 MHz).
Les résultats de ces implémentations sont présentés dans le tableau 3.5, dans la partie *Phase non linéaire*.

TAB. 3.5 – Ressources en ALMs

		ALMs	Freq. max.	Blocs DSP (9 bits)
Filtre RIF	2^{nd} étage actuel	1168	141 MHz	-
	2^{nd} étage actuel	185	260 MHz	2(16)
Phase non linéaire	IIR SOS	358	148 MHz	-
	IIR SOS	18	148 MHz	3(24)
	Lattice/Ladder	733	50 MHz	-
	Structure Allpass	239	98 MHz	-
Phase linéaire	IIR+ égaliseur	1400	97 MHz	-
	Structure Allpass modifiée	984	74 MHz	-
	Structure 2 passages	415	142 MHz	-
	Réduction du modèle	594	147 MHz	-
	Réduction du modèle	27	147 MHz	5(36)

Il est a noter que la fréquence maximale de fonctionnement de la structure lattice/ladder n'atteint pas la spécification. Aucun effort n'a été fait pour améliorer la situation puisque le nombre de ressources utilisées est déjà supérieur à celui d'une structure SOS classique.

La seconde étape consiste à comparer les différentes solutions présentées dans les paragraphes précédents permettant l'obtention d'une phase linéaire en utilisant un filtre RII. Le résultat des différentes implémentations est présenté en tableau 3.5 dans la partie *Phase linéaire*. Toutes respectent les spécification en amplitude et en phase du projet.
Concernant la fréquence maximale de fonctionnement de la structure composée de la cascade filtre RII et filtre égaliseur, du fait du grand nombre de ressources requises aucune optimisation n'a été effectuée. Dans le cas du filtre *two-path Allpass* modifié la fréquence de fonctionnement

souhaitée pour chaque branche est 62.5 MHz, une fréquence de 74 MHz est donc suffisante.
En reconsidérant toutes les solutions étudiées, la structure deux passages utilisant un filtre *two-path Allpass* est celle qui allie la caractéristique de phase linéaire avec le plus faible nombre de ressources nécessaires. Dans le cas du remplacement du filtre RIF « 2^{nd} étage » actuel par cette solution, les ressources DSPs non utilisées peuvent alors être réemployées dans le premier étage de filtrage permettant encore de diminuer la consommation de l'ensemble TFB (cf. tableau 3.6).

Tab. 3.6 – Ressources *DSP* du « premier étage »

	ALMs	Frequence max.	DSP blocks (9 bits)
2 Demi Bande	166	184 MHz	4(13)
Quart de Bande	147	261 MHz	2(8)

Cependant dans le cas où le second étage doit être capable de compenser une chute dans la bande d'intérêt induite par le premier étage (c'est le cas avec l'utilisation de la cascade CIC + quart de bande), le choix d'un filtre RII résultant d'une réduction de modèle est préférable (le point de départ étant un filtre RIF, plus « malléable »). Dans le cas contraire (lors de l'utilisation de la cascade CIC + 2HB pour premier étage par exemple), les structures basées sur les filtres *two-path Allpass* ou les cellules SOS peuvent alors être employées.

6 Filtrage numérique pour la Radioastronomie

L'interféromètre ALMA est l'un des premiers à être équipé d'un système de filtrage numérique. L'habitude des concepteurs des interféromètres radio en ce qui concerne la phase induite par le système est d'exiger sa linéarité dans la bande de mesure. Dans le cas de systèmes analogiques cela permet de réduire la complexité de la correction à appliquer post-corrélation et d'étalonner la phase instrumentale plus aisément en fonction de la fréquence. Cet étalonnage est nécessaire car les systèmes interférométriques analogiques n'offrent pas des performances de stabilité en phase parfaitement reproductibles. C'est pour cette raison que les filtres RIF à réponse symétrique (caratérisés par une phase linaire) sont préférés, a priori, pour la réalisation de tels instruments. Ainsi, une étude concernant les filtres RII (nécessitant un nombre de poids parfois 4 fois inférieur à celui d'un filtre RIF fournissant les mêmes caractéristiques) a été menée. Plusieurs méthodes permettant une implémentation à moindre coût avec l'obtention d'une phase linéaire ont été exposées. Dans le cas d'une phase que trop approximativement linéaire, on préfèrera l'emploi d'une structure RII classique pour laquel il faudra de toute façon calculer (compenser) la non-linéarité de phase
Cependant nous avons aussi étudié la possibilité de l'utilisation de filtre RII à phase non-linéaire pour la réalisation des systèmes de filtrage numérique. La phase étant reproductible indéfiniment d'une réalisation à l'autre (sous réserve d'employer la même structure avec les mêmes poids), elle peut être corrigée post-corrélation puisque parfaitement connue (à l'aide d'une table de correction par exemple). L'utilisation de tels filtres doit alors être prise en compte. De plus, le paramètre d'importance en interférométrie est la différence de phase entre 2 bras d'une paire interférométrique, dans ce cas les phase induites par le système de filtrage (qui sont identiques pour toutes réalisations) sont alors éliminées. Il peut tout de même apparaître un inconvénient à utiliser ce type de filtre dans certains cas : lors d'un calcul de puissance, par voie interférométrique par exemple, à travers la bande totale résultant de l'opération de filtrage, cette phase non linéaire

peut altérer le calcul de puissance.

7 Conclusion

Un état de l'art des structures de filtres RII et surtout de filtres RII à phase linéaire, ou approximativement linéaire, a été présenté. Une architecture pour chaque type de filtre (phase non-linéaire et approximativement linéaire – ondulation de 0.3°) s'est révélée très intéressante en termes d'utilisation de ressources logiques : le filtre composé d'un réseau de filtres allpass en cascade. L'utilisation d'une telle structure a tout de même un inconvénient : les filtres en résultant ne possède pas une grande flexibilité concernant leurs caratéristiques en fréquences et spécialement dans la bande passante (impossibilité de compenser une bande passante lors de l'emploi d'une struture sans multiplieur).
Il reste encore à éclaircir la potentielle utilisation d'un filtre RII à phase non linéaire dans un système destiné à la radioastronomie. Le problème lié au mélange des phases analoiques et numériques induites par les différentes parties (front-end et back-end) de l'instrument mérite approfondissement.
Il est à noter que si on se place dans le cas d'une architecture de filtrage à phase (approximativement) linéaire respectant l'ondulation maximale de la phase établie par le projet, la structure deux passages est la plus intéressante. Elle permet une diminution du nombre de ressources comparée à la structure existante, dans le cas où les blocs DSP ne sont pas utilisés.

Chapitre 4

Réalisations d'un Filtre Polyphase

1 Introduction

Après avoir étudié une solution alternative à chaque « sous-étage » du filtre TFB actuel, il est question dans ce chapitre de revoir completement l'architecture de filtrage. L'implémentation d'une structure de filtre polyphase a été envisagée dans cet esprit. Le filtre synthétisé doit permettre de diviser la bande de base en 32 sous bandes égales afin d'obtenir une résolution spectrale de base équivalente à celle obtenue avec le filtre TFB. Les caractéristiques spectrales de la SB restent identiques : 47 dB d'atténuation, 0.2 dB d'ondulation et une région de transition de 1/32.
Le principal incovénient d'une telle structure par rapport à celle du filtre TFB concerne la flexibilité du positionnnement des SBs. En effet, la position des SBs dans la BB d'une architecture polyphasée classique est fixée et définie par le module TFD (Transformée de fourier Discrète) de cette dernière. Une étude concernant cette dite « immobilité » et ses inconvénients (bande de fréquence non analysée entre les SBs) est menée dans une seconde partie visant à évaluer la complexité d'une structure polyphase plus souple.

2 Réseau de filtres polyphases

Un réseau de filtres polyphases [23] est composé de deux principaux blocs : la structure polyphasée du filtre et le module TFD permettant de placer chaque SBs synthétisée à un endroit précis de la BB. Le module assurant la translation en fréquence peut être réalisé de différentes façons qui sont exposées dans les sections suivantes. Dans un premier temps est exposée la structure polyphase permettant la réalisation du filtre lui-même qui permet d'extraire un motif spectral de la bande initiale d'entrée.

2.1 La structure polyphase

Soit $H(z) = \sum_{n=0}^{N-1} h(n)z^{-n}$, la fonction de transfert d'un filtre RIF d'ordre N. L'équation de la fonction de transfert de la structure polyphase [9, 20] résultante de cette derniére est donnée

par :

$$H(z) = \sum_{n=0}^{N-1} h(n)z^{-n} = \sum_{m=0}^{M-1} z^{-m} H_m(z^M) \tag{4.1}$$

$$\text{où } H_m(z^M) = \sum_{n=0}^{\lfloor \frac{N-1}{M} \rfloor} h(m+nM)z^{-nM}$$

L'architecture électronique retranscrivant cette équation est présentée en Figure 4.1.

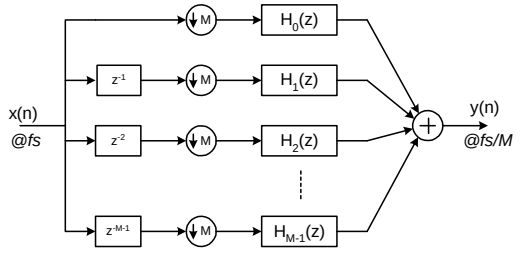


FIG. 4.1 – *Architecture électronique polyphase*

L'opération de décimation, qui permet d'extraire un motif spectral de largeur $\frac{f_{ech}}{M}$ (dans le cas d'un signal d'entrée complexe de largeur f_{ech}), a été déplacé en amont du filtre afin de simplifier la structure de ce dernier (cf. identités remarquanbles des systèmes multi-cadences en Figure 2.4 *a*) et*b*)). Cette structure permet l'obtention d'un filtre passe-bas permettant d'extraire une bande de largeur $\frac{f_{ech}}{M}$ de la BB complexe à partir d'une combinaison de filtres allpass se différenciant uniquement par leur phase respective. Le principe d'une structure polyphase est exposé en détail en [9].

2.2 Translation en fréquence

Maintenant que le principe du filtre polyphase est exposé, il faut s'intéresser à l'extraction des différentes SBs de la BB. En Figure 4.2 est présentée l'architecture permettant d'extraire une SB de largeur $\frac{f_{ech}}{M}$ à la position k ($k \in [0, M-1]$) dans la BB (SB centrée en $f = k\frac{f_{ech}}{M}$).

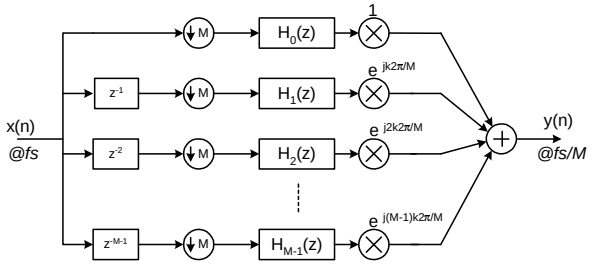


FIG. 4.2 – *Architecture électronique : filtre polyphase et translation en fréquence*

Cette structure répond à l'équation :

$$H_{PB} = \sum_{m=0}^{M-1} z^{-m} H_m(z^M) \cdot e^{j2\pi \frac{mk}{M}} \tag{4.2}$$

où H_m est défini en (4.1).
Afin d'extraire les M SBs composant la BB, l'architecture doit être adaptée. La structure résultant de cette adaptation est présentée en Figure 4.3.

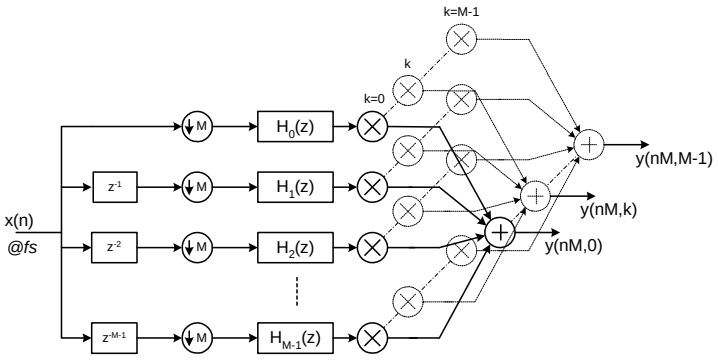
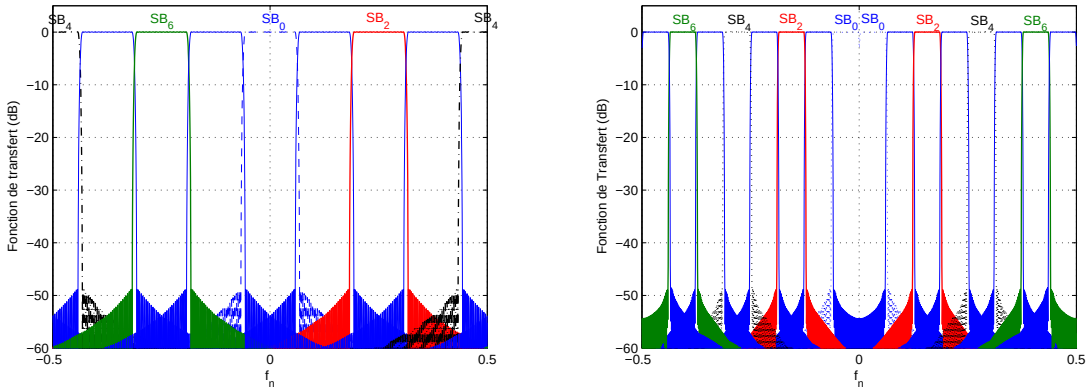


FIG. 4.3 – *Architecture électronique polyphasée à M SBs*

Contrairement au cas général présenté ici, le système de filtrage du projet ALMA traite des sinaux réels. Après application de cette structure à notre BB réelle, les signaux caractérisant chaque SB obtenus sont complexes et doivent, pour réaliser une découpe de la BB réelle de largeur $\frac{f_{ech}}{2}$ en M SBs égales, être convertis en signaux réels ou traités par un système « réel ». Les spectres des SBs en sortie des systèmes « complexe » et « réel » sont tracés en Figures 4.4.



(a) Sortie complexe

(b) Sortie réelle

FIG. 4.4 – *Spectre en sortie du réseau de filtres polyphases*

Les numéros des différentes SBs sont indiqués sur chacun des graphiques. En Figure 4.4(a), l'ensemble filtre + conversion fréquentielle utilisé possède une répartition paire des SBs (SB centrée en $\frac{k2\pi}{M}$, $k \in [0, M-1]$) tandis que l'ensemble exposé en Figure 4.4(b) possède une répartition impaire (SB centrée en $\frac{(k+0.5)2\pi}{M}, k \in [0, M-1]$).La structure exposée en Figure 4.3, dont le spectre en sortie est tracé en Figure 4.4(a) permet de séparer en M SBs (8 dans le cas présent) la bande de fréquences normalisées $[-0.5, 0.5]$. Cette structure est donc idéale dans le cas du traitement d'un signal complexe. Dans le cas d'un signal réel à traiter, pour diviser la bande de fréquences normalisées $[0, 0.5]$ en M SBs, une conversion complexe-réel du signal doit être effectuée en sortie de la structure du réseau polyphase (cas ALMA où la BB de 2 GHz, de

nature réelle, doit être subdivisée en 32 SBs de 62.5 MHz chacune).

3 Réalisation d'un réseau de filtres polyphases

Comme cela vient d'être présenté, un réseau de filtres polyphases est composé de 2 principaux éléments, le filtre lui-même décomposé en M sous-parties et un composant permettant le déplacement en fréquence des SBs considérées : un bloc TFD.
La partie d'intérêt ici concerne l'implémentation de ce bloc TFD, celle du filtre étant similaire aux implémentations utilisées précédemment. La formule permettant de réaliser la TFD sur N points d'un signal $x(k)$ est la suivante :

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{nk} \tag{4.3}$$

avec $W_N = e^{\frac{-2j\pi}{N}}$.
Une méthode appelée *FFT* (Fast Fourier Transform) [40] permet de réduire le nombre d'opérations requises pour le calcul d'une telle transformée. Le signal résultant d'une telle opération est complexe. Dans cette section est aussi évoqué l'utilisation d'une *DCT* (Discrete Cosine Transform) qui permet d'obtenir en sortie un signal réel.

3.1 Transformée de Fourier rapide

La TFD peut être représentée sous forme matricielle :

$$F_N = \begin{pmatrix} W_N^0 & W_N^0 & W_N^0 & \dots & W_N^0 \\ W_N^0 & W_N^1 & W_N^2 & \dots & W_N^{N-1} \\ \vdots & \vdots & \vdots & & \vdots \\ W_N^0 & W_N^{N-1} & W_N^{2(N-1)} & \dots & W_N^{(N-1)^2} \end{pmatrix} \tag{4.4}$$

En [41] est proposé un algorithme rapide permettant de calculer la DFT d'un signal. Le calcul direct d'une telle fonction, pour un signal réel, nécessite $2N^2$ multiplications et $2N(N-1)$ additions, N étant le nombre de points sur lequel la DFT est calculée. Dans le cas de la FFT, la complexité de calcul est réduite à $N/2\log_2(N)$ multiplications et $N\log_2(N)$ additions. L'idée de base de la FFT est de décomposer la TFD d'ordre N (avec $N = 2^i$) en m TFD d'ordre N_i avec :

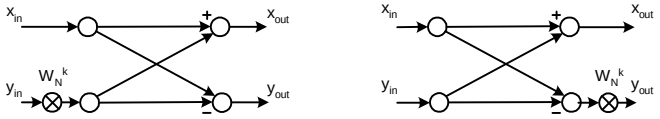
$$N = \prod_{i=1}^{m} N_i \tag{4.5}$$

N_i est appelé radix-i. La cellule de base de l'architecture d'une FFT-radix2, appelée « cellule papillon », est régie par les équations suivantes dans le cas d'une décimation en temps (DIT ou « partagée dans le temps ») :

$$\begin{aligned} x_{out} &= x_{in} + W_n^k y_{in} \\ y_{out} &= x_{in} - W_n^k y_{in} \end{aligned} \tag{4.6}$$

Une structure correspondant à une décimation en fréquence (DIF ou « partagée en fréquence ») peut aussi être utilisée, l'emploi de l'une ou de l'autre (l'une étant la fonction transposée de

l'autre) dépendant seulement de l'ordonnement des données souhaité en sortie ou en entrée. L'algorithme de la FFT DIT ré-arrange l'équation de la TFD (4.3) en deux parties : une somme sur les échantillons d'entrée $x(k)$ d'indice pair et une somme sur ceux d'indice impair. Dans le cas de la DIF, le partage est effectuée sur le calcul des échantillons de sortie $X(k)$. L'explication développée ici concerne la FFT-DIT. Ces structures de base sont représentées en Figure 4.5.



(a) DIT (b) DIF

FIG. 4.5 – *Structure radix 2 d'une FFT DIT et DIF*

En appliquant les identités remarquables caractérisant l'élément W ($W_n^{nN} = 1$, $W_N^{N/2} = -1$, $W_N^n = W_N^{n+N}$) à (4.4) ainsi qu'en décomposant (4.3) en deux parties, constituées des termes de rang pair et impair [40] ordonnées en utilisant l'inversion binaire, il est alors possible d'établir la structure FFT 8 points de la Figure 4.6.

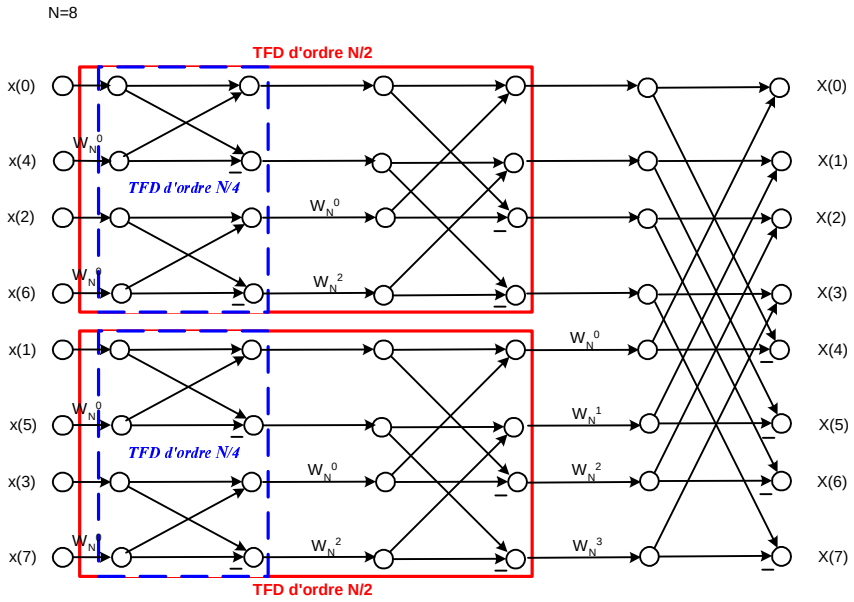


FIG. 4.6 – *Architecture d'une FFT 8 points*

La formule de la FFT est la suivante, elle est appliquable, récursivement, aux différents sous-blocs constituant la TFD principale sur N points ($N/2$, $N/4$ etc. . . jusqu'à obtention de la simple cellule de base sur 2 points) :

$$X(k) = \sum_{n=0}^{N/2-1} x(2n)W_{N/2}^{nk} + W_N^k \sum_{n=0}^{N/2-1} x(2n+1)W_{N/2}^{nk} \text{pour } k \in [0, N/2-1] \quad (4.7)$$

$$X(k) = \sum_{n=0}^{N/2-1} x(2n)W_{N/2}^{nk} - W_N^k \sum_{n=0}^{N/2-1} x(2n+1)W_{N/2}^{nk} \text{pour } k \in [N/2, N-1] \quad (4.8)$$

L'inversion binaire (opération réalisée sur $\log_2(N)$ bits) assure la permutation de l'ordre des donnée correspondant à l'entrée de la FFT N points : Pour $N = 8$, l'échantillon n°4 (100 en binaire) devient l'entrée de la voie n°1 (001 en binaire).
La structure se décompose en 3 étapes : le calcul de la TFD sur $N/4$ points, le calcul de la TFD sur $N/2$ points et enfin la TFD sur N points ($N = 8$), la cellule de base étant la cellule papillon de la Figure 4.5.
La FFT-radix2 « partagée en fréquence »possède une structure semblable à celle de la Figure 4.6 mais renversée.

Remarques : Seule la structure radix-2 a été étudiée dans ce chapitre, le but étant d'évaluer la complexité de différentes structures « filtre polyphase » basiques comparables entre elles. Les architectures radix-4 et split-radix n'ont pas été étudiées et ne sont pas abordées ici.

La structure présentée ci-dessus délivre un signal complexe qui dans certains cas (cas ALMA) doit être converti en un signal réel. Deux solutions s'offrent à nous pour palier ce problème : la cascade sur chacune des voies de sortie d'un étage « conversion complexe-réel » dont le composant principal est un filtre de Hilbert ou l'utilisation d'une DCT (Discrete Cosine Transform) en lieu et place d'une DFT. Ces points sont abordés plus loin dans ce chapitre.

3.2 La RFFT

Une remarque importante à propos de notre application concerne la nature du signal à traiter : c'est un signal réel. En [42] sont comparées différentes méthodes permettant de traiter, avec moins d'opérations logiques, une FFT sur N points appliquée à un signal réel. La méthode retenue, qui permet de reduire le nombre de multiplication et d'addition d'une façon non négligeable ($M/2$ multiplications et $M/2-2$ additions, M étant le nombre d'opérations nécessaires à la réalisation d'une FFT radix-2) est appelée RFFT (Real-valued FFT). En effet, si la séquence d'entrée d'une FFT est réelle et paire alors la partie réelle de la séquence obtenue par FFT est réelle symétrique et la partie imaginaire est impaire symétrique. Ce qui se traduit par les propriétés suivantes en sortie de la FFT :

$$X(0), X(\frac{N}{2}) = \text{réel} \tag{4.9}$$

$$X(k) = X^*(N-k) \forall k \in [1, N/2[\tag{4.10}$$

En figure 4.7 est représenté le flot de donné à travers les différents blocs de la RFFT.
La simplification de l'architecture apparait clairement sur cette figure, une sortie complexe sur deux peut être obtenue par sa paire complexe conjuguée. Il suffit alors de stocker la valeur complexe obtenue en sortie d'une la cellule papillon, d'en déduire son conjugué pour pouvoir l'appliquer à la cellule suivante. Ceci permet de sauver, dans le cas du 2^{eme} étage de la FFT la moitié des ressources d'un papillon ou un papillon entier sur deux dans le cas du 3^{eme} étage. En ce qui concerne le premier étage, qui traite des données réelles et dont le coefficient multiplicateur des cellules papillons est égal à 1, les signaux obtenus en sortie sont réels. Ceci permet de réduire le nombre de multiplication et d'addition de ce dernier de moitié. Le principe se répète pour des FFT effectuée sur un plus grand nombre de points.

En plus de l'implémentation des deux structures FFT présentées, il a aussi été décidé de réaliser celle d'une FFT utilisant l'algorithme proposé par Winograd [43]. Cet algorithme permet

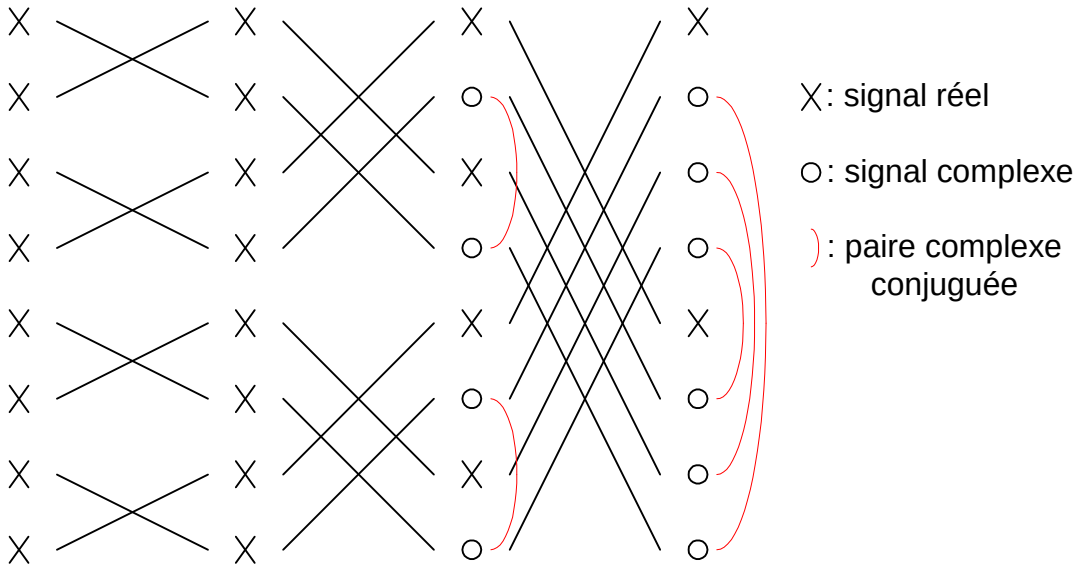


FIG. 4.7 – *Architecture d'une RFFT 8 points*

de réduire le nombre d'opérations nécessaires mais aboutit à une architecture plus complexe, ne possédant pas les caractéristiques de redondance et d'entrelacement de l'algorithme de Cooley-Tukey. L'implémentation d'une DFT sur 32 points est présentée en [44].

3.3 Transformée en cosinus discrète

La transformée en cosinus discrète est une transformation similaire à la TFD, basée sur l'analyse de Fourier, mais utilisant seulement des grandeurs réelles. La DCT est équivalente à une TFD de longueur double opérant sur des données réelles à symétrie paire (la transformé de Fourier d'une fonction réelle et paire est réelle et paire). Il existe 4 variantes communément utilisées de la DCT :

$$\text{DCT-I} : C_k^N = \frac{1}{2}(x_0 + (-1)^k x_{N-1}) + \sum_{n=1}^{N-2} x_n \cos\left[\frac{\pi}{N-1} nk\right], \; k \in [0, N-1] \quad (4.11)$$

$$\text{DCT-II} : C_k^N = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right) k\right], \; k \in [0, N-1] \quad (4.12)$$

$$\text{DCT-III} : C_k^N = \frac{1}{2} x_0 + \sum_{n=1}^{N-1} x_n \cos\left[\frac{\pi}{N} n \left(k + \frac{1}{2}\right)\right], \; k \in [0, N-1] \quad (4.13)$$

$$\text{DCT-IV} : C_k^N = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)\left(k + \frac{1}{2}\right)\right], \; k \in [0, N-1] \quad (4.14)$$

N étant le nombre de point sur lequel la DCT est calculée.
La DCT-II (4.12) est la forme la plus utilisée et est souvent nommée simplement DCT. Certains auteurs multiplient le terme X_0 par $\frac{1}{\sqrt{2}}$. La DCT-III, qui est l'inverse de la DCT-II, est simplement appelée DCT inverse ou IDCT.
La DCT est abordée dans différents papiers où plusieurs réalisations possibles sont proposées. En [45, 46] sont présentées des méthodes permettant l'obtention d'une DCT à partir du calcul

de la FFT présentée en section 3.1. Les premiers algorithmes de calcul rapides de la DCT-II et DCT-IV sont exposés en [47, 48], ils sont abordés dans la suite du paragraphe.

3.3.1 Construction d'une DCT à partir d'une FFT

Les premières études concernant la DCT étaient basées sur l'utilisation d'une FFT de longueur $2N$ aboutissant au calcul d'une DCT de longueur N. En [45] est donnée une méthode n'utilisant qu'une FFT de longueur N. Cette méthode est basée sur un ré-arrangement de la séquence d'entrée x_n $n \in [0, N-1]$ en groupant tout les termes pairs rangés dans l'ordre croissant suivis par les termes impairs rangés cette fois-ci dans l'ordre décroissant. On obtient la nouvelle séquence y_n :

$$\begin{cases} y_n = x_{2n} & n = 0, \ldots, \frac{N}{2} - 1 \\ y_{N-1-n} = x_{2n+1} & n = 0, \ldots, \frac{N}{2} - 1 \end{cases} \tag{4.15}$$

La DCT sur N points de la séquence x_n est alors obtenue en appliquant à y_n (4.12), en multipliant le premier terme par $\frac{1}{\sqrt{2}}$. Aprés quelques manipulations [46], on obtient l'équation de la DCT appliquée à y_n :

$$C_k^N = 2c(k)\Re\left\{ e^{-j\frac{\pi k}{2N}} F_k^N(y) \right\} \tag{4.16}$$

où $F_k^N(y)$ est la DFT sur N points du signal y_n avec $k \in [0, N-1]$ et

$$\begin{cases} c(k) = \frac{1}{\sqrt{2}} & \text{si } k = 0 \\ c(k) = 1 & \text{sinon} \end{cases}$$

En [46], une technique permettant de calculer 2 DCT sur N points à partir d'une FFT également sur N points est exposée. Ce type d'optimisation n'est cependant pas utilisé dans le cadre du travail présenté ici.

3.3.2 Algorithme rapide de calcul

Dans cette section vont être abordés deux algorithmes de calcul de la DCT aboutissant à des structures simplifiées avec un nombre d'opérations limité.
Le premier établi par Lee [47] permet de réduire le nombre de multiplication par deux en comparaison à une implémentation classique [45] et aboutit à une structure simplifiée (nommé *FCT* pour Fast Cosine Transform). La méthode est basée sur une découpe en deux parties égales d'une DCT N points. La structure est illustrée en Figure 4.8.
Après séparation des termes de rang pair et impair, une DCT de $N/2$ points est appliquée à chaque partie. Les équations retranscrivant l'opération à effectuer sont les suivantes :

$$g(k) = \sum_{n=0}^{N/2-1} \hat{x}(2n) C_{2(N/2)}^{(2k+1)n} \tag{4.17}$$

$$h(k) = \sum_{n=0}^{N/2-1} \left(\hat{x}(2n+1) + \hat{x}(2n-1)\right) C_{2(N/2)}^{(2k+1)n} \tag{4.18}$$

avec $\hat{x}(n) = e(n)x(n)$ où

$$\begin{cases} e(n) = \frac{1}{\sqrt{2}} & \text{si } n = 0 \\ c(k) = 1 & \text{sinon} \end{cases}$$

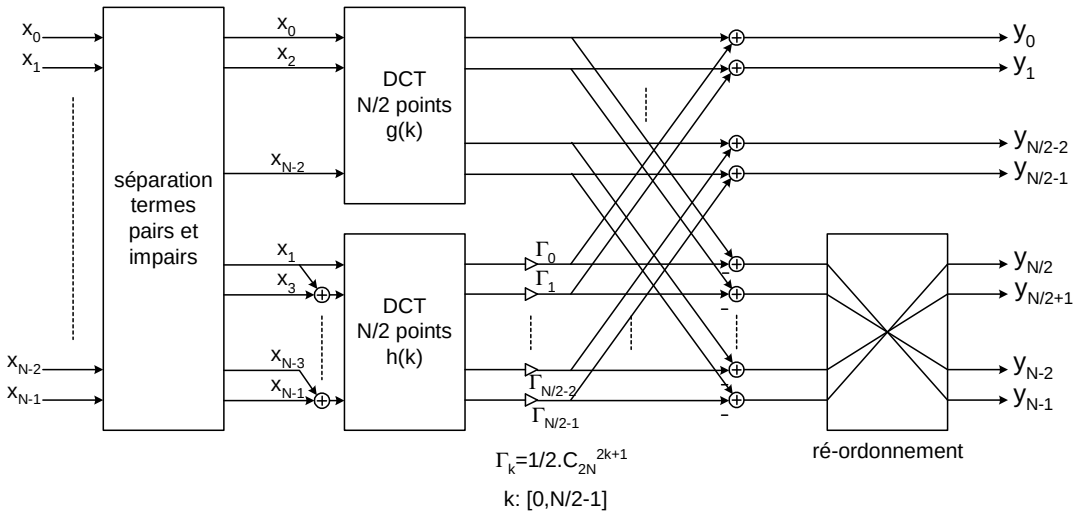


FIG. 4.8 – *DCT Lee, N points*

.
La sortie du bloc DCT est donnée par les relations :

$$y(k) = g(k) + \left(\frac{1}{C_{2N}^{2k+1}}\right) h(k) \tag{4.19}$$

$$y(N-1-k) = g(k) - \left(\frac{1}{C_{2N}^{2k+1}}\right) h(k) \text{ pour } k \in [0, N/2-1] \tag{4.20}$$

La DCT N points est donc décomposée en une somme de deux DCT sur $N/2$ points. Ce procédé peut être répété afin de décomposer encore la DCT jusqu'à obtention d'une cellule de base (cellule papillon). Le nombre de multiplication nécessaire à l'implémentation est donné par la formule suivante : $\mu(C_N^{II}) = (N/2)\log_2(N)$ et le nombre d'additions par : $\alpha(C_N^{II}) = (3N/2)\log_2(N) - N + 1$. Wang, en [49], a développé un algorithme nommé *SFCT* pour Simple Structured Fast Cosine Transform qui est moins sensible aux erreurs dues à une implémentation à virgule fixe que l'algorithme proposé par Lee. Cet algorithme fut le premier basé sur la factorisation de matrices de transformation. Plus récemment, Takala [50] proposa un algorithme basé sur le même décomposition matricielle mais aboutissant à une structure plus aisé à implémenter, reprenant le principe de Cooley-Tukey [41] concernant les propriétés du dit algorithme (redondance, entrelacement). L'équation de la DCT-II prend la forme suivante, aprés décomposition en matrices de transformation :

$$C_{2^k}^{II} = \frac{1}{\sqrt{2^k - 1}} D_c P_c \left[\prod_{s=1}^{k-1} (A'_{k-s}) P_{2^k,2}\right] A_0 P_{2^k,2}^T P_{2^k}^H \tag{4.21}$$

les coefficients de la DCT étant contenu dans la matrice diagonale par bloc A_{k-s} et les matrices P_i^H et $P_{i,2}$ étant des matrices de permutations nommées *Hadamard* (suivant l'ordre de Hadamard) et *perfect shuffle* respectivement [50]. L'équation n'est pas décrite en détail par souci de clarté.

Les coefficients sont calculés récursivement en appliquant les formules [49] :

$$\begin{aligned} d(1) &= \sqrt{0.5} \\ d(2i) &= \sqrt{0.5(1+d(i))} \\ d(2i+1) &= \sqrt{0.5(1-d(i))} \end{aligned} \tag{4.22}$$

Cet algorithme appliqué à une DCT-II 8 points résulte en la structure régulière de la Figure 4.9.

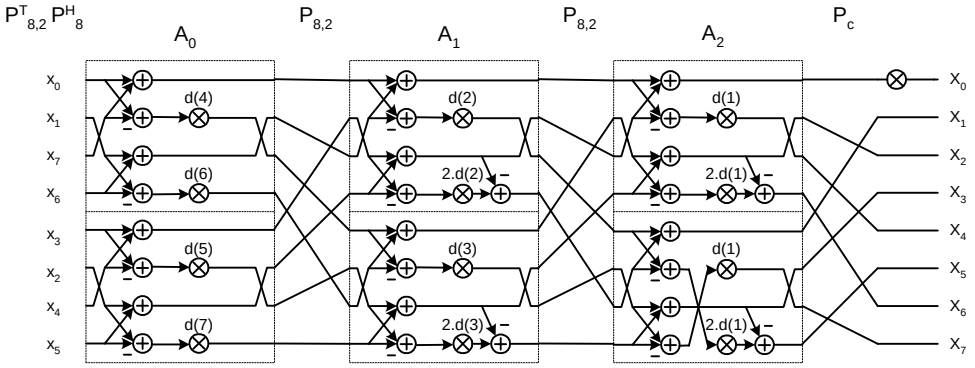


FIG. 4.9 – *DCT-II, 8 points*

Les deux algorithmes qui viennent d'être présentés permettent le calcul d'une DCT-II.

Nikara [51] propose des algorithmes permettant d'établir les DST et DCT de type II et IV basés sur le même principe de décomposition matricielle que Takala. La structure d'une DCT-IV 8 points obtenue en applicant cet algorithme est illustrée en Figure 4.10.

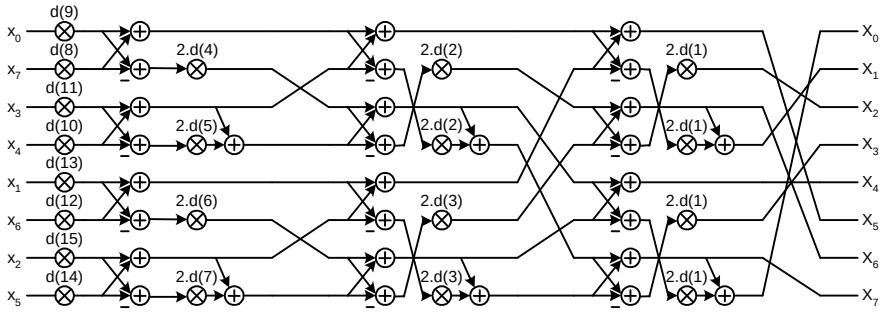


FIG. 4.10 – *DCT-IV, 8 points*

Elle utilise le même type d'architecture que celle présentée précédemment.
Dans le cas de la DCT-II, le nombre de multiplications et d'additions est identique à la structure proposée par Wang pour la DCT-II$\big(\mu(C_N^{II}) = (N/2)\log_2(N)+1$ et $\alpha(C_N^{II}) = (3N/2)\log_2(N) - N + 1\big)$, seule la caractéristique « répétitive » de la structure les différencie. Concernant la DCT-IV, un précédent travail de Wang [48] aboutissait à un nombre de multiplications $\mu(C_N^{IV}) = (3N/4)log_2(N)+ N/2$ et un nombre d'addition $\alpha(C_N^{IV}) = (7N/4)log_2(N)+N/2$. L'algorithme proposé par Nikara nécessite moins d'opérations logiques, $\mu(C_N^{IV}) = N(\log_2(N)/2+1)$ et $\alpha(C_N^{IV}) = (3N/2)\log_2(N)$, il est donc retenu.

L'étude de ces modules DCT a été effectuée dans le but de remplacer le module DFT composant le réseau de filtres polyphasés délivrant un vecteur de sortie de nature complexe, le Corrélateur ALMA traitant des données de nature réelle. Afin de réaliser une découpe en M SBs d'un signal en utilisant une structure polyphase couplée à une DCT, l'implémentation d'une DCT-IV est nécessaire [52] (répartition impaire des SBs). En Figure 4.11 est présentée l'achitecture à adopter afin de réaliser une telle découpe de la bande d'entrée.

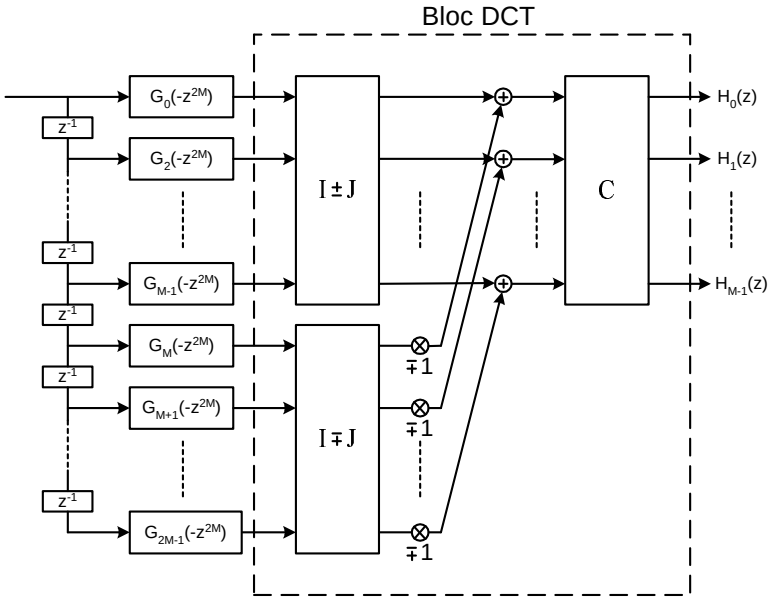


FIG. 4.11 – *Architecture Polyphasée couplée à une DCT*

Les éléments I et J sont les matrice unité et « anti-diagonale » unité. Cette structure nécessite seulement l'emploi d'une DCT-IV sur M points mais en contrepartie, la BB doit être découpée en $2M$ SBs.
En [53] est présentée une méthode permettant le passage d'une DCT-II à une DTC-IV par l'intermédiaire de matrices de transformation. C'est en cela que les papiers précédemment cités vont être utiles. L'implémentation de la DCT-IV peut alors être effectuée en appliquant à l'algorithme décrit par Takala (algorithme pour l'implémentation d'une DCT-II), les matrices suivantes :

$$C_{IV} = VC_{II}D \tag{4.23}$$

$$\text{avec } V = \begin{pmatrix} 0.5 & 0 & 0 & \cdots & 0 & 0 \\ -0.5 & 1 & 0 & \cdots & 0 & 0 \\ 0.5 & -1 & 1 & 0 & \cdots & \vdots \\ \vdots & & & & \ddots & 0 \\ 0.5 \cdots (-1)^{M+1} & \cdots & & 1 & -1 & 1 \end{pmatrix} \tag{4.24}$$

$$\text{et } D_{M*M} = diag\left(2\cos\left(\frac{\pi}{2M}(k+0.5)\right)\right)_{k=0\ldots M-1} \tag{4.25}$$

C_{IV} et C_{II} représentant respectivement les DCT-IV et DCT-II.

Une implémentation de cette structure (réseau de filtres polyphasés avec bloc DCT) et d'une structure polyphase couplée à une DFT suivie d'un réseau de filtre de Hilbert sont comparées en section 4

Remarque concernant « Pipelined FFT » (PFFT) et la DCT : Il est envisagée, dans les futurs mois, l'implémentation des structures FFT et DCT « pipelinée » [54, 51] qui devraient permettre de réduire considérablement le nombre de multiplications. Cette opération est réalisée en exploitant les propriétés de redondance et d'entrelacement des algorithmes utilisés pour la réalisation de la FFT[41] et de la DCT [51].

4 Implémentation

Les structures électroniques – DFT et DCT – présentées précédemment ont été implémentées afin de déterminer l'intérêt de leur utilisation potentielle. Une comparaison des ressources occupées par chacune d'entre elles est donnée dans le tableau 4.1.

TAB. 4.1 – Ressources en ALMs

	ALMs	Frequence max.	M512	Blocs DSP (9 bits)
Filtre Polyphase (32 voies)	13798	190 MHz	18	-
	8689	190 MHz	19	63(504)
DFT DIF	2866	182 MHz	6	-
	1708	168 MHz		27(210)
RFFT DIT	2266	127 MHz	1	-
	1850	127 MHz	-	6(46)
DFT-Winograd	1709	212 MHz	23	-
	1598	183 MHz	19	3(20)
Filtre de Hilbert	578	280 MHz	8	-
	330	280 MHz	8	2(12)
Filtre polyphase pour DCT (64 voies)	29875	183 MHz	29	-
	26424	184 MHz	26	63(504)
DCT-IV	4568	188 MHz	18	-
	3209	185 MHz	17	14(112)

Dans la suite de la section sont présentés les détails de l'implémentation de chacun des blocs constituant le banc de filtre.

4.1 Le bloc filtre

Le filtre utilisé pour la partie polyphase, associée à la DFT, est un filtre RIF d'ordre 1280 répondant aux specifications ALMA : atténuation >45 dB, ondulation dans la bande passante <0.2 dB, 1/32 de bande – les spécifiations en fréquence sont présentées dans le Tableau 4.2.

TAB. 4.2 – *Gabarit en fréquence du filtre associé à la DFT*

	Début	Fin
Bande passante	0	$30 \cdot f_{ech}/(64 \cdot 32)$
Bande d'arrêt	$34 \cdot f_{ech}/(64 \cdot 32)$	$f_{ech}/2$

La région de transition et la largeur de la bande passante correspondent à celle obtenue par la cascade des 2 étages de filtrage TFB. Ainsi la SB de largeur 62.5 MHz est obtenue après application du bloc DFT et du filtre de Hilbert assurant la conversion complexe-réel (qui permet de couvrir les fréquences positives avec les $M = 32$ SBs). Le processus de décimation par 32

appliqué en amont du filtre (Figure 4.1) est en fait intrinsèque au format d'entrée des données : entrée démultiplexée sur 32 voies. Le filtre est alors décomposé en 32 sous-filtres composés chacun de 40 poids (1280/32). Les coefficients du filtre sont encodés sur 10 bits et l'architecture électronique employée correspond à une implémentation classique d'un filtre RIF (cf. Annexe « Notions importantes de Traitement Numérique du Signal »). En figure 4.12 est tracée la réponse en fréquence du filtre.

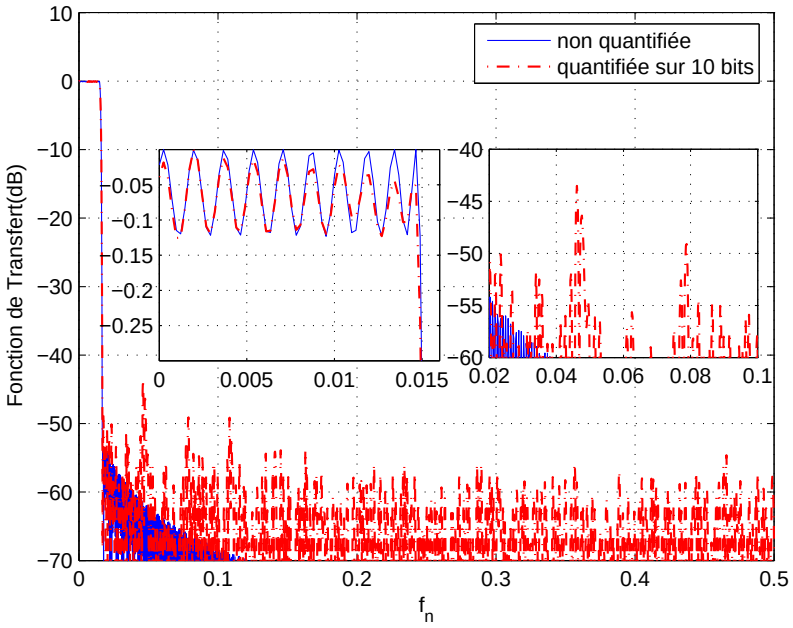


FIG. 4.12 – *Fonction de transfert du filtre RIF* 1280 *poids*

Le filtre associé au bloc DCT, lui, doit répondre à une spécification de largeur de bande 2 fois plus étroite (Tableau 4.3).

TAB. 4.3 – *Gabarit en fréquence du filtre associé à la DCT*

	Début	Fin
Bande passante	0	$15 \cdot f_{ech}/(64 \cdot 32)$
Bande d'arrêt	$17 \cdot f_{ech}/(64 \cdot 32)$	$f_{ech}/2$

Après application de la DCT les SBs générées sont de largeur 62.5 MHz. En effet l'apppication de la DCT comme indiqué en [52] entraine une répartition impaire des SBs autour de la fréquence centrale (Figure 4.4(b)). Le filtre permettant de remplir les spécifications est un filtre RIF d'ordre 2048 dont les coefficients sont aussi encodés sur 10 bits. La fonction de transfert est tracé en Figure 4.13.
En accord avec la méthode présentée en [52], la réponse impulsionnelle du filtre est découpée en 64 parties égales constituées chacune de 32 coefficients permettant l'obtention, après application de la DCT, de 32 SBs (Figure 4.11, pour $M = 32$). Afin de correspondre à l'entrée démultiplexée de ce nouveau filtre polyphase, le format de l'entrée délivrée par le système ALMA doit être adapté et subir un démultiplexage par 2.

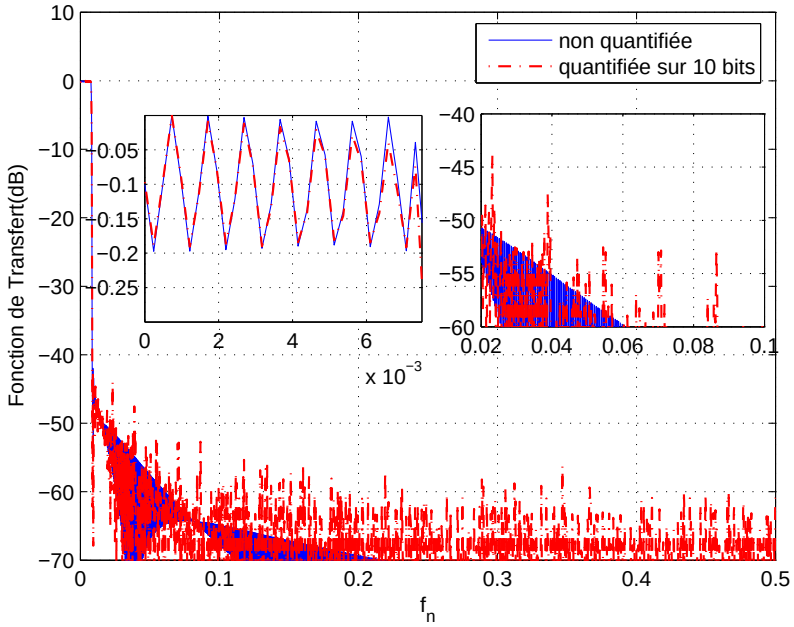


FIG. 4.13 – *Fonction de transfert du filtre RIF 2048 poids*

Concernant la troncature de la sortie du module « filtre », elle est réalisée sur 9 bits permettant de conserver une dynamique assez importante en sortie de ce dernier. Bien sur une modélisation mathématique de la structure de filtrage a été effectuée permettant de valider cette opération.

4.2 Le bloc de conversion en fréquence

Comme expliqué en sections 3.1, la structure électronique qui effectue cette conversion est décomposée en plusieurs blocs de traitement (ici 5 blocs puisque la DFT (DCT) est réalisée pour 32 points). Il a été décidé, après vérification du traitement par une modélisation mathématique, de tronquer chaque sortie de bloc à 9 bits. Les coefficients W_N $(d(n))$, eux, sont encodés sur 8 bits ce qui permet d'obtenir en sortie des blocs un signal sans dégradation notable.

Dans le cas d'une FFT-DIT, les équations implémentées sont les suivantes (Figure 4.5) :

$$x_{out} = (Re(x_{in}) + \Re(y_{in})) + j\,(\Im(x_{in}) + \Im(y_{in})) \quad (4.26)$$

$$\Re(y_{out}) = (\Re(x_{in}) - \Re(y_{in}))\cos\theta + (\Im(x_{in}) - \Im(y_{in}))\sin\theta \quad (4.27)$$

$$\Im(y_{out}) = (\Im(x_{in}) - \Im(y_{in}))\cos\theta - (\Re(x_{in}) - \Re(y_{in}))\sin\theta \quad (4.28)$$

En appliquant la méthode de la RFFT présentée en [42] (cf. section 3.2), suivant l'étage du bloc considéré, certaines simplifications de la structure dues aux paires conjuguées sont appliquées. En Figure 4.14 est présentée une vue schématique de la structure RFFT DIT implémentée.
3 différents blocs apparaissent sur les 5 étages de la FFT DIT, un délivrant 2 sorties réelles (rr_{fly}), un autre délivrant une sortie réelle et une autre complexe (rc_{fly}) et le dernier délivrant 2 sorties complexes (cc_{fly}).
Une structure dite mixte – association de 2 bloc Winograd et d'un étage FFT DIF – a aussi

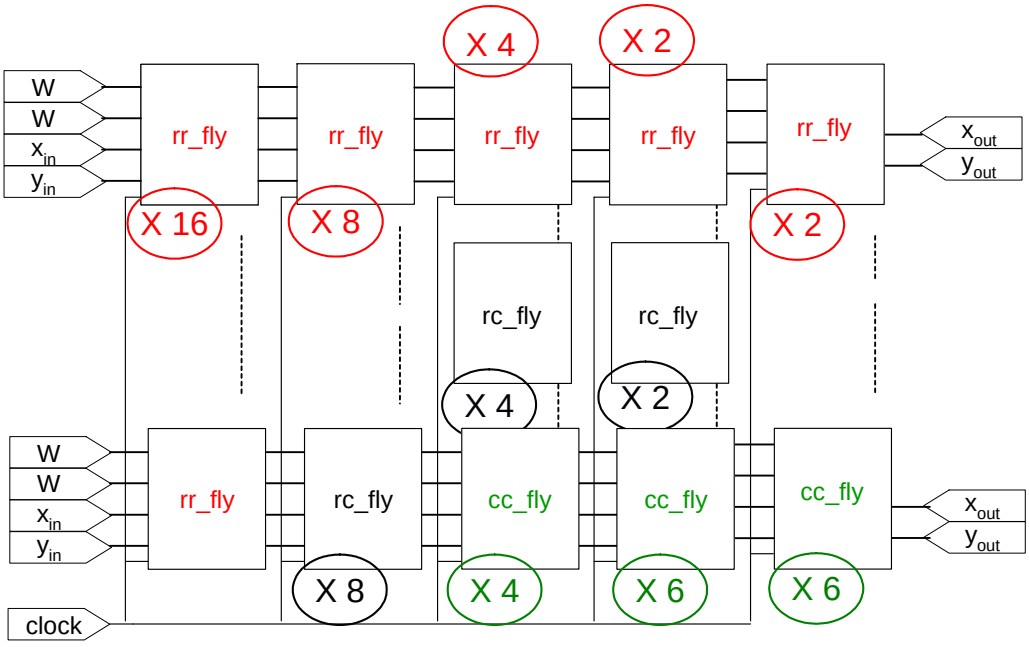


FIG. 4.14 – *Shématique de la FFT implémentée*

été implémentée [44]. C'est une méthode réputée pour son utilisation optimale des ressources. Elle est cependant caractérisée par une architecture complexe et non répétitive (impossibilité d'utiliser une telle structure pour une PFFT).
L'implémentation d'une architecture FFT DIF a aussi été réalisée, mais sans simplification de l'architecture. Cette structure fait figure d'étalon pour la comparaison avec les autres architectures étudiées et implémentées.

Remarque : L'utilisation d'un bloc DFT pout la réalisation de la conversion en fréquence des différentes SBs entrainent irrémédiablement l'emploi d'un étage de conversion complexe-réel basé sur l'utilisation d'un filtre de Hilbert. Il a donc été nécessaire de synthétiser et d'implémentater un tel filtre. Les spécifications que doit remplir ce filtre sont étroitement liées à celle de la SB finale, à savoir l'ondulation dans la bande passante de l'association de ce filtre et du précédent < 0.2 dB et une bande utile de l'ordre de 93.75% qui correspond à une SB de largeur 58.59 MHz (valeur obtenue après supression de 2 canaux de chaque coté de la bande avant reconstruction de la BB, cf. section 6.2.3 en 6.2.3 du Chapitre 1). En figure 4.15 est tracée la fonction de transfert du filtre synthétisé (ordre $n = 58$).
Au vu de la réponse impulsionnelle du filtre de Hilbert (un coefficient sur deux est nul), il est implémenté par une struture composée de 2 voies parallèles : l'une, voie des échantillons de rang pair, est simplement constituée d'un délai et l'autre, voie des échantillons de rang impair, est constituée d'une structure RIF classique à échantillons symétriques (du à la réponse antisymétrique des coefficients, les échantillons sont soustraits avant multiplication par le coefficient correspondant). La décimation par 2 est intrinsèque à l'architecture. En figure 4.16 est présenté la vue RTL de la structure utilisée afin de répondre à (1.7).

Dans le cas de l'emploi d'une DCT pour l'obtention de la conversion en fréquence, l'architecture

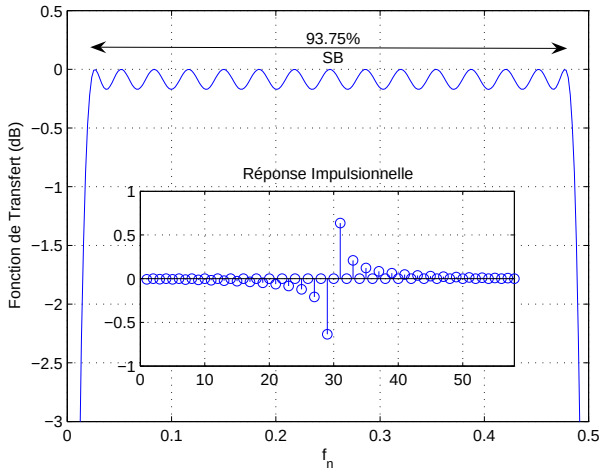


FIG. 4.15 – *Réponse impulsionnelle et réponse en fréquence du filtre de Hilbert*

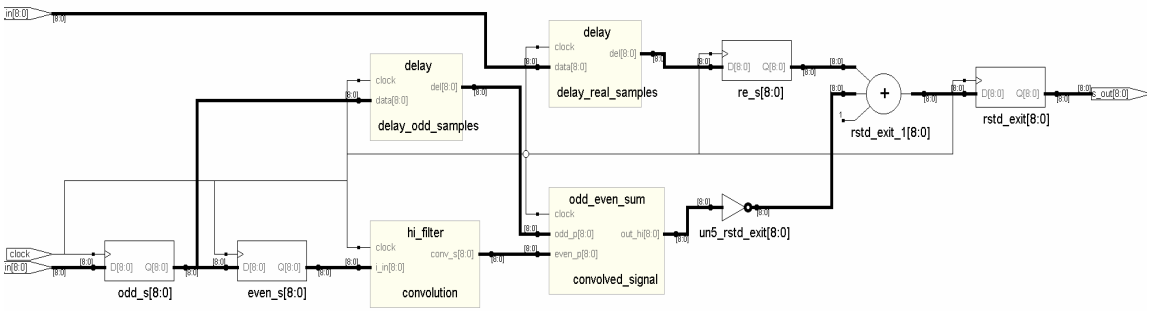


FIG. 4.16 – *Vue RTL du filtre de Hilbert*

est aussi composée de 5 étages. Dans l'optique de l'utilisation de la structure polyphase + DCT, une DCT-IV doit être utilisée ou une DCT-II adaptée comme expliqué en section 3.3.2. Deux structures ont été retenues (Figures 4.10 et 4.9 adaptées au cas $M = 32$), l'une pour la réalisation d'une DCT-II et l'autre pour la réalisation d'une DCT-IV. Seule la DCT-IV a été implémentée puisque directement applicable à la sortie du bloc filtre (la mise en œvre de l'adaptation d'une DCT-II étant assez lourde), chaque étage de la structure étant constitué de différents blocs apparaissant sur la figure correspondante.

5 Couverture complète de la bande de base

L'inconvénient de l'utilisation d'une structure polyphasée pour l'analyse d'une large bande provient des petites portions de bande de fréquence non analysées entre chaque SBs (les « gaps »). Cela est dû à la région de transition du filtre utilisé qui n'est jamais aussi étroite que voulue, la réponse idéale du filtre étant une fenêtre rectangulaire. Une solution consiste à contraindre cette région à une valeur très faible, afin d'obtenir une pente la plus raide possible. Cette action a cependant comme effet d'augmenter drastiquement la complexité du filtre synthétisé (augmentation du nombre de poids d'un facteur égal au facteur de réduction de la bande de transition) sans pour autant éliminer définitivement ces parties de bandes non analysées.

Dans le cas ALMA, le DDS du système TFB permet le chevauchement des SBs puisque la possibilité de placer chaque SB à n'importe quelle fréquence dans la BB est donnée. Ce déplacement en fréquence n'est pas permis avec une architecture polyphase à moins de revoir entièrement l'implémentation de cette dernière. Cela obligerait à utiliser une structure beaucoup plus complexe (le module FFT étant une structure optimisée) entrainant un besoin en ressources logiques incomparable. En [55] est présenté une structure de banc de filtre permettant de combler plus ou moins cette lacune du système polyphasé. L'amélioration du système de filtrage est basée sur la possibilité de passer d'une répartition paire des SBs à une répartition impaire (Figure 4.17).

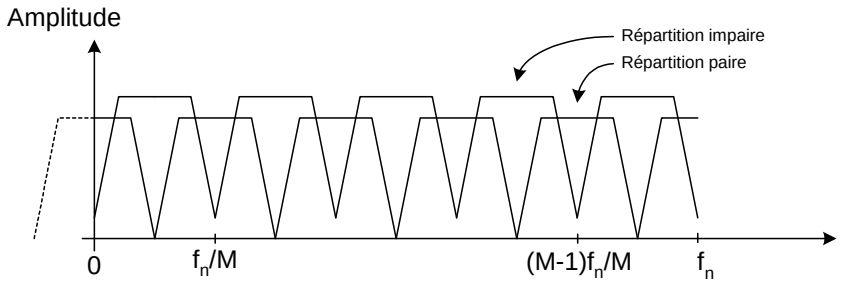


FIG. 4.17 – *Répartition paire et impaire des SBs*

Ceci est réalisé en modifiant la formule de la DFT :

$$y(k) = \sum_{n=0}^{N-1} x(n) W_N^{-(k+\nu)n},\ k \in [0, N-1] \tag{4.29}$$

où $W = e^{-k(2\pi/N)}$ et où $\nu = 0$ pour la répartition paire et $\nu = 0.5$ pour l'impaire.
Cette modification entraine une augmentation de la complexité de la DFT par 2 pour pouvoir utiliser les 2 répartitions à tout moment ; cependant il est alors possible de diminuer la complexité du filtre par 2 par exemple en spécifiant une région de transition 2 fois plus large (possibilité de spécifier une région de transition moins stricte puisque les SBs sont chevauchées de moitié).
En [52] est présentée une structure de banc de filtres qui permet d'obtenir la même fonctionnnalité avec un module DCT de longueur double (deux fois plus de SBs à synthétiser). Ici aussi, le filtre utilisé peut être de complexité moindre (facteur 2 envisageable). Une optimisation de l'implémentation de ce dernier est présentée dans ce même papier.
Ces solutions n'ont pu être implémentées dans le temps imparti à cette thèse et restent en suspens pour l'instant.

6 Conclusion

Ce chapitre avait pour but d'évaluer la complexité d'une architecture polyphasée et de comparer en terme de rapport complexité - efficacité (souplesse du traitement) cette dernière au système TFB actuel.
Deux architectures ont été étudiées : l'une basée sur un couple filtre polyphase - DFT (plus conversion complexe-réel) et l'autre sur un couple filtre polyphase - DCT. La première totalise un nombre de ressources d'environ 34600 ALMs (13798 filtre + 2266 RFFT + 32×578 Hilbert) et la deuxième d'environ 34400 ALMs (29875 filtre + 4568 DCT). Il est à noter que la complexité des 2 structures provient essentiellement de la partie « bloc de filtres »(et de la conversion complexe-réel dans le cas de la DFT).

Reste à implémenter les structures permettant de couvrir la bande totale à analyser en utilisant les répartition paires et impaires pour pouvoir réellement effectuer une comparaison en terme de ressources avec la structure TFB actuellement utilisée. Dans le pire cas, sans diminution de la complexité du filtre et sans optimisation de la structure permettant les deux répartitions possibles, cela multiplierait par deux le nombre d'ALMs nécessaire. Le travail est tout de même assez avancé pour pouvoir statuer sur le fait que la structure polyphasée complète sera beaucoup moins volumineuse que celle du TFB (16 puces de 10970 ALMs et 70 DSP blocks) mais au prix d'un manque de flexibilité, flexibilité qui fait l'originalité et l'efficacité du système ALMA.

Chapitre 5

Solution Retenue pour le Projet ALMA

1 Introduction

Après avoir présenté les solutions étudiées dans l'optique de l'optimisation d'une partie du système de filtrage ALMA (optimisation du premier étage et du second étage TFB) ou du système entier (architecture polyphase), le présent chapitre expose la solution retenue. Avant intégration du nouveau système de filtrage dans les puces FPGA, une étude permettant la validation du design tout au long de la conception a été effectuée. Le flot de conception mis en œuvre pour l'implémentation des différentes structures de filtrage est proposé dans la section qui suit. Dans une deuxième partie, ce nouveau système est présenté. Les tests et la validation sur carte de ce dernier y sont exposés.

2 Flot de conception

La réalisation et la validation d'une fonctionnalité destinée à être implémentée dans un circuit numérique (composant FPGA par exemple) se déroule en plusieurs étapes. En Figure 5.1 est présenté un algorigramme retraçant le flot de conception utilisé lors de l'élaboration des différentes fonctions utilisées dans le cadre de ce travail (l'étude des différents systèmes de filtrage envisagés).
Les spécifications du système à concevoir constituent le point de départ de la conception. De ces dernières découlent la description fonctionnelle de la fonction à réaliser.

2.1 Description fonctionnelle

Cette étape consiste à traduire la fonctionnalité (et l'interface) du système souhaité en une description plus ou moins hiérarchique ou plus ou moins détaillée (i. e. plus ou moins bas niveau et proche de l'electronique). Cette découpe hiérarchique et ce niveau de détail dépendent de la « confiance » que le concepteur accorde à l'outil de synthèse (certains étant plus ou moins performant) et au niveau de contrôle qu'il veut avoir sur cette synthèse. La description d'un module peut s'effectuer par saisie schématique (schéma bloc par exemple) ou par l'écriture d'un code spécialisé (VHDL pour Very high speed integrated circuit Hardware Description Language, ou Verilog - le langage utilisé par la suite est le VHDL). Ce code constitue en fait toujours la véritable source de description du système car la saisie schématique aboutit à la génération d'un code équivalent par les outils CAO.

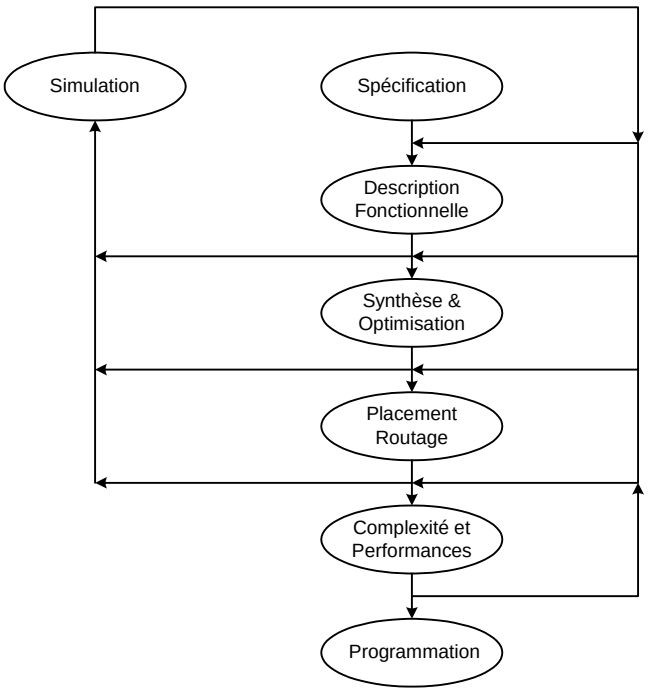


FIG. 5.1 – *Flot de conception*

L'utilisation du VHDL présente deux avantages essentiels : indépendance vis à vis de la technologie et optimisation de l'environnement de test.

2.2 Synthèse

La synthèse est effectuée à l'aide d'un outil CAO spécialisé, le synthétiseur qui est un compilateur particulier capable de générer une description structurelle du circuit à partir de la description VHDL. A travers le style d'écriture du code VHDL synthétisable (dit RTL pour Register Transfert Level), le synthétiseur reconnaît des « primitives » qu'il implante et interconnecte pour réaliser une fonctionnnalité équivalente à la description VHDL. Si le VHDL est indépendant de la technologie, la synthèse marque le passage vers une technologie donnée, voire un composant donné.
La description structurelle permet d'évaluer certaines caractéristiques du circuit physique. En effet la complexité est liée à la liste des primitives requises ; le taux de remplissage du FPGA peut alors être déterminé, les caractéristiques physiques des primitives étant connues par le fondeur. Les performances en vitesse du circuit peuvent aussi être estimées en associant à chaque primitive un retard fixe. Cependant cette première estimation n'est que grossière puisque les délais de routage entre primitives qui sont bien souvent critiques, ne sont pas encore connus.
Le synthétiseur utilisé dans le cadre de cette thèse permet l'optimisation des futures performances physiques du circuit en associant, avant synthèse, aux fichiers VHDL un fichier de contraintes. Ainsi, en plus de la contrainte de timing général (vitesse nominale de fonctionnement du design) habituelle, il est laissé au concepteur le choix du type de ressources (de primitives) à utiliser pour les différentes parties du design ou encore le choix de partionner ce même design en différents blocs soumis à divers timing. D'autres options permettant encore d'améliorer les caractéristiques physiques du circuit sont disponibles mais ne sont pas abordées ici. En fin de synthèse, un fichier

VQM (Verilog Quartus Mapping file) est généré. Il constitue le point d'entrée pour l'étape de placement routage.

2.3 Placement - Routage

Cette étape consiste, à partir de la description structurelle du circuit (fichier VQM obtenu après synthèse) à positionner physiquement les primitives sur le silicium et les interconnexions entre primitives et entre primitives et pins d'entrée - sortie. Cette procédure est complètement automatisée dans le cas du routage d'un FPGA.
Suite à cette étape, les performances réelles en vitesse du circuit ainsi que la complexité de la structure (le taux de remplissage du FPGA) sont disponibles.

2.4 Simulation

La simulation permet de valider les différentes étapes du flot de conception descendant. Afin de réaliser cette simulation, tout l'environnement de la fonctionnalité à implémenter dans le FPGA (entrée - sortie, horloge...) est décrit en VHDL (description non-synthétisable). Cette description VHDL, appelée « Test Bench » constitue la struture de plus « haut niveau » du nouveau système.
La première simulation effectuée concerne le code VHDL avant synthèse décrivant la fonctionnalité. Elle permet la validation de la fonctionnalité à implémenter. Ensuite le modèle structurel, obtenu après routage, est simulé ; les performances en vitesse du circuit sont alors évaluées. Un fichier SDF (Standard Delay File) contenant les retards des portes logiques est délivré par le simulateur.
Afin de tester et de valider la fonctionnalité, des vecteurs de tests de type radio-astronomique (signaux aléatoires similaires à ceux que le circuit physique traitera lorsqu'il opérera dans l'électronique du radio-telescope : bruit blanc gaussien associé à des sinusoïdes) sont générés et sont utilisés comme entrée du Test Bench (Figure 5.2).

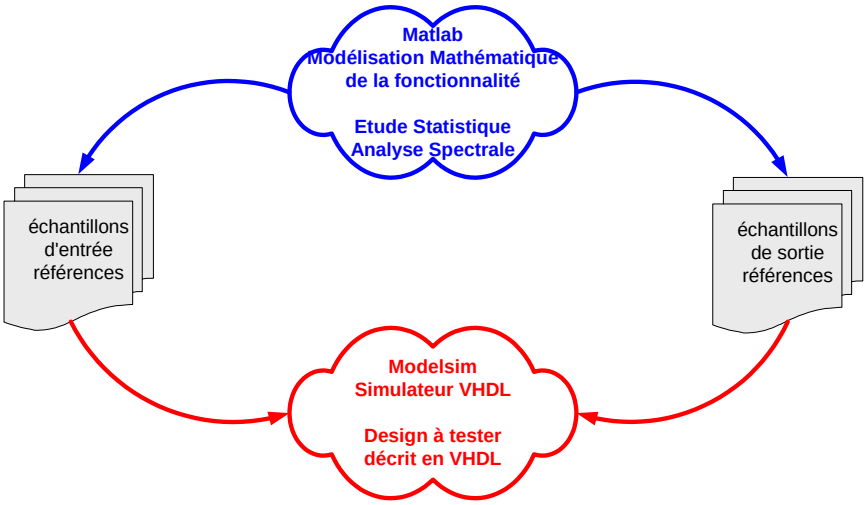


FIG. 5.2 – *Simulation fonctionnelle*

L'utilisation de tels signaux permet une analyse du système dans des conditions opérationnelles. Les estimations de consommaton effectuées aux cours des simulations, et basées sur le nombre de commutations dans le modèle structurel, devraient être en particulier très proche de la réalité.

Ces signaux de test sont générés à l'aide de Matlab (software de modélisation mathématique) et sont encodés sur un nombre de bits définis correspondant aux signaux délivrés au système de filtrage après numérisation. Une description mathématique de la fonctionnalité à réaliser est alors effectuée sous Matlab, nourrie par ces même vecteurs de test, permettant ainsi d'obtenir des vecteurs de test de sortie qui serviront de référence à la simulation VHDL (Figure 5.2).

3 Validation et test de la solution retenue

Compte tenu du manque de flexibilité de la structure polyphase présentée dans le chapitre précédent, cette solution est écartée des structures envisageables. Dans le tableau 5.1 est donc présenté un récapitulatif des solutions étudiées les plus intéressantes dans l'optique du remplacement des deux étages de filtrage TFB.

TAB. 5.1 – Récapitulatif des solutions étudiées

		ALMs	Fréq. max.	Blocs DSP (9 bits)
1^{er} étage	TFB 1^{er} étage	1775	180 MHz	-
	CIC non-rec QB	653	240 MHz	-
2^{eme} étage	2^{nd} étage actuel	185	260 MHz	2(16)
	(sans bloc DSP)	1168	141 MHz	-
Phase linéaire	Structure 2 passages	415	142 MHz	-
	Réduction du modèle	27	147 MHz	5(36)
Phase non linéaire	Structure Allpass	239	98 MHz	-
	IIR SOS	18	148 MHz	3(24)

Il est à noter que les fréquences maximales de fonctionnement indiquées concernent l'étage de filtrage synthétisé. Elles permettent seulement de s'assurer de la robustesse du design.
Dans la première partie du tableau sont donnés les résultats relatifs au 1^{er} étage de filtrage TFB. La solution optimale retenue, concernant cet étage, est composée d'un filtre CIC non-récursif cascadé avec un filtre quart de bande. Elle permet de diminuer presque d'un facteur 3 la complexité de la fonctionnalité « 1^{er} étage TFB ».
Le code VHDL, et donc la fonctionnalité du système, a été validé grâce à des patterns de tests générés sous Matlab en utilisant la méthode décrite en section 2.4.

Dans la seconde partie du tableau, ce sont les résultats relatifs au 2^{nd} étage TFB qui sont présentés (rappelons que l'implémentation actuelle est celle utilisant les blocs DSP) : les structures à phase linéaire et à phase non-linéaire y apparaissent. Les solutions engendrant une phase non-linéaire sont ici présentées à titre indicatif puisque l'étude de l'utilisation de telles structures en radioastronomie doit encore être approfondie. On peut tout de même remarquer leur intérêt potentiel en terme d'occupation de ressource en comparaison avec la solution actuelle.
Les solutions présentant la caractéristique de phase linéaire n'ont cependant pas été retenues pour l'utilisation dans le système de filtrage ALMA, le 2^{nd} étage actuel semblant optimal (dû à la méthode de récirculation des poids utilisée et à l'utilisation optimale des blocs DSP). La solution nommée « structure 2 passages » dont la vue RTL est donnée en Figure 5.3 n'aboutit pas à une diminution des ressources assez importante comparée à celle utilisée actuellement (comparaison effectuée avec l'implémentation sans bloc DSP) pour justifier une telle complexification de l'architecture du filtrage ALMA.
La solution dérivée de la méthode de réduction du modèle aboutit à une cascade de 6 cellules SOS. Au vue des résultats présentés dans le tableau, cette solution est plus volumineuse que l'actuelle.

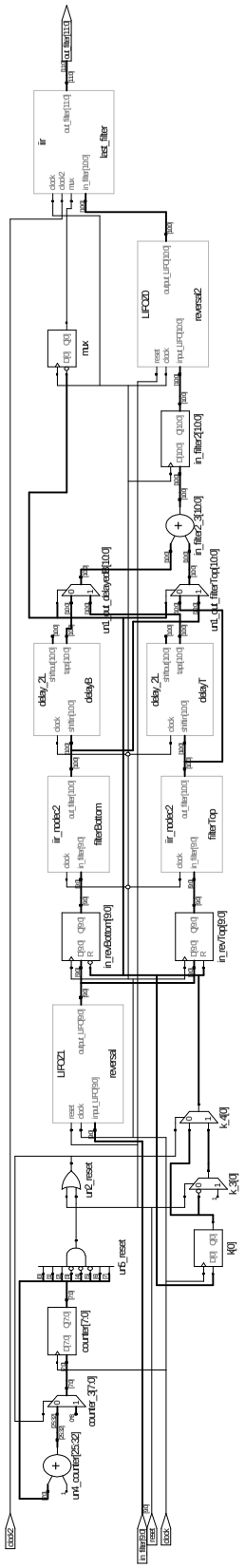
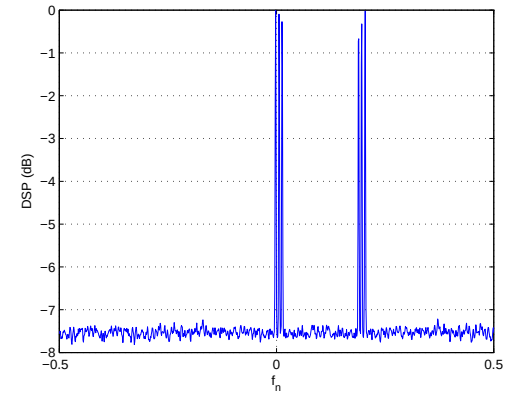


FIG. 5.3 – *Vue RTl de la structure 2 passages*

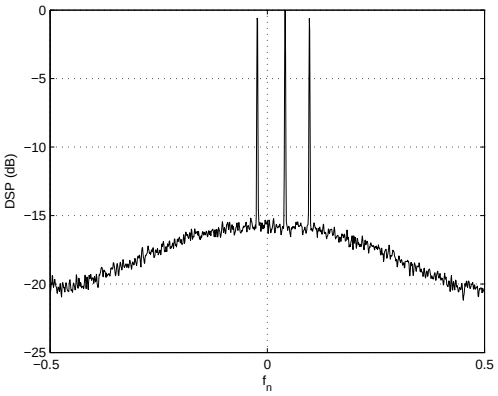
Le système de filtrage retenu se décompose donc en trois parties : un filtre CIC non-récursif ($D = 8$, $N = 2$), un filtre quart de bande d'ordre $n = 15$ et la structure du filtre RIF actuel du 2^{nd} étage TFB. Les poids de ce dernier doivent cependant être recalculés afin de compenser la chute dans la bande passante engendrée par la cascade CIC + QB.

3.1 Traitement du signal dans le nouveau bloc de filtrage

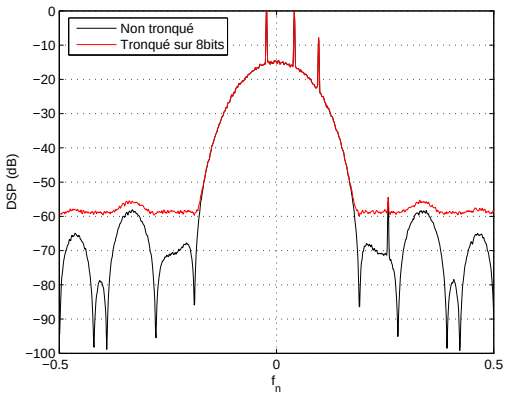
Le signal délivré par le DDS au premier étage de filtrage utilise la représentation décalée (cf. section 6.2.1 en p. 35 du Chapitre 1). Dans le cas d'un filtre CIC non-récursif, le passage à une représentation classique est effectué dans le premier bloc $\left(1 + z^{-1}\right)$ en ajoutant 1 à la somme des deux échantillons successifs. Le spectre du signal obtenu en sortie est celui de la Figure 5.4(b).



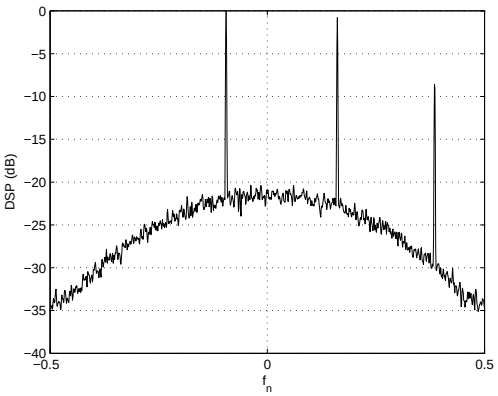
(a) Sortie du DDS



(b) Sortie du filtre CIC, après décimation



(c) Sortie du quart de bande

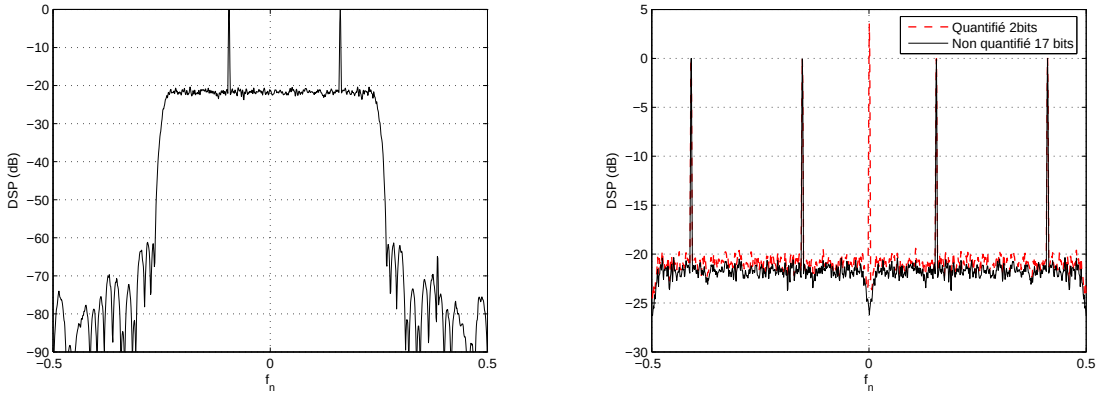


(d) Sortie du quart de bande après décimation

FIG. 5.4 – *Spectre de sortie de chaque sous-étage*

Le signal de sortie résultant de l'architecture CIC non-récursive est codée sur 12 bits (pleine échelle). Il n'est pas tronqué afin de ne pas biaiser le signal. Après application du filtre quart de bande, le spectre de sortie obtenu est tracé en Figure 5.4(c). Cette sortie est tronquée à 8 bits (12 LSBs supprimés) permettant d'assurer la compatibilité des signaux entre la sortie de ce filtre et l'entrée du second étage TFB. La dynamique en sortie, correspondant à 8 bits, est égale à approximativement 48 dB. Le spectre de la Figure 5.4(d) est celui obtenu aprés décimation ; apparait sur cette Figure la chute dans la bande passante induite par cette cascade CIC + QB.

La troncature effectuée en sortie de l'étage « filtre quart de bande » introduit l'apparition d'une composante continue et devra être compensée en entrée du second étage TFB.
Sur la Figure 5.5, sont présentés les spectres en sortie du « 2^{nd} étage TFB » ainsi que le spectre en sortie de l'étage conversion complexe-réel résultant de la modélisation mathématique.



(a) Sortie sur 17 bits avant décimation

(b) Sortie du convertisseur

FIG. 5.5 – *Sortie du filtre fixant la SB et sortie du convertisseur complexe-réel*

Après conversion et requantification du signal sur 9 bits (destiné au calcul de la puissance de sortie) puis requantification sur 2 bits redirigé vers le corrélateur (Figure 1.20), la composante DC amenée par les différentes requantifications se retrouve au centre du spectre, soit hors de la bande d'intéret (Figure 5.5(b)).
La chute dans la bande passante résultant de la cascade filtre CIC - filtre quart de bande a été compensée par celle du second étage TFB. En Figure 5.6 est présentée la fonction de transfert de ce « nouveau » 2^{nd} étage.

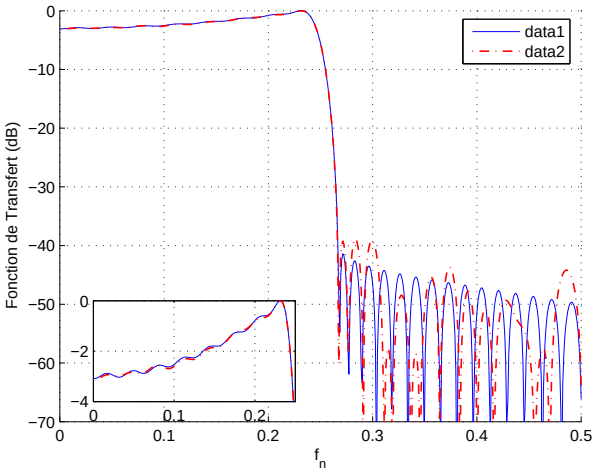


FIG. 5.6 – *Fonction de transfert du filtre fixant la SB à 62.5 MHz*

Un algorithme de minimisation – *amoeba simplex minimization* – ayant comme entrée la réponse en fréquence de l'étage à compenser est utilisé. La fonction de transfert de départ est générée par l'algorithme de Remez, les poids constituant le filtre sont ensuite ajustés individuellement

afin de répondre à la forme désirée. De ce fait, la spécification d'ondulation (à savoir $0,2$ dB) est respectée.

3.2 Perte associée à l'ensemble CIC + QB + 2^{nd} étage

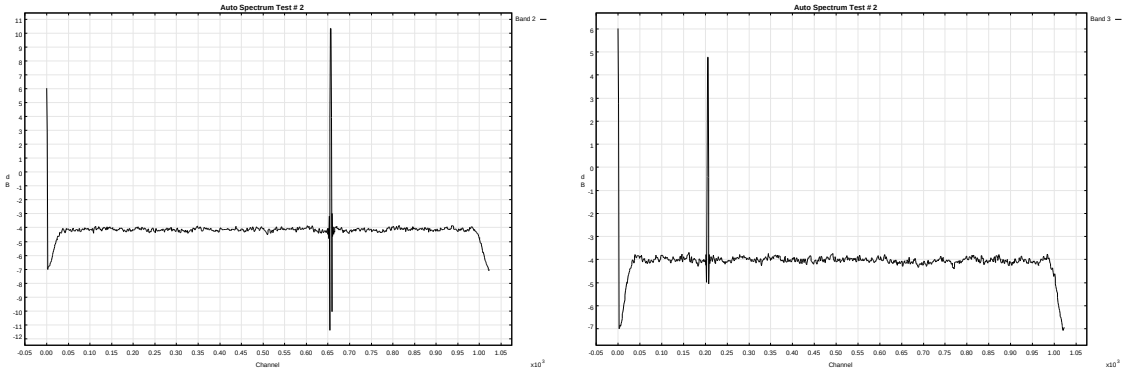
L'adoption d'une structure multi-étages conduit inévitablement à une perte de sensibilité lors du calcul de la fonction de corrélation du signal (due aux différentes requantifications du signal). Afin de limiter le nombre de ressources utilisées dans chacun des étages de filtrage constituant la cascade, une troncature des LSB est réalisée à chaque sortie. La perte d'information (liée à la suppression des LSBs) peut conduire à une dégradation du signal à analyser. Nous avons mené une étude semblable à celle présentée en section 6.2.2, p. 36 du Chapitre 1 afin d'apprécier cette dégradation.
Comme indiqué précédemment la sortie du filtre CIC non récursif est codée sur 12 bits (sans troncature). Dans le cas d'une troncature à 8 bits, comme il en a été question en section 4.2.1.2, p. 67 du Chapitre 2, le biais engendré est égal à $\frac{1}{2^5}$ (troncature de 4 LSBs), valeur non-négligeable. Après traitement du signal par l'étage quart de bande, ce biais se retrouve multiplié par le gain de ce nouvel étage ($\sum_{k=0}^{N-1} h(k) = 758$), à savoir : $\frac{758}{32}$. Ce biais est alors additionné à celui induit par la troncature en sortie du quart de bande : $= \frac{1}{1024}$ (9 LSBs sont tronqués, après suppression des MSBs, pour atteindre le format de sortie désiré). En comparaison avec le biais induit par le 1^{er} étage de filtrage TFB ($\frac{1}{1024}$), celui-ci est beaucoup trop important. C'est pour cette raison que la solution non tronquée de la structure CIC non-récursive est préférée (sous peine de dégradation de l'efficacité du traitment du signal ou de modification trop importante de l'étage de conversion complexe-réel où ce biais est compensé).
La cascade CIC-QB induit alors un biais égal à $\frac{1}{2^{13}}$ (12 LSBs tronqués pour l'obtention de la sortie codée sur 8 bits). Ce biais doit être multiplié par le gain du dernier étage de filtrage (le 2^{nd} étage TFB adapté à la cascade) égal à 496, on obtient alors le biais total engendré par cette nouvelle structure : 2^{-4}. Le fait de ne pas tronquer le signal en sortie du CIC induit une augmentation de la complexité de la structure QB (de faible importance tout de même, QB à entrée sur 8 bits : 195 ALMs, QB à entrée sur 12 bits : 295 ALMs). Cependant elle permet de limiter de façon importante le biais induit par la structure cascadé (correspondant à 91 dB) et donc la perte d'efficacité ; une compensation de ce dernier n'est alors pas nécessaire.

Concernant le surplus de bruit ajouté dans la bande d'intérêt après décimation (dû à une dynamique un peu faible, codage sur 8 bits), il peut toujours être compensé par une intégration plus importante ($\sigma \cong \frac{1}{\sqrt{B\tau}}$). Il est cependant important de limiter le plus possible son importance.

3.3 Intégration de la solution dans les puces FPGA

La solution retenue ici a été intégrée dans les puces FPGA de la carte ALMA TFB afin d'évaluer le gain en consommation et la diminution de la température de jonction des puces (cf. section 7.2 en 7.2 du Chapitre 1). Les spectres de la Figure 5.7 représentent deux sous-bandes adjacentes (n°2 et n°3 sur 32) de 62.5 MHz obtenues après filtrage par l'intermédiaire du Test Fixture et de son corrélateur.
Une raie, modélisant une information astronomique, a été placée dans chacune des deux sous-bande. Ces spectres permettent de confirmer le bon fonctionnement du filtre implémenté :

- le design a subi les tests de validation avec succès (cf. section 7.1.2 en p. 41 du Chapitre 1)
- aucune raie « fantôme » (due au repliement spectral) n'apparait dans la bande passante des sous-bandes



(a) Sous-bande n°2 (b) Sous-bande n°3

FIG. 5.7 – *Spectre de deux Sous-bandes adjacentes, obtenu par le Test Fixture*

– la bande passante résultante du filtrage complet est plate et respecte l'ondulation maximale acceptable, c'est à dire ±0.2 dB (Figure 5.8).

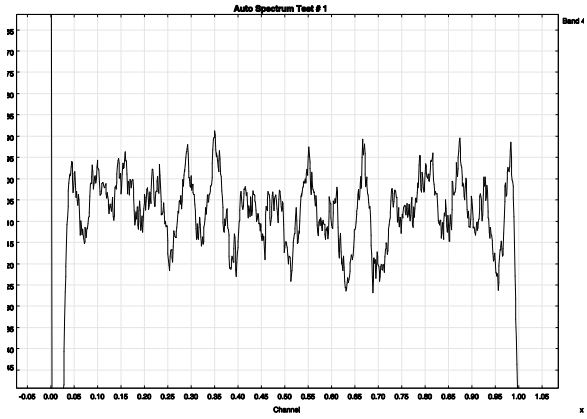


FIG. 5.8 – *Ondulation dans la bande passante*

3.4 Tests thermiques

La consommation obtenue avec ce nouveau design, en fonctionnement nominal (puces allumées, incrément du DDS chargé et test 7 lancé), avoisinne les 58 W. Comparé à l'ancien design qui consommait 75 W, cela représente une amélioration de l'ordre de 21 %. Interessons nous maintenant aux températures de jonction qui conditionnnent la durée de vie d'une puce. En Figure 5.9 sont tracés les évolutions dex température de jonction des 16 puces de la carte en fonction du temps, jusqu'au régime établi.

Ce relevé de températures a été effectué en fonctionnement nominal, sans refroidissement de la carte par un quelconque flot d'air. Le nouveau design permet donc aussi de diminuer la température de jonction des puces (à comparer avec le tracé en Figure 1.35). La température atteinte par la puce la plus exposée à la dissipation de ces voisines est de l'ordre de 100 °C, loin de la température maximale conseillée par Altera. Le graphique présenté en Figure 5.10 permet

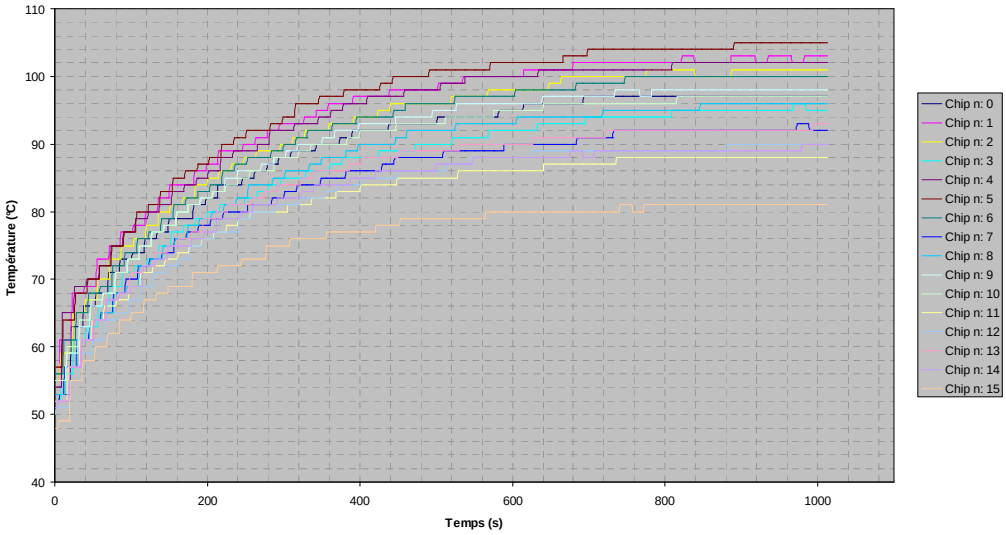


FIG. 5.9 – *Température de jonction de chaque puce en fonction du temps*

de comparer la température de jonction des puces pour les deux structures (l'originale : TFB et la nouvelle : CIC + QB) en fonction de la puissance du flot d'air ventillant la carte.

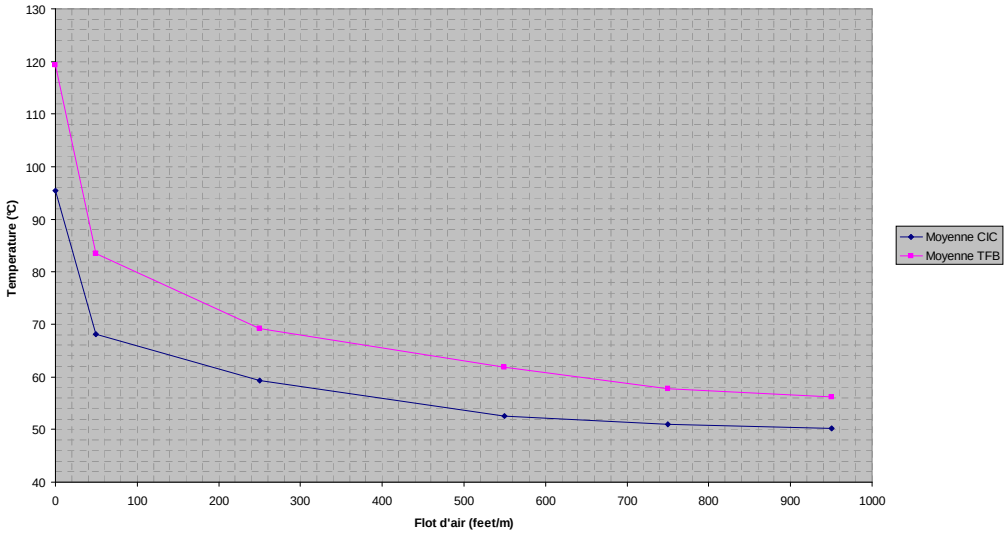


FIG. 5.10 – *Comparaison des 2 structures électroniques en terme de température de jonction*

Cinq mesures à différentes valeurs de flot d'air ont été effectuées. Le graphique en résultant est seulement voué à l'évaluation de la température de jonction moyenne des puces pour une certaine vitesse de ventillation et ne doit être, en aucun cas, pris comme référence. Ces mesures ont été effectuées dans un cadre bien particulier, avec un flot d'air très directif. Une conclusion peut tout de même être établie au vu des résultats : plus la vitesse de propulsion de l'air est élevé moins la différence entre les architectures « TFB » et « CIC » est flagrante en terme de température de jonction.

4 Conclusion

Ce chapitre présente la solution qui a été retenue dans le cadre du projet ALMA. Ce choix résulte de différentes études menées durant le déroulement de la thèse. L'objectif était la réduction de la consommation du système de filtrage actuel qui s'élevait à 75 W par carte de filtrage (rappelons que 512 cartes sont nécessaires pour traiter tous les signaux provenant des antennes). L'amélioration obtenue en terme de consommation en intégrant cette solution dans le design TFB est de l'ordre de 20 %. La puissance dissipée par la carte atteint désrormais 60 W. Ceci a des répercutions sur la configuration du système de refroidissement des racks du Corrélateur et va surement permettre de diminuer le nombre de ventilateurs utilisés dans ces racks sur le site d'Atacama à 5000 m.
Ce nouveau système de filtrage est en cours de test sur le site du NRAO à Charlottesville. Les cartes de filtrage seront ensuite intégrées dans un système composé de deux antennes ALMA situé à Socorro (Nouveau Mexique) afin de valider le fonctionnement complet du système de détection (filtrage et corrélation) avant d'être implanté sur le site d'Atacama.

Conclusions et Perspectives

Aujourd'hui, les projets de radioastronomie demandent l'utilisation de systèmes électroniques de plus en plus performants et donc nécessitant une architecture de plus en plus complexe. Le projet ALMA fait partie de ces nouveaux projets, les caractéristiques du Corrélateur (banc de filtre plus corrélation) parlant d'elles-mêmes : 512 cartes de filtrage au format $6U$ traitant chacune 2 GHz de bande et 512 cartes de corrélation au format $9U$ calculant chacune 64×4 klags. Toutes ces cartes sont rassemblées dans une pièce ventilée mais située à 5000 m d'altitude où l'air est raréfié (pression moyenne de 550 mbar). Dans ce cas la dissipation thermique du système total devient un paramètre crucial.

Les travaux de recherche menés durant cette thèse concernent en premier lieu l'amélioration de la consommation du système de filtrage ALMA (TFB). Une étude de différents types de filtre et différentes architectures électroniques a été effectuée. La première solution investiguée se rapporte à une modification de la structure TFB elle-même.
Le premier étage TFB a alors été remis en question. Ce premier étage permet d'effectuer une décimation par 32 de la bande d'entrée sans contrainte spécifique de la région de transition. A cet effet, le filtre Cascaded Integrator Comb (CIC) est le mieux adapté puisque possédant une architecture simple tout en délivrant un facteur de décimation important. Après adaptation de la structure au format du signal d'entrée du TFB, plusieurs structures ont été comparées : un filtre CIC à architecture récursive, à architecture non-récursive ou possédant une architecture à entrée démultiplexée. Après implémentation de ces dernières, les résultats obtenus sont loin de ceux escomptés. Il a donc été décidé d'utiliser une structure multi-étages permettant d'atteindre le facteur de décimation initial composée au moins d'un filtre CIC.
Cette solution s'est révélé fructueuse puisqu'elle permet la diminution du nombre de ressources logiques pour l'implémentation de la fonctionnalité. Plusieurs cas ont été étudiés, la cascade CIC - RIF quart de bande s'est avérée la plus intéressante.

Toujours dans l'optique d'améliorer la consommation du système, un recensement des structures de filtre RII à phase non-linéaire ainsi que de différentes méthodes permettant l'obtention de la phase linéaire – ou approximativement linéaire – dans la bande passante a été effectué. L'objectif éventuel était le remplacement du second étage de filtrage TFB.
La contrainte de phase linéaire a toujours été une caractéristique essentielle des systèmes de filtrage utilisés traditionnellement en analogique pour les projets de radioastronomie. L'étude du filtre RII s'explique par le fait de la faible complexité de sa structure (pour un même gabarit en fréquence, un filtre RII nécessite – environ – 10 fois moins de coefficients qu'un filtre RIF). L'inconvénient de ce type de filtre est sa phase non linéaire. La comparaison des méthodes résultant en une linéarisation de cette phase avec le filtre RIF utilisé dans le 2^{nd} étage TFB a révélé la structure « 2 passages » composée de filtres « two path allpass » comme étant la solution optimale.

La possibilité de l'emploi d'un filtre RII – à phase non linéaire – dans un projet de radioastronomie a été introduite. Du fait de son caractère parfaitement reproductible en numérique, un filtre possédant une phase non linéaire semblerait être une bonne option pour l'avenir, mais cette question reste pour l'instant en suspens et mériterait d'être étudiée de manière plus approfondie. Il serait plus facile d'utiliser ce type de filtre plutôt qu'un filtre RII dont la phase n'a été qu'« approximativement » linéarisée.

La dernière option que nous avons étudiée consiste à remplacer le système TFB complet par un système de filtres polyphases. Le but d'une telle étude était de comparer ces deux solutions en terme de complexité-flexibilité.
Deux structures de réseau de filtres polyphases, l'une basée sur l'utilisation d'une DFT et l'autre sur l'utilisation d'une DCT, ont été étudiées et implémentées. Afin de réaliser une couverture complète de la bande à analyser, les structures doivent subir une modification. En effet, par l'utilisation d'une structure polyphase clasique, des « trous » apparaissent entre chaque SB correspondant à des bandes de fréquences non analysées. Il est possible de combler ces « trous » spectraux en juxtaposant deux réalisations de filtres polyphases, l'une aboutissant à une répartition paire des sous-bandes et l'autre à une répartition impaire.
L'implémentation de cette solution est en projet et ne sera pas présentée dans ce document. Néanmoins, il est possible de conclure sur l'intérêt d'une telle structure. Du point de vue de l'utilisation des ressources, elle est en effet plus intéressante que celle du système TFB. Cependant le système TFB est beaucoup plus flexible et cette flexibilité est le principal atout du système de filtrage ALMA, elle en fait l'originalité. Dans le contexte ALMA, le système polyphase ne constitue donc pas une solution optimale.

Après étude et comparaison des différentes solutions envisageables dans le but de diminuer la consommation de la carte TFB, le choix final s'est porté sur une cascade CIC - RIF quart de bande - RIF demi-bande. Il a été décidé de conserver l'architecture du filtre demi-bande actuel (le 2^{nd} étage TFB) malgré l'amélioration en terme d'utilisation de ressources de la solution filtre RII « deux passages ». Cela ne justifie pas une telle complexification de l'architecture du filtre, l'implémentation de ce 2^{nd} étage TFB étant déja bien optimisée et premettant de synthétiser deux largeurs de bande (62.5 MHz et 32.25 MHz).
La structure de filtrage retenue permet de réduire la compléxité du filtre total et de descendre la consommation de la carte de 20 % : la consommation atteignait 75 W par carte TFB, elle est désormais de l'ordre de 58 W. Les températures de jonction ont, quant à elles, diminué d'une douzaine de degrés d'une version à l'autre. Ce nouveau système de filtrage, composé d'une cascade de 3 étages, a été validé mais est toujours en cours de test sur le site d'intégration du Corrélateur ALMA à Charlottesville ; notre nouveau design sera utilisé au final dans le Corrélateur ALMA sur le site de Chajnantor à 5000 m.

Ajoutons enfin que les différentes études que nous avons menées pour le projet ALMA en mettant l'accent sur la flexibilité du filtrage numérique et la faible consommation ont permis au laboratoire hôte de ce travail d'élargir ses connaisances en terme de filtrage numérique, le préparant ainsi à l'utilisation de tels systèmes dans le cadre de futurs projets.

Bibliographie

[1] B. Quertier and G. Comoretto et al, "Enhancing the baseline ALMA correlator performances with the second generation correlator digital filter system," *ALMA Memo*, no. 476, November 2003.

[2] B. Quertier, "Système de filtrage numérique pour le corrélateur de l'interféromètre ALMA," Ph.D. dissertation, Université Bordeaux I, 2004.

[3] A. R. Thompson, J. M. Moran, and G. w. Swenson, *Interferometry and Synthesis in Radio Astronomy*. Wiley-Interscience, 1986.

[4] C. Recoquillon and A. Baudry et al, "The ALMA 3-bit 4 Gsample/s, 2-4 GHz input bandwidth, flash analog-to-digital converter," *ALMA Memo*, no. 532, July 2005.

[5] J. Webber et al, "64 antenna correlator specifications and requirements," *ALMA System Document*, 2002.

[6] R. Escoffier et al, "The ALMA correlator," *A&A*, vol. 462, pp. 801–810, 2007.

[7] G. Comoretto, "Phase effects in an hybrid correlator," Tech. Rep., 2003, http://www.arcetri.astro.it/~comore/phase.pdf.

[8] T. Parks, L. Rabiner, and J. McClellan, "FIR digital filter design techniques using weighted chebyshev approximations," *Proc. IEEE*, vol. 63, pp. 595–610, April 1975.

[9] F. Harris, *Multirate Signal Processing for Communication Systems*. Prentice Hall PTR, 2004.

[10] F. C. Cooper, "Correlators with two-bit quantization," *Aust. J. Phys.*, vol. 23, pp. 521–7, August 1970.

[11] A. R. Thompson, "Quantization efficiency for eight or more sampling levels," MMA Memo 220, Tech. Rep., July 1998.

[12] G. Comoretto, "Data processing in the ALMA tunable filterbank," *ALMA MEMO*, vol. CORL-60.01.07.00-00X-A-MAN, 2005.

[13] ——, *Programming Manual for Tunable Filter Bank*, 2004, aLMA Document CORL-00.01.07.01-A-MAN.

[14] ——, "Sub-channel stitching and truncation errors in the ALMA tunable filterbank," *ALMA MEMO*, vol. CORL-60.01.07.00-012-A-DOC, 2005.

[15] F. R. Schwab, "Van vleck correction for the GBT correlator," *GBT Memorandum*, no. 250, August 2007.

[16] P. C. B. Quertier, "Pre-prototype TFB card test report," Tech. Rep., 2005, aLMA Document CORL-60.01.07.10-004-A-REP.

[17] J. Max, *Méthode et Techniques de Traitement du Signal et Application aux Mesures Physiques*, Masson, Ed., 1977.

[18] B. Quertier, *Preliminary Acceptance In-House (PAI) Procedure for the ALMA Tunable Filter Bank (TFB) Card*, 2005, aLMA Document CORL-60.01.07.05-007-A-PLA.

[19] E. B. Hagenauer, "An economical class of digital filters for decimation and interpolation," *IEEE Transactions in Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 2, pp. 155–162, April 1981.

[20] P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications : A tutorial," in *Proceedeings of the IEEE*, vol. 78, January 1990, pp. 56–93.

[21] F. Daneshgaran and M. Laddomada, "A novel class of decimation filters for $\Sigma\Delta$ A/D converters," *Wireless Communications and Mobile Computing*, vol. 2, pp. 867–882, November 2002.

[22] H.-K. Yang and W. M. Snelgrove, "High speed polyphase CIC decimation filters," *Circuits and Systems, 1996. ISCAS '96., 'Connecting the World'., 1996 IEEE International Symposium on*, vol. 2, pp. 229–232, May 1996.

[23] M. Bellanger, G. Bonnerot, and M. Coudreuse, "Digital filtering by polyphase network : Application to sample-rate alteration and filter banks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 2, pp. 109–114, April 1976.

[24] J. I. Yonghong Gao, Lihong Jia and H. Tenhunen, "A comparison design of comb decimators for sigma-delta adcs," *Analog Integrated Circuits and Signal Processing*, vol. 22, no. 1, pp. 51–60, 1999.

[25] G. Commoretto, "Notes on the implementation of a time demultiplexed comb filter," http://www.obs.u-bordeaux1.fr/electronique/Publications/Comoretto.pdf.

[26] S. Chu, "Multirate filter designs using comb filters," *IEEE Transactions on Circuits and Systems*, vol. cas-31, no. 11, pp. 913–924, November 1984.

[27] H. Ochi and H. Honma, "Low sensitivity realization of linear-phase FIR digital filters using complex arithmetic," *Circuits and Systems, 1991., IEEE International Sympoisum on*, vol. 1, pp. 172–175, June 1991.

[28] D. M. Rabrenovic and M. D. Lutovac, "Elliptic filters with minimal Q-factors," *Electronics Lettes*, vol. 30, no. 3, pp. 206–207, February 1994.

[29] L. D. Milic and M. D. Lutovac, "Design of elliptic IIR filters with a reduced number of shift-and-add operations in multipliers," *Proc. 3^{rd} IEEE Conf. Electron., Circuits, Syst. ICECS*, pp. 398–401, October 1996.

[30] M. D. Lutovac, D. V. Tosic, and B. L. Evans, "Emqf filter design in matlab," *4th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services*, vol. 1, pp. 125–128, 1999.

[31] L. D. Milic and M. D. Lutovac, "Design of computationally efficient elliptic IIR filters with a reduced number of shift-and-add operations in multipliers," *IEEE, Transactions on Signal Processing*, vol. 45, no. 10, pp. 2422–2430, October 1997.

[32] M. D. Lutovac and L. D. Milic, "Design of multiplierless elliptic IIR filters with a small quantization error," *IEEE, Transactions on Signal Processing*, vol. 47, no. 2, pp. 469–479, February 1999.

[33] S. Lawson, "A new direct design technique for ALP recursive digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 1, pp. 499–512, May 1993.

[34] C.-K. Lu, M. Anderson, and S. Summerfield, "Design of approximately linear-phase allpass-based QMF banks," *Proc. Int. Symposium DSP*, pp. 56–61, July 1996.

[35] A. Antoniou and R. Howell, "Design of phase equalizers for recursive filters," *Communications Computers and Signal Processing, IEEE Pac Rim '93*, vol. 1, pp. 104–107, May 1993.

[36] S. R. Powell and M. Chau, "A technique for realizing linear phase IIR filters," *IEEE Transactions on Signal Processing*, vol. 39, no. 11, pp. 2425–2435, November 1991.

[37] M. D. Lutovac and L. D. Milic, "Approximate linear phase multiplierless IIR halfband filter," *IEEE Transactions on Signal Processing Letters*, vol. 7, no. 3, pp. 52–53, March 2000.

[38] B. C. Moore, "Principal component analysis in linear systems : Controllability, observability and model reduction," *IEEE Transactions on Automatic Control*, vol. 26, no. 1, pp. 17–32, February 1981.

[39] C. Xiao, J. Olivier, and P. Agathoklis, "Design of linear phase IIR filters via weighted least-squares approximation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 6, pp. 3817–3820, 2001.

[40] M. Kunt, *Traitement Numérique des Signaux.* Dunod, 1981.

[41] J. Cooley and J. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics Computation*, vol. 19, pp. 297–301, 1965.

[42] H. V. Sorensen et al, "Real-valued fast fourier transform algorithms," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 6, pp. 849–863, June 1987.

[43] S. Winograd, "On computing the discrete fourier transform," *Proc. Nat. Acad. Sci.*, vol. 73, no. 4, pp. 1005–1006, April 1976.

[44] T. Toivonen, "Number theoretic transform based block motion estimation," Ph.D. dissertation, University of Oulu, Departement of Electrical Engineering, 2002.

[45] M. Narashima and A. Peterson, "On the computation of the discrete cosine transform," *IEEE Trans. Commun*, vol. 26, no. 6, pp. 934–936, June 1978.

[46] C. Diab et al, "A new IDCT-DFT relationship reducing the IDCT computational cost," *IEEE Transactions on Signal Processing*, vol. 50, no. 7, pp. 1681–1684, July 2002.

[47] B. Lee, "A new algorithm for the discrete cosine transform," *Speech, and signal processing*, vol. 32, no. 6, pp. 1243–1245, 1984.

[48] Z. Wang, "Fast algorithms for the discrete VV transform and for the discrete fourier transform," *IEEE Transactions on acoustics, Speech, and Signal Processing*, vol. 32, no. 4, pp. 803–816, August 1984.

[49] ——, "Pruning the fast discrete cosine transform," *IEEE transaction on communications*, vol. 39, no. 5, pp. 640–643, May 1991.

[50] J. Takala et al, "Constant geometry algorithm for discrete cosine transform," *IEEE Transaction on signal processing*, vol. 48, no. 6, pp. 1840–1843, June 2000.

[51] J. A. Nikara et al, "Discrete cosine and sine transforms, regular algorithms and pipeline architectures," *Signal Processing*, vol. 86, no. 2, pp. 230–249, February 2006.

[52] R. Koilpillai and P. Vaidyanathan, "Cosine-modulated FIR filter banks satisfying perfect reconstruction," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 770–783, 1992.

[53] M. F. Mansour, "Canonical transformations of the discrete cosine transform," *Elsevier - Signal Processing*, vol. 87, no. 6, pp. 1355–1362, June 2007.

[54] H. L. Groginsky and G. A. Works, "A pipeline fast fourier transform," *IEEE Transactions on Computers*, vol. 19, no. 11, pp. 1015–1019, November 1970.

[55] R. Brenan and T. Schneider, "A flexible filterbank structure for extensive signal manipulations in digital hearing aids," *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 6, pp. 569–572, May 1998.

[56] A. G. Dempster and M. D. Macleod, "Multiplier blocks and complexity of IIR structures," *Electronics Letters*, vol. 30, no. 22, pp. 1841–1842, October 1994.

[57] ——, "Constant integer multiplication using minimum adders," *IEE Proc.-Circuits Devices Syst.*, vol. 141, no. 5, pp. 407–413, October 1994.

[58] ——, "Use of minimum-adder multiplier blocks in FIR digital filters," *IEEE Transactions on Circuits and Systems II - Analog and Digital Signal Processing*, vol. 42, no. 9, pp. 569–577, September 1995.

[59] L. Gazsi, "Explicit formulas for lattice wave digital filters," *IEEE Transactions on Circuits and Systems*, vol. CAS-32, no. 1, pp. 68–88, January 1985.

[60] V. DeBrunner, *Recursive Digital Filters*, 1999, ch. Originally for the John Wiley Encyclopedia of Electrical and Electronics Engineering.

[61] L. Pernebo and L. M. Silverman, "Model reduction via balanced state space representations," *IEEE Transactions on Automatic Control*, vol. 27, no. 2, pp. 382–387, April 1982.

[62] B. S. Chen, B. W. Chiou, and S. C. Peng, "Minimum sensitivity IIR filter design using principal component approach," *IEE Proceedings-G*, vol. 138, no. 4, pp. 474–482, August 1991.

[63] V. Sreeram and P. Agathoklis, "Design of linear-phase IIR filters via impulse-response gramians," *IEEE Transactions on Signal Processing*, vol. 40, no. 2, pp. 389–394, February 1992.

[64] B. S. Chen, S. C. Peng, and B. W. Chiou, "IIR filter design via optimal hankel-norm approximation," *IEE Proceedings-G*, vol. 139, no. 5, pp. 586–590, October 1992.

[65] D. F. Enns, "Model reduction with balanced realizations : an error bound and a frequency weighted generalization," *Proceedings of 23rd Conference on Decision and Control*, pp. 127–132, December 1984.

[66] S. Holford and P. Agathoklis, "The use of model reduction techniques for designing IIR filters with linear phase in the passband," *IEEE Transactions on Signal Processing*, vol. 44, no. 10, pp. 2396–2404, October 1996.

[67] R. Fletcher, *Practical Methods of Optimization*, N. Y. J. W. . Sons, Ed., 1987.

Bibliographie Personnelle

Conférence

P. Camino et al, "An Economical Class of Digital Filters for Decimation and Interpolation",15th IMEKO TC4 International Symposium on Novelties in Electrical Measurements and Intrumentations, vol. 1, pp. 55-60, September 2007.

Revue Scientifique

R. Escofier et al, "The ALMA Correlator", A&A, vol. 462, pp. 801-810, 2007.

Document ALMA

B. Quertier and P. Camino, "Pre-Ptototype TFB Card Test Report", Technical Report, ALMA Document CORL-60.01.07.10-004-A-REP, 2005.

Annexes

Notions Importantes de Traitement Numérique du Signal

Cette introduction au traitement du signal n'est qu'une énumération des notions essentielles – mathématiques ou spécifiques au traitement numérique – utilisées dans la suite du document, le but n'étant pas de fournir un cours de traitement numérique du signal (TNS).

1 Les outils mathématiques

L'utilisation des techniques du traitement numérique du signal nécessite un certain nombre de connaissances et d'outils mathématiques [40]. Cette partie à pour but de lister et d'éclaircir ces différentes fonctions primordiales au TNS qui seront, dans la suite du document, appliquées à la radioastronomie.
Dans un premier temps, nous abordons les signaux élémentaires numériques permettant d'effectuer différentes opérations de TNS. Parmi eux, nous retiendrons la suite unité ou impulsion de Dirac et l'échelon unité ou échelon de Heaviside.
L'impulsion de Dirac, ou symbole de Kronecker $\delta(k)$ est défini comme suit :

$$\begin{cases} \delta(0) = 1 \\ \delta(n) = 0 \quad \text{pour n} \neq 0 \end{cases}$$

Quant à l'échelon unité il est défini ainsi :

$$\begin{cases} u(n) = 1 \quad \text{pour n} \geq 0 \\ u(n) = 0 \quad \text{pour n} < 0 \end{cases}$$

Une autre fonction remarquable découlant de l'impulsion de Dirac, utile entre autres lors de la discrétisation d'un signal, est le peigne de Dirac (de pas T_e ou période d'échantillonnage) :

$$\delta_{T_e}(t) = T_e \sum_{n=-\infty}^{+\infty} \delta(t - nT_e) \tag{1}$$

Voici les définitions de différentes fonctions mathématiques utilisées en TNS :

La fonction de convolution C'est une fonction symétrique. Elle est définie, sur un nombre d'échantillon N, par :

$$c(n) = x(n) * y(n) = \sum_{k=0}^{N} x(k)y(k-n) \tag{2}$$

La fonction de corrélation La fonction d'inter-corrélation de deux séquence $x(n)$ et $y(n)$ de longueur N est définie par :

$$R_{xy}(k) = \sum_{n=0}^{N} x(n)y(n+k) \tag{3}$$

R_{xx} est la fonction d'auto-corrélation de $x(n)$.

La transformée en z Pour une séquence $x(n)$ composée de N échantillons, elle est définie par :

$$X(z) = \sum_{n=0}^{N} x(n)z^{-n} \tag{4}$$

où $z = e^{j\omega T_e}$ est une variable complexe.

La Transformée de Fourier Discrète (TFD) La TFD sur N points de la séquence $x(n)$ est définie par :

$$X(k) = \sum_{n=0}^{N-1} x[n]e^{-2j\pi n\frac{k}{N}} \tag{5}$$

avec $k \in \mathbb{N}$ variant de $-\frac{N}{2}$ à $\frac{N}{2} - 1$.
Il est à noter que la TFD de la fontion d'auto-corrélation d'une sequence $x(n)$ est égale à sa Densité Spectrale de Puissance (DSP).

2 Les signaux radioastronomiques

On peut classer les signaux selon qu'ils sont réels ou complexes et en tenant compte de leur caractère prévisible ou non, on parlera alors de signaux déterministes ou aléatoires. Les signaux émis par les sources astronomiques sont de nature aléatoire. Plus exactement, ils sont constitués d'un bruit blanc gaussien véhiculant l'information utile. La notion de bruit blanc gaussien est équivalente à celle d'un processus alétoire, caractérisé par une moyenne (ou espérance mathématique E) nulle et un écart type σ ($\sigma = \sqrt{(Var)}$, Var étant la variance). La notion de blancheur implique que tous les échantillons sont décorrélés ($R_{BB}(k) = \sigma^2\delta_k$).
En radioastronomie, les phénomènes électromagnétiques captés par les antennes considérés comme des bruits pour l'ingénieur des télécommunications sont donc des signaux d'un grand intérêt. En effet, dans beaucoup de disciplines, ces bruits caractérisent tous phénomènes perturbateurs gênant la perception ou l'interprétation d'un signal.

3 Numérisation du signal

Afin d'être traités par des systèmes numériques, ces signaux reçus par les antennes doivent être numérisés. La Figure 1 présente les différentes étapes de cette numérisation.
La première étape consiste à appliquer au signal un filtre anti-repliement permettant de « borner » le signal à échnatilloner. En pratique aucun spectre, n'est rigoureusement borné et il y a donc toujours repliement après échantillonnage. Lorsqu'un recouvrement a lieu dans une zone spectrale donnée, l'information contenue dans cette zone est irrémédiablement altérée. Ce filtre permet de maintenir la bande de signal utile à convertir à une fréquence d'échantillonnage f_e à une largeur inférieure ou égale à $f_e/2$ (théorème de Shannon).

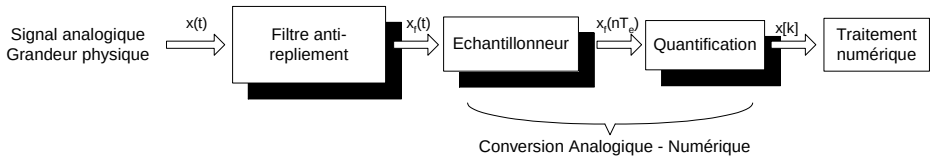


FIG. 1 – *Numérisation d'un signal*

L'échantillonnage est l'opération qui discrétise le signal à traiter. Soit un signal analogique noté $x(t)$ à spectre borné (i.e. le module de sa tranformée de Fourier $X(f)$ est nul pour $f > f_e/2$). l'échantillonnage de ce signal est opéré en lui appliquant un peigne de Dirac de pas T_e :

$$x_s(t) = x(t) \cdot \sum_{k=-\infty}^{+\infty} \delta(t - kT_e) \tag{6}$$

La Transformée de Fourier du signal $x_s(t)$ est alors une reproduction du spectre $X(f)$ à chaque multiple de la fréquence f_e :

$$X_s(f) = X(f) * \frac{1}{T_e} \sum_{n=-\infty}^{+\infty} \delta\left(f - \frac{n}{T_e}\right) = \frac{1}{T_e} \sum_{n=-\infty}^{+\infty} X\left(f - \frac{n}{T_e}\right) \tag{7}$$

Pour $f = \nu/T_e$, soit $\nu = f/f_e$ la fréquence normalisée, on a $X(\nu) = X_s(f)$. Le résultat de l'échantillonnage dans le domaine des fréquences $(X(\nu))$ est exposé en Figure 2.

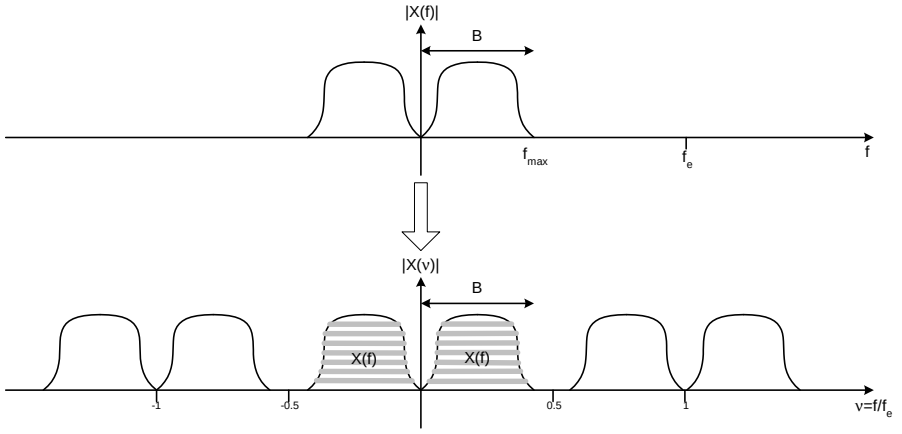


FIG. 2 – *Spectre du signal échantillonné*

On remarquera, qu'ici aucun recouvrement des motifs n'a lieu, $f_{max} < f_e/2$.

La dernière étape est la quantification du signal. Cette opération a pour objectif de représenter un signal par une échelle de valeurs discrètes représentées en électronique par un mot binaire de n bits. Le signal échantillonné est en fait comparé à des seuils référence V_{si} qui permettent de définir l'appartenance d'un échantillon à l'un des niveaux de codage. Le nombre de niveaux de codage, et donc la précision avec laquelle le signal est converti, résultent du nombre de bits n avec lequel les signaux sont encodés. L'imprécision qui résulte de cette opération est appelée erreur de quantification ou encore bruit de quantification.

4 Filtre numérique

Après l'étape de numérisation du signal, le traitement du signal peut être effectué. En radioastronomie, l'étape de traitement qui suit et la détection du signal. Elle est effectuée par un Corrélateur qui, dans le cas ALMA, est composé d'un système de filtrage numérique suivi d'une opération de corrélation. Afin d'obtenir les informations spectrales du signal étudié, une transformée de Fourier du résultat de la corrélation est ensuite réalisée.

Le filtre numérique est un système linéaire invariant. sa fonction de transfert est donnée par :

$$\frac{Y(z)}{X(z)} = \frac{\sum_{i=0}^{p} a_i z^{-i}}{1 + \sum_{j=1}^{q} b_j z^{-j}} = H(z) \tag{8}$$

où a_i sont les coefficients (ou poids) du numérateur et b_i ceux du numérateur. La fonction de transfert peut aussi s'écrire sous la forme suivante :

$$H(z) = \frac{\prod(z - z_i)}{\prod(z - p_j)} \tag{9}$$

qui fait apparaître ainsi les zéros (z_i) et pôles (p_i) de cette dernière.
La synthèse d'un filtre numérique consiste à déterminer une expression de $H(z)$ tel que $H(\nu)$ satisfasse un gabarit ($z = e^{j\omega T_e} = e^{j2\pi\nu T_e}$). Le plus souvent, ce gabarit définit une réponse en amplitude comme illustré sur la Figure 3.

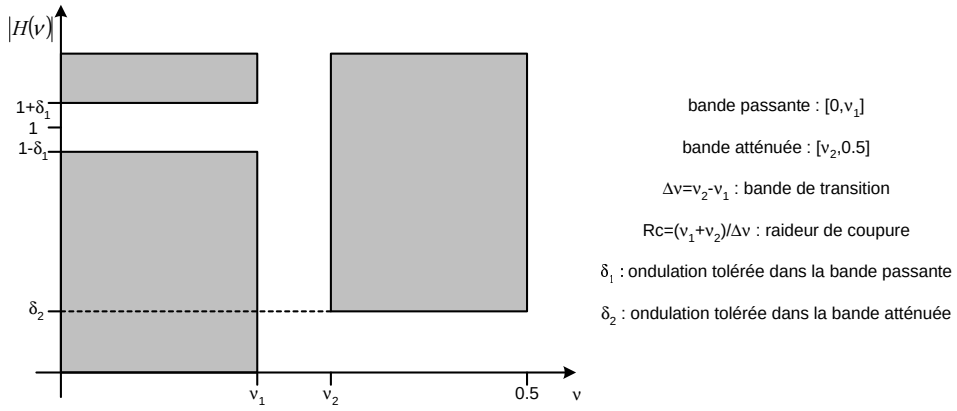


FIG. 3 – *Gabarit de filtre passe-bas pour la synthèse*

On utilise habituellement des grandeurs définies en décibels pour établir le gabarit :

- $A_p = 20\log\left(\frac{1+\delta_1}{1-\delta_1}\right)$ amplitude des ondulations dans la bande passante
- $A_a = 20\log(\delta_2)$ amplitude des ondulations dans la bande atténuée
- $A = 20\log\left(\frac{1-\delta_1}{\delta_2}\right)$ gain

où δ_1 et δ_2 sont définis en Figure 3.
La synthèse du filtre est réalisée, a partir de ces spécifications, à l'aide de différentes méthodes qui ne sont pas exposées ici.
Deux catégories de filtres peuvent être obtenus à partir de (8).

4.1 Les filtre RIF

Un filtre RIF (Réponse Impulsionnelle Finie) de réponse impulsionnelle $h(k)$ de longueur N est caractérisé de la manière suivante :

$$y(k) = \sum_{i=0}^{N-1} a_i x(k-i) = \sum_{i=0}^{N-1} h(i)x(k-i) \tag{10}$$

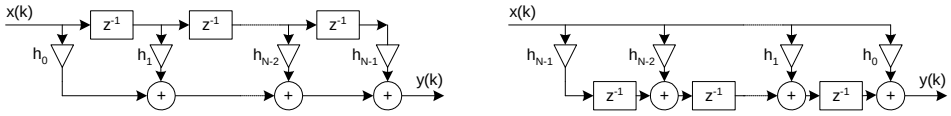
Il est à noter que les filtres RIF sont toujours stables du fait de leur réponse impulsionnelle. En effet $\sum_{i=0}^{N-1} h(i) < \infty$.
Parmi les propriétés spécifiques des filtres RIF, celle de la phase linéaire est particulièrement intéressante dans le domaine de la Radioastronomie. Définissons le retard de groupe, appelé aussi temps de propagation de groupe :

$$\tau_g = -\frac{d\phi(\omega)}{d\omega} \tag{11}$$

ϕ étant la phase de la fonction de transfert et ω la pulsation ($\omega = 2\pi f$).
Ce retard de groupe est représentatif du temps qu'une composante fréquentielle du signal met pour traverser le système. Lorsque $\phi(\omega)$ est linéaire, τ_g induit par le système est alors constant. La condition de phase linéaire est obtenue en definissant une réponse impulsionnelle symétrique (de longueur N) :

$$h(n) = \pm h(N-1-n) \tag{12}$$

En figure 4 sont présentées deux structures électroniques classiques permettant l'implémentation d'un filtre RIF.



(a) Forme directe (b) Forme directe transposée

FIG. 4 – *Structures électroniques*

4.2 Les filtres RII

Un filtre RII (Réponse Impulsionnelle Infinie) est caractérisé par une structure récursive :

$$y(k) = \sum_{j=1}^{M-1} b_j y(k-j) + \sum_{i=0}^{N-1} a_i x(k-i) \tag{13}$$

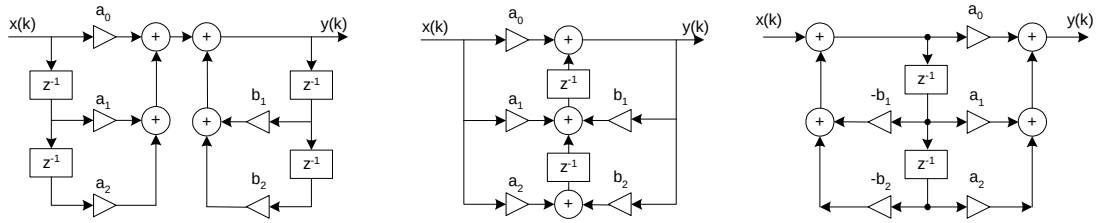
Contrairement au filtre RIF, les filtres RII ne sont pas à phase linéaire. Cependant différentes techniques mathématiques permettent l'obtention d'une phase linéaire dans la bande d'intérêt du filtre.
Dans le cas d'un filtre RII, la stabilité dépend de la position des pôles dans le plan z vis à vis du cercle unité). Pour garantir la stabilité d'un tel filtre, il est nécessaire que ces pôles soient

strictement situés à l'intérieur du cercle unité.

Il est usuel, dans le cas des filtres RII, de représenter la fonction de transfert sous une forme cascadée de cellule SOS (Second Order Section) définie par la relation suivante :

$$H(z) = \prod_{k=1}^{L} H_k(z) = \prod_{k=1}^{L} \frac{a_{0k} + a_{1k}z^{-1} + a_{2k}z^{-2}}{1 + b_{1k}z^{-1} + b_{2k}z^{-2}} \tag{14}$$

Elle résulte d'une factorisation de (8). Contrairement aux formes cascadées résultant de (14), les structures provenant de l'équation 8 sont dites à forme directe. L'opération de quantification des coefficients permettant l'implémentation est décisive quant au choix de la structure à adopter. En effet les différentes formes directes pouvant être obtenues sont particulièrement sensibles à cette quantification. La position des pôles et des zéros de la fonction de transfert dépend de tous les coefficients quantifiés. A performance équivalente, le nombre de bits nécessaire au codage des coefficients de la structure cascade (des coefficients de chaque SOS) est moins important que celui nécessaire au codage de ceux de la structure directe. En effet un coefficient a_0 de la structure directe codé sur n bits a pour équivalent le produit des k coefficients a_{0k} des SOS codés sur m bits, avec $m = \sqrt[k]{n}$. De cette façon, la structure SOS obtenue est plus stable en nécessitant moins de précision. Dans le cas d'un filtre numérique, les structures de réalisation correspondent aux schémas synoptiques de la Figure 5 (ici illustrées pour l'implémentation en SOS).



(a) Forme 1 (b) Forme 1 transposée (c) Forme 2

FIG. 5 – *Structures électroniques*

Le choix d'une de ces structures de réalisation va seulement être guidé par la simplicité matérielle ou logicielle de leur implémentation : le nombre d'opérateurs retard et d'additionneurs principalement.

Table des Polynômes Primitifs

n	XNOR from	n	XNOR from	n	XNOR from	n	XNOR from
3	3,2	45	45,44,42,41	87	87,74	129	129,124
4	4,3	46	46,45,26,25	88	88,87,17,16	130	130,127
5	5,3	47	47,42	89	89,51	131	131,130,84,83
6	6,5	48	48,47,21,20	90	90,89,72,71	132	132,103
7	7,6	49	49,40	91	91,90,8,7	133	133,132,82,81
8	8,6,5,4	50	50,49,24,23	92	92,91,80,79	134	134,77
9	9,5	51	51,50,36,35	93	93,91	135	135,124
10	10,7	52	52,49	94	94,73	136	136,135,11,10
11	11,9	53	53,52,38,37	95	95,84	137	137,116
12	12,6,4,1	54	54,53,18,17	96	96,94,49,47	138	138,137,131,130
13	13,4,3,1	55	55,31	97	97,91	139	139,136,134,131
14	14,5,3,1	56	56,55,35,34	98	98,87	140	140,111
15	15,14	57	57,50	99	99,97,54,52	141	141,140,110,109
16	16,15,13,4	58	58,39	100	100,63	142	142,121
17	17,14	59	59,58,38,37	101	101,100,95,94	143	143,142,123,122
18	18,11	60	60,59	102	102,101,36,35	144	144,143,75,74
19	19,6,2,1	61	61,60,46,45	103	103,94	145	145,93
20	20,17	62	62,61,6,5	104	104,103,94,93	146	146,145,87,86
21	21,19	63	63,62	105	105,89	147	147,146,110,109
22	22,21	64	64,63,61,60	106	106,91	148	148,121
23	23,18	65	65,47	107	107,105,44,42	149	149,148,40,39
24	24,23,22,17	66	66,65,57,56	108	108,77	150	150,97
25	25,22	67	67,66,58,57	109	109,108,103,102	151	151,148
26	26,6,2,1	68	68,59	110	110,109,98,97	152	152,151,87,86
27	27,5,2,1	69	69,67,42,40	111	111,101	153	153,152
28	28,25	70	70,69,55,54	112	112,110,69,67	154	154,152,27,25
29	29,27	71	71,65	113	113,104	155	155,154,124,123
30	30,6,4,1	72	72,66,25,19	114	114,113,33,32	156	156,155,41,40
31	31,28	73	73,48	115	115,114,101,100	157	157,156,131,130
32	32,22,2,1	74	74,73,59,58	116	116,115,46,45	158	158,157,132,131
33	33,20	75	75,74,65,64	117	117,115,99,97	159	159,128
34	34,27,2,1	76	76,75,41,40	118	118,85	160	160,159,142,141
35	35,33	77	77,76,47,46	119	119,111	161	161,143
36	36,25	78	78,77,59,58	120	120,113,9,2	162	162,161,75,74
37	37,5,4,3,2,1	79	79,70	121	121,103	163	163,162,104,103
38	38,6,5,1	80	80,79,43,42	122	122,121,63,62	164	164,163,151,150
39	39,35	81	81,77	123	123,121	165	165,164,135,134
40	40,38,21,19	82	82,79,47,44	124	124,87	166	166,165,128,127
41	41,38	83	83,82,38,37	125	125,124,18,17	167	167,161
42	42,41,20,19	84	84,71	126	126,125,90,89	168	168,166,153,151
43	43,42,38,37	85	85,84,58,57	127	127,126		
44	44,43,18,17	86	86,85,74,73	128	128,126,101,99		

Bloc Multiplieur

En [56], l'idée d'utiliser une même cellule de base (opération *shift-add*) pour les différents coefficients composant un filtre est abordée. Cette méthode est intéressante dans le cas où le filtre doit être implémenté sans aucun multiplieur, dans un souci de minimisation de la consommation de ce dernier par exemple.
Différents algorithmes sont employés permettant de synthétiser des blocs composés d'un ou plusieurs coefficients [57, 58]. Le terme utilisé pour désigner ce regroupement de coefficients est *bloc multiplieur*. Ces méthodes ne sont applicables que dans le cas de l'utilisation des structures cascade et parallèle. En Figure 1 apparait l'utilisation de ces blocs multiplieur dans une structure cascade.

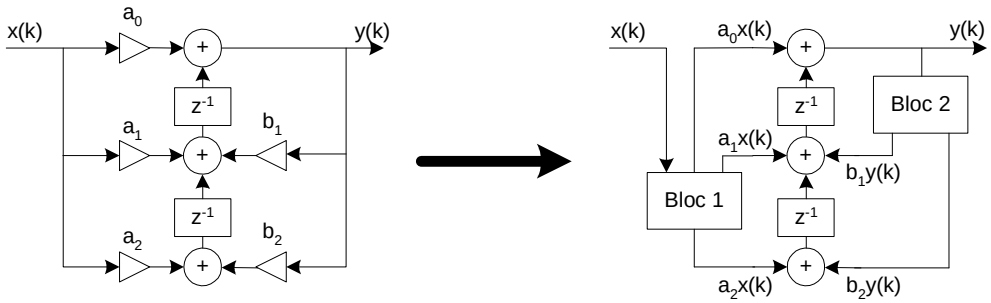


FIG. 1 – *Exemple d'utilisation du bloc multiplieur, forme canonique transposée*

La structure transposée, qui nécessite cependant deux fois plus d'additionneurs, permet de regrouper tous les coefficients et ainsi de réaliser l'implémentation à l'aide d'un seul bloc multiplieur (Figure 2).

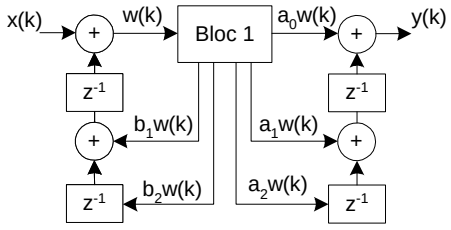


FIG. 2 – *Forme transposée, un seul bloc multiplieur*

Les méthodes sont basées sur l'utilisation de graphes induisant l'optimisation du dit bloc multiplieur, donc du nombre d'additionneurs utilisés pour l'implémentations des coefficients.

En Figure 3 est présentée le graphe représentant le contenu d'un bloc permettant de synthétiser 3 coefficients sans multiplication.

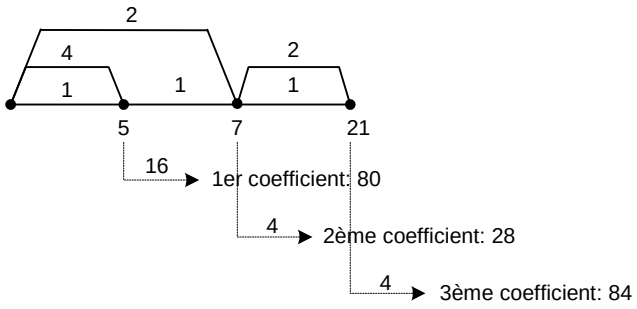


FIG. 3 – *Arbre des coefficients*

Chaque noeud du graphe représente une addition et chaque chiffre assigné aux lignes une opération de décalage. De cette manière les « fondamentaux » du graphe sont obtenus (valeurs aux différents noeuds). C'est à partir de ces derniers, qui sont à valeurs impaires, que les différents coefficients, à valeurs égales aux fondamentaux ou alors multiples de deux, sont déduits. De cette manière seulement 3 additions sont nécessaires à la réalisation de 3 valeurs de coefficients.
Deux algorithmes ont été mis en place par Dempster permettant la génération d'un graphe synthétisant un coefficient (algorithme MAG [57]) et d'un graphe synthétisant un ensemble de coefficients (algorithme RAG-n [58]), tout ceci à un coût optimal.

Filtre *Lattice*

Les filtres « lattice » (structure en treillis) constituent une alternative courante à la structure allpass présentée en section 2.2 du Chapitre 5. Ces filtres présentent les mêmes avantages concernant le nombre de coefficients nécessaire à l'implémentation de la structure (N coeffficients pour un ordre N) et les caractéristiques de filtrage obtenues.
Pour un filtre d'ordre N « all-pole » (ou filtre Auto-Regressif - AR), « allpass » ou « all-zero » (ou filtre RIF : filtre Moving-Average - MA) décrit par les coefficients $a(n), n \in [1, N+1]$, il existe N coefficients lattice $k(n)$ correspondant qui constituent la structure lattice. Les paramètres $k(n)$ sont appelés les coefficients de réflection du filtre.
Pour un filtre lattice RII (allpass ou all-pole), la structure lattice correspondante est celle décrite en Figure 1.

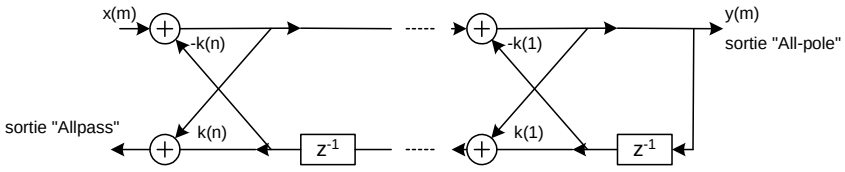


FIG. 1 – *Structure lattice*

L'implémentation d'un filtre RII classique (filtre ARMA) sous forme lattice, composé d'un numérateur $b(n)$ et d'un dénominateur $a(n)$, est présentée en Figure 2.

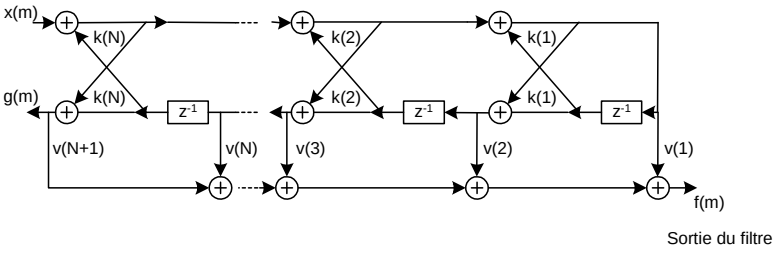


FIG. 2 – *Structure lattice d'un filtre RII*

Cette structure est appelée « lattice/ladder » et est constituée des coefficients lattice $k(n)$, représentant le dénominateur b, et des coefficients ladder $v(n)$, représentant le numérateur a. Les coefficients obtenus peuvent être encodés sur un plus faible nombre de bits qu'une structure classique du fait de la faible sensibilité qui caractérise cette structure.

De cette structure de filtre est dérivée une famille de filtres nommée *Wave Digital Filter* - WDF [59, 60].Ces structures WDF permettent d'obtenir des architectures à faible délai et

des filtres à faible sensibilité (de type passe-bas ou passe-haut). Une structure WDF-lattice est constituée de 2 voies parallèles contenant une cascade de filtres allpass de 2^{nd} ordre avec un filtre allpass d'ordre 1, dans le cas d'un filtre d'ordre impair. Les équations régissant les 2 voies parallèles sont les suivantes, A_1 étant d'ordre M impair et A_2 d'ordre N pair.

$$A_1(z) = \frac{-\gamma_0 + z^{-1}}{1 - \gamma_0 z^{-1}} \prod_{l=1}^{m} \frac{-\gamma_{2l-1} + \gamma_{2l}(\gamma_{2l-1} - 1)z^{-1} + z^{-2}}{1 + \gamma_{2l}(\gamma_{2l-1} - 1)z^{-1} - \gamma_{2l-1} z{-2}} \quad (1)$$

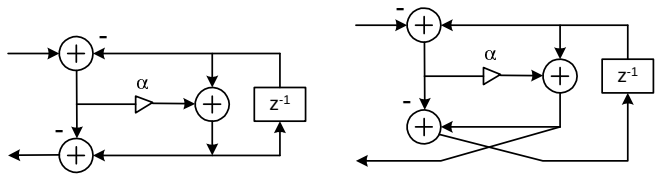
$$A_2(z) = \prod_{l=m+1}^{m+n} \frac{-\gamma_{2l-1} + \gamma_{2l}(\gamma_{2l-1} - 1)z^{-1} + z^{-2}}{1 + \gamma_{2l}(\gamma_{2l-1} - 1)z^{-1} - \gamma_{2l-1} z{-2}} \quad (2)$$

où $m = (M-1)/2$ et $n = N/2$. La valeur des coefficients est calculée de la manière suivante, semblable au calcul des coefficients de la première structure allpass (3.15) :

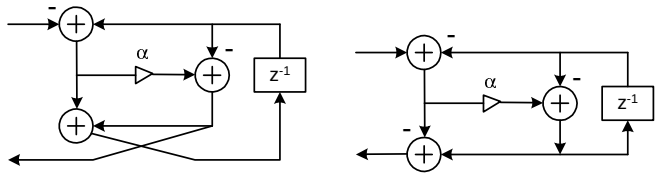
$$\gamma_0 = r_0 \quad (3)$$

$$\gamma_{2l-1} = -r_i^2,\ \gamma_{2l} = \frac{2r_l cos(\theta_l)}{1 + r_l^2} \text{ for } l = 1, 2, \ldots, m+n \quad (4)$$

avec $z = r_0$ pôle réel de A_1 et $z = r_l exp(\pm j\theta_1)$ les m et n pôles complexes conjugués de A_1 et A_2.
Cette structure est basée sur le même principe d'assemblage de cellules de faible ordre que la structure *two-path allpass* en (3.7)). Les structures WDF-lattice de premier ordre sont réalisées comme illustré en Figure 3.



(a) $\alpha = 1 - \gamma$ pour $\gamma \in]1/2, 1[$ (b) $\alpha = \gamma$ pour $\gamma \in]0, 1/2[$



(c) $\alpha = -\gamma$ pour $\gamma \in]-1/2, 0[$ (d) $\alpha = 1 + \gamma$ pour $\gamma \in]-1, -1/2[$

FIG. 3 – *Structures électronique d'une cellule allpass lattice-WDF d'ordre* 1

Ainsi, le coefficient α à implémenter est toujours compris entre 0 et 0.5. Concernant la sensibilité de la caractéristique en amplitude de la structure, le comportement est identique à la structure *two-path allpass* précédente.
En Figure 4 est représentée une structure lattice d'ordre 2 (cf. (2)) dérivé des Figures 3.

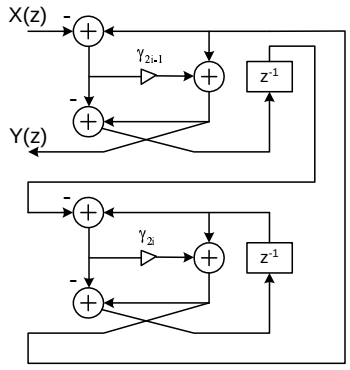


FIG. 4 – *Structure lattice d'ordre* 2, $\gamma \in]0, 1/2[$

Présentation des Algotithmes de Réduction

Les notions essentielles communes à tous les algorithmes qui vont être utilisés sont dans un permier temps présentées. ces prérequis ne sont que brièvement introduits, le but étant seulement d'exposer ces algorithmes dans les sections suivantes, pas d'en expliquer les moindres détails.

1 Généralités

Les gramians de controlabilité et d'observabilité sont utilisés afin de déterminer le modèle réduit. Le système (3.37) est assumé stable, controlable et observable, ce qui signifie que les gramians[5] sont non-singuliers. Ils sont donnés par :

$$\begin{aligned} W_c &= \sum_{k=0}^{\infty} A^k b b^T (A^T)^k \\ W_o &= \sum_{k=0}^{\infty} (A^T)^k c^T c A^k \end{aligned} \tag{1}$$

W_c étant le gramian de controlabilité et W_o celui d'observabilité [61].
Il peut être montré [38] qu'il est possible de trouver une matrice de transformation T qui rend les deux gramians égaux à une matrice diagonale positive Σ.

$$W_c = W_O = \Sigma \tag{2}$$

La représentation « espace d'état » du système résultante est alors dite *« balancée »* sur l'intervalle $[0, \infty]$. Les gramians qui sont tous deux égaux à Σ sont donnés par la solution unique aux équations de Lyapunov :

$$\begin{aligned} A\Sigma A^T - \Sigma &= -bb^T \\ A^T \Sigma A - \Sigma &= -c^T c \end{aligned} \tag{3}$$

Soit le système « balancé » partionné suivant :

$$\begin{aligned} \begin{pmatrix} x_1(n+1) \\ x_2(n+1) \end{pmatrix} &= \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} u(n) \\ y(n) &= \begin{pmatrix} c_1 & c_2 \end{pmatrix} \begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} + d\ u(n) \end{aligned} \tag{4}$$

[5] En théorie des système, le gramian G est une matrice à valeur réelle utilisé pour déterminer l'indépendance linéaire de fonctions. Ici, ils permettent d'évaluer si les paires de fonctions (W_o, W_c) sont controlables et-ou observables. Elles le sont si et seulement si G est non-singulier

En le placant dans le système d'équations 3 et en découpant la matrice Σ de la sorte $\Sigma = diag\,(\Sigma_1, \Sigma_2)$ avec $\sigma_{min}(\Sigma_1) > \sigma_{max}(\Sigma_2)$, on obtient alors le sous-système (A_{11}, b_1, c_1, d) controlable et observable [61] mais non balancé (caractéristique nécessaire à la stabilité du modèle). Certaines méthodes ont été établies afin de pallier ce problème. Ce sont ces dernières qui sont utilisées dans les algorithmes présentés dans les sections suivantes, aboutissant à un modèle réduit stable.
Suivent donc, 4 algorithmes basés sur différentes méthodes amenant à l'obtention du modèle réduit. Un algorithme d'optimisation est enfin présenté permettant d'améliorer les résultats des précédents.

2 Réduction du modèle par décomposition en élément singulier

Supposons un filtre RIF d'ordre m et de réponse impulsionnelle h dont la fonction de stransfert est la suivante :

$$H(z) = \sum_{i=0}^{m} h_i z^{-i} \tag{5}$$

Il va être utilisé comme base à la réalisation du filtre RII à phase linéaire et à sensibilité minimale. Soit $\phi(H)$ la matrice de Hankel associée à H(z) :

$$\phi(H) = \begin{pmatrix} h_1 & h_2 & \dots & & h_m \\ h_2 & h_3 & \dots & h_m & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ h_m & 0 & \dots & & 0 \end{pmatrix} \tag{6}$$

La décomposition en valeur singulière de cette dernière est la suivante :

$$\phi(H) = U\Sigma V^T \tag{7}$$

où Σ est la matrice diagonale comportant les valeurs singulières non-nulles de $\phi(H)$

$$\Sigma = diag(\sigma_1, \sigma_2, \dots, \sigma_m) \text{ avec } \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m \tag{8}$$

U et V sont les matrices des vecteurs singuliers et sont unitaires (i.e. $U^T U = I$). En [62], une méthode permettant obtenir une réalisation balancée du filtre $H(z)$ (5) est donnée. Elle aboutit aux équations suivantes :

$$\begin{aligned} A &= (U\Sigma^{1/2})^{-1}(U\Sigma^{1/2})^{\uparrow} \\ b &= \text{première colonne de } \Sigma^{1/2}V^T \\ c &= \text{première ligne de } U\Sigma^{1/2} \\ d &= 0 \end{aligned} \tag{9}$$

où $X^{\uparrow}$ correspond à la matrice X décalée d'une ligne vers le haut et complétée par des 0. Le système (A, b, c) est alors complètement stable, controlable, observable et balancé.
La deuxième étape est la réduction de l'ordre du modèle. Après application de la décomposition présentée en (4), une troncature du modèle obtenue (i.e. (A_{11}, b_1, c_1)) n'aboutit pas à une réalisation balancée. Quelques manipulations sont nécessaires à l'obtention de cette dernière. L'ordre optimal r du modèle réduit est obtenu pour [62] :

$$||\phi(H) - \phi_r(H)|| \leq \left[\sum_{i=r+1}^{m} \sigma_i^2\right]^{1/2} \text{ : norme-Frobenius minimale} \tag{10}$$

avec $\phi_r(H) = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ (u et v étant les vecteurs singuliers). La somme de droite, elle, représente l'erreur acceptable à atteindre.
En découpant les matrices (A, b, c) en sous-matrices $A_{11} \in \mathbb{R}^{r*r}$, $A_{12} \in \mathbb{R}^{r*(m-r)}$, $A_{21} \in \mathbb{R}^{(m-r)*r}$, $A_{22} \in \mathbb{R}^{(m-r)*(m-r)}$, $c_1 \in \mathbb{R}^{1*r}$, $c_2 \in \mathbb{R}^{1*(m-r)}$, $b_1 \in \mathbb{R}^{r*1}$ et $b_2 \in \mathbb{R}^{(m-r)*1}$ et en appliquant les équations suivantes, on obtient le modèle reduit (de taille r) balancé, controlable et observable (A', b', c', d'), de faible sensibilité aux variations des coefficients.

$$\begin{aligned} A' &= A_{11} + A_{12}(I - A_{22})^{-1} A_{21} \\ b' &= b_1 + A_{12}(I - A_{22})^{-1} b_2 \\ c' &= c_1 + c_2(I - A_{22})^{-1} A_{21} \\ d' &= c_2(I - A_{22})^{-1} b2 \end{aligned} \tag{11}$$

3 Gramian de la réponse impulsionnelle

En considérant le système d'équation 3.37, le gramian de la réponse impulsionnelle du dit système est défini comme suit [63] :

$$P = \sum_{k=0}^{\infty} \begin{pmatrix} h_{k+1}^2 & \dots & h_{k+1}h_{k+n} \\ \vdots & \vdots & \vdots \\ h_{k+1}h_{k+n} & \dots & h_{k+n}^2 \end{pmatrix} \tag{12}$$

avec h la réponse impulsionnelle du système.
La méthode permettant d'obtenir le modèle réduit s'applique en trois étapes. La première consiste à calculer P en tant que solution à l'équation Lyapunov suivante :

$$P - \hat{A}^T P \hat{A} = \hat{c}^T \hat{c} \tag{13}$$

avec $(\hat{A}, \hat{b}, \hat{c}, \hat{d})$ le système originel sous sa forme de réalisation canonique controlable.
Dans un deuxième temps, il est nécessaire de calculer la matrice orthogonale L qui diagonalise P :

$$L^T P L = \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_n \end{pmatrix} \tag{14}$$

avec toujours la même relation $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$ (σ_i étant les valeurs propres de P).
Enfin, la dernière étape, qui transforme et permet de partionner les matrices SS (State Space, ou espace d'état), est présentée ci-dessous.

$$\begin{aligned} A_d &= L^{-1}\hat{A}L = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \\ b_d &= L^{-1}\hat{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ c_d &= \hat{c}L = \begin{pmatrix} c_1 & c_2 \end{pmatrix} \end{aligned} \tag{15}$$

On obtient le modèle d'ordre réduit (A_r, b_r, c_r, d_r) en appliquant $A_r = A_{11}$, $b_r = b_1$, $c_r = c_1$ et $d_r = \hat{d}$. Pour trouver l'ordre optimal du modèle réduit, il suffit de faire décroite ce dernier tant

que les spécifications d'atténuation et de phase sont respectées.

La première étape est de définir un filtre RIF. Ce dernier peut être représenté sous la forme canonique controlable afin d'appliquer la méthode décrite précédemment.

$$\begin{aligned} \hat{A} &= \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \\ \hat{b} &= \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}^T \\ \hat{c} &= \begin{pmatrix} h_1 & h_2 & \dots & h_N \end{pmatrix} \\ \hat{d} &= h_0 \end{aligned} \tag{16}$$

Ce filtre, défini par le système d'équations 16, possède quelques propriétés intéressantes qui simplifient grandement le calcul du modèle réduit (deux de ces proriétés sont données ici) :

$$\begin{aligned} W_c &= I \\ W_0 &= P = H^T H \end{aligned} \tag{17}$$

4 Norme de Hankel

Cette méthode est basée sur l'approximation de la norme Hankel, c'est à dire trouver une matrice de Hankel $\hat{\Gamma}$ de rang r tel que la norme-L^2 $||\phi - \hat{\phi}||$ soit minimisée, avec ϕ la matrice de Hankel du système original. L'algorithme [64] permettant de trouver l'approximation d'un filtre RIF a comme point d'entrée la réalisation balancée du système (9). Cependant, dans le cas présent la matrice diagonale Σ (7) est ordonnée différemment :

$$\Sigma = diag(\sigma_1, \sigma_2, \dots, \sigma_r, \sigma_{r+2}, \dots, \sigma_m, \sigma_{r+1}) \tag{18}$$

toujours avec $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq \sigma_{r+1} \geq \sigma_{r+2} \geq \dots \geq \sigma_m > 0$, r étant l'ordre du modèle réduit et m celui du modèle original.
En s'inspirant de la décomposition présentée en (4), avec cette fois-ci $\Sigma = diag(\Sigma_1, \sigma_{r+1})$ et donc $A_{11} \in \mathbb{R}^{(m-1)*(m-1)}$, on obtient les composants A_H, b_H, c_H :

$$\begin{aligned} A_H &= \left\{ \left[A_{11} - b_1 \left(b_2\right)^{-1} A_{21} \right]^{-1} \right\}^T \\ B_H &= \left\{ - \left(b_2\right)^{-1} A_{21} \left[A_{11} - b_1 \left(b_2\right)^{-1} A_{21} \right]^{-1} \right\}^T \\ C_H &= - \left(c_2^T\right)^{-1} A_{12}^T \Gamma \left\{ \left[A_{11} - b_1 \left(b_2\right)^{-1} A_{21} \right]^{-1} \right\}^T \end{aligned} \tag{19}$$

avec $\Gamma = \Sigma_1^2 - \sigma_{r+1}^2 I_{n-1}$.
Le modèle réduit d'ordre r est obtenu en ne gardant que la partie stable du système (A_H, b_H, c_H, d). En d'autres termes, l'approximation de la norme-Hankel est obtenue ainsi :

$$\hat{H(z)} = \left[c_H (zI - A_H)^{-1} b_H \right]_+ + d \tag{20}$$

où $[H(z)]_+$ représente la partie stable de $H(z)$ possédant tous ses pôles à l'intérieur du cercle unité.
L'erreur entre la réponse du filtre obtenue après réduction et celle de l'original est égale à σ_{r+1}. Si l'ordre du filtre RII à obtenir est donné, on peut tout de suite connaitre l'erreur engendrée ε par l'approximation. Dans le cas contraire, si l'erreur ε maximale est définie avant la réalisation de la réduction du modèle, on peut choisir l'ordre du filtre en conséquence avec la relation $\sigma_{r+1} < |\varepsilon|$.

5 Réduction du modèle par pondération en fréquence

Pour obtenir le filtre RII de plus faible ordre respectant les spécifications du filtre RIF original, il est nécessaire de maintenir l'erreur très faible pour toutes les fréquences et en particulier pour celles de la bande de transition du filtre. La méthode de pondération des fréquences [65] a pour but la diminution de l'erreur induite à certaines fréquences en cascadant un filtre pondérateur au système originel. En fixant la bande passante du filtre pondérateur à l'endroit de la bande de transition du filtre originel, on maintient à de faibles valeurs, lors de la réduction du modèle, les erreurs dans cette bande de transition. En [66], une méthode utilisant cette technique est présentée.
On va donc pondérer un système original (RIF à phase linéaire) dans sa forme balancé (A, b, c) avec un filtre (A_i, b_i, c_i) aussi sous sa forme balancé. Le système à réduire est maintenant le suivant :

$$\tilde{A} = \begin{pmatrix} A & bc_i \\ 0 & A_i \end{pmatrix} \quad \tilde{b} = \begin{pmatrix} 0 \\ b_i \end{pmatrix} \quad \tilde{c} = \begin{pmatrix} c & 0 \end{pmatrix} \tag{21}$$

Un algorithme aboutissant à la réduction de ce modèle est donné en [66].

6 Approximation des moindres carrés pondérée

Considérons un système régi par les équations 3.37 qui est balancé, observable et controlable d'ordre m. L'approximation « weighted least square » de $H(z)$ (3.38) consiste à trouver pour un ordre r donné ($r < m$), un système d'ordre réduit $((A_r, b_r, c_r, d)$, avec $A_r \in \mathbb{R}^{r*r}$, $b_r \in \mathbb{R}^{r*1}$, $c_r \in \mathbb{R}^{1*r}$) de fonction de transfert $H_r(z)$ qui minimise la fonction suivante [39] :

$$J_{ew} = || \left(H_r(z) - H(z)\right) H_w(z)||^2 \tag{22}$$

où $H_w(z)$ est la fonction de pondération (cf. section 5) décrite par le système (A_w, b_w, c_w) de taille $A_w \in \mathbb{R}^{k*k}$, $b_w \in \mathbb{R}^{k*1}$, $c_w \in \mathbb{R}^{1*k}$.
Soit $H_{ew}(z)$ la fonction d'erreur pondérée définie par $H_{ew}(z) = \left(H_r(z) - H(z)\right) H_w(z)$. Elle peut être exprimée de la sorte :

$$H_{ew} = c_{ew}(zI - A_{ew})^{-1} b_{ew} \tag{23}$$

avec

$$A_{ew} = \begin{pmatrix} A_r & 0 & b_r c_w \end{pmatrix} \quad b_{ew} = \begin{pmatrix} b_r d_w \\ b d_w \\ b_w \end{pmatrix} \tag{24}$$

$$c_{ew} = \begin{pmatrix} c_r & -c & 0 \end{pmatrix}$$

Un algorithme récursif permettant le calcul du modèle réduit optimal est présenté en [39]. La minimisation de la function J_{ew} en (22) est réalisée par l'intermédiaire de l'algorithme *BFGS* (Broyden-Fletcher-Goldfarb-Shanno [67]).