Architectures Reconfigurables Dynamiquement pour les Systèmes Embarqués

Gilles SASSATELLI[1], Lionel TORRES[1], Gaston CAMBON[1], Jérôme GALY[1]

[1]LIRMM, UMR 5506, Université Montpellier II 161 Rue ADA 34392 Montpellier CEDEX 5

Tel. (33)(0)4-67-41-85-85 Fax (33)(0)4-67-41-85-00

sassate@lirmm.fr

Résumé – L'avènement des réseaux de troisième génération, tels l'UMTS (Universal Mobile Telecommunication System) autorisera pour la première fois sur un réseau de téléphonie numérique mobile des bandes passantes importantes (jusqu'à 2 Mbit/s). Des services nouveaux, comme la diffusion de musique qualité CD, de vidéo ou encore la visioconférence deviendront envisageables. Les solutions architecturales actuelles retenues par les concepteurs de terminaux mobiles ne seront alors plus à même de répondre aux contraintes de plus en plus fortes, car aux impératifs actuels de coût, consommation viendront s'ajouter ceux de puissance de traitement. Une architecture dynamiquement reconfigurable vouée à répondre à ces contraintes est proposée et détaillée; ainsi que des résultats comparatifs sur un algorithme de visioconférence implémenté sur différentes architectures est également exposé.

1 Introduction

Les terminaux mobiles de demain seront résolument tournés vers l'Internet. Ils ne seront plus seulement des téléphones, mais auront des fonctionnalités évoluées que l'on considère aujourd'hui être l'exclusivité des PDA (Personal Data Assistant), voire des ordinateurs de bureau.

La technologie de commutation actuelle de type circuit, utilisée dans les réseaux mobiles de seconde génération (GSM) évoluera progressivement vers le mode paquet, et plus précisément l'IP (Internet Protocol). Son efficacité en terme de partage de ressource physique, jointe aux importantes bandes passantes (2 Mbit/s) disponibles sur les systèmes de troisième génération (UMTS) autorisera l'avènement de services nouveaux, comme le transfert de fichiers multimédia (MP3, MPEGx…), la commande à distance ou encore la visioconférence.

2 Les Systèmes sur puce

Les densités d'intégration actuelles autorisent la réalisation de véritables systèmes sur puce (SoC : System on Chip). Plusieurs blocs (on parle de coeurs) d'origine diverses sont alors intégrés sur le même substrat. On peut distinguer trois axes principaux de recherches basés sur cette approche.

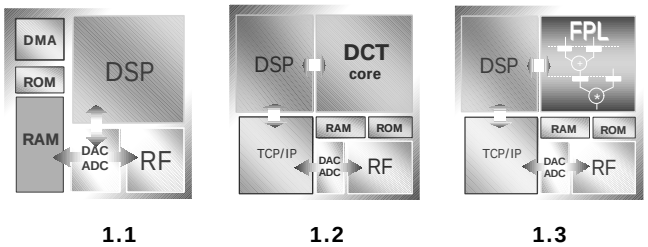

FIG. 1 : Les trois approches

2.1 L'approche logicielle

C'est la plus commune, c'est celle retenue dans nombre de téléphones cellulaires actuels. L'ensemble des fonctionnalités (hors radio et canal) sont prises en charge par le microprocesseur ou le DSP intégré (Figure 1.1). Mais les exigences en termes de puissance de calcul deviennent telles qu'il sera difficilement possible d'intégrer des microprocesseurs toujours plus gros, pour d'évidentes raisons de coût et de consommation.

2.2 L'approche dédiée

Celle-ci consiste à identifier le futur champ d'application visé, et à utiliser un cœur dédié (Figure 1.2) au traitement des parties communes à toutes les application, idéalement considéré comme chronophage (partie la plus exigeante en temps de traitement). Par exemple, si l'on identifie un champ d'application orienté autour de l'image et de la vidéo, notamment des protocoles JPEG et MPEG, on choisira d'intégrer sur silicium un coeur de IDCT (Inverse Direct Cosinus Transform), cette opération commune étant connue comme la partie critique de ces deux algorithmes. Cette approche est intéressante, mais n'autorise certes pas la souplesse de l'approche logicielle ; les images codées au format JPEG2000 utilisant la transformée en ondelettes par exemple ne pourront pas profiter du coeur IDCT ; l'éventail des applications envisageables est donc restreint.

2.3 L'approche reconfigurable

Ici, la solution consiste à intégrer un cœur constitué d'un réseau reconfigurable (Figure 1.3), de type FPGA par exemple [1].

L'utilisation d'applications faisant un usage intensif de JPEG conduira à synthétiser un coeur d'IDCT cablé, mais aussi des parties dépendantes de l'application, comme le codage de Huffman, ou la quantification. A l'inverse, une application

orientée autour du MPEG nous autorisera à conserver le cœur IDCT, mais il sera également possible de synthétiser l'estimation de mouvement (Motion Estimation), également connue comme chronophage.

3 Architecture du 'Systolic Ring'

3.1 Principe de l'approche

Un œil attentif sur les applications qui seront disponibles sur les réseaux de demain révèle une caractéristique importante : Ces applications sont à dominante multimédia, c'est à dire qu'elles sont orientées vers le traitement de flots de données, peu de contrôle (opérations conditionnelles) n'est nécessaire sur les flots de données à traiter, l'impératif essentiel étant un volume important d'opérations arithmétiques à réaliser en temps réel. Les FPGA [1] classiques, de par leur architecture, sont adaptés au traitement de données au niveau bit (granularité fine), de ce fait l'implémentation d'applications multimédia orientées donnée implique la synthèse d'opérateurs arithmétiques (multiplieurs, additionneurs), exigeants en terme de place occupée, et peu performants de par leur architecture fortement combinatoire.

Notre approche consiste à implémenter un réseau reconfigurable à granularité déplacée [3]. Au lieu d'utiliser des blocs configurables (type CLB :Configurable Logic Block) aptes à faire des manipulations au niveau bit, nous définissons (Figure 2) pour le 'Systolic Ring' une cellule de base appelée Dnode (Data node), apte à effectuer un éventail d'opérations principalement arithmétiques [3].

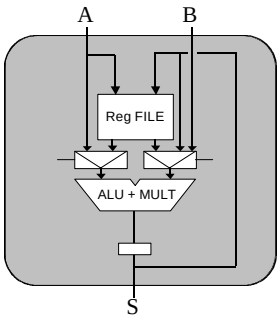

FIG. 2 : L'architecture d'un Dnode

A l'image d'un FPGA, chaque Dnode possède un certain nombre de bits de configuration qui définissent sa fonctionnalité à un instant donné. L'ensemble des configurations de tous les Dnodes du réseau reconfigurable sont stockées dans le plan mémoire, qui est une RAM. Il est donc possible, si l'on est capable de modifier le contenu de cette mémoire en cours de traitement, de changer la fonctionnalité de tout ou partie du réseau : C'est le principe de reconfiguration dynamique. Nous utilisons à cet effet un contrôleur de configuration, qui est un processeur type RISC (Figure 3) avec un jeu d'instructions adapté, qui aura pour tache principale de gérer l'évolution de la fonctionnalité du réseau. Cette architecture n'est pas vouée à prendre en charge toutes les fonctionnalités du système, mais se conçoit plutôt dans une approche système sur puce comme un cœur dédié à la prise en charge des parties chronophages des applications, délestant ainsi le processeur central.

D'un point de vue fonctionnel, on distingue deux phases :

- Dans une première phase le processeur central, qui répond à une sollicitation extérieure charge l'application visée. Celle-ci, préalablement écrite pour une exécution mixte, comporte une partie de son code qui est destiné au contrôleur de configuration du réseau reconfigurable; le processeur central envoie donc directement cette portion de code objet dans la mémoire programme de celui-ci.
- Une fois le code objet voué à la gestion dynamique de la configuration du réseau reconfigurable chargé, celui-ci s'exécute sur le RISC, le processeur central peut alors commencer à envoyer les données à traiter vers notre architecture, et récupérer celles-ci en fin de traitement.

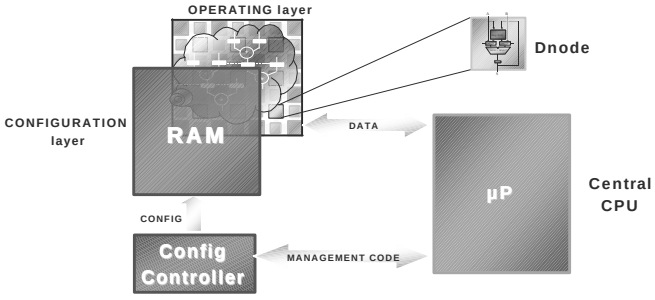

FIG. 3 : L'architecture globale

3.2 Architecture de la couche opérative

La couche de traitement, ou couche opérative à une architecture particulière. Nous adoptons ici une disposition particulière facilitant le transfert des donnée pour les applications orientées flots de données. Les Dnodes sont organisés en couches, chaque couche étant reliée avec la précédente et la suivante par le biais d'un composant spécialisé : le Switch (Figure 4). Il se charge d'aiguiller les données, et est également configurable dynamiquement. L'ensemble de la structure prends la forme d'un anneau, car rebouclée sur elle-même, le flot de données est « tournant ».

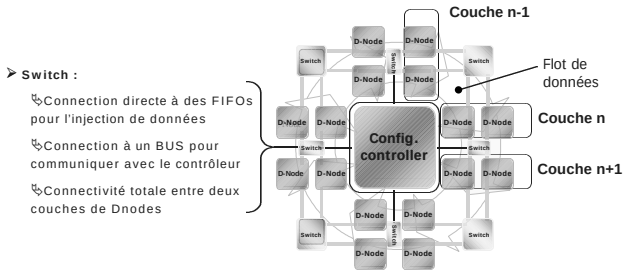

FIG. 4 : L'architecture en anneau

Ce système, si il facilite le transit des flots de données, par son architecture ne solutionne cependant pas le problème classique rencontré dans nombre de flots de données : Ceux-ci comportent de manière quasi systématique des rétropropagations de données, comme illustré figure 5. Ces impératifs de traitement sont à l'origine des problèmes de

routage des données rencontrés usuellement dans les architectures reconfigurables classiques. L'avènement de technologie d'intégration toujours plus performantes autorise la création de réseau de capacités toujours plus grande, où ce problème de routage devient incontournable.

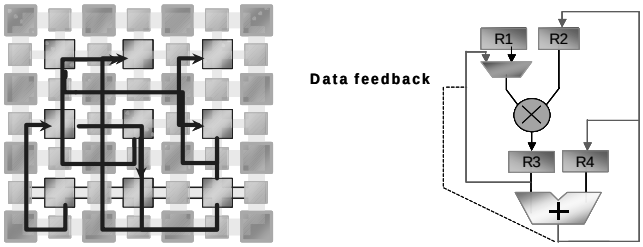

FIG. 5 : Le problème de 'data feedback'

La solution retenue pour répondre à ces problèmes consiste en l'emploi d'un second chemin de donnée, allant à contresens du premier. C'est ce que l'on appelle le réseau de rétropropagation ; chaque switch possède son propre pipeline dans lequel il écrit de manière systématique les données traitées n provenance de la couche de Dnodes amont. Tous les autres switchs de l'architecture ont un point de lecture sur ce pipeline, et ont donc la possibilité de récupérer les données issues de ce stage (Figure 6).

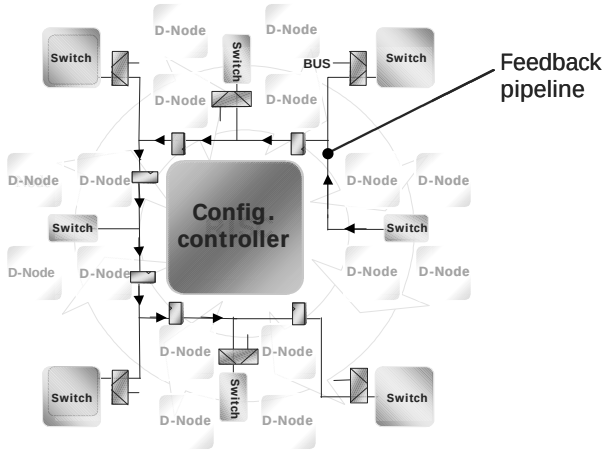

FIG. 6 : Le réseau de rétropropagation

4 Résultats

L'implantation d'un algorithme d'estimation de mouvement (Motion Estimation), dont font un usage intensif nombre d'applications video [2], et notamment de visioconférence (H.261) donne des résultats intéressants. Une version à 16 Dnodes de notre structure est comparée en termes de nombres de cycles à deux implantations du même algorithme sur ASIC [2], et sur Pentium MMX.

TAB. 1: Comparaison de performances

Systolic Ring	ASIC[2]	MMX[4]
3757	*581*	*28900*

Notre architecture, si elle est moins performante que l'ASIC, l'est quand –même bien plus que l'implémentation MMX, en conservant un avantage majeur : celui de la flexibilité. Il est à noter qu'une version à 64 Dnodes obtiendrait des performances comparables à l'ASIC.

Une version à 8 Dnodes, décrite en VHDL à été entièrement validée en simulation, et prototypée avec succès sur une plate-forme à base de FPGAs à fortes capacités. La bande passante de cette version appelée Ring8 est évaluée à 3 Go/s à la fréquence de fonctionnement typique de 200 MHz.

Cette même architecture à également été synthétisée sur technologie ST 0.18µm et 0.25µm, et à donné les résultats exposés tableau 2.

TAB. 2: Evaluations post synthèse

	0.25µm	**0.18µm**
Dnode	0.06 mm^2	0.04 mm^2
Coeur	0.9 mm^2	0.7 mm^2
Fréquence	180 MHz	200 MHz

Dans un contexte de système sur puce, cette architecture pourrait prendre place aux cotés d'un cœur de processeur éprouvé du commerce comme l'ARM7, l'ensemble constituerait une solution de choix, disposant des capacités de traitement optimisées du Systolic Ring jointe à l'utilisation d'un cœur de processeur apte à faire tourner un grand éventail de systèmes d'exploitation et d'applications ; le tout à un coût raisonnable.

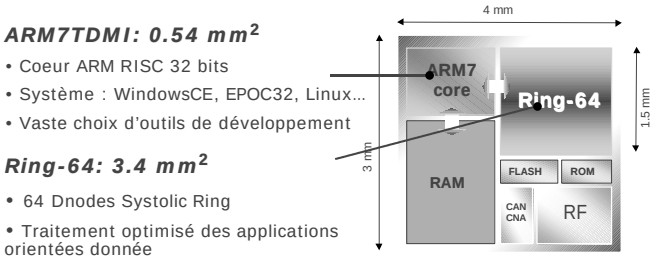

FIG. 7 : Une architecture envisageable de SoC

5 Conclusion

Nous avons ici exposé une alternative aux architectures traditionnelles pour l'embarqué. En lieu et place des solutions classiques consistant à utiliser un unique cœur de processeur ou DSP pour gérer l'ensemble des fonctionnalités au niveau numérique, nous proposons une approche coprocesseur avec un cœur dédié au traitement des algorithmes chronophages. A chacun de ceux-ci correspond un programme s'exécutant sur le contrôleur de configuration, et gérant dynamiquement la configuration de notre réseau. Dans un contexte applicatif orienté téléphonie de troisième génération, on peut imaginer que le terminal intégrant cette architecture soit évolutif, ne se limitant donc pas à la bibliothèque de programmes initiale, et, consécutivement au lancement d'une nouvelle application

(visioconférence par exemple) se connecte directement à un centre serveur, et télécharge le code objet correspondant. Notre réseau, dynamiquement reconfigurable dédiée aux application orientées donnée prouve donc son efficacité non seulement en termes de performances, mais aussi en termes de coût : Une version à 8 Dnodes de notre architecture à été synthétisée en technologie ST 0.18µm, avec une surface sur Silicium de l'ordre de 0.7 mm^2 ; ce qui la situe clairement dans un contexte de système sur puce, destiné aux applications fortement contraintes telles que celles pour l'embarqué.

Références

[1] Stephen Brown and J. Rose, "Architecture of FPGAs and CPLDs: A Tutorial," IEEE Design and Test of Computers, Vol. 13, No. 2, pp. 42-57, 1996

[2] A.Bugeja and W. Yang, "A Re-configurable VLSI Coprocessing System for the Block Matching Algorithm", IEEE Trans. On VLSI systems, vol. 5, September 1997.

[3] W. H. Mangione-Smith et al, "Seeking Solutions in Configurable Computing," IEEE Computer, pp. 38-43, December 1997

[4] Intel Application Notes for Pentium MMX, http://developer.intel.com/