Minimisation du temps total de traitement pour optimiser le fonctionnement d'un contrôleur de buffers pour les systèmes de vision embarquée

Khadija Hadj Salem[1], Yann Kieffer[1], Stéphane Mancini[2]

[1] LCIS - EA 3747, 50 rue Barthélémy de Laffemas BP54, 26902 VALENCE Cedex 09, France
{khadija.hadj-salem, yann.kieffer}@lcis.grenoble-inp.fr

[2] TIMA - UMR N°515 , 46 avenue Félix Viallet, 38031 GRENOBLE Cedex, France
stephane.mancini@imag.fr

Mots-clés : *Accès mémoires, Ordonnancement, Temps total de traitement*

1 Contexte applicatif : l'atelier MMOpt

La gestion des accès mémoires a un impact significatif sur la performance et la consommation d'énergie des systèmes de vision embarquée en particulier la conception des circuits de traitement d'image. Cet article traite la problématique de l'optimisation des accès mémoires dans une nouvelle génération de système de cache, dénommé *Memory Management Optimisation* (MMOpt), qui a été proposé par Mancini et al. [2]. La stratégie du MMOpt consiste à décomposer les données d'entrée/sortie d'un traitement de type "noyau", ainsi que les espaces d'itération en "tuiles", puis d'ordonnancer les transferts de données et le séquencement des calculs pour optimiser le système.

MMOpt produit une architecture générique, dénommée *Tile Processing Unit (TPU)* . Cette dernière se compose d'une unité de traitement qui implémente le noyau, d'un contrôleur de buffers où sont stockées des tuiles d'entrées et d'une unité de préchargement qui charge des tuiles d'entrée depuis la mémoire centrale. Ces unités fonctionnent en parallèle de façon à recouvrir les calculs et les préchargements. Le séquencement des calculs, des transferts depuis la mémoire centrale et le placement des données en mémoire locale sont calculés par MMOpt à partir de la matrice d'incidence "tuiles à calculer/tuiles requises". Les premiers travaux sur MMOpt [2] ont proposé une méthode d'optimisation qui s'avère intéressante, mais l'analyse fine des résultats montre qu'il est encore possible d'optimiser certaines métriques.

2 Problème de minimisation du temps total de traitement

L'optimisation du fonctionnement des TPUs générés par l'outil MMOpt engendre une problématique d'optimisation riche. Nous la modélisons comme un problème d'ordonnancement multi-objectif, dénommé *3-objective Process Scheduling and Data Prefetching Problem* (3-PSDPP), qui considère l'optimisation de trois critères : le nombre total de préchargements N qui représente le trafic depuis la mémoire centrale et la consommation d'énergie du circuit produit, la quantité de buffers Z qui représente en grande partie la surface du circuit (encombrement et coût de production) et le temps total de traitement Δ (performance et/ou contrainte temps-réel).

Pour les deux premiers critères, des méthodes de résolution ont été proposées [1]. Par ailleurs, le troisième critère sera abordé dans cette étude. Nous considérons alors le problème *Minimum Completion Time of 3-PSDPP* (MCT-PSDPP), qui consiste à déterminer un séquencement des préchargements de tuiles de l'image d'entrée et leurs emplacements dans les mémoires locales

(buffers), et un séquencement des calculs des tuiles de sortie de manière que Δ soit minimal (voir Figure 1).

Le problème MCT-PSDPP peut donc être formulé ainsi : on considère un jeu $\mathcal{X}$ de X tuiles d'entrée qui doivent être pré-chargées de la mémoire externe vers des buffers, un ensemble $\mathcal{Y}$ de Y tuiles de sortie à calculer, une durée α de préchargement et une durée β de calcul. Le calcul d'une tuile de sortie nécessite en général plusieurs tuiles de l'image d'entrée noté $\mathcal{R}_y$, où $\mathcal{R}_y \subseteq \mathcal{X}, \forall\ y \in \mathcal{Y}$. Il est impératif que ces tuiles soient accessibles dans les buffers internes du système au moment de ce calcul. Un modèle mathématique, formalisant les données d'entrée, les variables à déterminer, les contraintes et l'objectif à minimiser, a été alors proposé afin de caractériser le problème.

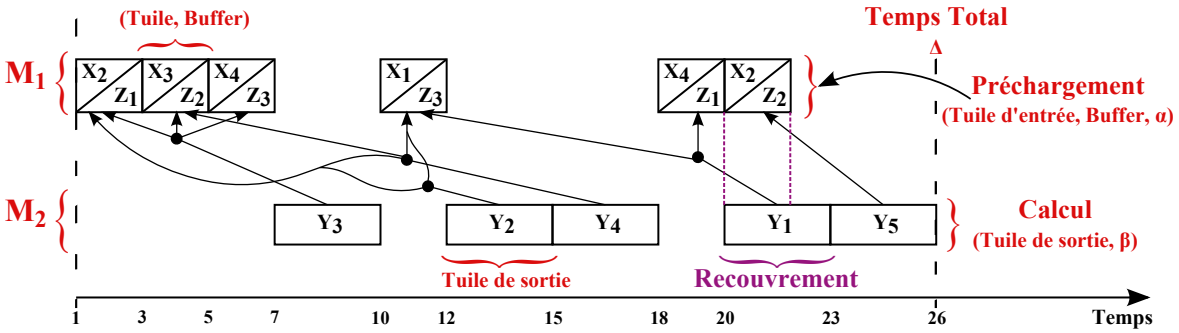

FIG. 1 – Séquencement type (Préchargements & Calculs) du TPU

3 Complexité et résolution du problème MCT-PSDPP

A notre connaissance, le problème d'ordonnancement MCT-PSDPP n'a pas été étudié avant dans la littérature de recherche opérationnelle. Par ailleurs, l'analyse de complexité de certains cas particuliers du problème a été établie afin de prouver la NP-complétude de ce problème. En effet, de nouvelles classes d'instances, liées au choix de α par rapport à β, ainsi qu'au choix de Z (fixé en entrée ou à déterminer), ont été identifiées. De plus, deux bornes inférieures ont été développées.

Des méthodes exactes et approchées sont alors proposées pour aborder la résolution du MCT-PSDPP. En effet, dans le cas où Z est fixé comme donnée d'entrée, un algorithme basé sur la notion de "Classe C", qui regroupe l'ensemble de tuiles de sortie où le nombre total de tuiles d'entrée requises ne dépasse pas le nombre de buffers Z, a été considéré. En revanche, lorsqu'on doit également déterminer Z, un autre algorithme basé sur la notion du "Groupe G", qui définit l'ensemble de tuiles de sortie pouvant être calculées successivement en réutilisant les tuiles requises par la première tuile calculée, a été proposé. En outre, le problème a été modélisé en programme linéaire en nombres entiers (PLNE), où nous considérons plusieurs versions.

Des expérimentations numériques, associées aux différentes méthodes proposées, ont été effectuées sur un ensemble d'instances réelles générées par Mancini et al. [2].

Références

[1] K. Hadj Salem, Y. Kieffer, and M. Mancini. Formulation and practical solution for the optimization of memory accesses in embedded vision systems. In *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, Gdańsk, Poland, September 11-14, 2016*, pages 609–617, 2016.

[2] S. Mancini and F. Rousseau. Enhancing non-linear kernels by an optimized memory hierarchy in a high level synthesis flow. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 1130–1133. EDA Consortium, 2012.