Parallélisation de la classification spectrale

Dans le cadre de la segmentation d'images couleurs ou hyperspectrales, nous avons affaire à un grand flot de données dont il faut extraire des classes sans connaissance a priori.
Les méthodes non supervisées sont alors privilégiées et notamment une des plus connues, la classification spectrale. Une stratégie de parallélisation de la méthode basée sur une décomposition en sous-domaines a été définie. Elle s'avère très efficace pour traiter des problèmes de segmentations d'images par exemple. Des résultats que nous allons présenter à la conférence PACBB'12 (http ://www.pacbb.net) montrent une très bonne scalabilité de notre stratégie parallèle.

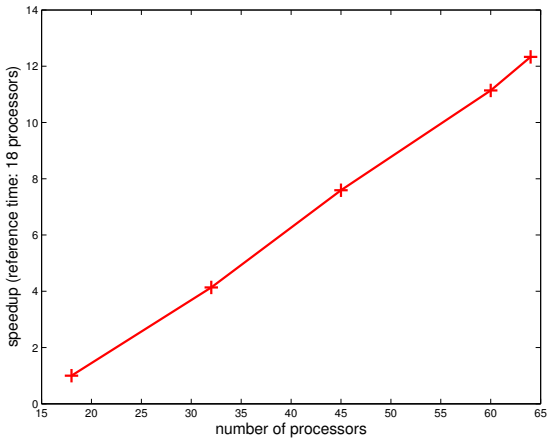

FIGURE 1 – Speedup en prenant comme référence le temps avec 18 processeurs

Cependant il reste une limitation liée à la capacité mémoire de stockage des processus ; ce qui explique d'ailleurs pourquoi le temps de référence est pris avec 18 processeurs dans la figure 1 : l'image choisie comporte trop de pixels pour être répartie sur moins de 18 processeurs.

En effet, le stockage de la matrice affinité pleine, même si cette matrice est construite pour les données de chacun des sous-domaines et non plus pour l'ensemble des données de l'image complète, limite la méthode pour le traitement de grand volume de données.

Une étude préliminaire de seuillage de la matrice affinité a été menée : l'idée est de ne garder que certaines composantes de cette matrice suivant un certain critère. Nous avons montré, sur des données simples, que nous pouvons nous contenter de ne garder qu'un petit nombre de composantes sans perdre au niveau de la qualité de la solution (ie le nombre de clusters détectés). Les résultats de cette étude préliminaire ont été soumis à la conférence VECPAR'12 (http ://nkl.cc.u-tokyo.ac.jp/VECPAR2012/)

La matrice d'affinité n'est donc plus considérée comme pleine mais stockée sous forme creuse : nous gagnons ainsi en espace de stockage et nous sommes

susceptibles de traiter de plus grands ensembles de données. Cela nous permet aussi d'utiliser des codes de calcul adaptés à ces structures creuses.

Pour valider cette étude de seuillage de matrice et l'adaptation des codes parallèles aux structures creuses, nous souhaiterions disposer d'un accès à des super-calculateurs disposant d'importantes capacités mémoires.

Cela permettrait de :

- comparer avec les codes pleins sur des problèmes de grandes tailles (segmentation d'images hyperspectrales...), problèmes qui ne passent pas avec nos codes pleins sur les ordinateurs auxquels nous avons accès actuellement et vérifier que notre approche avec seuillage donne des résultats conformes ;
- expérimenter sur des problèmes encore plus gros (images 3D) qui ne passeront qu'avec des codes creux.