

Analyse de performance et optimisation de code

Pierre AYOUB – Océane FLAMANT

8 février 2019



**INSTITUT DES SCIENCES ET
TECHNIQUES DES YVELINES**

Résumé

La simulation numérique est un procédé informatique visant à modéliser un phénomène par ordinateur, s'agissant le plus souvent d'un phénomène physique. Cette modélisation prend forme par des systèmes d'équations décrivant l'état du système physique représenté à chaque instant. De nombreux domaines scientifiques convergent vers la simulation informatique, tel que certaines branches de la physique, de l'analyse et de l'optimisation mathématique, ou encore le calcul haute performance en informatique. Enfin, la simulation trouve naturellement de nombreuses applications concernant des sujets variés, tel que la simulation du climat et des événements météorologiques, la simulation d'essais nucléaires, de l'effet d'un médicament sur un corps, ou encore des astres et de l'univers. Ce rapport s'articulera donc autour de la simulation de fumée, phénomène impliquant les lois de la mécanique des fluides. Notre travail portera sur l'aspect du calcul haute performance de cette simulation.

Table des matières

1	Introduction	4
2	Analyse du code	5
3	Protocole expérimental	7
3.1	Théorie	7
3.2	Pratique	7
4	Optimisations et mesures	9
4.1	Mise en place de la compilation	11
4.2	Inlining	14
4.3	Mots-clés appropriés	15
4.4	Modification des calculs	16
4.4.1	Multiplication	16
4.4.2	Division	17
4.4.3	Invariants	18
4.4.4	Résultat	18
4.5	Mémoire cache	20
4.6	Déroulage de boucle	22
4.7	Vectorisation et parallélisation	22
5	Conclusion	24

1 Introduction

Le projet que nous vous présentons aujourd'hui consiste à analyser puis, grâce à nos mesures, optimiser un code de simulation numérique. Ce dernier nous offre une interface graphique permettant d'ajouter de la fumée dans un espace confiné et, ainsi, d'en observer le comportement. Nous pouvons influencer la quantité de fumée et sa vitesse dans l'espace. De plus, l'application nous donne le contrôle sur la résolution de la simulation, cela revient à dire sur sa précision, qui détermine principalement la performance du programme.

Le déroulement du projet s'est effectué en plusieurs étapes distinctes :

Analyse du code Cette phase consiste à analyser le programme d'un point de vue mathématique et informatique. De cette première approche, il s'agira de comprendre les opérations du programme sur les équations qui régissent le système physique. De l'autre approche, il convient d'étudier l'architecture logicielle de l'application, ainsi que les choix mis en œuvre afin d'implémenter le ou les algorithmes nécessaires.

Protocole expérimental Une fois l'analyse effectuée, nous pouvons en déduire le moyen le plus adapté afin de mesurer les performances de notre implémentation. Nous allons donc mettre en avant les critères théoriques à atteindre dans nos mesures, puis nous exposerons la manière dont nous avons mis ceci en pratique.

Optimisations et mesures Grâce au protocole mis en place, nous pouvons quantifier la performance du programme. De ce fait, nous serons en mesure d'expérimenter différentes techniques d'optimisation sur le programme et d'en calculer l'accélération.

2 Analyse du code

L'analyse du code est la première étape à effectuer afin d'optimiser un code. Il s'agit d'identifier les différentes sections et fonctions qui pourraient être critiques, ainsi que de comprendre mathématiquement quelles fonctions remplissent quels rôles.

L'interface du programme et l'interaction avec l'utilisateur est géré par les fichiers « demo.html », « FluidSolver.java » et « WebStart.java ». À priori, ces fichiers nous importe peu dans notre processus d'analyse et d'optimisation. Le cœur de la simulation se déroule dans le fichier « fluid.c », dans lequel se situe les fonctions de calcul des phénomènes physique.

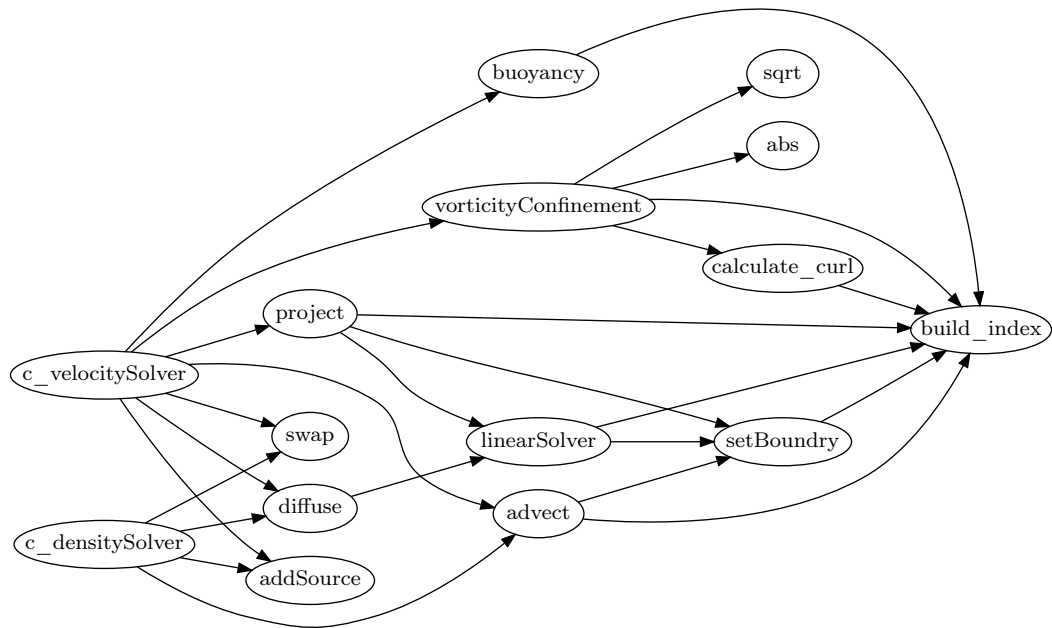


FIGURE 1 – Graphe d'appel du fichier *fluid.c*

Avant d'observer et d'analyser attentivement le code du fichier « fluid.c », nous avons utilisé **Cflow** afin d'obtenir un graphe d'appel pour avoir une vue d'ensemble et observer les liens entre les fonctions. Le résultat est présenté dans la figure 1. Nous décrivons ici les fonctions principales. On remarque très rapidement les deux points d'entrées du programme : « *c_densitySolver()* » et « *c_velocitySolver()* ». Ces deux fonctions sont utilisées conjointement pour assurer les deux fonctionnalités de notre simulation : la première permet de calculer la position d'une particule de fumée ajouté par la souris, la deuxième permet de calculer leur déplacement en fonction des différentes vitesses dans l'espace. Nous observons ensuite deux fonctions qui sont ici pour assurer des opérations physique : « *project()* » et « *diffuse()* ». Le nom laisse sous entendre que la première permet de faire une projection (des coordonnées d'un système à un autre ?) et la seconde permet de diffuser (une particule dans l'espace ?), mais nos connaissances en mécanique des fluides nous arrêtent ici pour les suppositions. Cependant, il est intéressant d'observer que ces deux fonctions font appel à une nouvelle fonction : « *linearSolver()* ». Cette fonction, comme le nom le laisse entendre, est le cœur de calcul de l'application : elle permet de résoudre

un système d'équation linéaire. Il ne fait aucune doute que c'est dans cette fonction que de nombreuses optimisations seront possibles ! Enfin, on observe quelques autres fonctions de physique (« `vorticityConfinement()` », « `advect()` », « `buoyancy()` ») ou des fonctions de mathématiques (« `sqrt()` », « `abs()` ») qui gravitent autour des fonctions précédemment décrites. Enfin, une fonction semble être appelée par quasiment toutes les autres de manière récursive : la fonction « `build_index()` ». Cette fonction, très courte, permet de passer d'un couple de coordonnées (i, j) identifiant une donnée à un offset correspondant à son emplacement mémoire par rapport à la base de la matrice. Cette fonction, présente à de nombreux endroits dans le code, devrait aussi avoir droit à une attention particulière.

3 Protocole expérimental

La mise en place d'un protocole expérimental de mesure est une étape nécessaire et cruciale dans toute optimisation de code. D'une part, le but de ce protocole est de mettre en lumière les points chauds du programme, c'est-à-dire les parties du code qui ralentissent considérablement l'exécution de la simulation. Ces points chauds seront les cibles de nos optimisations. D'autre part, après chaque tentative d'optimisation, le protocole doit nous permettre de mesurer l'impact de cette dernière, qu'il soit positif ou négatif, et enfin de le quantifier.

3.1 Théorie

Le protocole se doit d'être le plus représentatif possible d'une utilisation typique du programme. Pour cela, on ne doit pas uniquement mesurer la simulation exécutant une fonctionnalité particulière, mais l'ensemble des fonctionnalités. Il est bon de noter que ceci est vrai dans le cas d'un unique protocole, comme ici : en effet, on pourrait imaginer plusieurs protocoles différents pour optimiser chaque fonctionnalité indépendamment des autres. Ici, le programme nous offre deux fonctionnalités : ajouter de la fumée dans un espace confiné et ajouter une vitesse (c'est-à-dire un vecteur vitesse) modifiant ainsi la trajectoire de la fumée. Notre protocole doit donc faire intervenir ces deux possibilités dans notre expérience.

Lors de nos expériences, il ne faut pas oublier que le hasard ou l'aléa des mesures peuvent biaiser un résultat. Afin d'éviter cela, il faut donc utiliser une valeur moyenne ou une valeur médiane.

3.2 Pratique

Pour effectuer nos mesures, il nous faut utiliser un outil de d'analyse de performance. Plusieurs possibilités s'offrent à nous, parmi lesquels nous pouvons notamment citer : **Performance Counters for Linux (perf)**, **OProfile**, **Gprof**, **MAQAO**, **Callgrind/Kcachegrind**, et bien d'autre. Parmi ces choix-là, nous avons choisis d'utiliser **perf**. Nous pouvons lui citer plusieurs avantages qui nous ont intéressés : intégration en ligne de commande, outil bas-niveau, développement stable, accès aux compteurs hardware. Nous pouvons donc en déduire que c'est un outil tout à fait adapté à l'analyse de code bas-niveau (C, assembleur).

Le programme de simulation utilise une interface graphique pour rentrer en interaction avec l'utilisateur. De ce fait, nous ne pouvons pas lui donner un jeu de données défini à l'avance qui serait constant entre les différentes mesures. Nous devons alors définir des actions à réaliser de la manière la plus similaire possible entre les exécutions. De manière « arbitraire », nous définissons un run de la manière suivante : une fois l'applet lancé, nous traçons une ligne de fumée en diagonale, du point en bas à droite au point en haut à gauche [clic gauche enfoncé], puis nous refaisons la même opération en appliquant une vitesse sur ce trajet [clic droit enfoncé], dans le même sens.

Cependant, même en appliquant les mêmes actions entre chaque expérience, les temps d'exécution seront irrémédiablement différents. Dès lors, pour mesurer la performance d'une fonction, nous n'allons pas regarder le nombre de cycles passés dans cette fonction mais son pourcentage de cycle par rapport à l'ensemble du programme. En effet, étant donné que le pourcentage est un rapport sur le temps de l'ensemble du programme, ce dernier peut varier tout en gardant un pourcentage similaire pour une même fonction, car le nombre de cycles passé dans cette fonction augmentera proportionnellement.

Pendant nos expériences, nous choisirons une grille de 225x225 car c'est le premier seuil où l'application devient trop lente pour être utilisée sur notre machine de test. Cela nous permet de mieux nous rendre compte de nos optimisations.

Lors de nos mesures préliminaires, on observe au maximum un écart de $\pm 5\%$ entre les différentes mesures successives. Nous choisissons donc, pour chaque résultat, d'effectuer 5 tests et de prendre la valeur médiane : cela nous semble suffisant.

4 Optimisations et mesures

Le compilateur utilisé pour ce projet est GNU Compiler Collection (GCC) v8.2.0. La mise en place de la compilation consiste à déterminer les bons flags de compilation sur lesquels nous allons partir pour optimiser notre code. Nous ajoutons tout d’abord « -g3 » afin d’inclure dans le binaire les symboles et informations sur le code source. Nous nous assurons que ce flag n’as pas d’incidence sur les performances. Ci-dessous les résultats des analyses du projet au départ, avec GCC.

Samples: 79K of event 'cycles:ppp', Event count (approx.): 63306420665

Overhead	Command	Shared Object	Symbol
46,86%	appletviewer	libfluid.so	[.] linearSolver
23,99%	appletviewer	libfluid.so	[.] build_index
7,86%	appletviewer	libfluid.so	[.] build_index@plt
2,91%	appletviewer	libfluid.so	[.] advect
2,09%	appletviewer	libfluid.so	[.] project
1,11%	appletviewer	libfluid.so	[.] vorticityConfinement
0,92%	appletviewer	libfluid.so	[.] SWIG_JavaArrayInFloat
0,71%	appletviewer	libfluid.so	[.] SWIG_JavaArrayArgoutFloat
0,70%	appletviewer	libc-2.28.so	[.] __memmove_avx_unaligned_erms
0,68%	appletviewer	libfluid.so	[.] calculate_curl
0,65%	appletviewer	libfluid.so	[.] addSource
0,65%	appletviewer	libfluid.so	[.] swap

FIGURE 2 – Profilage global avec GCC en « -O0 »

Comme nous pouvions l’imaginer, la Figure 2 nous confirme que les fonctions les plus chaudes sont « linearSolver() » et « build_index() », la première de par sa nature de calcul intensif, la deuxième du fait qu’elle soit appelée énormément de fois. On remarque que toutes les autres fonctions sont d’une importance mineure dans le temps d’exécution du programme : inférieure à 3% pour la plus chaude. À elle seule, la fonction « linearSolver() » prend presque 50% du temps d’exécution du programme : c’est la plus importante.

La Figure 3 nous permet d’analyser quelles instructions de la fonction « linearSolver() » ont un impact important sur le temps d’exécution. La première image nous montre un lot successif d’instructions prenant chacune 1 à 4% du temps d’exécution : des *add*, *mul*, *mov*, *lea*. La deuxième image met en évidence une instruction *div* qui prends à elle seule 16% du temps de la fonction. On identifie donc parfaitement les deux sources de ralentissement de la fonction : la première étant un trop grand nombre d’instructions pour effectuer les calculs et la seconde étant la division.

Le temps passé dans la fonction « build_index() » est divisé équitablement en *mov/add* d’après la Figure 4, permettant de faire les deux additions et la multiplication que doit faire la fonction. On remarque tout de même que 32% du temps de la fonction est passé dans la mise en place de la stack frame de la fonction (2 premières instructions *mov*). Enfin, on remarquera que 7% du temps de la simulation est passée dans la fonction « build_index@plt ». Pour comprendre d’où vient cette fonction, il faut revenir sur les bases du fonctionnement du chargement dynamique de code. Afin d’avoir des bibliothèques partagées qui puissent être chargées qu’une seule fois en mémoire mais à des emplacements indépendants pour deux processus différents, on utilise un mécanisme de code à position indépendante (Position-Independent

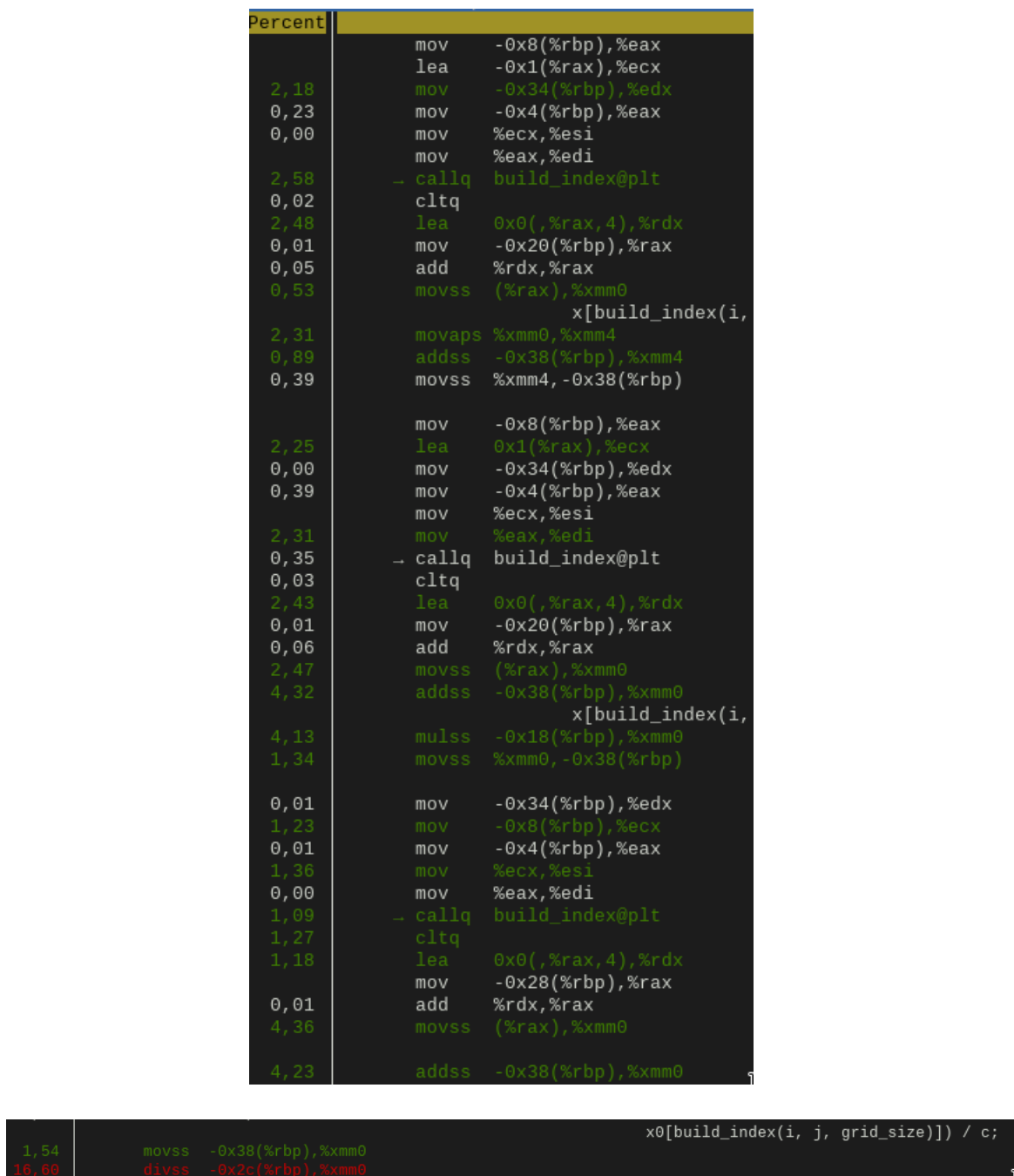


FIGURE 3 – Profilage de « linearSolver() » avec GCC en « -O0 »

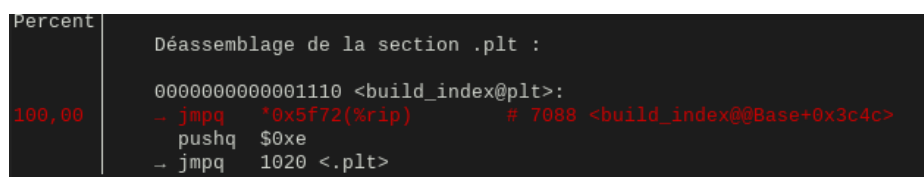
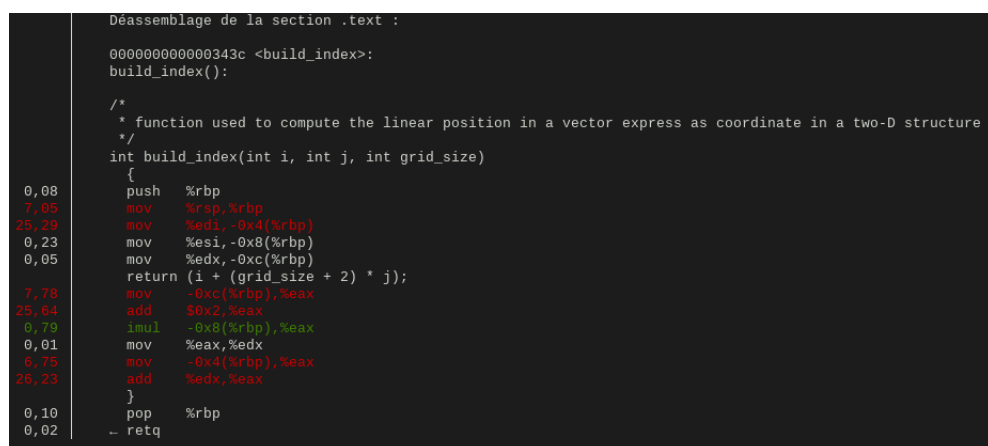


FIGURE 4 – Profilage de « build_index() » et « build_index()@plt » avec GCC en « -O0 »

Code (PIC)). Ce système fait recours à une Global Offset Table (GOT) contenant les adresses des variables et fonctions dont l'emplacement n'est pas connu au moment de la compilation. Cette GOT est consultée par les fonctions contenues dans la Procedure Linkage Table (PLT), une table contenant des fonctions *stubs* (wrapper) qui sont appelées à la place de la vraie fonction. Cette fonction stub (@plt) permet d'aller consulter la GOT, de remplir la GOT lors du premier appel (*lazy loader*), enfin d'exécuter la vraie fonction. Toutes les prochaines fois ou la fonction stub sera appelée, alors on va directement charger l'adresse contenu dans la *GOT* pour appeler la fonction désirée.

4.1 Mise en place de la compilation

Premièrement, nous avons testé les flags suivants activant des options de lot d'optimisation : « -O1 », « -O2 », « -O3 » (agressivité au détriment de la taille du code généré), « -Os » (optimisation de la taille du code). L'utilisation d'un de ces flags nous apporte une légère amélioration de fluidité visuelle, cependant il se trouve que nous n'avons pas observé de différence majeure entre ces 4 différentes flags, que ce soit en termes de code généré ou de performances et répartition du temps entre les fonctions. Ci-dessous les images des résultats avec GCC et « -O2 ».

Au premier abord de la Figure 5, nous pourrions être étonnés que la fonction « linearSolver() » soit passée de 46% à 57%, comme si l'optimisation par le compilateur l'avait rendu plus lente. En fait il n'en est rien : on passe en effet plus de temps dans la fonction « linearSolver() » car on passe moins de temps dans les autres fonctions, ces dernières étant elles aussi optimisées. En effet, on voit que le temps passé dans la

Samples: 66K of event 'cycles:ppp', Event count (approx.): 52703678032

Overhead	Command	Shared Object	Symbol
57,93%	appletviewer	libfluid.so	[.] linearSolver
10,67%	appletviewer	libfluid.so	[.] build_index@plt
10,58%	appletviewer	libfluid.so	[.] build_index
2,47%	appletviewer	libfluid.so	[.] advect
1,97%	appletviewer	libfluid.so	[.] project
1,21%	appletviewer	libfluid.so	[.] vorticityConfinement
0,95%	appletviewer	libc-2.28.so	[.] __memmove_avx_unaligned_erms
0,51%	appletviewer	libfluid.so	[.] calculate_curl
0,41%	appletviewer	libfluid.so	[.] buoyancy
0,39%	appletviewer	libfluid.so	[.] SWIG_JavaArrayInFloat
0,36%	appletviewer	libfluid.so	[.] setBoundry

FIGURE 5 – Profilage global avec GCC en « -O2 »

2,58	0,12	-- callq build_index@plt	
		movss 0xc(%rsp),%xmm0	
			x0[build_index(i, j, grid_size))] / c;
0,02		mov %ebx,%esi	
0,07		mov %ebp,%edx	
			x[build_index(i, j-1, grid_size))] + x[build_index(i, j+1, grid_size))] +
2,51		cltq	
			x0[build_index(i, j, grid_size))] / c;
0,09		mov %r13d,%edi	
			x[build_index(i, j-1, grid_size))] + x[build_index(i, j+1, grid_size))] +
3,66		addss (%r12,%rax,4),%xmm0	
			x[build_index(i, j, grid_size))] = (a * (x[build_index(i-1, j, grid_size))] + x[build_index(i+1, j, grid_size))] +
11,31		uiss 0x20(%rsp),%xmm0	
2,79		movss %xmm0,0xc(%rsp)	
			x0[build_index(i, j, grid_size))] / c;
0,01		-- callq build_index@plt	
		mov 0x18(%rsp),%rcx	
0,02			x[build_index(i, j-1, grid_size))] + x[build_index(i, j+1, grid_size))] +
		mov %ebx,%esi	
		mov %ebp,%edx	
0,17			x[build_index(i, j-1, grid_size))] + x[build_index(i+1, j, grid_size))] +
			x0[build_index(i, j, grid_size))] / c;
2,49		cltq	
			x[build_index(i, j-1, grid_size))] + x[build_index(i, j+1, grid_size))] +
0,26		movss 0xc(%rsp),%xmm0	
			x[build_index(i, j, grid_size))] = (a * (x[build_index(i-1, j, grid_size))] + x[build_index(i+1, j, grid_size))] +
0,02		mov %r13d,%edi	
0,05		mov %r15d,%ebx	
			x[build_index(i, j-1, grid_size))] + x[build_index(i, j+1, grid_size))] +
14,58		addss (%rcx,%rax,4),%xmm0	
2,88		movss %xmm0,0xc(%rsp)	
			x[build_index(i, j, grid_size))] = (a * (x[build_index(i-1, j, grid_size))] + x[build_index(i+1, j, grid_size))] +
0,01		-- callq build_index@plt	
			x0[build_index(i, j, grid_size))] / c;
2,75		movss 0xc(%rsp),%xmm0	
33,55		divss 0x24(%rsp),%xmm0	
			x[build_index(i, j, grid_size))] = (a * (x[build_index(i-1, j, grid_size))] + x[build_index(i+1, j, grid_size))] +
0,27		cltq	
2,73		movss %xmm0, (%r12,%rax,4)	

FIGURE 6 – Profilage de « linearSolver() » avec GCC en « -O2 »

fonction « build_index() » est passé de 24% à 10%. Enfin, il en va de même pour le temps qui a augmenté de 3% pour « build_index@plt » : puisque « build_index() » est plus courte, alors on passe autant de temps à appeler la fonction qu'à effectuer les calculs.

Concernant la fonction « linearSolver() » de la Figure 6, on voit que le nombre d'instructions a nettement été réduit. On est passé de beaucoup d'instruction prenant chacune 1 à 4% à quelques petites instructions et 3 instructions principales de 10 à 33%. Ce nouveau schéma permettra sans doute de mieux cibler les instructions qui vont avoir besoin d'être optimisé.

Comme nous pouvons le constater sur la Figure 7, le code a été considérablement réduit et optimisé par le compilateur pour la fonction « build_index() ». En effet, une grande partie des instructions ont été supprimés car on utilise un passage de valeur par registres (plus rapide) et non plus par la pile. De plus, on utilise l'instruction *lea* spécialisé dans le calcul d'adresse. Finalement, il nous reste *lea* qui prend 15% de la fonction et le retour de la fonction qui prend 85% du temps.

Nous avons aussi testé les performances du programme avec « Clang/Low Level

Percent	
	Déassemblage de la section .text :
	000000000001770 <build_index>:
	build_index():
	/*
	* function used to compute the linear position in a vector express as coordinate in a two-D structure
	*/
	int build_index(int i, int j, int grid_size)
	{
	return (i + (grid_size + 2) * j);
0,25	add \$0x2,%edx
0,04	imul %esi,%edx
15,13	lea (%rdx,%rdi,1),%eax
	}
04,58	retq

FIGURE 7 – Profilage de « build_index() » avec GCC en « -O2 »

Samples: 189K of event 'cycles:ppp', Event count (approx.): 151193835371			
Overhead	Command	Shared Object	Symbol
79,73%	appletviewer	libfluid.so	[.] linearSolver
2,69%	appletviewer	libfluid.so	[.] advect
1,38%	appletviewer	libc-2.28.so	[.] __memmove_avx_unaligned_erms
1,14%	appletviewer	libfluid.so	[.] project
0,98%	appletviewer	libfluid.so	[.] vorticityConfinement
0,65%	appletviewer	libfluid.so	[.] c_velocitySolver
0,52%	appletviewer	libfluid.so	[.] Java_fluidJNI_c_1velocitySolver
0,44%	appletviewer	libawt.so	[.] AnyIntSetRect
0,40%	appletviewer	libawt.so	[.] Java_sun_java2d_loops_FillRect_FillRect
0,31%	appletviewer	[kernel.kallsyms]	[k] clear_page_erms
0,25%	appletviewer	libfluid.so	[.] Java_fluidJNI_c_1densitySolver
0,21%	appletviewer	[kernel.kallsyms]	[k] i8042_interrupt
0,17%	appletviewer	[kernel.kallsyms]	[k] native_irq_return_iret
0,16%	appletviewer	libjvm.so	[.] 0x00000000007b3410
0,16%	appletviewer	libawt.so	[.] Region_GetBounds
0,15%	appletviewer	[kernel.kallsyms]	[k] __list_del_entry_valid
0,14%	appletviewer	perf-21725.map	[.] 0x00007fedc12bd07f
0,13%	appletviewer	[kernel.kallsyms]	[k] error_entry

FIGURE 8 – Profilage global avec Clang en « -O3 »

Virtual Machine (LLVM) », et les flags d’optimisation les plus agressifs (« -O3 »). Ci-dessous les images des résultats avec Clang et « -O3 ».

Illustré par la Figure 8, avec Clang et les options de compilation les plus agressives, on voit que le temps passés dans la fonction « linearSolver() » est passé de 57% avec GCC à 80%. Encore une fois, cela veut dire que les autres fonctions ont été encore plus radicalement minimisées. On notera aussi les fonctions « build_index() » et « build_index@plt » ont complètement été supprimés.

Par la Figure 9, nous observons que le code de « linearSolver() » a été extrêmement réduit en 5 instructions principales, introduites par Streaming SIMD Extensions (SSE). C’est vers ce genre de code que devrait converger nos optimisations manuelles.

Nous observons donc un net gain à utiliser Clang par rapport à GCC, qu’il soit visuel ou factuel. Cependant, pour nous permettre d’appliquer des optimisations vu en cours à la main, nous allons opter pour la version que produit GCC comme base de travail.

Enfin, nous noterons que nous avons aussi essayé d’utiliser l’option d’optimisation « -Ofast » avec GCC. Ce fut un succès visuel puisque l’accélération était nettement présente, cependant nous avons préféré ne pas laisser cette option activée : ce flag

```

2,89      x[build_index(i, j, grid_size)] = (a * ( x[build_index(i-1, j, grid_size)] + x[build_index(i+1, j, grid_size)] +
c0:  lea    (%r12,%rsi,1),%r10d
0,01      add    $0x2,%r10d
0,11      movslq  %r10d,%rbp
0,02      movss   (%rbx,%rbp,4),%xmm2
3,83      lea    0x4(%r14,%rsi,1),%ebp
0,01      movslq  %ebp,%rbp
0,28      addss  (%rbx,%rbp,4),%xmm2
15,55     addss  (%rbx,%r11,1),%xmm2
12,05     addss  (%rcx,%rdi,1),%xmm2
12,41     mulss  %xmm0,%xmm2
12,67     addss  (%r8,%rdx,1),%xmm2
34,35     divss  %xmm1,%xmm2
3,23      movss  %xmm2,(%rbx,%r9,1)
0,01      add    %r13,%r11
0,08      add    %r13,%r8
3,04      add    %r12,%esi

```

FIGURE 9 – Profilage de « linearSolver() » avec Clang en « -O3 »

permet d'activer des optimisations mathématiques, notamment sur les flottants, en autorisant des arrondis et le non-respect de certains standards de l'Institute of Electrical and Electronics Engineers (IEEE). N'ayant aucun moyen de vérifier par nous même si les calculs sont encore de la précision voulue avec ce flag, nous préférons ne pas tenter une opération qui pourrait remettre en cause la validité de la simulation.

4.2 Inlining

Comme vu dans la section précédente, la fonction « build_index() » était très consommatrice en cycle (malgré l'optimisation du passage de paramètres par les registres plutôt que par la pile) à cause de 2 phénomènes :

1. L'appel de la fonction, qui passe par deux indirections du fait de la PLT, nous fait exécuter des instructions supplémentaires.
2. Les branchements dans le code ne favorise pas le principe de localité des caches.

Une optimisation bien connue consiste donc à *inliner* une fonction : cela revient à remplacer automatiquement lors de la compilation, dans le code généré du binaire, tous les appels à ladite fonction par le corps de la fonction. Cela a pour avantage de réduire considérablement le temps d'utilisation d'une fonction (plus de branchement ni de passage de paramètre). Cependant, cela peut grossir de manière drastique la taille du code si la fonction inliner est une longue fonction : ce qui n'est pas le cas ici, de toute évidence.

Pour procéder à l'inlining, cela est très simple puisqu'il suffit de l'indiquer au compilateur avec le mot clé approprié. On passe donc du code suivant (Listing 1) à la ligne 1 à celui de la ligne 2.

```

1 int build_index(int i, int j, int grid_size)
2 inline int build_index(int i, int j, int grid_size)

```

Listing 1 – Fonction « build_index() » avant et après inlining

En termes de gain, la fluidité visuelle est grandement améliorée et tout de suite remarquable grâce à l'inlining. Par le profilage globale (Figure 10), on peut observer

Samples: 51K of event 'cycles:ppp', Event count (approx.): 40762790708

Overhead	Command	Shared Object	Symbol
74,63%	appletviewer	libfluid.so	[.] linearSolver
2,27%	appletviewer	libfluid.so	[.] advect
1,50%	appletviewer	libc-2.28.so	[.] __memmove_avx_unaligned_erms
1,07%	appletviewer	libfluid.so	[.] project
0,72%	appletviewer	libfluid.so	[.] vorticityConfinement
0,63%	appletviewer	libfluid.so	[.] SWIG_JavaArrayArgoutFloat
0,54%	appletviewer	libfluid.so	[.] SWIG_JavaArrayInFloat

FIGURE 10 – Profilage global après inlining de la fonction « build_index() »

```

70:  mov    %rsi,%rax
      for (j = 1; j <= grid_size; j++) {
      mov    $0x1,%edx
      nop
      x[build_index(i, j, grid_size)] = (a * ( x[build_index(i-1, j, grid_size)] + x[build_index(i+1, j, grid_size)] +
0,04  80:  movss  -0x4(%rbx,%rax,1),%xmm2
3,19      addss  0x4(%rbx,%rax,1),%xmm2
      x[build_index(i, j-1, grid_size)] + x[build_index(i, j+1, grid_size)]) +
0,05      add    $0x1,%edx
      x[build_index(i, j, grid_size)] = (a * ( x[build_index(i-1, j, grid_size)] + x[build_index(i+1, j, grid_size)] +
21,70  addss  (%r12,%rax,1),%xmm2
      x[build_index(i, j-1, grid_size)] + x[build_index(i, j+1, grid_size)]) +
12,34  addss  (%r14,%rax,1),%xmm2
      x[build_index(i, j, grid_size)] = (a * ( x[build_index(i-1, j, grid_size)] + x[build_index(i+1, j, grid_size)] +
12,45  mulss  %xmm0,%xmm2
      x[build_index(i, j-1, grid_size)] + x[build_index(i, j+1, grid_size)]) +
12,49  addss  (%r10,%rax,1),%xmm2
      x0[build_index(i, j, grid_size))] / c;
34,35  divss  %xmm1,%xmm2
      x[build_index(i, j, grid_size)] = (a * ( x[build_index(i-1, j, grid_size)] + x[build_index(i+1, j, grid_size)] +
3,28  movss  %xmm2, (%rbx,%rax,1)
0,04  add    %rbp,%rax
      for (j = 1; j <= grid_size; j++) {

```

FIGURE 11 – Profilage de « linearSolver() » après inlining de la fonction « build_index() »

que les fonctions « build_index() » et « build_index@plt » ont été complètement supprimées et que l'on est passé d'un temps de 56% dans la fonction « linearSolver() » à maintenant 74%, ce qui est bon signe : il faut maintenant se concentrer sur l'optimisation de cette fonction.

Concernant le code de la fonction « linearSolver() », on observe dans la Figure 11 qu'il a été considérablement réduit et compacté, il ressemble maintenant fortement au code créer par Clang, décrit dans la section précédente.

4.3 Mots-clés appropriés

Les mots-clés peuvent souvent faire une différence capitale dans le processus d'optimisation du compilateur. En effet, de nombreuses optimisations ne peuvent être effectuées que sous réserve de certaines conditions, par exemple : qu'une variable ne soient pas modifié par un autre thread ou encore qu'une variable soit constante. Nous disposons, en C, des mots-clés (entre autres) « static » et « const » qui sont très utiles pour indiquer ce genre de cas et bien d'autres encore. Après avoir appliqué le mot-clé « static » à toutes les fonctions (hormis les 2 points d'entrée de la bibliothèque) afin d'indiquer qu'elles sont locales au module compilé (améliorant par exemple l'*inlining* automatique) et le mot-clé « const » aux variables constantes de la fonction « linearSolver() », nous obtenons le résultat de la Figure 12 au profilage.

Le suffix « .isra.constprop » signifie que GCC a automatiquement effectué une propagation de constante et/ou a retiré des paramètres inutiles. À partir de cette information, nous nous sommes rapidement rendu compte qu'il y avait des variables

Samples: 36K of event 'cycles:uppp', Event count (approx.): 27984766957			
Overhead	Command	Shared Object	Symbol
75,51%	appletviewer	libfluid.so	[.] linearSolver.isra.0.constprop.2

FIGURE 12 – Profilage global après ajouts des mots-clés appropriés

inutiles passées à « `linearSolver()` » : « dt » n'était jamais utilisé, et « b » était toujours égal à 0. Nous avons donc nous-mêmes supprimé ces variables dans la fonction.

Il n'y a pas de gain de performance visible ou mesurable, cependant cela éclairci le code source et facilitera peut-être l'analyse du compilateur pour de prochaines optimisations.

4.4 Modification des calculs

Comme nous pouvions le voir dans les sections précédentes, ils nous fallait maintenant nous concentrer sur le cœur de calcul de « `linearSolver()` ». L'instruction la plus problématique était la division, suivi de la multiplication puis des additions. Analysons la situation : pour cela, nous avons déterminé les différentes valeurs que pouvait prendre les variables de la fonction, puis nous avons affiché en temps réel les valeurs des variables avec un compteur permettant de savoir combien de fois quelle variable a pris quelle valeur.

4.4.1 Multiplication

Nous avons d'abord analysé la variable « a », celle qui se retrouve dans la multiplication. Il se trouve que cette variable ne prend que les valeurs 1 ou 0 : il s'agit en réalité d'un booléen, bien que le type soit un float. Cette variable multiplie une suite de 3 additions : si « a » est à 1, alors on garde les 3 additions ; si la variable est à 0, alors il ne faut pas faire les additions. De toute évidence, cette variable n'est là que pour contrôler le fait de garder ou non les additions et n'est pas là pour effectuer une mise à l'échelle par un facteur précis. Dès lors, si la variable est à 0, notre programme va faire les 3 additions pour ensuite les multiplier avec 0 : c'est une grosse perte de temps. On conseille souvent d'éviter les branchements dans une boucle, mais il se trouve qu'ici, avoir un branchement (quelques cycles, 2 à 4) permettrait d'éviter :

1. Pire des cas : 1 multiplication (quelques cycles, 2 à 4).
2. Meilleur des cas : 1 multiplication et 3 additions (plus d'une dizaine de cycles, 10 à 15).

Les cycles ont été pris pour un processeur *Sandy Bridge* sur les documents d'**Agner Fog**. Un branchement permet donc clairement d'être gagnant s'il est assez prédictible.

Nous avons donc procéder de la manière suivante, visible sur le Listing 2 : nous faisons un *NOT* de la variable dans une autre variable, déclaré constante. Nous prenons garde à stocker le résultat dans un *int* et à faire un *cast* au préalable : de cette manière, GCC va générer une instruction *test* plutôt qu'une instruction *ucomiss* (plus coûteuse) pour effectuer la comparaison. Nous stockons le *NOT* dès le début

```

1 const int inv_a = !(int)a;
2 [...]
3 ((inv_a ? 0 :
4   ((x[build_index(i-1, j, grid_size)] + x[build_index(i+1, j, grid_size)]) +
5     (x[build_index(i, j-1, grid_size)] + x[build_index(i, j+1, grid_size)]))
6  ) + x0[build_index(i, j, grid_size)]
7 ) / c;

```

Listing 2 – Résultat de la modification de la multiplication

de la fonction pour ne pas avoir à le calculer à chaque itération de la boucle, et nous voulons le *NOT* afin de faire arriver le cas où « a » est à 0 en premier. En effet, nous avons testé un grand nombre de scénarios pour trouver dans quelle configuration le code généré et la prédiction de branchement fonctionnait au mieux :

1. Tentative de faire intervenir le cas le plus fréquent en premier, puis en second.
2. Utilisation de *if-else* ou d'opérateur ternaire.
3. Séparation du résultat dans une variable temporaire, ou calcul inliner.
4. Utilisation d'un *builtin* afin d'indiquer à GCC la branche la plus prise, ou de laisser totalement la main au processeur.

Il se trouve que finalement, la combinaison qui fonctionnait de manière la plus optimale possible se trouve être :

1. Le cas le plus fréquent en premier.
2. Utilisation d'un opérateur ternaire.
3. Calcul *inliner*.
4. Ne pas indiquer à GCC la branche la plus souvent prise (le CPU se débrouille très bien tout seul!).

À noter que pour les 3 additions, nous les avons regroupés depuis la forme $(a + b + c + d)$ vers la forme $((a + b) + (c + d))$. Cela permet d'explicitier la non-dépendance des opérations dans cette chaîne, et ainsi de laisser le CPU commencer le calcul de $(c + d)$ avant la fin de $(a + b)$.

4.4.2 Division

La deuxième optimisation majeure est la suppression de la division. Comme vu sur la Figure 11, la division prenait un peu plus de 30% du temps d'exécution de notre fonction, ce qui est énorme. Nous avons donc analysé la variable « c » : il se trouve que cette variable prends la valeur 1 avec une probabilité de $\frac{2}{3}$ et la valeur 4 avec une probabilité de $\frac{1}{3}$. Diviser par 1 revient à ne pas diviser, et diviser par 4 revient à multiplier par 0,25. Nous calculons donc l'inverse de la variable au début de la fonction, et nous multiplions le tout par cette variable plutôt que d'effectuer une division coûteuse à chaque itération : soit on multiplie par 1, soit on multiplie par 0.25. Ici, l'utilisation d'un branchement pour ne pas multiplier lorsque « c » est égal à 1 était tout aussi coûteux que de multiplier par 1, donc inutile.

```

1 const float inv_c = 1. / c;
2 [...]
3 ((inv_a ? 0 :
4   ((x[build_index(i-1, j, grid_size)] + x[build_index(i+1, j, grid_size)]) +
5     (x[build_index(i, j-1, grid_size)] + x[build_index(i, j+1, grid_size)]))
6  ) + x0[build_index(i, j, grid_size)]
7  ) * inv_c;

```

Listing 3 – Résultat de la suppression de la division

```

1 const float tmp1 = -0.5f * (1. / grid_size);
2 const float tmp2 = 0.5f * grid_size;
3 [...]
4 for (j = 1; j <= grid_size; j++) {
5     div[build_index(i, j, grid_size)] = tmp1 * (x[build_index(i+1, j,
6       grid_size)] - [...]);
7 }
8 [...]
9 for (j = 1; j <= grid_size; j++) {
10    x[build_index(i, j, grid_size)] -= tmp2 * [...];
11    y[build_index(i, j, grid_size)] -= tmp2 * [...];
12 }

```

Listing 4 – Calcul préalable d’invariants de boucle

4.4.3 Invariants

Nous avons procédé à quelques légères optimisations dans la fonction « `project()` » : en effet, des invariants de boucle étaient présents et pouvaient être pré-calculé. Un invariant de boucle est, ici, un groupe de calcul calculé à chaque itération inutilement alors qu’il pourrait être calculé avant ladite boucle. Vous pouvez observer l’optimisation sur le Listing 4.

4.4.4 Résultat

Tout d’abord, d’expérience utilisateur, ces optimisations sont un franc succès. Le mode de résolution « 255x255 » peut enfin être utilisé sur notre machine de test de manière tout à fait confortable, avec environ 60 images par secondes (ressentit).

Sur la Figure 13, on observe un net changement de répartition de temps de calcul : « `linearSolver()` », qui prenant auparavant 75% du programme, n’en prends maintenant que 54%! On obtient donc un gain d’environ 20%, ce qui est non-négligeable.

Sur la Figure 14, on peut observer nos changements effectués : un *mul* a disparu, le *div* a été remplacé par un *mul*, un *test* et un *je* sont apparu... De nouvelles

Samples: 35K of event 'cycles:uppp', Event count (approx.): 25047808192			
Overhead	Command	Shared Object	Symbol
54,39%	appletviewer	libfluid.so	[.] linearSolver
4,64%	appletviewer	libfluid.so	[.] advect
3,37%	appletviewer	libfluid.so	[.] c_velocitySolver
2,08%	appletviewer	libfluid.so	[.] project.isra.1
1,68%	appletviewer	[unknown]	[k] 0xffffffffba000a27
1,58%	appletviewer	libc-2.28.so	[.] __memmove_avx_unaligned_erms
1,09%	appletviewer	libawt.so	[.] Java_sun_java2d_loops_FillRect_FillRect
1,09%	appletviewer	libawt.so	[.] AnyIntSetRect
1,08%	appletviewer	libfluid.so	[.] SWIG_JavaArrayArgoutFloat
0,89%	appletviewer	libfluid.so	[.] SWIG_JavaArrayInFloat
0,81%	appletviewer	libfluid.so	[.] calculate_curl
0,56%	appletviewer	libjvm.so	[.] 0x00000000007b3410

FIGURE 13 – Profilage global après modifications des calculs

```

0,02      for (j = 1; j <= grid_size; j++) {
0,01      mov     $0x1,%ecx
0,01      nop
          x[build_index(i, j, grid_size)] =
2,03 90:      add     $0x1,%ecx
          ((inv_a ? 0 :
5,78      movaps  %xmm4,%xmm0
0,07      test    %r13d,%r13d
1,82      je      cf
          return (i + (grid_size + 2) * j);
0,06      mov     %edx,%esi
          ((x[build_index(i-1, j, grid_size)] +
0,02      lea     -0x1(%rdx),%r8d
3,78      lea     0x1(%rdx),%edi
0,06      sub     %ebp,%esi
0,07      movslq  %edi,%rdi
0,03      movslq  %r8d,%r8
4,02      movss  (%rbx,%r8,4),%xmm0
0,10      addss  (%rbx,%rdi,4),%xmm0
          (x[build_index(i, j-1, grid_size)] +
0,04      lea     (%r9,%rsi,1),%edi
0,04      movslq  %esi,%rsi
3,89      movslq  %edi,%rdi
0,11      movss  (%rbx,%rdi,4),%xmm1
15,39     addss  (%rbx,%rsi,4),%xmm1
          ((inv_a ? 0 :
15,10     addss  %xmm1,%xmm0
          ) + x0[build_index(i, j, grid_size)]
17,84  cf:      addss  (%r15,%rax,1),%xmm0
1,97      add     %ebp,%edx
          ) * inv_c;
19,86     mulss  %xmm3,%xmm0
          x[build_index(i, j, grid_size)] =
5,92      movss  %xmm0, (%rbx,%rax,1)
0,05      add     %r12,%rax
          for (j = 1; j <= grid_size; j++) {

```

FIGURE 14 – Profilage de « linearSolver() » après modifications des calculs

Samples: 112K of event 'cycles:uppp', Event count (approx.): 84705775418			
Overhead	Command	Shared Object	Symbol
62,45%	appletviewer	libfluid.so	[.] linearSolver
6,60%	appletviewer	libfluid.so	[.] advect
4,13%	appletviewer	libfluid.so	[.] c_velocitySolver
2,81%	appletviewer	libfluid.so	[.] project.isra.1
1,69%	appletviewer	[unknown]	[k] 0xffffffffba000a27
1,64%	appletviewer	libc-2.28.so	[.] __memmove_avx_unaligned_erms
1,12%	appletviewer	libfluid.so	[.] calculate_curl
1,08%	appletviewer	libfluid.so	[.] SWIG_JavaArrayArgoutFloat
0,89%	appletviewer	libfluid.so	[.] SWIG_JavaArrayInFloat
0,87%	appletviewer	libawt.so	[.] Java_sun_java2d_loops_FillRect_FillRect
0,57%	appletviewer	libfluid.so	[.] c_densitySolver

FIGURE 15 – Profilage global avant opitmsiation du cache

Samples: 140K of event 'cycles:uppp', Event count (approx.): 106073766713			
Overhead	Command	Shared Object	Symbol
55,17%	appletviewer	libfluid.so	[.] linearSolver
7,27%	appletviewer	libfluid.so	[.] advect
4,57%	appletviewer	libfluid.so	[.] c_velocitySolver
3,05%	appletviewer	libfluid.so	[.] project.isra.1
1,92%	appletviewer	[unknown]	[k] 0xffffffffba000a27
1,79%	appletviewer	libc-2.28.so	[.] __memmove_avx_unaligned_erms
1,28%	appletviewer	libfluid.so	[.] calculate_curl
1,24%	appletviewer	libawt.so	[.] Java_sun_java2d_loops_FillRect_FillRect
1,18%	appletviewer	libfluid.so	[.] SWIG_JavaArrayArgoutFloat
0,94%	appletviewer	libfluid.so	[.] SWIG_JavaArrayInFloat
0,64%	appletviewer	libfluid.so	[.] c_densitySolver
0,54%	appletviewer	perf-16242.map	[.] 0x00007fa0dd26fc43

FIGURE 16 – Profilage global après opitmsiation du cache

instructions qui ont la même fonction que les anciennes, mais qui consomment moins de temps.

4.5 Mémoire cache

Maintenant que la simulation fonctionne de manière très satisfaisante avec une résolution de « 225x225 », nous passons nos mesures sur la résolution « 450x450 », qui n'est toujours pas fluide. Nous pouvons observer le résultat de la mesure avant la prochaine optimisation sur la Figure 15 : notre fonction « linearSolver() » prends 62% de notre temps d'exécution.

Dans cette partie, nous avons essayé d'optimiser l'utilisation de la mémoire cache en améliorant la localité des accès à la mémoire. Pour cela, deux techniques ont été mises en œuvres :

1. Inversion des deux boucles.
2. Cache Blocking.

Pour la première technique, il s'agit d'inverser l'ordre des boucles intérieur et extérieur (i et j) afin de parcourir les matrices par lignes et non par colonnes. Plus la matrice sera grande (et ne rentrera plus entièrement dans la mémoire cache), plus l'amélioration sera visible.

Pour la seconde technique, il s'agit de traiter la matrice par bloc carré plutôt que la traiter complètement d'un seul coup. Le schéma de la Figure 17 nous explique le principe très intuitivement. Le but de cette manœuvre est de faire tenir le bloc traité dans la mémoire cache et ainsi d'avoir un minimum de *cache miss*. Si on traite

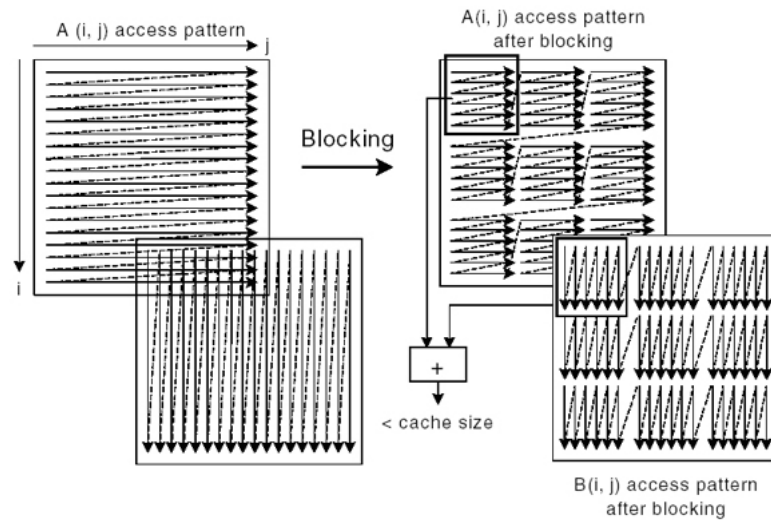


FIGURE 17 – Explication de l’effet du cache blocking (source : Intel.com)

la matrice ligne par ligne mais qu’une ligne est plus grosse que la mémoire cache, alors à chaque changement de ligne nous aurions des *cache miss* inutiles. Une chose important dans le cache blocking est la taille d’un bloc. Le déterminer de manière peut-être une tâche assez ardue dû au décalage entre la théorie et la pratique. Dans notre cas, nous l’avons déterminé de manière empirique en effectuant plusieurs tests sur notre machine.

Ces deux techniques nous ont permis d’obtenir un léger gain en « 450x450 », illustré par la Figure 16 : nous pouvons voir que nous gagnons à peu 7% de temps passé dans la fonction grâce à l’optimisation de la mémoire. Si les matrices traités étaient beaucoup plus grosses, alors le gain serai beaucoup plus visible. En revanche, si les matrices deviennent très petite, alors cette optimisation nous fait perdre du temps dans les contrôles de boucle.

Le résultat final de l’optimisation se résume par le code du Listing 5.

```

1 #define BLOCK_SIZE 8
2 [...]
3 for (j = 1; j <= grid_size; j += BLOCK_SIZE) {
4     for (i = 1; i <= grid_size; i += BLOCK_SIZE) {
5         for (jj = j; jj <= j + BLOCK_SIZE && jj <= grid_size; jj++) {
6             for (ii = i; ii <= i + BLOCK_SIZE && ii <= grid_size ; ii++) {
7                 x[build_index(ii, jj, grid_size)] = [...]

```

Listing 5 – Cache blocking et inversion des boucles

```

1 for (ii = i; ii <= i + BLOCK_SIZE && ii <= grid_size ; ii+=2) {
2     x[build_index(ii, jj, grid_size)] =
3         ((inv_a ? 0 :
4             ((x[build_index(ii-1, jj, grid_size)]
5                 + x[build_index(ii+1, jj, grid_size)]) +
6                 (x[build_index(ii, jj-1, grid_size)]
7                     + x[build_index(ii, jj+1, grid_size)])))
8             ) + x0[build_index(ii, jj, grid_size)]
9         ) * inv_c;
10    x[build_index(ii+1, jj, grid_size)] =
11        ((inv_a ? 0 :
12            ((x[build_index(ii, jj, grid_size)] +
13                x[build_index(ii+2, jj, grid_size)]) +
14                (x[build_index(ii+1, jj-1, grid_size)] +
15                    x[build_index(ii+1, jj+1, grid_size)])))
16            ) + x0[build_index(ii+1, jj, grid_size)]
17        ) * inv_c;
18 }

```

Listing 6 – Boucle intérieure déroulée par un facteur de 2

4.6 Déroutage de boucle

Nous avons tenté d'effectuer un déroulage de boucle manuel sur la boucle intérieure du calcul de « `linearSolver()` ». Le principe de l'*unrolling* est d'effectuer, au sein d'une seule itération de la boucle, deux ou plusieurs itérations de notre boucle initiale. Cela a pour avantage d'optimiser l'utilisation du pipeline et de réduire le nombre de comparaisons/jump effectués au total. Cependant, dans notre cas, la consommation de cycles des calculs sur nombres flottants est tellement grande que ces avantages sont invisibles dans notre application.

Le Listing 6 présente le déroulage de la boucle par un facteur 2. Nous avons aussi essayé avec un facteur 4, mais ce n'était pas plus satisfaisant. Finalement, soit il n'y avait aucun speed-up, soit il y avait un léger ralentissement.

4.7 Vectorisation et parallélisation

La vectorisation ou parallélisation de la boucle de « `linearSolver()` » nous semble de premier abord plutôt délicat. En effet, à chaque itération, on souhaite calculer $x_{i,j}$, qui dépend lui-même des 4 cases adjacentes de la matrice : $x_{i-1,j}$, $x_{i,j-1}$, $x_{i+1,j}$, $x_{i,j+1}$. Cela revient à dire que toutes les cases sont dépendantes entre elles, ce qui rend l'algorithme, sous cette forme, exécutable uniquement de manière itérative.

Notre analyse de l'algorithme a été assez succincte faute de temps, elle est donc à prendre avec des pincettes. Même dans le cas où elle s'avère exact, il y a sans doute

un moyen de transformer l'algorithme pour le rendre parallélisable ou vectorisable.

Pour le parallélisme, nous aurions pu utiliser l'API *OpenMP*. Pour la vectorisation, nous aurions pu utiliser le jeu d'instruction SSE ou AVX disponibles sur nos processeurs *Intel*.

5 Conclusion

Pour conclure, nous aurons tout de même réussi à apporter des optimisations que nous jugeons intéressantes à notre programme de simulation. Au départ, le mode de résolution « 225x225 » était complètement inutilisable, celui en « 150x150 » était lent, et il nous était obligé d'utiliser l'application en « 90x90 » pour avoir une utilisation confortable. Maintenant, le mode « 225x225 » est aussi fluide que le « 90x90 », et le mode « 450x450 », bien qu'étant saccadé, est envisageable.

Nous aurons été confronté à de nombreux petits soucis éparses pendant notre démarche, tel que la problématique de comment mesurer les performances d'une application interactive, ou des problèmes de cohérences entre les mesures. Ces problèmes de cohérences pouvaient avoir des explications diverses : utilisation du CPU par un autre processus, erreurs dans le code, mauvais paramétrage de l'outil de mesure... D'où l'importance de faire plusieurs *runs* par expérience.

Pour conclure, nous pensons que ce projet nous aura apporté une bonne méthode d'analyse et d'optimisation, et nous aura permis de mettre en application réel les techniques et sujets vus en cours. Une très bonne première approche !

Acronymes

GCC GNU Compiler Collection

LLVM Low Level Virtual Machine

PIC Position-Independent Code

GOT Global Offset Table

PLT Procedure Linkage Table

SSE Streaming SIMD Extensions

AVX Advanced Vector Extensions

SSE Streaming SIMD Extensions

IEEE Institute of Electrical and Electronics Engineers

CPU Central Processing Unit