

Analyse de performance et optimisation de code

AYOUB Pierre – BONNAFOUS Camille – FLAMANT Océane

22 avril 2019



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 l'analyse et de l'optimisation d'un code de calcul, cœur des simulations numériques présentés ci-dessus.

Table des matières

1	Introduction	4
2	Analyse du code	6
3	Protocole expérimental	7
3.1	Driver	7
3.2	Théorie	8
3.2.1	Cache L1	8
3.2.2	Cache L2	8
3.2.3	RAM	9
3.2.4	Analyse de sensibilité	9
3.3	Pratique	9
3.3.1	Cache L1	9
3.3.2	Cache L2	10
3.3.3	RAM	10
4	Optimisations et mesures	13
4.1	Phase 1	13
4.1.1	Cache L1	13
4.1.2	Cache L2	15
4.1.3	RAM	17
4.2	Phase 2	18
4.2.1	Cache L1	18
4.2.2	Cache L2	22
4.2.3	RAM	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 calcul, appelé kernel. Les mesures doivent s'effectuer à l'aide de l'instruction *x86 RDTSC*, et des deux outils d'analyse de performance suivant : *MAQAO* et *LIKWID*. *RDTSC* nous permet de mesurer le nombre de cycles entre deux instants, *MAQAO* rends possible l'exécution d'analyses statiques (*CQA*) et dynamiques (*LPROF*) d'un binaire, présentées par un rapport haut niveau à l'aide de *ONE-VIEW*, enfin *LIKWID* permet d'obtenir un grand nombre de métriques très précises concernant, notamment, l'usage de la mémoire.

Afin d'étudier les différents niveaux de la hiérarchie mémoire, chaque membre du groupe analysera un niveau qu'il se verra assigné. Ci-dessous la liste des assignations :

Pierre Cache L1 :

- Intel Core i7-6600U @ 2.8 GHz, Skylake 6^{ème} génération, 14nm, 2 cœurs 4 threads (Hyper-Threading)
- 32 KiB L1i, 32 KiB L1d (par coeur)
- 256 KiB L2 (par coeur)
- 4096 KiB L3 (partagé)

Océane Cache L2 :

```
-----
CPU name:      Intel(R) Core(TM) i7-5500U CPU @ 2.40GHz
CPU type:      Intel Core Broadwell processor
CPU stepping:  4
*****
Hardware Thread Topology
*****
Sockets:      1
Cores per socket: 2
Threads per core: 2
-----
HWThread      Thread      Core      Socket      Available
0              0          0          0            *
1              0          1          0            *
2              1          0          0            *
3              1          1          0            *
-----
Socket 0:      ( 0 2 1 3 )
-----
Cache Topology
*****
Level:         1
Size:          32 kB
Cache groups:  ( 0 2 ) ( 1 3 )
-----
Level:         2
Size:          256 kB
Cache groups:  ( 0 2 ) ( 1 3 )
-----
Level:         3
Size:          4 MB
Cache groups:  ( 0 2 1 3 )
-----
*****
NUMA Topology
*****
NUMA domains:  1
-----
Domain:        0
Processors:    ( 0 2 1 3 )
Distances:     10
Free memory:   1639.13 MB
Total memory:  5832.04 MB
-----
```

FIGURE 1

Camille RAM :

— Voir le fichier « .odt » annexe envoyé par Camille.

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 d'architecture informatique. Il convient d'étudier les choix mis en œuvres afin d'implémenter le ou les calculs 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

Présentons notre kernel par son prototype, que nous observons sur le Listing 1. Nous voyons qu'il y a 3 variables qui sont manipulées :

- n correspond à la taille de nos tableaux.
- a est un tableau de *float* à deux dimensions dont la taille en fonction de n : $4n^2$ Bytes.
- b est un tableau de *double* à une dimension dont la taille en fonction de n : $8n$ Bytes.

```
1 void baseline(unsigned n, float a[n][n], double b[n])
```

Listing 1 – Prototype du kernel non-optimisé

Nous sommes face à un code de calcul très simple en apparence, illustré dans le Listing 2 : deux boucles imbriquées, un branchement, un calcul mêlant multiplication et exponentiel. Mais plusieurs éléments remarquables qui risquent de poser problème au niveau de la rapidité d'exécution apparaissent alors : les boucles impliquent qu'il faut prêter attention au sens de parcours des tableaux, le branchement nous laisse penser qu'il faudrait essayer de le supprimer, enfin l'exponentiel et la multiplication sont des opérations lourdes.

```
1 for (j = 0; j < n; j++) {  
2     for (i = 0; i < n; i++) {  
3         if (j == 0)  
4             b[i] = 1.0;  
5         b[i] *= exp(a[i][j]);  
6     }  
7 }
```

Listing 2 – Kernel non-optimisé

3 Protocole expérimental

La mise en place d'un protocole expérimental de mesure est une étape nécessaire et cruciale dans tout processus d'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 des calculs : 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 Driver

Le code de notre environnement de mesure est séparé en deux parties : le *driver* et le *kernel*. Le kernel contient la fonction de calcul à proprement dite, sur laquelle nos optimisations se porteront. Le driver est le code qui nous permet d'effectuer nos mesures. Le paragraphe suivant sera consacré à une rapide explication de son fonctionnement, le code étant accessible dans le fichier « *driver.c* ».

Les premières lignes du driver servent à récupérer les options de mesures passées en argument à l'application. Vient ensuite une boucle qui englobe toute l'expérience de mesure, elle correspond à l'exécution des méta-répétitions. Une méta-répétition est une répétition qui comprend l'expérience dans sa globalité. L'utilité d'avoir plusieurs méta-répétitions vient du fait que plusieurs mesures sont nécessaires pour être représentatives, car une mesure isolée pourrait être biaisée. On prend la mesure médiane issues des différentes méta-répétitions. Les premières lignes de la boucle des méta-répétitions correspondent à l'allocation et l'initialisation des tableaux utilisés par le kernel. Il faut faire attention à bien utiliser la mémoire lors de l'initialisation, sinon *Linux* pourrait ne pas vraiment allouer le tableau en mémoire (initialisation paresseuse). Ensuite, nous entrons dans une petite boucle qui effectue le warmup, c'est-à-dire la « mise en chauffe » (terme vague mais expliquant le nom) du processeur. Derrière cette appellation grossière se cache un remplissage de la mémoire cache avec nos tableaux, puisque l'on effectue plusieurs fois la fonction de calcul « dans le vide », sans effectuer de mesure. Cette phase de warmup permet de passer le régime transitoire, où le temps de calcul s'améliore à chaque itération, pour arriver dans le régime permanent, où le temps de calcul est stable. Une fois le warmup terminé, on passe à la mesure proprement dite, effectuée par l'instruction *RDTSC*. Le choix s'est porté sur cette instruction pour son efficacité : en effet, elle appelle directement une instruction assembleur x86 correspondante, et son imprécision n'est que de quelques dizaines de cycles seulement. Entre nos deux instructions *RDTSC* (start/stop) se trouve une boucle de répétition d'appel du kernel. Cette boucle de répétition est nécessaire afin d'avoir une mesure précise : il se trouve que si on effectuait un seul appel au kernel, il se pourrait que les instructions de mesure (*RDTSC*) et de contrôle prennent autant ou plus de temps que le code de calcul, ce qui biaiserait complètement les résultats. Il faut donc faire plusieurs répétitions afin d'avoir un temps de calcul conséquent par rapport au reste des instructions.

3.2 Théorie

Lors de nos expériences, de nombreux paramètres peuvent être sujets à des variations aléatoires ou à des erreurs de mesure, ainsi un résultat peut être biaisé. Afin d'éviter cela, il est impératif d'utiliser une valeur représentative de nos différentes mesures : une valeur moyenne ou une valeur médiane. La valeur médiane semble être un meilleur candidat contrairement à la moyenne, car cette dernière peut-être fortement modifiée par une valeur extrême qui n'as pas lieu d'être. C'est donc la valeur médiane que nous prendrons des résultats de nos mesures successives issues des méta-répétitions.

Pour que chaque membre de l'équipe puisse travailler sur son niveau de mémoire cache, il nous faut trouver la taille des données d'entrée à utiliser. Selon le prototype présenté dans le Listing 1, la taille totale de nos données d'entrées est de $4n^2 + 8n$, avec n la taille entrée en paramètre du programme.

3.2.1 Cache L1

Ci-dessous les paramètres des mesures :

- (1) Taille pour tenir dans L1 : $n = 88$. En effet, $(n^2 * 4) + (n * 8) = 31680B$, sachant que la taille du cache est de $32kiB = 2^{15} = 32768B$, et que si l'on prend $n = 90$, on obtient $33120B$, on a bien : $n = 88 < L1 < n = 90$ avec une bonne marge de sécurité.
- (2) Nombre de répétitions du warm-up : 1000. Ce nombre est suffisant pour avoir ensuite des mesures stables dans tous nos tests car les caches sont remplis avec nos tableaux, choisi par observation. Il permet donc de passer le régime transitoire.
- (3) Nombre de répétitions des mesures : on choisit un nombre qui nous permet d'avoir un temps minimum représentatif par méta-répétition. Ce nombre varie en fonction de la taille de notre tableau. Avoir plus ou moins 1 seconde de mesures permet d'avoir une faible marge d'erreur de mesures des cycles avec un *RDTSC*. Par exemple, pour une taille de 89, on peut choisir 10000.

3.2.2 Cache L2

Pour que les deux tableaux entrent entièrement dans le cache L2, il faut que la formule respecte les contraintes suivantes :

- La taille totale doit être plus grande que la taille du cache L1 (1). Pour plus sécurité il a été décidé que la taille totale devait être au moins trois fois plus grande que celle du L1.
- L2 partage sa mémoire pour stocker à la fois les instructions, les données et ce qui tourne en background, on ne peut donc en utiliser approximativement que 90% (2).

Ces deux contraintes peuvent être transformées sous forme d'inéquation :

- (1) : $3 * TL1 < 4n * n + 8n$,
- (2) : $4n * n + 8n \leq 0,9 * TL2$

Après la résolution de ces équations on obtient $n = 156$ comme minimum et $n = 242$ comme maximum.

3.2.3 RAM

Voir le fichier « .odt » annexe envoyé par Camille.

3.2.4 Analyse de sensibilité

Une fois que l'on connaît la taille des données à fournir en entrée, il faut effectuer une analyse de sensibilité pour les autres paramètres.

Nombre de méta-répétition des mesures Il nous est donné à 31. C'est le nombre donné dans la consigne, qui est suffisant pour avoir un nombre de mesure significatives.

Le nombre de warmup Il doit se situer entre 1 et 1000. Pour le déterminer, il doit être le seul paramètre que l'on ait fait varier. On fait plusieurs exécutions et, avec les valeurs obtenues, on fait une courbe pour voir à partir de quelle valeur cela devient stable. Il faut aussi vérifier que $\frac{\text{médiane} - \text{minimum}}{\text{minimum}}$ est inférieur à 5%.

Le nombre de répétitions On le trouve de la même manière que le nombre de warmup.

3.3 Pratique

Lors de nos premières tentatives pour trouver les paramètres, nous avons remarqué que ce qui prenait le plus de temps dans notre noyau de calcul était l'exponentiel. Afin de pouvoir vérifier si nos paramètres sont corrects, nous avons donc modifié le fichier *kernel.c* pour que ce soit le temps de récupération des données qui soit le plus grand. Cette version n'est utilisée que pour tester la véracité des paramètres trouvés dans la section précédente, et non pas pour évaluer les performances des optimisations ou des compilateurs.

3.3.1 Cache L1

On peut observer une nette différence de performance entre un $n = 88$ et un $n = 90$, qui se traduit par le fait de tenir ou de ne pas tenir en cache *L1*. On assigne le programme de calcul au cœur n°2 de la machine avec *taskset*, sur son numéro de thread physique, dans le but de limiter les changements de contexte et de flush de la mémoire cache. Les tests sont effectués en rescue/failsafe mode, sans interface graphique, permettant d'avoir un minimum de tâches tournant en arrière-plan. De plus, le gouverneur du processeur est sélectionné sur le mode performance et la fréquence est fixée. Toutes les mesures effectuées dans le cache L1 sont automatisées par un script (*report/L1/P{1,2}/bench.sh*). Ce script permet d'aisément exécuter

l'ensemble des tests dans un environnement idéal, ainsi que d'assurer la reproductibilité du protocole expérimental, critère important d'une méthode scientifique. Les compilateurs utilisés sont *GCC v8.3.0*, *Clang v7.0.1*, *ICC v19.0.2.187*, tournants sur un *Debian v10 Buster (testing)* composé du noyau *Linux v4.19.0 x86_64*.

3.3.2 Cache L2

Avant de pouvoir exécuter et mesurer quelque chose il faut s'assurer que la machine n'utilise pas son temps de calcul ou d'accès à la mémoire pour autre chose que notre programme. Pour cela il faut passer en mode root qui n'utilise pas d'interface graphique et n'est pas connecté au réseau. On connecte la machine au secteur pour qu'elle ne se mette pas en mode économie d'énergie. Il faut ensuite fixer la fréquence à l'aide de *cpupower frequency-set* et enfin lorsque l'on compile le programme il faut aussi fixer le cœur afin d'éviter que le programme s'exécute sur plusieurs et qu'il faille transférer des données entre les différents cœurs ce qui ferait perdre du temps. Pour fixer le cœur on utilise *taskset*.

Pour vérifier le calcul théorique de la taille des données j'ai utilisé *likwid-perfctr* afin de voir si les données transitaient bien par le cache L2. Après avoir compilé avec *gcc* uniquement j'ai exécuté l'exécutable avec *likwid* et voici les résultats obtenus :

n	Data Volume (GByte)
100	4,24
150	14,06
220	37,9
235	34,4

On observe que ces résultats sont en corrélation avec les résultats théoriques : en dessous de 156 pratiquement aucune donnée ne passe par le cache L2, et quand on se rapproche de 242 une partie des données ne semble plus passer dans L2. Je suppose donc que ces données vont directement dans le cache L3. Au vu de ces informations, j'ai choisi de prendre 220 comme taille de données.

Voici ci-contre le graphique obtenu pour trouver le bon nombre de warmup. On peut observer que le nombre de cycles semble se stabiliser au alentour de 100 warmup. Pour plus de sécurité j'ai choisi 150 pour le nombre de warmup.

J'ai ensuite vérifié avec le calcul de l'exponentiel et on obtient bien le même résultat.

Pour trouver le bon nombre de répétitions j'ai uniquement fait les tests avec l'exponentiel. Comme vous pouvez le voir, on peut remarquer que l'ensemble est stable, les variations sont minimales. J'ai choisi comme nombre de répétitions 1200.

3.3.3 RAM

Voir le fichier « .odt » annexe envoyé par Camille.

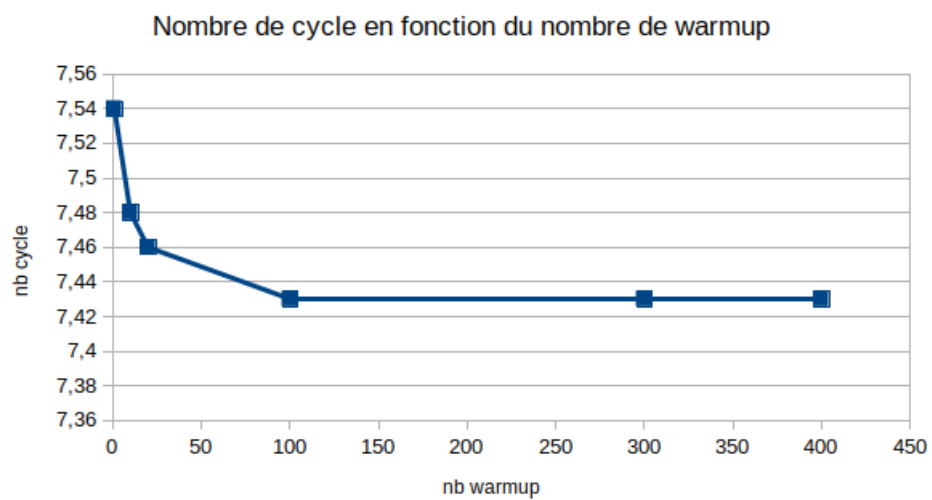


FIGURE 2 – Sans l'exponentiel

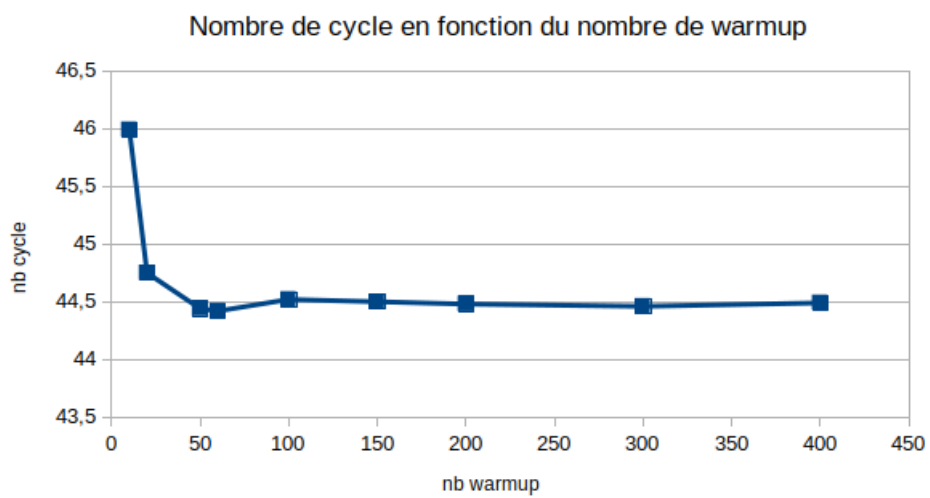


FIGURE 3 – Avec l'exponentiel

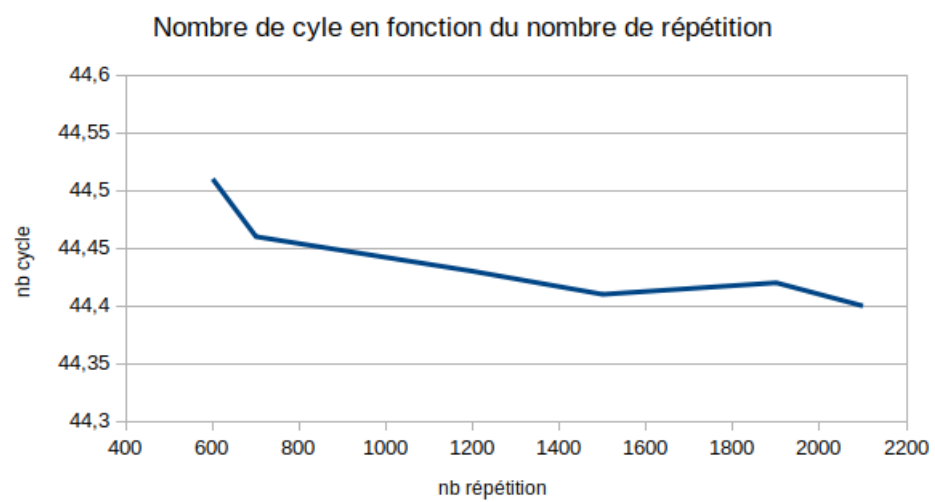


FIGURE 4

4 Optimisations et mesures

Dans cette section, nous présentons les résultats des mesures des différentes tentatives d’optimisation du code. La phase 1 correspond, pour résumer, à identifier les points chauds et tester différentes configurations de compilation. La phase 2 correspond, notamment, à une optimisation active du code en y apportant des modifications.

4.1 Phase 1

4.1.1 Cache L1

Pour cette première phase de test sur un jeu de donnée dans le niveau de cache L1, nous avons testé 3 compilateurs (« gcc », « icc », « clang ») avec différents jeux de flags de compilation. Nous précisons que l’intégralité de nos résultats sont disponibles sous formes brut dans les fichiers/répertoires suivants : « compil.txt », « likwid_{ref,opt} », « maqao_{ref,opt} ».

En plus des flags qui sont donnés dans la consigne, nous avons également testé le flag « -Ofast » qui permet d’activer des optimisations mathématiques qui ne respecte pas les standards en vigueur. Un programme qui ne requiert pas une stabilité numérique très précise obtiendra des gains considérable avec cette option, cependant, cela peut être dangereux de l’activer sans possibilité de vérifier les résultats des calculs du kernel. Nous avons ici pris le pari de l’activer.

Nous avons aussi testés d’autres options ciblées : fonctions inline, optimisations sur les boucles, sur les fonctions mathématiques ou encore sur les branchements. La liste ci-dessous présente les flags qui n’auront pas apporté de gain, ou pire, auront provoqué une accélération négative par rapport à « -Ofast -march=native » : « -faggressive-loop-optimizations », « -fbranch-probabilities », « -fdelayed-branch », « -fexpensive-optimizations », « -finline-functions », « -floop-block », « -floop-interchange », « -floop-unroll-and-jam », « -funsafe-math-optimizations ». Cependant, le flag « -funroll-all-loops », permettant de forcer l’unrolling des boucles, nous aura octroyé un léger gain systématique.

Dans la Table 1 est présenté la liste des résultats sur les flags obligatoires et les flags apportant un gain (les flags inutiles ou ralentissant ne sont pas inclus pour des soucis de visibilité). Nous pouvons ainsi voir que c’est « gcc », couplé à certaines options, qui est le plus rapide face à « clang » et « icc ». Nous notons tout de même l’efficacité redoutable de la génération de code spécialement pour l’architecture hôte (« -march=native »), permettant d’utiliser les instructions x86 les plus récentes, et des optimisations mathématiques agressives (« -Ofast »).

Nous avons ensuite utiliser les outils *MAQAO* et *LIKWID* pour expliquer les différences de performances entre deux versions du code. Après nos tests avec notre script permettant de détecter les flags permettant d’avoir le meilleur speed-up, nous allons étudier les différences de performances entre la version de référence « gcc -O2 »

Compiler	Flags	Time (s)
gcc	-O2	22.90
gcc	-O3	22.58
gcc	-Os	47.99
gcc	-O3 -march=native	22.77
gcc	-Ofast	9.59
gcc	-Ofast -march=native	4.64
gcc	-Ofast -march=native -funroll-all-loops	4.42
clang	-O0	30.38
clang	-O2	23.18
clang	-O3	26.04
clang	-Os	23.11
clang	-O3 -march=native	26.21
clang	-Ofast	18.18
clang	-Ofast -march=native	25.97
icc	-O0	23.05
icc	-O2	16.25
icc	-O3	16.27
icc	-Os	15.96
icc	-O3 -xHost	17.08
icc	-Ofast	16.25
icc	-Ofast -xHost	16.41
icc	-Ofast -xHost -funroll-loops -unroll-aggressive	16.39

TABLE 1 – Benchmarks des compilateurs et flags

et la version la plus rapide, « gcc -Ofast -march=native -funroll-all-loops ».

Procédons tout d’abord à une analyse rapide avec *LIKWID*, les résultats étant présentés dans la Figure 5. Nous pouvons expliquer la différence de performance par les métriques suivantes concernant la mémoire : on observe que la version optimisé à fait un nombre significativement moins important que la version de référence d’éviction de données du cache L1 (2.051.211 vs. 16.140.154), ainsi qu’un ratio de miss bien plus faible dans le cache L2 (0.0001 vs. 0.0259).

Enfin, passons à l’étude avec MAQAO. Sur le page « Global » présenté par la Figure 6, nous pouvons déjà avoir une très bonne idée des différences entre les deux binaires, justifiant d’une telle accélération ($\text{acc} = \frac{212}{41} = 5.2$). Premièrement, nous observons que sur le binaire optimisé, nous passons deux fois plus de temps dans la boucle que dans la version de référence : cela signifie que les fonctions mathématiques (multiplication, mais surtout l’exponentiel) ont été considérablement optimisées. Ensuite, nous voyons que la version de référence présente deux chemins (Flow Complexity) dans la boucle, tandis que la version optimisée ne présente qu’un chemin d’exécution possible. Nous notons aussi que l’efficacité d’accès aux données (Array Access Efficiency) a été augmenté de 20% dans la version optimisée, sûrement par modifications des boucles imbriquées. Enfin, nous pouvons imaginer une tentative de vectorisation de la part du compilateur pour la version optimisée.

likwid_opt.txt	likwid_ref.txt
<pre> 37 ----- 1 Group 1: L2 2 +-----+ 3 Event Counter Core 1 4 +-----+ 5 INSTR_RETIRED_ANY FIXC0 1052131496852 6 CPU_CLK_UNHALTED_CORE FIXC1 340267528996 7 CPU_CLK_UNHALTED_REF FIXC2 282970729197 8 L1D_REPLACEMENT PMC0 506082090 9 L1D_M_EVICT PMC1 16140154 10 ICACHE_64B_IFTAG_MISS PMC2 1585542 11 +-----+ 12 13 +-----+ 14 Metric Core 1 15 +-----+ 16 Runtime (RDTS) [s] 102.0052 17 Runtime unhaltd [s] 121.1787 18 Clock [MHz] 3376.5501 19 CPI 0.3234 20 L2D load bandwidth [MBytes/s] 317.5256 21 L2D load data volume [GBytes] 32.3893 22 L2D evict bandwidth [MBytes/s] 10.1266 23 L2D evict data volume [GBytes] 1.0330 24 L2 bandwidth [MBytes/s] 328.6470 25 L2 data volume [GBytes] 33.5237 26 +-----+ 27 28 Group 2: L2CACHE 29 +-----+ 30 Event Counter Core 1 31 +-----+ 32 INSTR_RETIRED_ANY FIXC0 1035367651370 33 CPU_CLK_UNHALTED_CORE FIXC1 334016970889 34 CPU_CLK_UNHALTED_REF FIXC2 278504062902 35 L2_TRANS_ALL_REQUESTS PMC0 1034707587 36 L2_RQSTS_MISS PMC1 26790341 37 +-----+ 38 39 +-----+ 40 Metric Core 1 41 +-----+ 42 Runtime (RDTS) [s] 100.4011 43 Runtime unhaltd [s] 119.2732 44 Clock [MHz] 3376.7572 45 CPI 0.3235 46 L2 request rate 0.0010 47 L2 miss rate 2.587520e-05 48 L2 miss ratio 0.0259 49 +-----+ 50 </pre>	<pre> 37 ----- 1 Group 1: L2 2 +-----+ 3 Event Counter Core 1 4 +-----+ 5 INSTR_RETIRED_ANY FIXC0 175026938120 6 CPU_CLK_UNHALTED_CORE FIXC1 67198687447 7 CPU_CLK_UNHALTED_REF FIXC2 55629506790 8 L1D_REPLACEMENT PMC0 636961928 9 L1D_M_EVICT PMC1 2051211 10 ICACHE_64B_IFTAG_MISS PMC2 74656 11 +-----+ 12 13 +-----+ 14 Metric Core 1 15 +-----+ 16 Runtime (RDTS) [s] 19.8344 17 Runtime unhaltd [s] 23.9312 18 Clock [MHz] 3391.9663 19 CPI 0.3839 20 L2D load bandwidth [MBytes/s] 2055.2976 21 L2D load data volume [GBytes] 40.7656 22 L2D evict bandwidth [MBytes/s] 6.6187 23 L2D evict data volume [GBytes] 0.1313 24 L2 bandwidth [MBytes/s] 2062.1572 25 L2 data volume [GBytes] 40.9016 26 +-----+ 27 28 Group 2: L2CACHE 29 +-----+ 30 Event Counter Core 1 31 +-----+ 32 INSTR_RETIRED_ANY FIXC0 158843843600 33 CPU_CLK_UNHALTED_CORE FIXC1 60994318553 34 CPU_CLK_UNHALTED_REF FIXC2 50493481857 35 L2_TRANS_ALL_REQUESTS PMC0 846618030 36 L2_RQSTS_MISS PMC1 70863 37 +-----+ 38 39 +-----+ 40 Metric Core 1 41 +-----+ 42 Runtime (RDTS) [s] 18.0006 43 Runtime unhaltd [s] 21.7217 44 Clock [MHz] 3391.9545 45 CPI 0.3840 46 L2 request rate 0.0053 47 L2 miss rate 4.461174e-07 48 L2 miss ratio 0.0001 49 +-----+ 50 </pre>

FIGURE 5 – A gauche, la version de référence. A droite, la version avec les flags d’optimisation.

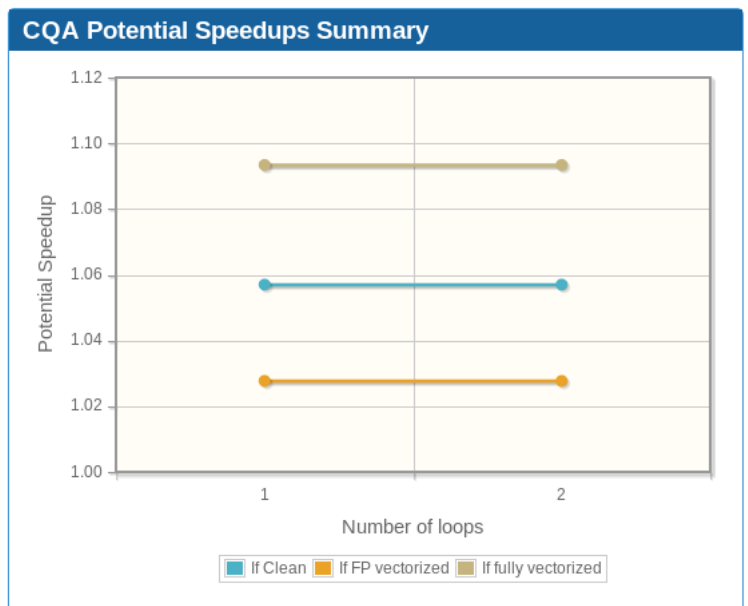
Pour finir avec l’analyse de *MAQAO*, sur la Figure 7, nous pouvons observer ce qu’à concrètement fait le compilateur. La fonction exponentielle, qui prenait 14% du temps, à été remplacé par une version optimisée « fini », ne prenant plus que 1.46% du temps. Nous pouvons voir que le linkage de la bibliothèque mathématique (*libm*) a été remplacé par sa version vectorisée (*libmvec*). Enfin, nous observons que notre unique boucle à bien été déroulée car nous trouvons l’ajout d’une *tail loop*.

4.1.2 Cache L2

On remarque avec ces résultats que *icc* rend l’exécution plus rapide que *gcc*, néanmoins, les différentes options d’optimisation n’ont pas un impact considérable contrairement à *gcc* qui entre O2 et O3 gagne quelques secondes.

À l’aide de *LIKWID*, on peut observer que le volume de data est différent entre l’exécution avec *icc* et *gcc*, il est d’environ 37 GBytes avec *gcc* et 7 GBytes avec *icc*. Le miss rate est équivalent entre les deux compilateur.

Global Metrics ?		
Total Time (s)		212.11
Time in loops (%)		10.79
Compilation Options		binary: -funroll-loops is missing.
Flow Complexity		2.00
Array Access Efficiency (%)		50.00
Clean	Potential Speedup	1.06
	Nb Loops to get 80%	1
FP Vectorised	Potential Speedup	1.03
	Nb Loops to get 80%	1
Fully Vectorised	Potential Speedup	1.09
	Nb Loops to get 80%	1



Global Metrics ?		
Total Time (s)		41.03
Time in loops (%)		20.11
Compilation Options		OK
Flow Complexity		1.00
Array Access Efficiency (%)		31.99
Clean	Potential Speedup	1.00
	Nb Loops to get 80%	1
FP Vectorised	Potential Speedup	1.00
	Nb Loops to get 80%	1
Fully Vectorised	Potential Speedup	1.08
	Nb Loops to get 80%	1

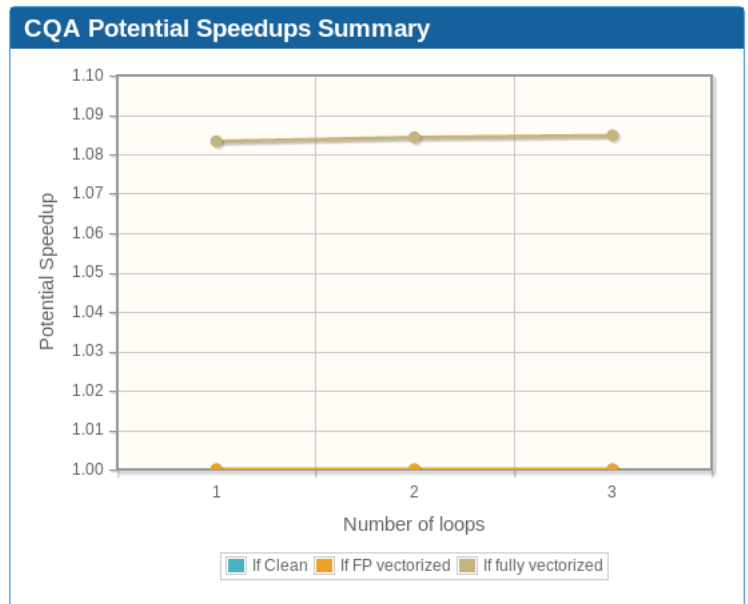


FIGURE 6 – Au-dessus, la version de référence. En dessous, la version avec les flags d’optimisation.

Functions and Loops ?						
► Filters ?						
Name	Module	Coverage (%)	Time (s)	Nb Threads	Deviation (coverage)	
○ f64xsubf128	libm-2.28.so	63.65	135.01	1	0.00	
○ exp	libm-2.28.so	14.61	30.99	1	0.00	
▼ baseline	baseline	10.88	23.07	1	0.00	
▼ Loop 7 - kernel.c:13-19 - baseline		10.79	22.89			
○ Loop 6 - kernel.c:16-19 - baseline		10.79	22.89			

Functions and Loops ?						
► Filters ?						
Name	Module	Coverage (%)	Time (s)	Nb Threads	Deviation (coverage)	
○ _ZGVeN16vvv_sincosf	libmvec-2.28.so	74.29	30.48	1	0.00	
▼ baseline	baseline	24.08	9.88	1	0.00	
▼ Loop 8 -- baseline		20.06	8.23			
▼ Loop 6 -- baseline		19.84	8.14			
○ Loop 7 -- baseline		19.84	8.14			
○ Loop 9 -- baseline		0.22	0.09			
○ _ZGVdN4v__exp_finite	baseline	1.46	0.6	1	0.00	

FIGURE 7 – Au-dessus, la version de référence. En dessous, la version avec les flags d’optimisation.

Numéro	Commande	Résultat (médiane)
1	gcc -O2	39,93
2	gcc -O3	37,62
3	gcc -O3 -march=native	37,79
4	icc -O2	17,79
5	icc -O3	17,77
6	icc -O3 -xHost	17,74

4.1.3 RAM

Voir le fichier « .odt » annexe envoyé par Camille.

4.2 Phase 2

La phase 2 constituait en un exercice d'application d'optimisation, en ceci que nous avons recommencé à utiliser notre version du programme par défaut (*gcc -O2*) pour appliquer des optimisations manuelles au niveau du code source. Dans les sections suivantes, nous allons vous présenter, pour chaque niveau de mémoire étudiée, les optimisations ayant été faites. Pour des raisons de visibilité, le code ne sera pas inclus dans ce rapport : toutes les optimisations vont visibles dans le fichier « *kernel.c* ».

4.2.1 Cache L1

La Figure 8 nous permet de comparer les différents temps d'exécution des différentes optimisations. Nous observons d'emblée qu'une optimisation nous fait perdre des performances (OPT-OMP), tandis que les autres optimisations apportent des améliorations minimales. Enfin, on observe qu'à l'optimisation de la fonction exponentielle (OPT-EXP), les gains de performances sont très nets. Nous allons détailler les optimisations dans les prochains paragraphes. On remarquera qu'il n'y a pas d'analyse des rapports de LIKWID dans les sections suivantes : en effet, une fois en L1, la mémoire n'est pas un facteur limitant des performances de notre code comparé aux calculs. Par conséquent, LIKWID ne nous a pas été utile à ce niveau.

Accès séquentiel au tableau (OPT-LOOP-STRIDE-1) MAQAO nous indique, dans la Figure 9, que l'efficacité d'accès aux tableaux est mauvais (50% d'efficacité). En inversant la boucle intérieure et extérieure, on améliore l'efficacité à 58%, comme illustré par la Figure 10. Si on enlève en plus le *if* (une autre optimisation), cette efficacité passe alors à 75%. Cependant, malgré cette meilleure efficacité, le gain de performances est négligeable : l'accès à la mémoire n'est pas le facteur limitant de notre kernel.

Suppression du *if* (OPT-IF-HOISTING) La Figure 11 nous montre que le fait d'avoir deux chemins possible dans la boucle limite les performances, de plus MAQAO nous indique un *Flow Complexity* de 2. Une fois le *if* isolé dans une boucle appart qui se fait avant la boucle principale, nous avons un *Flow Complexity* de 1 et plus d'avertissement concernant de multiples chemins à l'intérieur de la boucle. On observe un très léger gain au niveau des performances, mais encore une fois, deux chemins possibles (empêchant un pipelining optimale) n'est pas un facteur très limitant par rapport au reste.

Déroulage de boucle (OPT-LOOP-UNROLLING) MAQAO nous indiquait une possibilité d'unroller notre boucle si notre nombre d'itération est significativement plus élevée que le facteur d'unrolling, ce qui est le cas ici (Figure 12). Une fois l'unrolling de la boucle intérieure effectué avec un facteur de 2, nous observons un

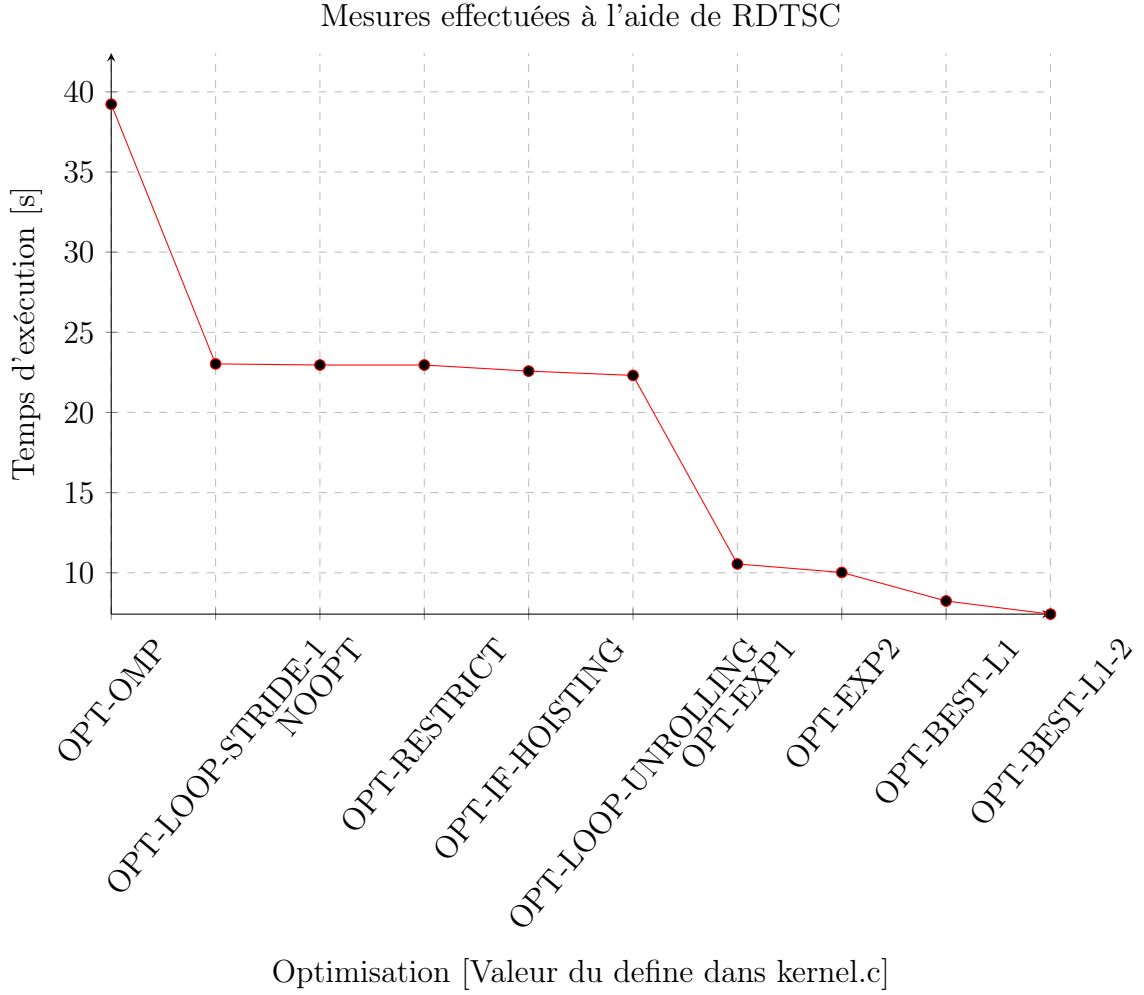


FIGURE 8 – Temps d'exécution en fonction de l'optimisation

très léger gain de performances puisque l'on peut commencer à effectuer la prochaine itération avant la fin de la première. Encore une fois, le gain n'est pas significatif.

Pointeurs restreints (OPT-RESTRICT) Nous avons utilisé le mot-clé *restrict* pour indiquer que les pointeurs pointaient sur des zones de mémoire distinctes et indépendantes entre elles. Cela permet notamment au compilateur de plus facilement vectoriser le code et ré-arranger l'ordre des instructions, puisqu'il y a moins de « précaution » à prendre. Ici, c'est une optimisation extrêmement légère, mais qui peut parfois révéler de bonnes surprises.

Utilisation de l'exponentielle sur un float (OPT-EXP-1) La bibliothèque standard propose plusieurs versions de l'exponentielle, dont une pour un double (par défaut) et une pour un float. Celle pour un float est naturellement moins précise, mais permet en revanche d'être extrêmement plus rapide. En effet, MAQAO nous indiquait que la majorité du temps de notre kernel était passé dans la bibliothèque de

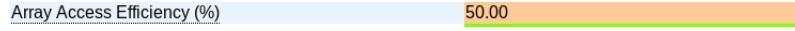


FIGURE 9 – Rapport de MAQAO sur NOOPT : accès non-stride 1.

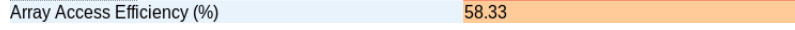


FIGURE 10 – Rapport de MAQAO sur OPT_LOOP_STRIDE_1 : augmentation de l'efficacité.

mathématiques, voir la Figure 13. En utilisant la fonction exponentielle sur un float, on améliore significativement les performances : on obtient un speed-up d'environ 2.3.

Utilisation d'une fonction exponentielle custom (OPT-EXP-2) Afin d'améliorer la rapidité de la fonction exponentielle, nous avons décidé de programmer notre propre version. Cette version est précise pour un x faible (< 5), et permet d'obtenir un speed-up considérable par rapport à l'exponentielle originale. Par rapport à l'exponentielle sur float, elle apporte un petit gain de performances. Nous savons qu'utiliser cette approximation n'est pas un problème puisque les valeurs de notre tableau a sont comprises entre -1 et 1. L'explication de l'optimisation est la suivante. La fonction exponentielle est définie par la série entière suivante :

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots$$

En appliquant le développement de série entière puis la formule du binôme de Newton, on peut exprimer la fonction exponentielle de cette manière :

$$\exp(x) = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n$$

L'astuce consiste à calculer cette limite pour un petit n , tout en gardant une bonne approximation : cela permet d'avoir un calcul rapide. Nous avons choisi $n = 256$, qui permet d'avoir une bonne approximation pour un $x < 5$, ce qui nous convenait. Avec un $n = 256$, il nous suffisait de multiplier la valeur $1 + x/256$ par elle-même 8 fois, puisque $x^{256} = x^{2^{128}} = x^{2^{2^{64}}} = \dots$. Effectuer cette multiplication 8 fois est très rapide et ne nécessite pas de boucle. Enfin, nous avons inliné la fonction pour supprimer le coût de l'appel.

Parallélisation de la boucle avec OpenMP (OPT-OMP) Notre boucle ne présentant pas de dépendances inter-itérations, il est évident d'effectuer une parallélisation. Cependant, n'ayant pas un gros volume de données à traiter puisque nous sommes en L1, le surcoût de la parallélisation a considérablement ralenti le programme, ce qui était prévisible. OpenMP n'est intéressant que pour la parallélisation massive d'un traitement de gros volumes de données sur des clusters de calcul, pas pour dispatcher des petits traitements sur un petit processeur d'ordinateur de bureau ne possédant que 4 cœurs software.

The structure of this loop is probably <if then [else] end>.

The presence of multiple execution paths is typically the main/first bottleneck.

Try to simplify control inside loop: ideally, try to remove all conditional expressions, for example by (if applicable):

- hoisting them (moving them outside the loop)
- turning them into conditional moves, MIN or MAX

FIGURE 11 – Rapport de MAQAO sur NOOPT : chemins multiples.

Unroll opportunity
Loop is data access bound.
Workaround
Unroll your loop if trip count is significantly higher than target unroll factor and if some data references are common to consecutive iterations. This can be done manually. Or by recompiling with -funroll-loops and/or -floop-unroll-and-jam.

FIGURE 12 – Rapport de MAQAO sur NOOPT : possibilité d'unrolling.

Conversion de type Comme exposé dans la Figure 14, MAQAO nous informe qu'une instruction complexe est exécuté une fois dans notre boucle, ce qui réduit les performances. Cette instruction, *CVTSS2SD*, convertit un *float* vers un *double* au moment de stocker le résultat de l'exponentielle. N'utiliser que des *doubles* n'apportait pas de performances, car l'exponentielle sur un *double* est plus longue qu'une exponentielle sur un *float*. Enfin, n'utiliser que des *floats* apportait un léger gain de performances, mais entraînait une perte de précision sur les très petites valeurs stockées dans *b*. Par conséquent, nous avons délibérément choisis de ne pas appliquer cette optimisation.

Toutes les optimisations (OPT-BEST-L1) Toutes les optimisations (sauf la conversion de type et la parallélisation, n'apportant aucun gain) ont été réunies dans un seul code. Aucune modification supplémentaire n'as été apporté par rapport aux optimisations décrites au-dessus. On observe au final un gain non-négligeable par rapport à la version originale (speed-up de 2.9).

Toutes les optimisations avec flags (OPT-BEST-L1-2) Ce benchmark correspond à la version du code étant la mieux optimisée, mais mesurée avec les options de compilation donnant le meilleur résultat dans la phase 1. On observe qu'on retrouve un léger gain par rapport à cette version du code sans ces options de compilation (paragraphe précédent), mais qu'elle est moins rapide que la version générée dans la phase 1, sans modification du code : en effet, on obtient 7.5 cycles par itération sur cette version modifiée, contrairement à 4.5 cycles par itération sur la version originale. On en conclue que sur ce code, le compilateur aura fait mieux que nous. La différence de performance s'explique par le fait que nous n'avons pas tiré parti de la vectorisation, et que nous forçons le compilateur à utiliser notre approximation de l'exponentielle. En revanche, en lui laissant le libre arbitre, il a pu utiliser (dans la version originale du code) une version vectorisée de l'exponentielle, ce qui fait toute la différence! Cette différence n'est pas visible dans le code assembleur pour la fonction exponentielle car c'est un appel de fonction, elle n'a pas été inlineé (*call exp_finite*). Cependant, MAQAO détecte très bien la différence de vectorisation,

Name	Module	Coverage (%)	Time (s)	Nb Threads	Deviation (coverage)
o f64xsubf128	libm-2.28.so	64.2	13.72	1	0.00
o exp	libm-2.28.so	14.27	3.05	1	0.00

FIGURE 13 – Rapport de MAQAO sur NOOPT : fonction exponentielle limitante.

Conversion instructions
Detected expensive conversion instructions (typically from/to single/double precision).
Details
<ul style="list-style-type: none"> • CVTSS2SD: 1 occurrences
Workaround
Use double instead of single precision only when/where needed by numerical stability and avoid mixing data with different types. In particular, check if the type of constants is the same as array elements. In C/C++, FP constants are double precision by default and must be suffixed by 'f' to make them single precision.

FIGURE 14 – Rapport de MAQAO sur NOOPT : instruction complexe détectée.

comme illustré sur la Figure 15 qui nous montre le degré de vectorisation du code originale avec les bonnes options de compilation et la Figure 16 qui nous montre le degré de vectorisation du code optimisé. En effet, on voit que dans notre code, nous pourrions obtenir un gain considérable en vectorisant notre fonction exponentielle et notre boucle : ici se trouve le goulet d'étranglement de notre programme.

Vectorisation Comme identifier dans les paragraphes précédents, pour obtenir un gros speed-up, il nous fallait vectoriser notre code. Nous avons alors commencé à tenter une vectorisation en utilisant des *intrinsics*. Nous avons aussi analysé la version d'exponentielle vectorisée d'Agner Fog dans son manuel d'optimisation. Cependant, nous n'avons pas réussi à venir à terme de notre code : entre les *intrinsics* de GCC et d'ICC, entre les différents jeux d'instructions SIMD (SEE/2/3/4/4, AVX/512), entre les codes pour le C et ceux pour le C++, et enfin entre les versions customs de fonction vectorisée ou l'utilisation de bibliothèques mathématiques vectorisées (VML, IPP, Sleef, SVML, VCL...), nous nous sommes un peu perdus. Je pense que nous aurions dû commencer par vectoriser un code plus simple pour nous entraîner, peut-être directement en assembleur pour avoir une compréhension maximale. Quoi qu'il en soit, à cette heure, nous n'avons plus le temps de retenter l'expérience : mais nous savons comment nous lancer et quel chemin suivre pour obtenir un speed-up maximal.

4.2.2 Cache L2

4.2.3 RAM

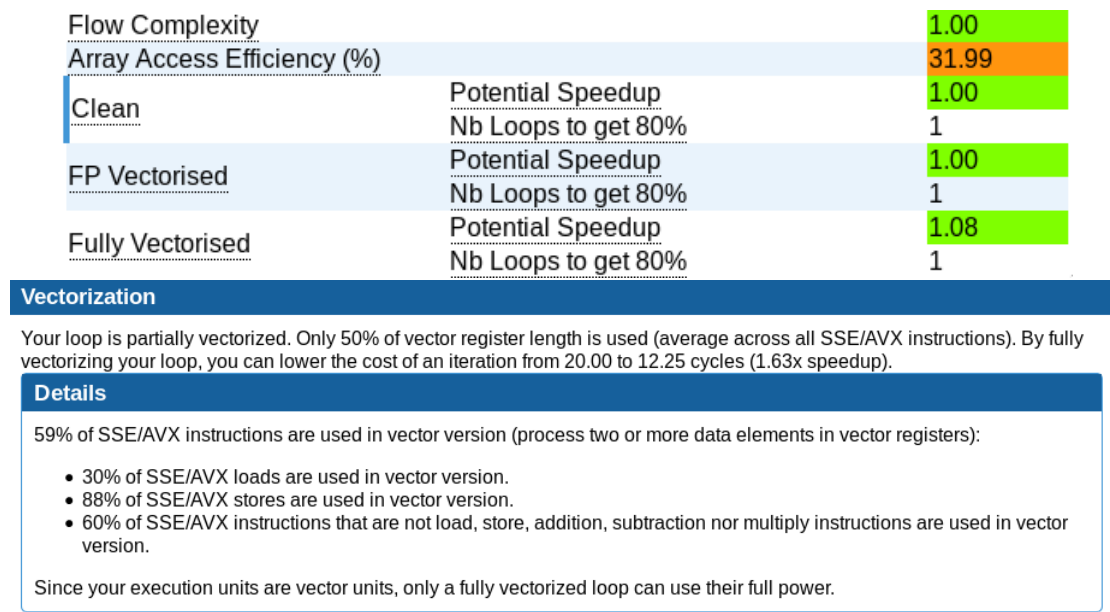


FIGURE 15 – Rapport de MAQAO sur NOOPT : vectorisation avec les bons flags

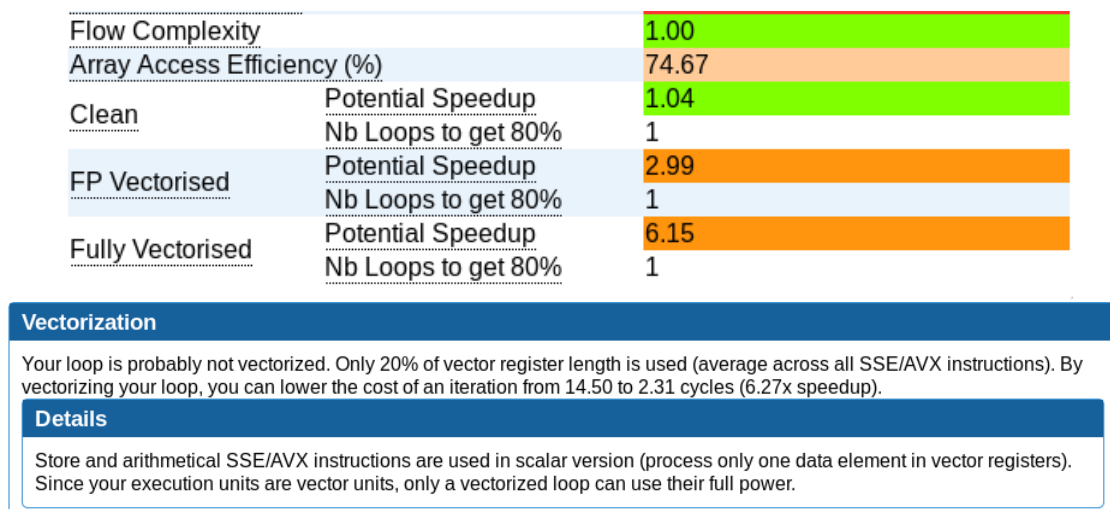


FIGURE 16 – Rapport de MAQAO sur BEST-L1 : vectorisation

5 Conclusion

Pour conclure, nous aurons tout de même réussis à apporter des optimisations que nous jugeons intéressantes à notre kernel. Au départ, le temps de calcul d'une itération de notre boucle (une fonction exponentielle, une multiplication et une conversion de type float vers double) était d'environ 23s. Après nos optimisations manuelles, une itération du kernel prenait environ 8s, soit presque 3 fois plus rapide ! Malheureusement, nous ne sommes pas parvenus à atteindre la rapidité de calcul obtenue par les meilleures options du compilateur, à savoir 4,5s. Nous pensons qu'avec la vectorisation de notre fonction exponentielle et de notre boucle intérieure du kernel, cela aurait pu être possible de faire mieux ! Nous aurons été confrontés à de nombreux petits soucis épars pendant notre démarche, classique pour des étudiants découvrant les outils de mesures de performances, résultants parfois à 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 la rigueur de mesure apprise en TD, détaillée dans le rapport.

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 réelle les techniques et sujets vus en cours. Une très bonne première approche !