

浙江大学



《手写SIMD向量化》

lab2 实验报告

题 目：	手写SIMD向量化
上课时间：	2023暑假
授课教师：	ZJUSCT
姓 名：	杜宗泽
学 号：	3220105581
组 别：	个人
日 期：	7月9日

标题

- 1 Lab Description
- 2 Introduction Knowledge(可以跳过不看)
 - 2.1 SIMD
 - 2.2 向量化指令
 - 2.3 汇编语言
- 3 Lab Design & Test Result
 - 3.1 Code Design
 - 3.2 汇编比较
 - 3.2.1 SIMD的汇编
 - 3.2.2 普通的汇编
 - 3.2.3 源代码中普通表达式的汇编
 - 3.3 实验结果
- 4 Discussion

1 Lab Description

使用这些函数 API 需要 include 对应的头文件，不同 SIMD 指令集需要的头文件不同，具体需要参考 Intel 相关文档。

```
1 #include <smmintrin.h>
2 #include <emmintrin.h>
3 #include <immintrin.h>
```

本次实验的实验目标：通过使用手写 SIMD 向量化的方式对这个循环进行优化

```
1 /* 可以修改的代码区域 */
2 // -----
3 for (int i = 0; i < MAXN; ++i)
4 {
5     c[i] += a[i] * b[i];
6 }
7 // -----
8
```

详情请看[实验指导](#)

2 Introduction Knowledge(可以跳过不看)

2.1 SIMD

单指令流多数据流（英语：Single Instruction Multiple Data，[缩写](#)：SIMD）是一种采用一个[控制器](#)来控制多个[处理器](#)，同时对一组数据（又称“[数据向量](#)”）中的每一个分别执行[相同](#)的操作从而实现空间上的[并行性](#)的技术。

在[微处理器](#)中，单指令流多数据流技术则是一个[控制器](#)控制多个平行的[处理微元](#)，例如Intel的[MMX](#)或[SSE](#)，以及AMD的[3D Now!](#)指令集。

[图形处理器](#)（GPU）拥有强大的并发处理能力和可编程流水线，面对单指令流多数据流时，运算能力远超传统CPU。[OpenCL](#)和[CUDA](#)分别是目前最广泛使用的开源和专利[通用图形处理器](#)（GPGPU）运算语言。

2.2 向量化指令

现代处理器一般都支持向量化指令，x86 架构下 Intel 和 AMD 两家的处理器都提供了诸如 SSE，AVX 等 SIMD 指令集，一条指令可以同时操作多个数据进行运算，大大提高了现代处理器的数据吞吐量。

现代编译器在高优化等级下，具有自动向量化的功能，对于结构清晰，循环边界清晰的程序，编译器的自动向量化已经可以达到很优秀的程度了。然而，编译器的优化始终是保守的，很多情况下编译器无法完成使用 SIMD 指令进行向量化的工作，为了追求性能，高性能计算领域经常需要手写 SIMD 代码进行代码优化。

显然直接手写汇编指令过于困难，在 C 语言环境下，Intel 提供了一整套关于 SIMD 指令的函数封装接口和指令相关行为的参照手册，可以在[参考资料](#)中找到。

2.3 汇编语言

汇编语言（英语：assembly language）[\[注 1\]\[1\]](#)是任何一种用于[电子计算机](#)、[微处理器](#)、[微控制器](#)，或其他可编程器件的[低级语言](#)。在不同的设备中，汇编语言对应着不同的[机器语言指令集](#)。一种汇编语言专用于某种[计算机系统结构](#)，而不像许多[高级语言](#)，可以在不同系统平台之间移植。

使用汇编语言编写的源代码，然后通过相应的汇编程序将它们转换成可执行的机器代码。这一过程被称为[汇编过程](#)。

关于汇编：

1. 学习内容十分推荐去看[CSAPP](#)
2. 有个在线的[编译器](#)能够有助于你理解!
3. 当然我也建议大家可以在自己本地的电脑上通过gcc的一些参数选项自己来了解汇编。

3 Lab Design & Test Result

3.1 Code Design

My Code:

```
1  for (int n = 0; n < 20; ++n)
2  {
3      /* 可以修改的代码区域 */
4      // -----
5      __m256 vecA, vecB, vecC, vecAB;
6      for (int i = 0; i < MAXN; i += 8)
7      {
8          vecA = _mm256_loadu_ps(a + i);
9          vecB = _mm256_loadu_ps(b + i);
10         vecC = _mm256_loadu_ps(c + i);
11
12         // vecAB = _mm256_mul_ps(vecA, vecB);
13         // vecC = _mm256_add_ps(vecAB, vecC);
14         vecC = _mm256_fmadd_ps(vecA, vecB, vecC);
15         _mm256_storeu_ps(c + i, vecC);
16     }
17 }
```

ps: 这段代码也没有考虑内存对齐问题，如果你的数据是对齐的，你当然可以使用 `_mm256_load_ps` 和 `_mm256_store_ps` 来提高性能。

3.2 汇编比较

3.2.1 SIMD的汇编

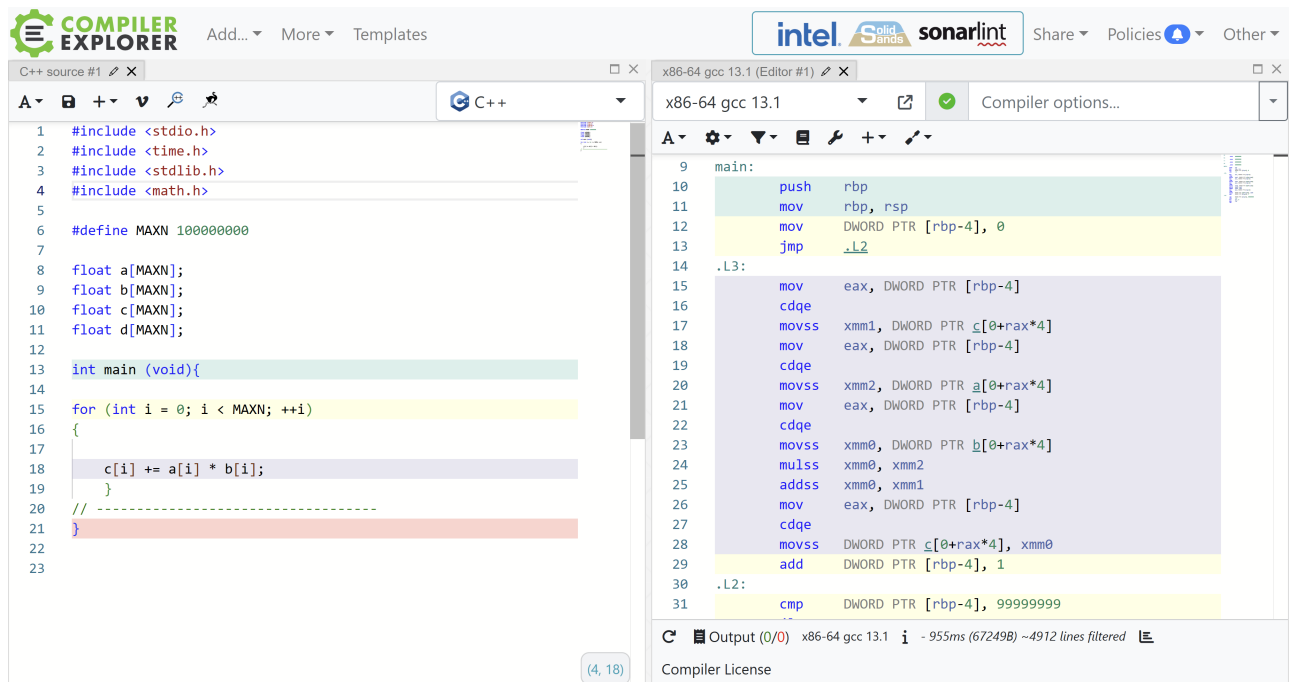
```
C++ source #1
11 float c[MAXN];
12 float d[MAXN];
13
14 int main (void){
15     for (int n = 0; n < 20; ++n)
16     {
17         /* 可以修改的代码区域 */
18         // -----
19         __m256 vecA, vecB, vecC, vecAB;
20         for (int i = 0; i < MAXN; i += 8)
21         {
22             vecA = _mm256_loadu_ps(a + i);
23             vecB = _mm256_loadu_ps(b + i);
24             vecC = _mm256_loadu_ps(c + i);
25
26             vecC = _mm256_fmadd_ps(vecA, vecB, vecC);
27             _mm256_storeu_ps(c + i, vecC);
28         }
29     }
30     for (int i = 0; i < MAXN; ++i)
31     {
32         c[i] += a[i] * b[i];
33     }
34 }
35 // -----
36 }
37 }
38
```

```
x86-64 gcc 13.1 (Editor #1)
x86-64 gcc 13.1 -march=native -mavx
16 .L11:
17     mov     DWORD PTR [rsp+192], 0
18     jmp     .L3
19 .L8:
20     mov     eax, DWORD PTR [rsp+192]
21     cdqe
22     sal     rax, 2
23     add     rax, OFFSET FLAT:a
24     mov     QWORD PTR [rsp+112], rax
25     mov     rax, QWORD PTR [rsp+112]
26     vmovaps ymm0, YMMWORD PTR [rax]
27     vmovaps YMMWORD PTR [rsp+136], ymm0
28     mov     eax, DWORD PTR [rsp+192]
29     cdqe
30     sal     rax, 2
31     add     rax, OFFSET FLAT:b
32     mov     QWORD PTR [rsp+104], rax
33     mov     rax, QWORD PTR [rsp+104]
34     vmovaps ymm0, YMMWORD PTR [rax]
35     vmovaps YMMWORD PTR [rsp+104], ymm0
36     mov     eax, DWORD PTR [rsp+192]
37     cdqe
38     sal     rax, 2
39     add     rax, OFFSET FLAT:c
40     vfmadd231ps ymm0, ymm0, YMMWORD PTR [rsp+24]
41     nop
42     vmovaps YMMWORD PTR [rsp+72], ymm0
43     mov     eax, DWORD PTR [rsp+192]
44     cdqe
45     sal     rax, 2
46     add     rax, OFFSET FLAT:c
47     mov     QWORD PTR [rsp+64], rax
48     vmovaps ymm0, YMMWORD PTR [rsp+72]
49     vmovaps YMMWORD PTR [rsp+8], ymm0
50     vmovaps YMMWORD PTR [rsp+56], ymm0
51     vmovaps ymm1, YMMWORD PTR [rsp+88]
52     vfmadd231ps ymm0, ymm1, YMMWORD PTR [rsp+24]
53     nop
54     vmovaps YMMWORD PTR [rsp+72], ymm0
55     mov     eax, DWORD PTR [rsp+192]
56     cdqe
57     sal     rax, 2
58     add     rax, OFFSET FLAT:c
59     mov     QWORD PTR [rsp+64], rax
60     vmovaps ymm0, YMMWORD PTR [rsp+72]
61     vmovaps YMMWORD PTR [rsp+8], ymm0
62     vmovaps ymm0, YMMWORD PTR [rsp+8]
63     mov     rax, QWORD PTR [rsp+64]
64     vmovups YMMWORD PTR [rax], ymm0
65     nop
66     add     DWORD PTR [rsp+192], 8
67 .L3:
68     cmp     DWORD PTR [rsp+192], 99999999
69     jle     .L8
70     mov     DWORD PTR [rsp+188], 0
71     jmp     .L11
Output (0/0) x86-64 gcc 13.1 i - 1019ms (77856B) ~5480 lines filtered
Compiler License
```

```
C++ source #1
11 float c[MAXN];
12 float d[MAXN];
13
14 int main (void){
15     for (int n = 0; n < 20; ++n)
16     {
17         /* 可以修改的代码区域 */
18         // -----
19         __m256 vecA, vecB, vecC, vecAB;
20         for (int i = 0; i < MAXN; i += 8)
21         {
22             vecA = _mm256_loadu_ps(a + i);
23             vecB = _mm256_loadu_ps(b + i);
24             vecC = _mm256_loadu_ps(c + i);
25
26             vecC = _mm256_fmadd_ps(vecA, vecB, vecC);
27             _mm256_storeu_ps(c + i, vecC);
28         }
29     }
30     for (int i = 0; i < MAXN; ++i)
31     {
32         c[i] += a[i] * b[i];
33     }
34 }
35 // -----
36 }
37 }
38
```

```
x86-64 gcc 13.1 (Editor #1)
x86-64 gcc 13.1 -march=native -mavx
48     vmovaps ymm0, YMMWORD PTR [rsp+72]
49     vmovaps YMMWORD PTR [rsp+88], ymm0
50     vmovaps ymm1, YMMWORD PTR [rsp+56]
51     vmovaps ymm0, YMMWORD PTR [rsp+88]
52     vfmadd231ps ymm0, ymm1, YMMWORD PTR [rsp+24]
53     nop
54     vmovaps YMMWORD PTR [rsp+72], ymm0
55     mov     eax, DWORD PTR [rsp+192]
56     cdqe
57     sal     rax, 2
58     add     rax, OFFSET FLAT:c
59     mov     QWORD PTR [rsp+64], rax
60     vmovaps ymm0, YMMWORD PTR [rsp+72]
61     vmovaps YMMWORD PTR [rsp+8], ymm0
62     vmovaps ymm0, YMMWORD PTR [rsp+8]
63     mov     rax, QWORD PTR [rsp+64]
64     vmovups YMMWORD PTR [rax], ymm0
65     nop
66     add     DWORD PTR [rsp+192], 8
67 .L3:
68     cmp     DWORD PTR [rsp+192], 99999999
69     jle     .L8
70     mov     DWORD PTR [rsp+188], 0
71     jmp     .L11
Output (0/0) x86-64 gcc 13.1 i - 1019ms (77856B) ~5480 lines filtered
Compiler License
```

3.2.2 普通的汇编



The screenshot shows the Visual Studio Code interface with two panels. The left panel displays a C++ source file with the following code:

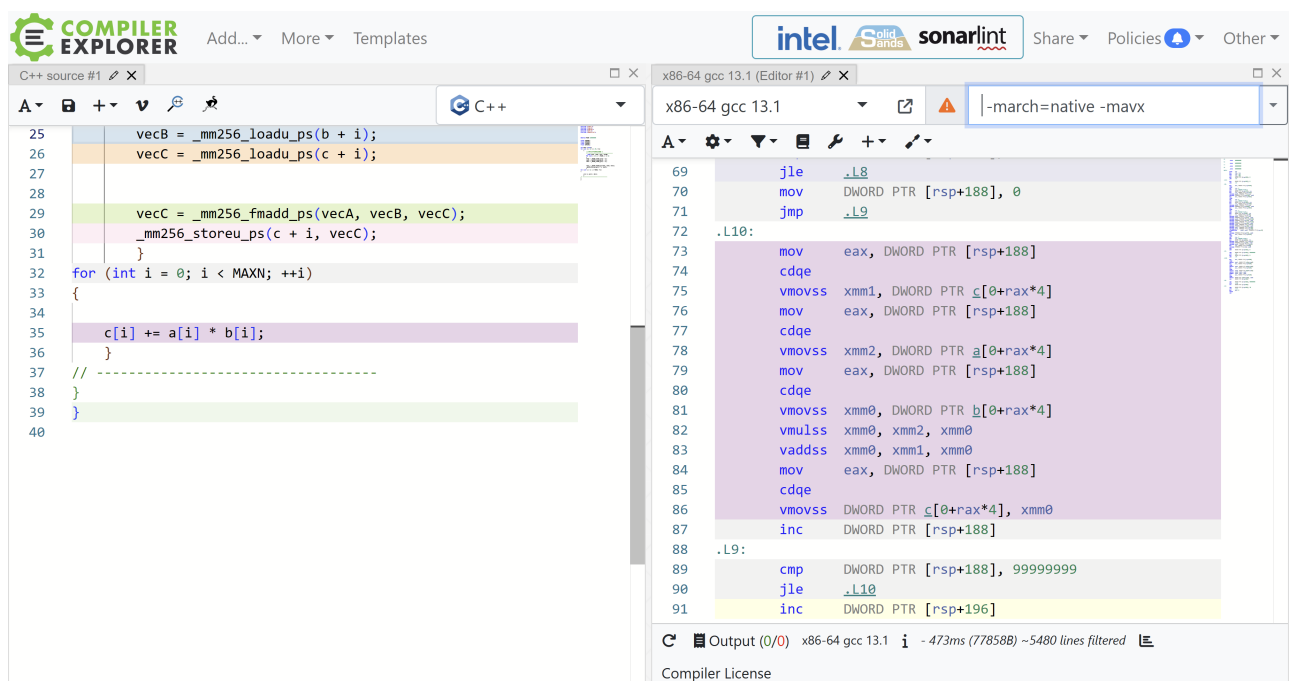
```
1 #include <stdio.h>
2 #include <time.h>
3 #include <stdlib.h>
4 #include <math.h>
5
6 #define MAXN 100000000
7
8 float a[MAXN];
9 float b[MAXN];
10 float c[MAXN];
11 float d[MAXN];
12
13 int main (void){
14
15     for (int i = 0; i < MAXN; ++i)
16     {
17         c[i] += a[i] * b[i];
18     }
19     // -----
20 }
21
22
23
```

The right panel shows the assembly output for x86-64 gcc 13.1. The assembly code is as follows:

```
9 main:
10     push    rbp
11     mov     rbp, rsp
12     mov     DWORD PTR [rbp-4], 0
13     jmp     .L2
14
15 .L3:
16     mov     eax, DWORD PTR [rbp-4]
17     cdq     xmm1, DWORD PTR [0+rax*4]
18     mov     eax, DWORD PTR [rbp-4]
19     cdq     xmm2, DWORD PTR [0+rax*4]
20     mov     eax, DWORD PTR [rbp-4]
21     cdq     xmm0, DWORD PTR [0+rax*4]
22     movss   xmm0, xmm2
23     mulss   xmm0, xmm1
24     addss   xmm0, xmm1
25     mov     eax, DWORD PTR [rbp-4]
26     cdq     xmm0, DWORD PTR [0+rax*4], xmm0
27     add     DWORD PTR [rbp-4], 1
28
29 .L2:
30     cmp     DWORD PTR [rbp-4], 99999999
31     jmp     .L3
```

3.2.3 源代码中普通表达式的汇编

我们不难看出他的处理也已经有SIMD（如：`vmovss`、`vaddss`）等的优化了



The screenshot shows the Visual Studio Code interface with two panels. The left panel displays a C++ source file with the following code:

```
25 vecB = _mm256_loadu_ps(b + i);
26 vecC = _mm256_loadu_ps(c + i);
27
28
29 vecC = _mm256_fmadd_ps(vecA, vecB, vecC);
30 _mm256_storeu_ps(c + i, vecC);
31 }
32 for (int i = 0; i < MAXN; ++i)
33 {
34     c[i] += a[i] * b[i];
35 }
36 // -----
37 }
38 }
39 }
40
```

The right panel shows the assembly output for x86-64 gcc 13.1 with the compiler options `-march=native -mavx`. The assembly code is as follows:

```
69     jle     .L8
70     mov     DWORD PTR [rsp+188], 0
71     jmp     .L9
72
73 .L10:
74     mov     eax, DWORD PTR [rsp+188]
75     cdq     xmm1, DWORD PTR [0+rax*4]
76     vmovss  xmm1, xmm2, xmm0
77     mov     eax, DWORD PTR [rsp+188]
78     cdq     xmm2, DWORD PTR [0+rax*4]
79     vmovss  xmm2, xmm1, xmm0
80     vaddss  xmm0, xmm1, xmm0
81     mov     eax, DWORD PTR [rsp+188]
82     cdq     xmm0, DWORD PTR [0+rax*4], xmm0
83     vaddss  xmm0, xmm1, xmm0
84     mov     eax, DWORD PTR [rsp+188]
85     inc     DWORD PTR [rsp+188]
86
87 .L9:
88     cmp     DWORD PTR [rsp+188], 99999999
89     jle     .L10
90     inc     DWORD PTR [rsp+196]
```

3.3 实验结果

ps:记得添加 `gcc` 的编译选项，否则可能不通过！

```

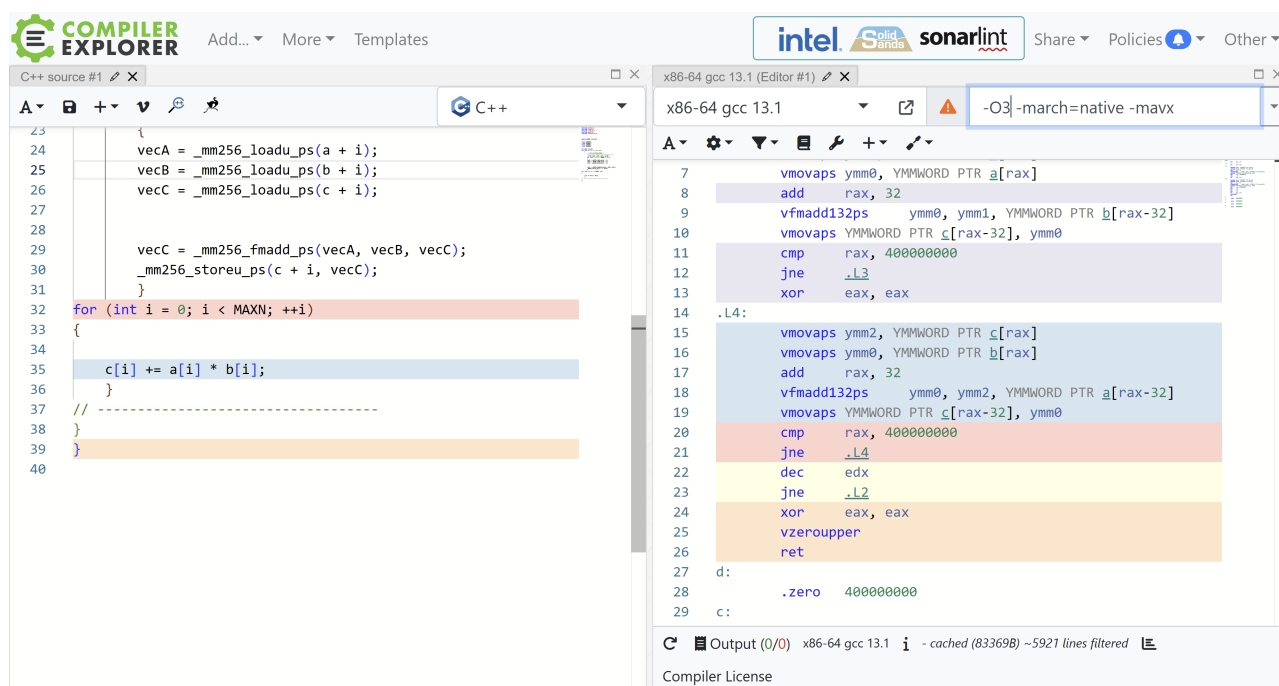
yaoyaoling@localhost ~/c/Z/H/lab2.5> gcc add.cpp -march=native -mavx
yaoyaoling@localhost ~/c/Z/H/lab2.5> ./a.out
normal time=5.352277
SIMD time=2.747737
Speed Up: 1.947885Check Passed
yaoyaoling@localhost ~/c/Z/H/lab2.5> ./a.out
normal time=5.252379
SIMD time=2.649900
Speed Up: 1.982105Check Passed

```

结论：最后的加速比将近2.

4 Discussion

通过手写SIMD的分装函数实现汇编上的优化。在学习的过程中我还看到，现代编译器都十分智能，我们甚至可以通过例如像：GCC 编译器：使用 `-O3` 选项启用所有优化，并使用 `-ftree-vectorize` 选项启用自动向量化。（如下图）当然有的时候可能因为分支复杂的原因无法进行优化，我们也可以添加编译指导语句例如：`\#pragma omp parallel for reduction(+:sum) private(x)` 来提示编译器帮我们优化



实际自动优化的加速比已经非常接近手动优化的了：

```

yaoyaoling@localhost ~/c/Z/H/lab2.5> gcc add.cpp -O3 -march=native -mavx
yaoyaoling@localhost ~/c/Z/H/lab2.5> ./a.out
normal time=1.198046
SIMD time=1.099501
Speed Up: 1.089627Check Passed

```

此外，这些封装函数的底层实现还是改变汇编语言。这当然得益于现代计算机高速发展使得cpu上能实现更多的指令加速工作。向量化的思想以及这个函数的整体构造，其实跟我们在CUDA 编程课上分配threads去完成加法工作一样十分相似。包括一些MPI的设计结构也是如此。通过向量化（亦或是封装统一处理的思想）来实现high performance！

最后的最后我想说的是：正如实验指导和上课讲过的那样：“深入到这个级别的优化已经开始需要考虑具体处理器的体系结构细节了，如某个架构下某条指令的实现延时和吞吐量是多少，处理器提供了多少向量寄存器，访存的对齐等等。这种时候编译器具体产生的汇编代码能比 C 语言代码提供更多的信息，你能了解到自己使用了多少寄存器，编译器是否生成了预期外的代码等等。但这一定是你**最后再去考虑**去通过SIMD优化代码”。