2º curso / 2º cuatr. Grado Ing. Inform.

**Dobles Grados** 

### Arquitectura de Computadores (AC)

Cuaderno de prácticas. Bloque Práctico 4. Optimización de código

Estudiante (nombre y apellidos): Grupo de prácticas y profesor de prácticas:

Denominación de marca del chip de procesamiento o procesador (se encuentra en /proc/cpuinfo y se lista con lscpu): AMD Ryzen 7 4700U with Radeon Graphics (8) @ 2.000GHz

Sistema operativo utilizado: *Kali GNU/Linux Rolling x86\_64* 

Versión de gcc utilizada: gcc version 11.3.0 (Debian 11.3.0-3)

Volcado de pantalla que muestre lo que devuelve lscpu en la máquina en la que ha tomado las medidas:

```
(icaro⊕kali)-[~]
Architecture:
                           x86 64
 CPU op-mode(s):
                           32-bit, 64-bit
 Address sizes:
                           48 bits physical, 48 bits virtual
 Byte Order:
                           Little Endian
CPU(s):
 On-line CPU(s) list:
                           0-7
endor ID:
                           AuthenticAMD
 Model name:
                           AMD Ryzen 7 4700U with Radeon Graphics
    CPU family:
                           23
    Model:
                           96
    Thread(s) per core:
    Core(s) per socket:
Socket(s):
    Stepping:
    Frequency boost:
                           enabled
    CPU max MHz:
                           2000,0000
    CPU min MHz:
                           1400,0000
    BogoMIPS:
                           3992.21
                           fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush mmx fxsr
    Flags:
                           se sse2 ht syscall nx mmxext fxsr_opt pdpe1gb rdtscp lm constant_tsc rep_good nopl nonstop
                            tsc cpuid extd_apicid aperfmperf rapl pni pclmulqdq monitor ssse3 fma cx16 sse4_1 sse4_2_
                           movbe popent aes xsave avx f16c rdrand lahf_lm cmp_legacy svm extapic cr8_legacy abm sse4
                            misalignsse 3dnowprefetch osvw ibs skinit wdt tce topoext perfctr_core perfctr_nb bpext
                            erfctr_llc mwaitx cpb cat_l3 cdp_l3 hw_pstate ssbd mba ibrs ibpb stibp vmmcall fsgsbase b
                           i1 avx2 smep bmi2 cqm rdt_a rdseed adx smap clflushopt clwb sha_ni xsaveopt xsavec xgetbv: xsaves cqm_llc cqm_occup_llc cqm_mbm_total cqm_mbm_local clzero irperf xsaveerptr rdpru v
                           bnoinvd arat npt lbrv svm_lock nrip_save tsc_scale vmcb_clean flushbyasid decodeassists pa
                           usefilter pfthreshold avic v_vmsave_vmload vgif v_spec_ctrl umip rdpid overflow_recov succ
                           or smca
Virtualization features:
 Virtualization:
                           AMD-V
Caches (sum of all):
                           256 KiB (8 instances)
 L1d:
                           256 KiB (8 instances)
  L1i:
 L2:
                           4 MiB (8 instances)
                           8 MiB (2 instances)
 L3:
NUMA:
 NUMA node(s):
 NUMA node0 CPU(s):
                           0-7
Vulnerabilities:
  Itlb multihit:
                           Not affected
                           Not affected
  L1tf:
                           Not affected
                           Not affected
 Meltdown:
  Spec store bypass:
                           Mitigation; Speculative Store Bypass disabled via prctl and seccomp
                           Mitigation; usercopy/swapgs barriers and __user pointer sanitization
Mitigation; Full AMD retpoline, IBPB conditional, IBRS_FW, STIBP disabled, RSB filling
  Spectre v1:
 Spectre v2:
                           Not affected
  Srbds:
                           Not affected
  Tsx async abort:
```

1. Modificar el código secuencial para la multiplicación de matrices disponible en SWAD (solo el trozo que calcula la multiplicación) para reducir el tiempo de ejecución. Justificar los tiempos obtenidos (usando siempre - O2) a partir de la modificación realizada. Incorporar los códigos modificados en el cuaderno.

#### MODIFICACIONES REALIZADAS (al menos dos modificaciones):

#### Modificación A) -explicación-:

He intercambiado los bucles j y k que calculan el producto de matrices , de forma que se pueda salvar provecho a la contigüidad de los datos en memoria (Aprovechar la localidad espacial)

#### Modificación B) –explicación-:

En el calculo del producto de matrices he realizado un desenrollado de bucle en el que por cada iteracion se va a calcular el valor de cuatro elementos de la matriz resultado.

Con el desenrollado se elimina la dependencia RAW con los retardos provocados por los LOAD/STORES consecutivos disminuyen

#### CÓDIGOS FUENTE MODIFICACIONES

A) Captura de pmm-secuencial-modificado\_A.c

Capturas de pantalla (que muestren la compilación y que el resultado es correcto):

```
(icaro® kali)-[~/.../Practica/Seminario4/CODIGO/bp4]

$ gcc -02 pmm-secuencial-modificado-a.c -0 pmm-secuencial-a -lrt

(icaro® kali)-[~/.../Practica/Seminario4/CODIGO/bp4]

$ ./pmm-secuencial-a 1000

Tiempo: 0.731830570 Tamaño: 1000 m3[0][0]: 1098.900000 m3[N-1][N-1]: 1099998.900000

(icaro® kali)-[~/.../Practica/Seminario4/CODIGO/bp4]

$ ./pmm-secuencial 1000

Tiempo: 1.001622608 Tamaño: 1000 m3[0][0]: 33283350.000000 m3[N-1][N-1]: 99899933383350.000000
```

B)

```
(icaro⊕ kali)-[~/.../Practica/Seminario4/CODIGO/bp4]

$ gcc -02 pmm-secuencial-modificado-b.c -0 pmm-secuencial-b -lrt

(icaro⊕ kali)-[∯/.../Practica/Seminario4/CODIGO/bp4]

$ ./pmm-secuencial-b 1000

Tiempo: 0.273129798 Tamaño: 1000 m3[0][0]: 33283350.000000 m3[N-1][N-1]: 9989993383350.000000

(icaro⊕ kali)-[~/.../Practica/Seminario4/CODIGO/bp4]

$ ./pmm-secuencial 1000

Tiempo: 1.010135195 Tamaño: 1000 m3[0][0]: 33283350.000000 m3[N-1][N-1]: 9989993383350.000000
```

TIEMPOS: para n = 1000

Modificación	Breve descripción de las modificaciones	-O2	
Sin modificar		1.001622608	
Modificación A)	Intercambio los bucles j y k	0.731830570	
Modificación B)	Desenrollo el bucle, calculo de 4 elementos/iteración	0.273129798	

## **COMENTARIOS SOBRE LOS RESULTADOS Y JUSTIFICACIÓN DE LAS MEJORAS EN TIEMPO:** Los codigos se han implementado tomando como ejemplo n = 1000.

Los codigos se nan implementado tomando como ejempio n – 100

#### **Modificacion A:**

He realizado un cambio en la forma de acceder a los datos "Localidad de acceso". Ahora se aprovecha mejor como el compilador los almacena, aprovecha la localidad. Podríamos entender una matriz como si fuese un array para el cual existen una serie de punteros donde indican donde comienza cada fila. Con la modificación realizada lo que conseguimos es movernos por las columnas de la matriz y sacar provecho de la localidad de estos datos Si se hubiese mantenido como en el original, lo que conseguiría sera un mayor numero de cambios de filas y por consiguiente un tiempo mayor para obtener cada resultado.

#### **Modificacion B:**

Reducimos el numero de saltos/iteraciones aplicando el desenrollado de bucles. Las variables als incremento en 4 en 4 . Por un lado el tamaño del codigo aumenta, pero por otro lado al eliminar las dependencias RAW, los stalls (atascos memoria) provocado por los load y store localizados cerca disminuyen.

2. Usar en este ejercicio el programa secuencial disponible en SWAD que utiliza como base el código de la Figura 1. Modificar en el programa el código mostrado en la Figura 1 para reducir el tiempo de ejecución. Justificar los tiempos obtenidos (usando siempre -O2) a partir de la modificación realizada. En las ejecuciones de evaluación usar valores de N y M mayores que 1000. Incorporar los códigos modificados en el cuaderno.

**Figura 1** . Código C++ que suma dos vectores. M y N deben ser parámetros de entrada al programa, usar valores mayores que 1000 en la evaluación.

```
struct nombre {
    int a;
    int b;
} s[N];

main()
{
    ...
    for (ii=0; ii<M;ii++) {
        X1=0; X2=0;
        for(i=0; i<N;i++) X1+=2*s[i].a+ii;
        for(i=0; i<N;i++) X2+=3*s[i].b-ii;

        if (X1<X2) R[ii]=X1 else R[ii]=X2;
    }
    ...
}</pre>
```

#### MODIFICACIONES REALIZADAS (al menos dos modificaciones):

#### Modificación A) – explicación -:

Descartamos el segundo for que realiza el calculo de X2 deido a que se ejecuta las mismas veces que el de X1 y no hay dependencias entre esas dos variables, de forma que junto el mismo bucle y calculo.

Se unen los bucles por que el struct esta definido con a y b dispuestos en memoria de manera correlativa (menos fallos de caché)

#### Modificación B) – explicación -:

Aplico un desenrollado de bucle. Reducimos el tiempo de acceso a los datos aprovechando su localidad espacial y accediendo a aquellos que están más contiguos. Se sacrifican el tamaño del codigo, pero los resultados son mejores, elimino la dependencia RAW.

#### CÓDIGOS FUENTE MODIFICACIONES

A) Captura figura1-modificado\_A.c

Capturas de pantalla (que muestren la compilación y que el resultado es correcto):

```
(icaro® kali)-[~/.../Practica/Seminario4/CODIGO/bp4]
$ ./figura-original 1000 1000
Tiempo: 0.002782018
Elemento 0 y 999 de R: -202208, -844581

(icaro® kali)-[~/.../Practica/Seminario4/CODIGO/bp4]
$ ./figura-modificado-a 1000 1000
Tiempo: 0.002184905
Elemento 0 y 999 de R: -198844, -896524
```

B)

```
(icaro® kali)-[~/.../Practica/Seminario4/CODIGO/bp4]

$ gcc -02 figura1-modificado-b.c -0 figura-modificado-b -lrt

(icaro® kali)-[~/.../Practica/Seminario4/CODIGO/bp4]

$ ./figura-modificado-b 1000 1000

Tiempo: 0.000672279

Elemento 0 y 999 de R: -201902, -898422
```

```
// ZONA A MODIFICAR
int X1_1,X1_2,X1_3;
int X2_1,X2_2,X2_3;

clock_gettime(CLOCK_REALTIME,&cgt1);
    for (ii=0; ii<M;ii++){
        X1-X1_1=X1_2=X1_3=0; X2=X2_1=X2_2=X2_3=0;

        //for(i=0; i<N;i++) X1 += 2*s[i].a + ii;
        //for(i=0; i<N;i++) X2 += 3*s[i].b - ii;

        for(i=0; i<N;i+=b){
            X1 += 2*s[i].a + ii;
            X1_2+= 2*s[i+2].a + ii;
            X1_3+= 2*s[i+3].a + ii;
            X1_3+= 2*s[i+3].a + ii;
            X1_3+= 2*s[i+3].b - ii;
            X2 += 2*s[i].b - ii;
            X2_1+= 2*s[i+1].b - ii;
            X2_2+= 2*s[i+3].b - ii;
            X2_3+= 2*s[i+3].b - ii;
            X2_3+= 2*s[i+3].b - ii;
            X2_3+= 2*s[i+3].b - ii;
            X2_5+= 2*s[i+3].b - ii;
            X3_5+= 2*s[i+3].b - ii;
            X3_5
```

#### TIEMPOS:

I ILIVII OU.		
Modificación	Breve descripción de las modificaciones	-O2
Sin modificar		0.002782018
Modificación A)	Calcular X1 y X2 en un mismo bucle	0.002184905
Modificación B)	Desenrollado de bucle	0.000672279

#### COMENTARIOS SOBRE LOS RESULTADOS Y JUSTIFICACIÓN DE LAS MEJORAS EN TIEMPO:

#### **Modificacion a:**

Unimos ambos bucles para sacar partido a la localidad espacial de los accesos. La forma en la que esta declarado el struct permite que los datos a y b estén dispuestos en memoria de manera correlatica permitiendo que se den menos fallos de cache y sea optimo un acceso a ambos bajo el mismo bucle

#### **Modificacion b:**

Desarrollo un desenrollado de bucle como en pmm-secuencial y con ello reduzco el numero de saltos, aquí he necesitado variables locales  $X1_1$  .... para evitar las dependencias RAW y mejorar el rendimiento

3. El benchmark Linpack ha sido uno de los programas más ampliamente utilizados para evaluar las prestaciones de los computadores. De hecho, se utiliza como base en la lista de los 500 computadores más rápidos del mundo (el Top500 Report). El núcleo de este programa es una rutina que opera con flotantes de doble precisión denominada DAXPY (*Double precision- real Alpha X Plus Y*) que multiplica un vector por una constante y los suma a otro vector (Lección 3/Tema 1):

for 
$$(i=0; i< N; i++)$$
  $y[i]= a*x[i] + y[i];$ 

A partir del programa DAXPY disponible en SWAD, generar los programas en ensamblador para cada una de las siguientes opciones de optimización del compilador: -O0, -Os, -O2, -O3. Explique las diferencias que se observan en el código justificando al mismo tiempo las mejoras en velocidad que acarrean. Incorporar los códigos al cuaderno de prácticas y destacar las diferencias entre ellos. Sólo se debe evaluar el tiempo del núcleo DAXPY. N deben ser parámetro de entrada al programa.

CAPTURA CÓDIGO FUENTE: daxpy.c

Tiempos ejec.	-O0	-Os	-O2	-O3
Longitud vectores	0.157637456	0.11322604	0.09477849	0.08806331
N = 107374182		8	3	6

**CAPTURAS DE PANTALLA** (que muestren la compilación y que el resultado es correcto):

```
—(icaro® kali)-[~/.../Practica/Seminario4/CODIGO/bp4
-$ gcc -00 <u>daxpy.c</u> -0 <u>daxpy</u> -lrt
   -(icaro®kali)-[~/.../Practica/Seminario4/CODIGO/bp4]
  $ ./daxpy 107374182 1000
Tiempo:0.157637456
                              / Tamaño Vectores:33554432
                                                                      / alpha*x[0]+y[0]=z[0](1000.000000*0.000000+0.000985=0.00098
  -(<mark>icaro⊕kali</mark>)-[~/.../Practica/Seminario4/CODIGO/bp4]
 -$ gcc -Os <u>daxpy.c</u> -o <u>daxpy</u>
   -(<mark>icaro®kali</mark>)-[~/.../Practica/Seminario4/CODIGO/bp4]
 -$ ./daxpy 107374182 1000
                              / Tamaño Vectores:33554432
Tiempo:0.113226048
                                                                      / alpha*x[0]+y[0]=z[0](1000.000000*0.000000+0.000985=0.00098
  -(icaro⊛kali)-[~/.../Practica/Seminario4/CODIGO/bp4]
 -$ gcc -02 <u>daxpy.c</u> -0 <u>daxpy</u>
  —(<mark>icaro⊛kali</mark>)-[~/…/Practica/Seminario4/CODIGO/bp4]
./daxpy 107374182 1000
Tiempo:0.094778493 / Tamaño Vectores:33554432 / alpha*x[0]+y[0]=z[0](1000.000000*0.000000+0.000985=0.00098
5) / / alpha*x[33554431]+y[33554431]=z[33554431](1000.000000*0.037781+0.434614=38.215512) /
   -(icaro®kali)-[~/.../Practica/Seminario4/CODIGO/bp4]
 -$ gcc -03 <u>daxpy.c</u> -0 <u>daxpy</u>
  —(icaro⊕ kali)-[~/.../Practica/Seminario4/CODIGO/bp4]
-$ ./daxpy 107374182 1000
                              / Tamaño Vectores:33554432
                                                                      / alpha*x[0]+y[0]=z[0](1000.000000*0.000000+0.000985=0.00098
 iempo:0.088063316
   //alpha*x[33554431]+y[33554431]=z[33554431](1000.000000*0.037781+0.434614=38.215512)/
```

#### https://www.rapidtables.com/code/linux/gcc/gcc-o.html

#### COMENTARIOS QUE EXPLIQUEN LAS DIFERENCIAS EN ENSAMBLADOR:

En comparación con la optimización O2 donde el compilador intenta aumentar el rendimiento del código sin comprometes el tamaño sin tomar demasiado tiempo para compilar, la bandera -O0 se caracteriza por no realizar ninguna optimización exhibiendo un código ensamblador bastante denso y con un alto tiempo de ejecución.

La optimización Os realiza mejora en lo referente a la longitud del código ensamblador obtenido, sin embargo no optimiza nada relacionado con el tiempo de ejecución.

Finalmente el flag -03 activa optimizaciones que implican una alta demanda en uso de memoria. No garantiza mejoras en el rendimiento, y de hecho, el sistema se puede resentir debido al uso abusivo de memoria.

El conclusión, el flag mas optimo para realizar una optimización de código es -O2.

**CÓDIGO EN ENSAMBLADOR** (no es necesario introducir aquí el código como captura de pantalla, ajustar el tamaño de la letra para que una instrucción no ocupe más de un renglón):

# (PONER AQUÍ SÓLO LA ZONA DEL CÓDIGO ENSAMBLADOR DONDE ESTÁ EL CÓDIGO EVALUADO, USE COLORES PARA DESTACAR LAS DIFERENCIAS)

```
gcc -O0 -S daxpy.c -o daxpy0.s -lrt
gcc -Os -S daxpy.c -o daxpyOs.s -lrt
gcc -O2 -S daxpy.c -o daxpyO2.s -lrt
gcc -O3 -S daxpy.c -o daxpyO3.s -lrt
```

daxpy00.s	daxpy0s.s	daxpy02.s	daxpy03.s
daxpy00.s  call clock gettime@PLT movl \$0, -4(%rbp) jmp .L10  .L11: movl -4(%rbp), %eax cltq leaq 0(,%rax,8), %rdx leaq x(%rip), %rax movsd (%rdx,%rax), %xmm0 movapd %xmm0, %xmm1 mulsd -16(%rbp), %xmm1 mulsd -16(%rbp), %eax cltq leaq 0(,%rax,8), %rdx leaq y(%rip), %rax movsd (%rdx,%rax), %xmm0 addsd %xmm1, %xmm0 movl -4(%rbp), %eax cltq leaq 0(,%rax,8), %rdx leaq z(%rip), %rax movsd (%rdx,%rax), %xmm0 addsd %xmm0, (%rdx,%rax) addl \$1, -4(%rbp)  .L10: movl -4(%rbp), %eax cmpl -8(%rbp), %eax jl .L11	daxpyOs.s  call clock gettime@PLT xorl %eax, %eax leaq z(%rip), %rdx leaqx (%rip), %rcx leaqy (%rip), %rsi  .L8: cmpl %eax, %ebp jle .L19 movq %r14, %xmm0 mulsd (%rcx,%rax,8), %xmm0 addsd (%rsi,%rax,8), %xmm0 movsd %xmm0, (%rdx, %rax,8) incq %rax jmp .L8  .L19: leaq 16(%rsp), %rsi xorl %edi, %edi call clock_gettime@PLT	daxpy02.s  call clock gettime@PLT xorl %eax, %eax .p2align 4,,10 .p2align 3  .L8: movsd 8(%rsp), %xmm0 mulsd (%rbx,%rax,8), %xmm0 addsd (%r12,%rax,8), %xmm0 movsd %xmm0, 0(%r13,%rax,8) addq \$1, %rax cmpl %eax, %ebp jg .L8 leaq 32(%rsp), %rsi xorl %edi, %edi call clock gettime@PLT	daxpy03.s  call  clock_gettime@PLT
novsd (%rdx,%rax), %xmm0	.L19:	1	divsd .LC8(%rip), %xmm0
movl -4(%rbp), %eax cltq leaq 0(,%rax,8), %rdx	xorl %edi, %edi		%xmm1, %xmm0 cmpl \$10, %r12d
movsd %xmm0, (%rdx, %rax)			.L34
movl -4(%rbp), %eax cmpl -8(%rbp), %eax			.LC7(%rip), %rdi movl
jl .L11 leaq -64(%rbp), %rax movq %rax, %rsi movl \$0, %edi			xorl %r13d, %r13d call printf@PLT
call clock gettime@PLT			leaq .LC5(%rip), %r14 .p2align
			4,,10 .p2align 3 .L14:
			movsd (%r15,%r13,8), %xmm3 movl
			%r13d, %esi movl

	0/	r13d, %ecx	
			movl
	%	r13d, %edx	
			movsd
		(%rbp,%r13,8	movsd
	(	%rbx,%r13,8)	
			movq
	%	%r14, %rdi	movl
	\$	64, %eax	IIIOV L
			movsd
	8	(%rsp), %xmm	
	 	i1, %r13	addq
	Ψ		call
	p	rintf@PLT	
			cmpl
	%	%r13d, %r12d	jg
	.	L14	78
		L29:	
			addq
		56, %rsp	.cfi_remem
	b	er_state	1011_10
			.cfi_def_c
	  f	a_offset 56	vorl
	%	Seax, %eax	xorl
			popq
	%	Srbx	c
		a_offset 48	.cfi_def_c
			popq
	%	Srbp	
	_	a officet 40	.cfi_def_c
		a_offset 40	popq
	%	5r12	
		5.5	.cfi_def_c
		a_offset 32	popq
	%	5r13	рорч
			.cfi_def_c
	f	a_offset 24	nong
	%	5r14	popq
			.cfi_def_c
	f	a_offset 16	
	9%	5r15	popq
		0	.cfi_def_c
	f	a_offset 8	
		L34:	ret
		LJ4.	.cfi_resto
	r	e_state	
			leal
	-	1(%r12), %ed	x pushq
			Pacind

	9	%rcx	
			.cfi_def_c
		fa_offset 120	
			movsd
		z(%rip), %xmm₄	
			movl
		%r12d, %esi	may . a ] a
			movslq
		%edx, %rax	movsd
	,	y(%rip), %xmm:	
			movl
		%edx, %r8d	
			movl
	9	%edx, %ecx	
			pushq
		(%r15,%rax,8)	
			.cfi_def_c
	-	fa_offset 128	
			movsd
		24(%rsp), %xmr	
			leaq
		.LC6(%rip), %	
			movsd
	1	0(%rbp,%rax,8)	
			movsd
		(%rbx,%rax,8),	, %xiiiilo movsd
		x(%rip), %xmm2	
			movapd
		%xmm5, %xmm1	ovapa
			movl
	:	\$8, %eax	
			call
		printf@PLT	
			popq
	9	%rsi	
			.cfi_def_c
		fa_offset 120	
		%rdi	popq
	-		.cfi_def_c
		fa_offset 112	.cri_uer_c
			jmp
		.L29	J P
		.L7:	
			xorl
		%edi, %edi	
			leaq
		16(%rsp), %rs:	
			leaq
		z(%rip), %r15	
			call
		clock_gettime(	@PLT
 		·	