Tema: Ejercicio de ingeniería inversa.
Habilidad Personal: Habilidad para implementar *software* en un contexto ético.

Program Outcomes:
CDIO:
2.5. PROFESSIONAL SKILLS AND ATTITUDES
2.5.1. Professional Ethics, Integrity, Responsibility and Accountability
4.5. IMPLEMENTING
4.5.3. Software Implementing Process
ABET:
(2.) Develop an ability to apply engineering design to produce solutions that meet specified needs with consideration of public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors.
(4.) an understanding of professional and ethical responsibility.

1. OBJETIVO:

Inferir "el código fuente" en Lenguaje C, a partir de su código máquina.

> **Nota**: Los códigos máquinas proporcionados como insumo en el anexo de este taller se generaron para una *CPU* de la arquitectura x86 de 64 bits. Se usó el compilador "clang" versión 11.0.1 sin especificar opciones especiales de compilación.

Conocer la información previa es fundamental y de mucha ayuda. La razón de ello se debe a que el código fuente de un único programa en Lenguaje C puede ser compilado para generar código máquina, literalmente, en centenares de arquitecturas. Inclusive, con un único código fuente y una misma CPU objetivo, se puede generar gran variedad de versiones de código máquina, usando diferentes compiladores y opciones de compilación. Todas ellas completamente funcionales y distintas entre sí.

Consultar:

https://www.geeksforgeeks.org/software-engineering-reverse-engineering/

2. LA ÉTICA DE LA DECOMPILACIÓN DE PROGRAMAS:

Traducido de https://www.cl.cam.ac.uk/teaching/2001/OptComp/local/ethics.html

Si la decompilación es posible hasta cierto punto, ¿Es este proceso permitido? La decompilación se utiliza por varias razones, entre ellas:

- Recuperación del código fuente perdido (por accidente),
- Migración de aplicaciones a una nueva plataforma hardware,
- Traducción de código escrito en lenguajes obsoletos que no son compatibles con las herramientas de compilación actuales,
- Existencia de virus o código malicioso en el programa, y
- Recuperación del código fuente de otra persona (para determinar un algoritmo, por ejemplo).

Sin embargo, no todos los usos de los decompiladores son usos legales.

Los programas de computadora están protegidos por la ley de derechos de autor. Los derechos de autor protegen la expresión de una idea en forma de programa, por lo que protegen la propiedad intelectual del desarrollador (o de la empresa) sobre el software. La ley de derechos de autor proporciona un paquete de derechos exclusivos al desarrollador de software, entre otros, el derecho a reproducir y hacer adaptaciones al programa informático desarrollado. Es una violación de estos derechos la realización de reproducciones y adaptaciones sin permiso del titular de los derechos de autor. Además, los acuerdos de licencia también pueden obligar al usuario a operar el programa de cierta manera y evitar el uso de técnicas de decompilación o desensamblaje en ese programa.

Cada país tiene excepciones a los derechos del propietario de los derechos de autor o se ha establecido un precedente en procedimientos judiciales. Esto significa que estos usos están permitidos por ley. Los más comunes son:

- Decompilación/desensamblado con fines de interoperabilidad (con otra pieza de software o hardware) cuando la especificación de la interfaz no está disponible,
- Decompilación/desmontaje con el fin de corregir errores cuando el propietario de los derechos de autor no esté disponible para realizar la corrección, y
- Para determinar partes del programa que no están protegidas por derechos de autor (por ejemplo, algoritmos), sin violar otras formas de protección (por ejemplo, patentes o secretos comerciales).

No todos los países implementan las mismas leyes, es necesario consultar los aspectos legales en caso de dudas.

Para mas información, por favor consulte:

https://ethics.csc.ncsu.edu/intellectual/reverse/study.php

3. INSTRUCCIONES PARA EL DESARROLLO DE ESTE TALLER:

Este taller se propone con fines académicos y como una forma de conocer la arquitectura x86-64 y su conjunto de instrucciones máquina. Los códigos que se proponen decompilar son programas que no infringen derechos de autor y no tiene ningún tipo de uso comercial.

En equipos de trabajo conformados, a lo sumo, por tres integrantes, se deben tomar los cuatro códigos máquina que se comparten al final de esta guía, y mediante un proceso de ingeniería inversa se debe inferir el código fuente en Lenguaje C que genera cada uno de estos códigos.

Los códigos máquina se comparten en una representación numérica hexadecimal. En la columna izquierda se indica la posición de memoria del primer byte de la instrucción en representación hexadecimal. Por facilidad, se escribe cada instrucción de bajo nivel en una línea horizontal. Cada instrucción consta de 1 a 7 *bytes* y cada *byte* se escribe mediante una pareja de números hexa. En estos *bytes* se presenta de primero, el *opcode* (contracción del inglés de las palabras: *operation code*) y luego el operando. En la arquitectura x86, hay instrucciones máquina con *opcode* de uno solo *byte* que pueden no tener operando. Existen otras con *opcode* de dos *bytes*. De igual manera, se pueden encontrar instrucciones con un operando de extensión de más de un *byte*. Por esta razón, se separa cada instrucción de código máquina (*opcode* + operando) en una línea independiente.

El código máquina de insumo se presenta en una tabla de dos secciones coloreadas. Los códigos en la sección amarilla corresponden a las instrucciones máquina de la función en C que se desea inferir. Finalmente los códigos en verde corresponden a los de la función “main()” que hace el llamado a las funciones objeto de este taller (en caso que aplique pues puede haber códigos en “main()” exclusivamente). La inclusión del “main()” en la mayoría de los casos es fundamental porque justo antes del llamado de la función (cuando aplique) se hacen copias de los parámetros en el “*stack*”, ya sea que estas copias sean de los valores o de las direcciones de dichos parámetros. Tanto los códigos máquina en amarillo, como los códigos del “main()” en verde deben ser decompilados.

A modo de guía, para solucionar el problema de ingeniería inversa propuesto se sugiere el siguiente procedimiento:

- Buscar los mnemónicos en lenguaje de ensamble de los respectivos *opcodes* según la tabla de la arquitectura x86 que se adjunta a esta guía y que está disponible en:

https://i.stack.imgur.com/VTxd0.jpg

Cada arquitectura de *CPU* tiene asociado su propio conjunto de instrucciones que está en capacidad de ejecutar.

- La mayoría de mnemónicos en este taller se pueden extraer de esta tabla; sin embargo, en algunos casos será necesario, acudir a otras fuentes de información mas especializadas para identificar sobre qué registros actuará cada *opcode*.
- Se sugiere usar cualquiera de los desensambladores en línea disponibles en la WEB. A continuación se mencionan unos pocos de los muchos disponibles:

  https://disasm.pro

  https://defuse.ca/online-x86-assembler.htm#disassembly2

  http://shell-storm.org/online/Online-Assembler-and-Disassembler/

- Después de que se tenga una propuesta preliminar de los mnemónicos de las instrucciones en forma de lenguaje de ensamble, es fundamental estudiar las instrucciones de la arquitectura x86. Aunque la arquitectura x86 consta de varios cientos de instrucciones, se sugiere focalizar el estudio sobre las que se usan en este taller. Este estudio conlleva a conocer: (1) qué hace cada instrucción, (2) sobre qué registros de la CPU actúa cada instrucción y (3) cuál es el posible efecto en las banderas del "*Status Register*". Esto dirige la atención a revisar algunos aspectos básicos de la arquitectura x86 y a identificar si la instrucción corresponde a una de (1) control de flujo, (2) operación aritmético-lógica, (3) de transferencia de datos entre CPU y memoria y (4) entrada y salida.

  https://www.felixcloutier.com/x86/

  https://shell-storm.org/x86doc/

  https://en.wikipedia.org/wiki/X86_instruction_listings

  https://docs.oracle.com/cd/E19253-01/817-5477/817-5477.pdf

- Se recomienda empezar a identificar los saltos (condicionales y no condicionales). Esta sugerencia lleva a inferir de primero, las secuencias de control de flujo que pudieron haber sido usadas en el código fuente original de Lenguaje C. Si todos los saltos son hacia adelante, el control de flujo en alto nivel se llevó a cabo mediante secuencias de selección con "if-else" o incluso "switch-case". Si por lo menos se encuentra un salto hacia atrás, existe por lo menos una secuencia de iteración en alto nivel que hizo uso de las secuencias "for-do-while". Mediante un análisis cuidadoso es relativamente fácil deducir las secuencias usadas en la función, según las plantillas sugeridas por su profesor al comienzo del curso.

  http://unixwiz.net/techtips/x86-jumps.html

- A continuación se recomienda identificar el uso, que el código en amarillo le da al "*stack*" de memoria y especialmente a entender cómo se crea y usa cada "*stack frame*". Este estudio lleva a identificar el uso de las variables automáticas, los parámetros pasados a la función ya sea por valor o por referencia, así como el "*calling convention*" usado, y otros elementos similares en el código que hacen uso del "*stack*" y que cada estudiante ya debe dominar en este punto del curso. Se compartió el código máquina del "main()" porque, como recordará, es en esta función que se hace la escritura y copiado de los parámetros de la función al "*stack*" antes de un llamado a subrutina en código máquina.

  https://eli.thegreenplace.net/2011/09/06/stack-frame-layout-on-x86-64

  https://textbook.cs161.org/memory-safety/x86.html

- A continuación, tras conocer las secuencias y las variables sobre las que se opera, se recomienda seguir las instrucciones paso a paso para inferir la lógica de la función y en consecuencia el propósito de la misma.

- Luego, se sugiere hacer una propuesta de la función y del "main()" en Lenguaje C. Para ello puede editar dicha propuesta de código fuente, para verificar en el sitio WEB de "*Compiler Explorer*" con las opciones apropiadas, si el código inferido genera el código máquina objeto de este taller:

  https://godbolt.org/

  Esto es muy importante porque el código máquina en representación hexadecimal de este taller fue generado en dicha arquitectura, y con la versión de compilador apropiado, sin ninguna opción especial para su compilación. *Esto provee un mismo punto de referencia para todos*.

- Es muy probable que se deba iterar sobre la metodología aquí sugerida para refinar el proceso de inferencia de un código fuente. Con esto no se invita a un proceso de prueba y error sino a una búsqueda inteligente que conduzca de manera eficiente a inferir dicho código que minimice la prueba y error. Ser consciente de este proceso de meta-cognición, permitirá refinar una propia metodología que se pudiera usar a futuro para analizar y entender otras arquitecturas a partir de las instrucciones de bajo nivel y su relación con el código en C que las genera.

4. CRITERIO DE CALIFICACIÓN DEL TALLER:

La solución a este taller y los respectivos entregables, consisten de dos partes. La primera parte corresponde a la propuesta del código fuente en Lenguaje C de una función cuyo código máquina en representación hexadecimal, corresponde de manera exacta al de la sección amarilla, así como la parte del "main()" que corresponde a la sección en verde. La segunda parte corresponde a un muy corto informe de no más de una (1) página explicando y justificando brevemente el proceso para llegar a las soluciones. La primera parte que consiste en el resultado final de los códigos fuente en Lenguaje C vale el 80% (es decir

20% por cada código en lenguaje C, tanto para la función como para el "main()" que llama a la función) y la justificación del proceso vale el 20%.

> **Nota**: Justificar el proceso es importante desde el punto de vista formativo; sin embargo, se valora más el resultado final pues no tiene sentido documentar y justificar un código fuente que sea incorrecto.

Mis mejores deseos en su aprendizaje.

Código 1:

```
  0:   55
  1:   48 89 e5
  4:   31 c0
  6:   c7 45 fc 00 00 00 00
  d:   c7 45 f8 ff ff 00 00
 14:   c7 45 f4 02 00 00 00
 1b:   8b 4d f4
 1e:   ba 01 00 00 00
 23:   d3 e2
 25:   83 f2 ff
 28:   23 55 f8
 2b:   89 55 f8
 2e:   5d
 2f:   c3
```

Nota: Este código corresponde a la función “main()”. En este caso el “main()” no hace llamado a ninguna función.

Código 2:

```
 0:  55
 1:  48 89 e5
 4:  48 89 7d f8
 8:  48 89 75 f0
 c:  48 8b 45 f8
10:  8b 08
12:  89 4d ec
15:  48 8b 45 f0
19:  8b 08
1b:  48 8b 45 f8
1f:  89 08
21:  8b 4d ec
24:  48 8b 45 f0
28:  89 08
2a:  b8 01 00 00 00
2f:  5d
30:  c3
31:  66 2e 0f 1f 84 00 00 00 00 00
3b:  0f 1f 44 00 00

40:  55
41:  48 89 e5
44:  48 83 ec 10
48:  c7 45 fc 00 00 00 00
4f:  c7 45 f8 aa 00 00 00
56:  c7 45 f4 55 00 00 00
5d:  48 8d 7d f8
61:  48 8d 75 f4
65:  e8 00 00 00 00
6a:  31 c9
6c:  89 45 f0
6f:  89 c8
71:  48 83 c4 10
75:  5d
76:  c3
```

Nota: Recuerde “main()” en verde y una función en amarillo, que es llamada desde el “main()”.

Código 3:

```
0:  55
1:  48 89 e5
4:  89 7d fc
7:  c7 45 f8 01 00 00 00
e:  c7 45 f4 01 00 00 00
15: 8b 45 f4
18: 3b 45 fc
1b: 0f 8f 18 00 00 00
21: 8b 45 f8
24: 0f af 45 f4
28: 89 45 f8
2b: 8b 45 f4
2e: 83 c0 01
31: 89 45 f4
34: e9 dc ff ff ff
39: 8b 45 f8
3c: 5d
3d: c3
3e: 66 90
40: 55
41: 48 89 e5
44: 48 83 ec 10
48: c7 45 fc 00 00 00 00
4f: c7 45 f8 04 00 00 00
56: c7 45 f4 00 00 00 00
5d: 8b 7d f8
60: e8 00 00 00 00
65: 31 c9
67: 89 45 f4
6a: 89 c8
6c: 48 83 c4 10
70: 5d
71: c3
```

Nota: Recuerde “main()” en verde y una función en amarillo, que es llamada desde el “main()”.

Código 4:

```
0:      55
1:      48 89 e5
4:      48 83 ec 10
8:      89 7d f8
b:      83 7d f8 00
f:      0f 85 0c 00 00 00
15:     c7 45 fc 00 00 00 00
1c:     e9 32 00 00 00
21:     83 7d f8 01
25:     0f 85 0c 00 00 00
2b:     c7 45 fc 01 00 00 00
32:     e9 1c 00 00 00
37:     8b 45 f8
3a:     8b 4d f8
3d:     83 e9 01
40:     89 cf
42:     89 45 f4
45:     e8 b6 ff ff ff
4a:     8b 4d f4
4d:     0f af c8
50:     89 4d fc
53:     8b 45 fc
56:     48 83 c4 10
5a:     5d
5b:     c3
5c:     0f 1f 40 00

60:     55
61:     48 89 e5
64:     48 83 ec 10
68:     c7 45 fc 00 00 00 00
6f:     c7 45 f8 04 00 00 00
76:     c7 45 f4 00 00 00 00
7d:     8b 7d f8
80:     e8 00 00 00 00
85:     31 c9
87:     89 45 f4
8a:     89 c8
8c:     48 83 c4 10
90:     5d
91:     c3
```

Nota: Recuerde “main()” en verde y una función en amarillo, que es llamada desde el “main()”.