**3 Machine-Level Representation of Programs**

# 英语

|  |  |
| --- | --- |
| 1. shield屏蔽 2. reference引用 3. consistent一致的 4. inefficiency低效率 5. malware恶意软件 6. involve涉及、调用 7. thereby从而 8. recursive循环的，iterative迭代的 9. delude oneself自欺欺人 10. arcane神秘的 11. feasible可行的   economically feasible and technically desirable经济可行、技术可实现   1. compromise妥协 2. safeguards保障措施】 3. elementary基本的 4. aggregate聚合的 5. guildeline指导方针 6. strip away剥去 7. annotated带解释的 8. ambiguity歧义 9. owing to由于 10. impose提出 | 1. specify指定 2. manipulate操作 3. underlying潜在的 4. get a sense of感觉到 5. infest侵扰 6. nuance细微差别 7. eliminate消除 8. prerequisite先决条件 9. corporation公司 10. proceed继续前进 11. legacy遗产 12. colloquially通俗地讲 13. overall最初的 14. elaborate复杂的 15. dictated被……制定的 16. whereas虽然 17. ever-changing不断改变的 18. indented缩进的 19. explanatory解释性的 20. dedicated专门的 21. enthusiasts爱好者 22. compact最紧凑的 23. reminiscent令人怀念的 |

**|** *Best of all, a program written in a high-level language can be compiled and executed on a number of different machines, whereas assembly code is highly machine specific.*

*When programming in a hign-level language such as C,and even more so in Java,we are shielded from the detailed machine-level implementation of our program*.*In contrast, when writing programs in assembly code (as was done in the early days of computing) a programmer must specify the low-level instructions the program uses to carry out a computation.*

# 3.1 *A Historical Perspective*

**|** 摩尔定律

预测芯片上的晶体管数量每一年半翻一番

# 3.2 *Program Encodings*

假设有两个C语言文件：p1.c和p2.c，然后使用gcc命令行进行编译生成可执行文件

gcc -0g -o p p1.c p2.c

命令gcc表示使用gcc这种C编译器进行编译，而且因为Linux上的默认C编译器就是gcc所以也可以简单地使用cc进行调用

-0g表示使用一种优化方式，产生一种遵守最初的原始C代码的机器代码。

为什么选择-0g是因为使用更高级别的优化将会生成经过大量转换的代码，以致于生成的机器代码和原始源代码之间的关系难以理解。因此，我们将使用 -Og 优化作为学习工具，然后看看当我们提高优化级别时会发生什么

但是在实践中，一般是使用更高级别的优化，比如-01，-02

和之前提到的编译系统一致，gcc命令经过“**预处理、编译、汇编、链接**”四步来产生最终的可执行程序

预处理：扩展源代码中的#include所引入的文件和#define所定义的宏，.c→.i

编译：将修改后的C文件.i使用编译器生成汇编文件p1.s和p2.s

汇编：使用汇编器将汇编文件转换为二进制可重定向文件

可重定向文件是机器代码的一致形式，它包含着所有指令的二进制形式，但是并没有填充全局变量的地址

链接：将多个可重定向文件p1.o和p2.o以及所调用的库函数功能重定向文件合并并生成最后的可执行文件p(-o p)

可执行文件是机器代码的第二种形式，是处理器执行的代码的确切形式。

## 3.2.1 *Machine-Level Code*

和1.9.3所描述的类似，计算机系统采用几种不同形式的抽象注释1，通过使用更简单的抽象模型来隐藏实现的细节

对于机器级编程而言，这些抽象层次中有两个是比较重要的

* 1. 指令集体系结构ISA：定义了处理器的状态，指令格式以及每条指令对状态的影响。大多数ISA定义了程序的指令序列是串行执行的
  2. 虚拟存储：提供了一个非常大的字节数组的存储模型

x86-64机器代码和C语言原始代码差别很大，一些对于C程序员来说不可见的处理器状态对于x86-64机器代码来说是可见的

* 1. PC：程序计数器，**在x86-64中标识为%rip，其中存储着下一条要执行的指令的地址**
  2. **整数寄存器文件包含着存储着64位值的16个命名位置**注释2
  3. **条件码寄存器保存着最近执行的算术或者逻辑指令的状态信息**
  4. 一组向量寄存器可以存放一个或多个整数或浮点数值

虽然C语言提供了一种模型，可以在内存中声明和分配各种数据类型的对象，但是机器代码只是简单地将内存看成一个很大的、按字节寻址的数组。C语言中的聚合数据类型，例如数组和结构，在机器代码中用一组连续的字节来表示。即使是对标量数据类型，汇编代码也不区分有符号或无符号整数，不区分各种类型的指针，甚至于不区分指针和整数。

在任何时候，只有有限的一部分虚拟地址被认为是合法的。例如x86-64虽然是64位的虚拟地址，但是在实践中高16位必须设置为0，从而实际能够访问的是范围内的一个字节

## 3.2.2 Code Examples

**|** gcc -0g -S mstore.c中的"-S"是表示只执行到编译阶段生成汇编代码文件

**|** gcc -0g -c mstore.c中的"-c"是表示只执行到汇编阶段生成可重定位的目标文件

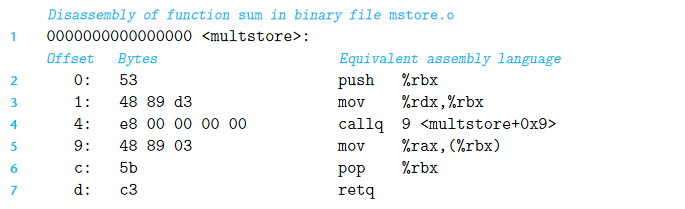
#### 反汇编器 disassembler

为了查看该机器代码文件的内容，需要使用一类称为反汇编器的程序

反汇编器将会从机器代码中生成一种和汇编代码格式类似的文件

在Linux/Windows系统中，使用命令objdump和-d命令参数来对机器代码文件进行解析。类似的解析结果如下：

objdump -d mstore.o



在左侧，有 14 个十六进制字节值，按前面显示的字节序列列出，分为 1 到 5 个字节的组。这些组中的每一个都是一条指令，其汇编语言等效项显示在右侧

可以从中总结到几个特征：

* 1. x86-64的指令长度从1-15字节不等，采用哈夫曼编码，使用频度高/操作数少的指令字节数少；使用频度低/操作数多的指令字节长
  2. 设计指令格式的方式是：**对于某种指令都给一个字节唯一用来标识此处开始为该指令**。例如上图中“push以53起始，mov以48起始，pop以5b起始，retq以c3起始，callq以e8起始”
  3. 反汇编器只是基于程序的机器代码文件中的字节序列来确定，不需要访问程序的源文件和汇编文件
  4. 反汇编器生成的汇编代码和编译器生成的汇编代码在指令的命名规则上有细微差别，例如指令后缀加不加“q”

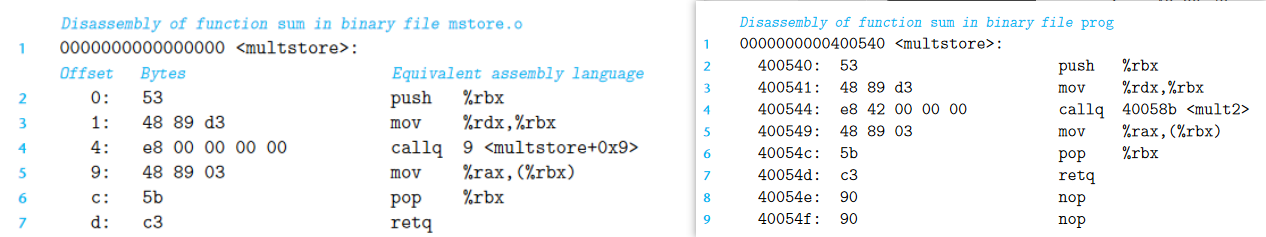
#### 链接main文件

生成实际可执行的代码需要对一组目标代码文件注释3运行链接器，而这一组目标代码文件中必须含有一个main函数

gcc -0g -o prog main.c mstore.c即预处理、编译、汇编、链接main和mstore两文件，最后生成prog可执行程序

#### 链接后反汇编与编译后反汇编的对比

链接之后的可执行文件的反汇编和编译之后的可重定向目标文件的反汇编示例如下：



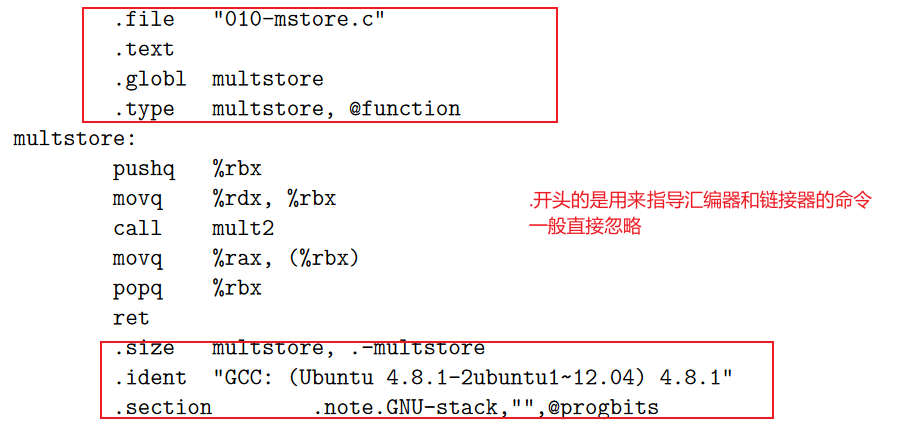
不同的点主要有：

* 1. 链接后的反汇编代码还包括用于启动和终止程序以及与操作系统交互的代码
  2. 链接后的反汇编代码已经给代码分配好了具体的地址范围，并直接指明了要跳转到的地址
  3. 插入的nop只是为了将函数代码变为16字节注释4，就存储器系统性能而言可以更好地放置下一个代码块

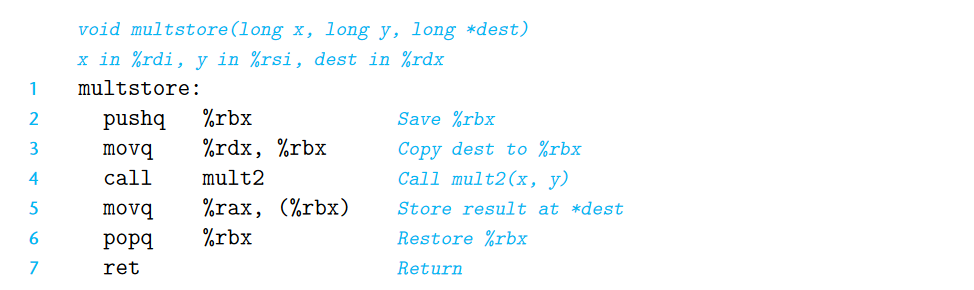
## 3.2.3 *Notes on Formatting*

经GCC编译器生成的汇编代码如下，并不容易阅读。有两方面的原因：

* 1. 该汇编代码中包含一些程序员不太需要关注的信息
  2. 该汇编代码并未提供任何关于程序或是它怎么工作的描述



下面是带有注释的、省略掉"."指令的汇编代码



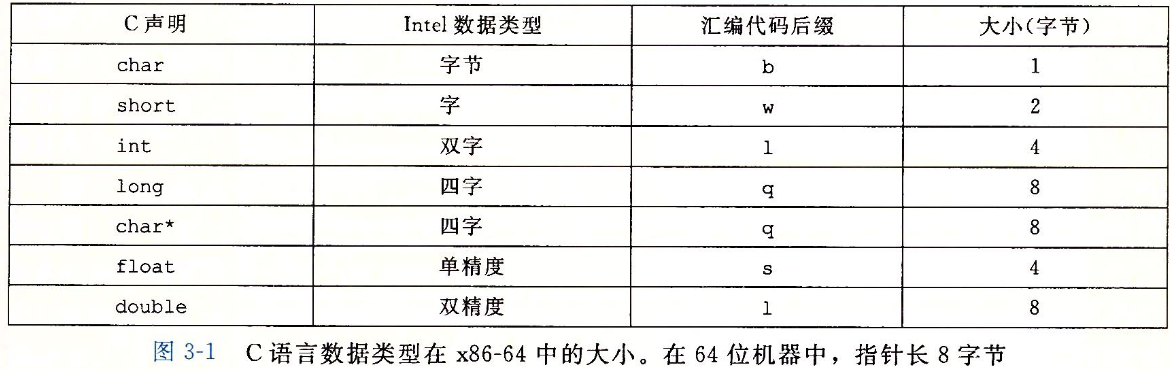
🐳对于一些应用程序，程序员必须用汇编代码来访问机器的低级特性。

一种方法是用汇编代码编写整个函数，在链接阶段把它们和C函数组  
合起来。  
另一种方法是利用GCC的支持，利用asm伪指令直接在C程序中嵌入汇编代码。

# 3.3 *Data Formats*

由于x86-64起源为 16 位架构，后来扩展到 32 位架构，Intel 使用术语“字Word”来指代 16 位数据类型。基于此，将 32 位数据量称为“双字DW”，将 64 位数据量称为“四字QW”

下图是C语言基本数据类型对应的x86-64表示



int占4字节，指针8字节，long占8字节

**大多数GCC生成的汇编代码指令都有一个字符后缀“b/w/l/q”，表示操作数的大小**。例如，数据传送指令有4个变种：movb、movw、movl、movq。注意float、double后缀是s、l。这不会引起歧义，因为浮点数与整数使用不同的指令和寄存器文件

# 3.4 *Accessing Information*

x86-64 中央处理单元 (CPU) 包含一组 16 个通用寄存器，用于存储 64 位值。这些寄存器用于存储整数数据和指针。如下图所示：



这16个寄存器名均以“%r”起始，但是对于32位、16位、8位的寄存器数据则有不同的命名规则。这是由于指令集的变革

* 1. 刚开始8086只有8个16位的寄存器，即ax~sp，根据它们的实际作用进行命名——高位h低位l
  2. 之后扩展到IA32，寄存器命名为eax~esp表示扩展到32位
  3. 再之后扩展到IA64，前8个寄存器命名为rax~rsp。并添加了剩下8个寄存器，命名r8~r15，32位为双字加后缀d，16位为字加后缀w，8位为字节加后缀b

因为指令允许对这16个寄存器的不同数据量位注释5进行操作，那么当将这些寄存器作为目的寄存器且操作数据大小少于8字节，就会遇到高位字节怎么办的情况：

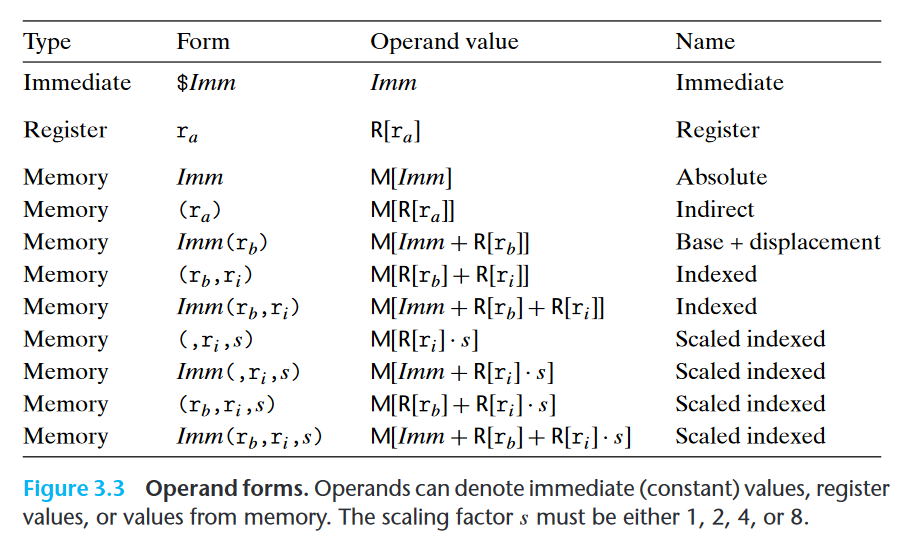
* 1. 操作1字节和2字节（字）的指令保持高位字节数据不变
  2. 操作4字节（双字）的指令，赋值高位4字节为0注释6

Figure3.2右侧的注释表示不太的寄存器在程序中充当的不同的作用

除了%rsp专用表示运行时栈的结束位置，并且有专门的指令来读写这个寄存器。其他15个寄存器的使用都较为灵活

## 3.4.1 *Operand Specifiers*

大多数指令都有一个或多个操作数指示符来指出执行一个操作所需要的源数据值以及结果要存放的目标位置。x86-64支持如下图所示的操作数格式：



各种不同操作数的可能性可分为三种类型：

* 1. 立即数

立即数表示常数值，书写格式为：$+C语言标准所表示的整数

eg:$12,$0x12

* 1. 寄存器数

寄存器数表示的是数据来源于x86-64的16个通用寄存器，且数据大小可以是字节、字、双字、四字。书写格式为：

* 1. 存储器数

根据存储器地址的表示方法可以分为：存储器立即寻址、寄存器间接寻址、寄存器相对寻址，变址寻址（带立即数和不带立即数），比例变址寻址(4种)

计算出的存储器地址称之为有效地址

🐳在这些寻址方式中，最常用的形式是其中**s必须是1、2、4、8，必须是64位的寄存器**。这个形式经常用来引用数组元素，其余寻址形式都可以省略这个形式的某部分来形成。

## 3.4.2 *Data Movement Instructions*

将许多不同的指令分为指令类，其中类中的指令执行相同的操作，但操作数大小不同

#### MOV类

|  |  |
| --- | --- |
|  | MOV类指令可以不需要任何转变的数将据从源位置复制到目的位置  MOV类包括movb、movw、movl、movq四类指令，它们执行的效果是相同的唯一的区别在于操作的数据大小不同：1、2、4、8 |

🐳源位置可以是一个立即数、寄存器数或者存储器数；目的位置可以是寄存器数或者存储器数。但是不允许源位置、目的位置都是存储器数

因此从存储器某位置复制数据到另一个位置需要使用两条指令：先mov到某寄存器，再由该寄存器mov到主存

mov指令使用的寄存器操作数可以是16个通用寄存器中的任意一个，但是需要保证使用寄存器片段的大小和指令的最后一个字符指令的大小相同(b、w、l、q)。且**对b、w、q指令的情况，mov指令只会更新目标操作数指示的特定区域，除了movl指令**注释7**会使得寄存器高4B为0**

下面是mov类指令的一些示例：

movl $0x4050,%eax %Immediate--Register,4 bytes  
movw %bp,%sp %Register--Register,2 bytes  
movb (%rdi,%rcx),%al %Memory--Register,1 byte  
movb $-17,(%esp) %Immediate--Memory,1 byte  
movq %rax,-12(%rbp) %Register--Memory,8 bytes

NASM

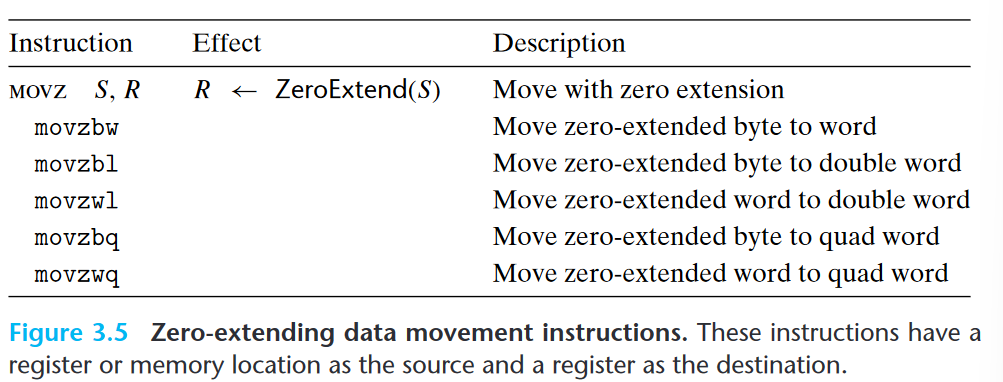
**|** Figure3.4中的最后一条指令movabsq是用来处理64位立即数的注释8。它可以直接使用任意的**64位立即数值**作为源操作数，但是**只能使用寄存器作为目的地址**

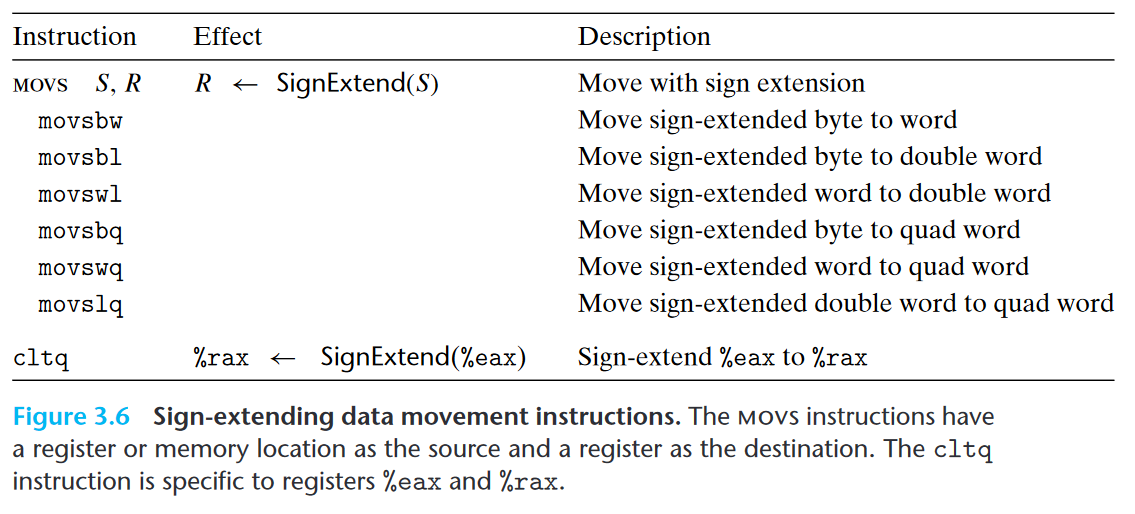
#### MOVZ、MOVS

🐳MOVZ和MOVS指令是用来将一个较小的数移动到一个较大的目的位置上。源位置可以是寄存器或者存储器，但是目的位置只能是寄存器。

1. MOVZ指令是将目的位置上的剩余字节填充0  
   MOVS指令是将目的位置上的剩余字节填充源操作数的MSB
2. 对于b、w、l、q四类数据大小，总共的指令类别有种

但是由于x86-64既有标准——对于任何产生32位结果的指令，若将结果放置在64位的目的处，则高位部分需设置为0。因此不存在movzlq



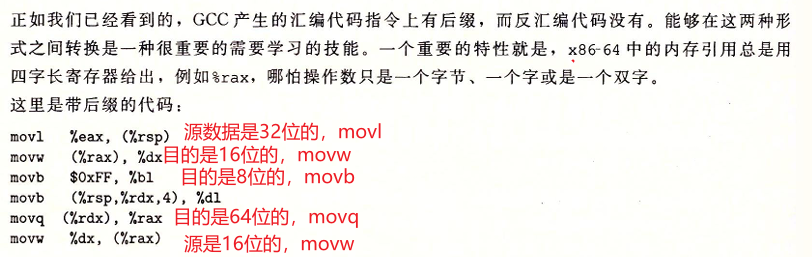


**|** cltq指令

cltq指令没有显式指定操作数，它隐含源操作数%eax以及目的操作数%rax，执行的功能是。等价于

但是相较于movslq指令需要指明操作数，cltq并不需要因此指令字长更紧凑

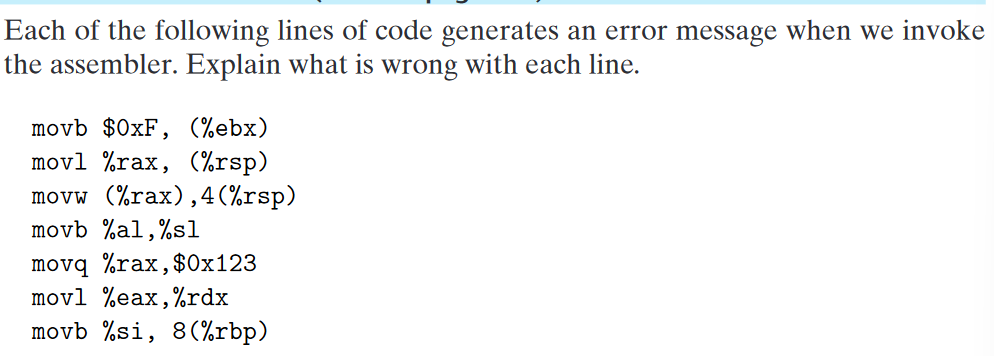
**|** 根据源、目数据类型填充mov指令



需要注意的是，**内存地址总是用64位的**

判断依据：**若目的位置是寄存器操作数，则需要根据寄存器所使用的大小决定mov的后缀；但如果使用的是存储器操作数，则需要根据源位置操作数大小来决定**

**|** 错误的指令示例：



1. 内存引用需要使用64位
2. movl是32位，但是%rax是64位
3. mov指令不能用于内存和内存之间的数据移动
4. 没有%sl这个寄存器，应该是%sil
5. 不能将立即数作为目的寄存器
6. movl是332位的，但是%rdx为64位
7. %si寄存器为16位，但是使用的是movb指令

## 3.4.3 *Data Movement Example*

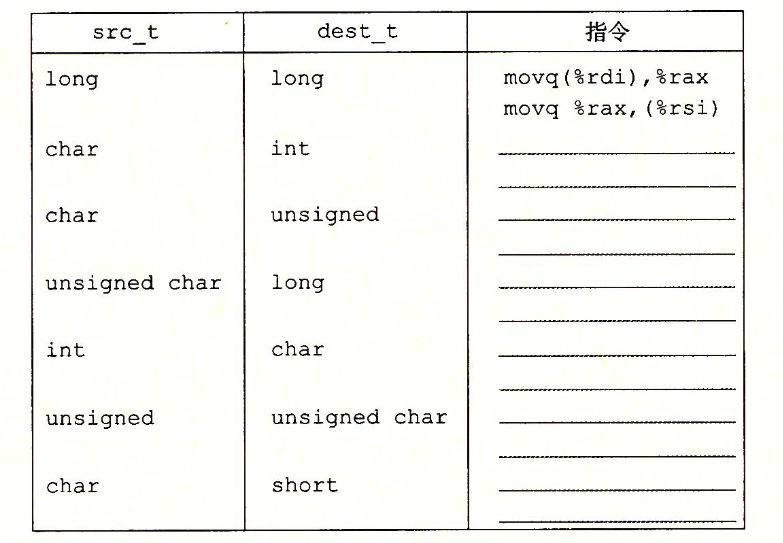
以Figure3.7中的交换值的代码示例为例进行分析

|  |  |
| --- | --- |
| 函数中的局部变量是被保存在寄存器中而不是内存中的。因为访问寄存器远快于访问主存 | 1. 返回值保存在%rax中；参数保存在%rdi、%rsi中   %rdi中存储的是xp指针的值 %rsi中存储的是y的值   1. 指针的解引用过程如下：    1. 首先将内存中指针的值赋给某个寄存器    2. 通过寄存器间接寻址的方式形成内存的有效地址，从而获得对应的值 |

**|** 使用MOV、MOVZ、MOVS类指令完成类型强制转换

\**dp=(dest\_t)*\*sp,sp存放在%rdi中，dp存放在%rsi中

中间变量使用%rax、%eax、%ax、%al



1. long→long：64位到64位

读数据：movq (%rdi),%rax

写数据：movq %rax,(%rsi)

1. char→int：8位到32位

读数据：movsbl (%rdi),%eax 因为写数据要求使用%eax则存数据也要使用%eax，所以需要进行符号扩展，使用movsbl指令

写数据：movl %eax,(%rsi) 这里必须要使用%eax因为数据大小由源来决定

1. char→unsigned：8位到32位

对于既要进行位扩展也要符号转换的情形：**C语言规定先进行位扩展再进行符号转换**

因此实际的调用指令同2

1. unsigned char→long：8位到64位

无符号数进行零扩展使用movzbl指令

读数据：movzbl (%rdi),%eax

写数据：movq %rax,(%rsi)

1. int→char：32位到8位

读数据：movl (%rdi),%eax

写数据：movb %al,(%rsi)

1. unsigned→unsigned char：32位到8位

同5

1. char→short：8位到16位

读数据：movsbw (%rdi),%ax

写数据：movw %ax,(%rsi)

## 3.4.4 *Pushing and Popping Stack Data* %rsp

|  |  |
| --- | --- |
|  | 栈顶采用“高到低”**向下增长**的形式且**栈顶有数据**  入栈：先减栈顶指针8，然后入数据 出栈：先取数据，再加栈顶指针8 |

|  |  |
| --- | --- |
|  | 在x86-64中，栈存放在内存中的某个区域。采用**向下增长的原则使得栈顶元素的地址是所有栈中元素地址最低的**——但是根据惯例，画栈时栈顶在最下端，采用向上增长 |

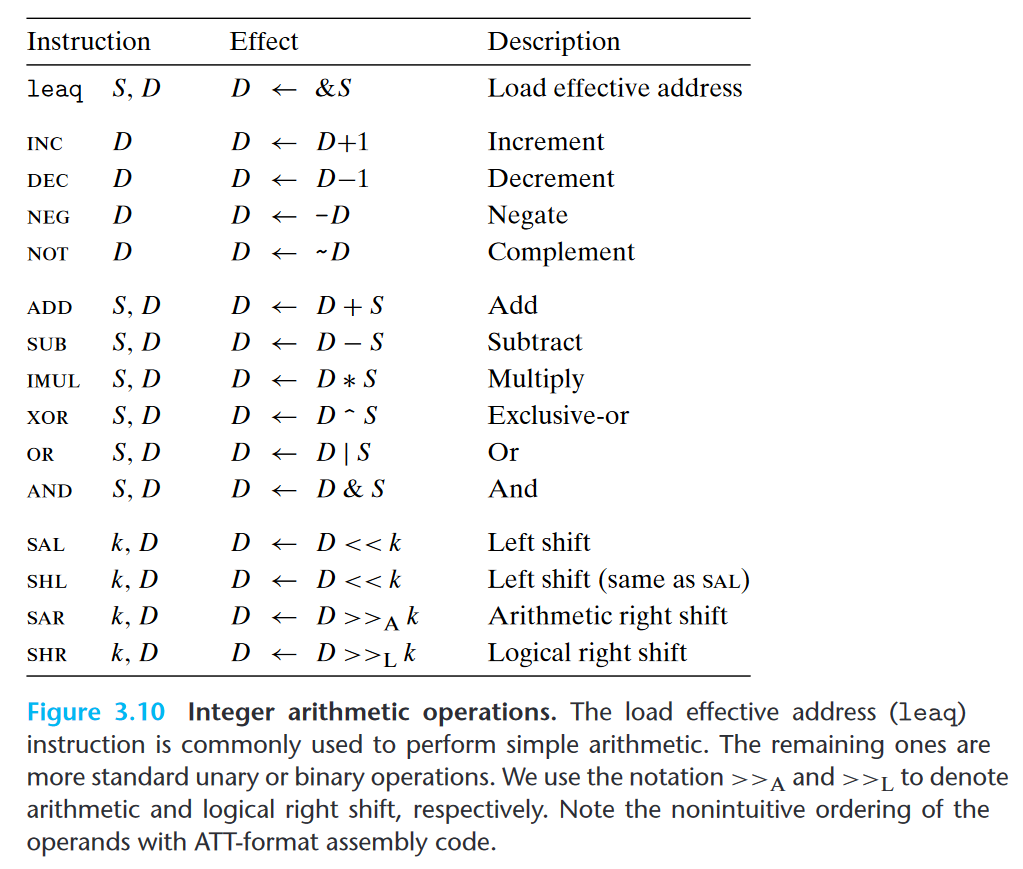
**出栈后，只是栈顶%rsp的值发生了变化，但是原栈顶位置上的元素还在直到被重新push覆盖**

**因为栈也存放在内存中，所以可以直接使用寄存器间接寻址的方式访问栈元素**

比如movq 8(%rsp),rdx则是将栈顶的下一个四字元素复制到%rax中

# 3.5 *Arithmetic and Logical Operations*

算术逻辑运算指令根据操作类型分为：加载有效地址、一元操作、二元操作和移位

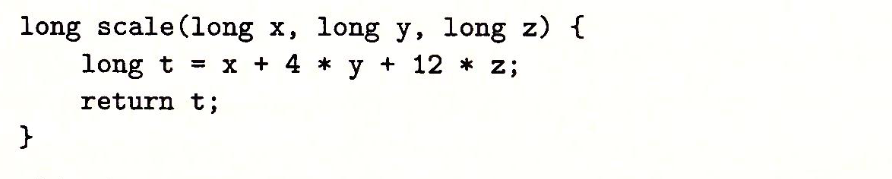


## 3.5.1 *Load Effective Address*

加载有效地址leaq指令实际上是movq的变形，它限制了源操作数必须是个内存引用，目的操作数必须是寄存器，但是并不从内存指定位置读取数据，而是将内存引用的有效地址写寄存器。leaq指令可以得到内存引用指针的值也可以进行一些利用缩放变址寻址方式实现一些简单的计算

eg：若%rdx值为x，则leaq 7(%rdx,%rdx,4),%rax即%rax=5x+7

因此编译器可以使用leaq指令来完成一些优化操作，如下：



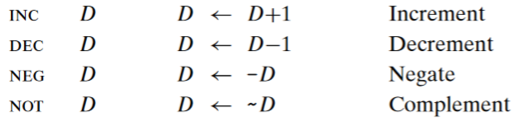
则若*x in %rdi,y in %rsi,z in %rdx*，那么x+4yi即leaq (%rdi,%rsi,4),%rax  
12z即3z\*4，leaq (%rdx,%rdx,2),%rdx,leaq(%rax,%rdx,4),%rax

🐳一定要注意leaq指令能够执行加法和有限形式的乘法，在编译一些简单的表达式时是十分有用的

## 3.5.2 *Unary and Binary Operations*

### 3.5.2.1 Unary Operations

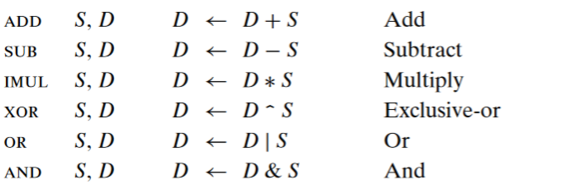
一元操作只有一个操作数，这个操作数既是源操作数也是目的操作数，**操作数的类型只能是寄存器或者存储器**。例如C语言中的自增、自减运算符



### 3.5.2.2 Binary Operations

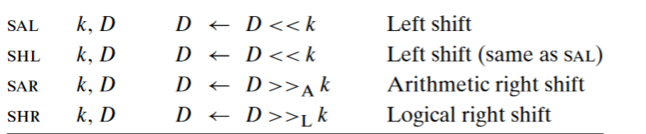
二元操作有两个操作数，其中**第二个操作数既是源操作数也是目的操作数**。例如subq %rax,%rdx执行的功能是

二元操作的第一个源操作数可以是立即数、寄存器或存储器；但是第二个操作数只能是寄存器/存储器注释9



## 3.5.3 *Shift Operations*

移位操作是二元操作：第一个操作数是移位量，移位量的表示可以使用立即数也可以使用%cl；第二个操作数是要移位的值，类型是寄存器或者存储器

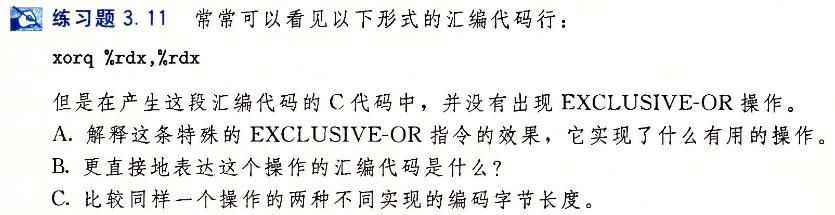


🐳原则上虽然使用%cl表示移位量，范围为0~255。但是在x86-64中，对w位长的数据值进行移位操作时，移位量是由%cl寄存器的低m位决定的，**也就是移位量%w**

例如当%cl=0xff时，salb会移动111b位，salw会移动1111b位，sall会移动11111b位，而salq会移动111111b位

## 3.5.4 *Special Arithmetic Operations*

**|** **使用异或指令实现寄存器的赋0**



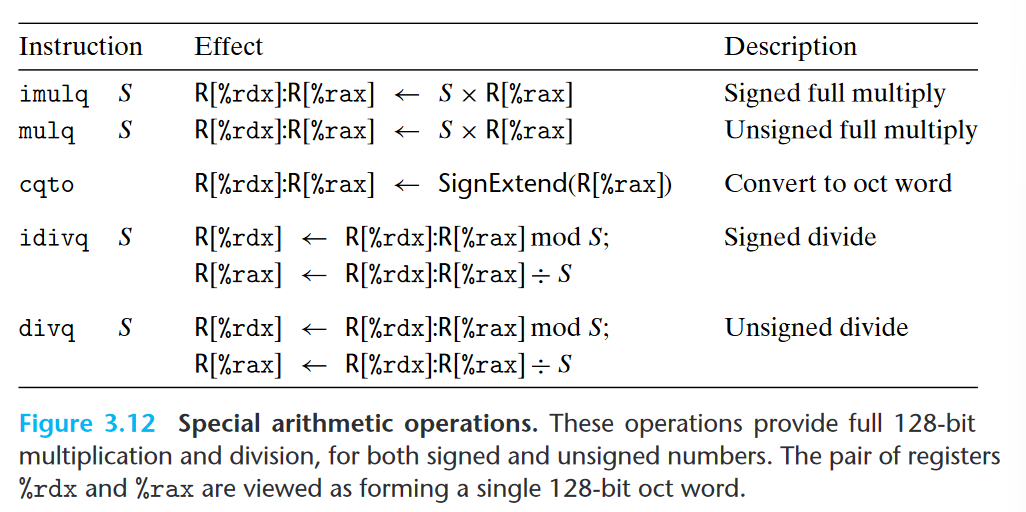
A.这条指令可以将%rdx置为0

B.movq $0,%rdx

C.使用XOR编码长度更少

正如我们在 2.3 节中看到的，两个 64 位有符号或无符号整数相乘可以产生需要 128 位来表示的乘积。 x86-64 指令集对涉及 128 位（16 字节）数字的操作提供有限的支持。继续采用字（2 字节w）、双字（4 字节l）和四字（8 字节q）的命名约定，**Intel 将 16 字节数量称为八进制字**o

下图描述了x86-64指令集中支持128位的整数乘法和除法



# 3.6 *Control*

[注释1] 虚拟机、进程(ISA+虚拟内存)、虚拟内存、文件

[注释2] x86-64只有16个整数通用寄存器

[注释3] main文件、库函数文件

[注释4] 2的n次幂

[注释5] B、W、DW、QW

[注释6] x86-64的惯例：任何为寄存器生成 32 位值的指令也会将该寄存器的高位部分设置为 0

[注释7] x86-64规定任何产生32位结果的指令，若要将结果放置在64位寄存器中，均需要将寄存器的高位设置为0

[注释8] 因为movq只能处理32位立即数，这个立即数然后会被符号扩展为64位

[注释9] 因为第二个操作数也要做目的操作数