The Programmer’s Guide to Numeric Conversions in X86

Scope: This guide explains the conversion of single numbers using a single instruction for each conversion. Vector conversions are not explained here. A vector conversion is the act of converting a collection of numbers (usually 2, 4, or 8 numbers) to another representation using a single X86 instruction to perform all the conversion concurrently.

Fixed numeric example

Select a number for specific examples. Let’s use 1970. There are many ways to write the mathematical value expressed by 1970. Here is a list of some ways to express that value.

1970	Decimal system, integer
1970.0	Decimal system, real
0x409EC80000000000	IEEE754 (64 bits)
0x44F64000	IEEE754 (32 bits)
0x67B2	IEEE754 (16 bits)
0x0000 0000 0000 07B2	Two’s complement (64 bits)
0x0000 07B2	Two’s complement (32 bits)
0x07B2	Two’s complement (16 bits)

All of the above express the same mathematical value, namely: one thousand nine hundred seventy. The first two are (mostly) not used in computer systems. The lower six are used in computer systems.

Examples of conversions also known as casting in the language of C++

1. Convert 64-bit long to 32-bit float
   cvtsi2ss xmm8, rcx
   Let rcx = 0x0000 0000 0000 07B2
   then the low half of xmm8 = 0x44F 64000

2. Convert 32-bit int to 32-bit float
   cvtsi2ss   xmm8, ecx
   Let ecs = 0x0000 07B2
   then the low half of xmm8 = 0x44F6 0000

3. Convert 64-bit long to 64-bit double
   cvtsi2sd   xmm8, rcx
   Let rcx = 0x0000 0000 0000 07B2
   then xmm8 = 0x409E C800 0000 0000

4. Convert 32-bit int to 64 bit double
   cvtsi2sd   xmm8, ecx
   Let ecx = 0x0000 07B2
   then xmm8 = = 0x409E C800 0000 0000

5. Convert 32-bit float to 64-bit long using rounding
   cvtss2si   rcx, xmm9
   Let the low half of xmm9 = 0x44F64000
   then rcx = 0x0000 0000 0000 07B2

6. Convert 32-bit float to 32-bit int using rounding
   cvtss2si   ecx. xmm9
   Let the low half of xmm9 = 0x44F6 4000
   then ecx = 0x0000 07B2

7. Convert 64-bit double to 64-bit long using rounding
   cvtsd2si   r12, xmm10
   Let r12 = 0x0000 0000 0000 07B2
   then xmm10 = 0x409E C800 0000 0000

8. Convert 64-bit double to 32-bit float using rounding
   cvtsd2si   xmm0,   xmm15
   Let xmm15 = 0x409E C800 0000 0000
   then the low half of xmm0 = 0x0000 07B2

To the 240 class:

The previous page show 8 different types of conversions (casts). There are a 4 more not recorded above. Those are conversion using truncation in place of rounding.

So witch ones are important? In my life-time I have only used two of the above conversions. Those two are repeated below.

3. Convert 64-bit long to 64-bit double

    cvtsi2sd   xmm8, rcx

    Let rcx = 0x0000 0000 0000 07B2
    then xmm8 = 0x409E C800 0000 0000

7. Convert 64-bit double to 64-bit long using rounding

    cvtsd2si   r12, xmm10

    Let r12 = 0x0000 0000 0000 07B2
    then xmm10 = 0x409E C800 0000 0000

Take it for what its worth. Those are the two conversion I have needed to use in the past. It is true that my personal programming is overwhelmingly 64-bit programming, but nevertheless those are the only two I have needed to use.

On the next page is a table I found somewhere on the Web. I forgot where I got, but here it is anyway.

Conversion Instructions (SSE)

The SSE conversion instructions convert packed and individual doubleword integers into packed and scalar single-precision floating-point values.

Table 3–32 Conversion Instructions (SSE)

Solaris Mnemonic	Intel/AMD Mnemonic	Description	Notes
cvtpi2ps	CVTPI2PS	convert packed doubleword integers to packed single-precision floating-point values	
cvtps2pi	CVTPS2PI	convert packed single-precision floating-point values to packed doubleword integers	
cvtsi2ss	CVTSI2SS	convert doubleword integer to scalar single-precision floating-point value	
cvtss2si	CVTSS2SI	convert scalar single-precision floating-point value to a doubleword integer	
cvttps2pi	CVTTPS2PI	convert with truncation packed single-precision floating-point values to packed doubleword integers	
cvttss2si	CVTTSS2SI	convert with truncation scalar single-precision floating-point value to scalar doubleword integer	