Closing Up Computer Arithmetic

3-5-2003

Opening Discussion

- What did we talk about last class?
- Have you seen anything interesting in the news?
- Let's redo an FP multiplication example because I botched the in class example. Don't try to put decimal places in the partial results.

Rounding

- For actually doing the arithmetic, they keep two extra bits. These bits are used give them information on rounding.
- They need two because a multiplication can have a leading zero and must be shifted for normalization.
- There is also a "sticky bit" that is set if there are any bits on beyond the round bit.

Round To Even

- There is still the question of where to round if there is only one bit on and it is the first rounding bit (guard). You have probably been taught to always round 0.50 up, but for best accuracy you should round up half the time and down the other half.
- They do this by rounding up if the last real bit is 1 and down if it is zero.

Special Forms

- Zero exponent is used for zero and for denormalized numbers. Not all computers support denormalized forms because they can slow things down.
- Having the maximum value in the exponent (255 for single precision) is used for ±inf (zero mantissa) and NaN (nonzero mantissa)

Bits Have No Meaning

- If I just hand you a 32-bit word from a MIPS machine, you have no way of telling if it is an integer, a single precision floating-point value, or even an instruction.
- All of the meaning is given by context and the way in which we read it and what we do with it.

Fused Multiply-Add

- The PowerPC processor is virtually identical to MIPS as far as this chapter is concerned, except that they have an extra instruction for floating-point values that does a multiply followed by an add.
- This can be helpful for many types of calculations and having it as a single instruction increases speed and accuracy (less rounding).

80x86 Floating-Point

- The 8087 uses a stack type of architecture for doing floating-point operations. Also uses an 80-bit internal representation. The extra accuracy comes in handy for some applications.
- This approach has typically been slow. AMD did better with it in the Athlon than Intel has, but both lag behind other processors. That is part of the motivation for SSE and SSE2 instructions.

Accuracy of Floating-Point

- Floating-point addition is not associative. When you subtract numbers of almost the same size you lose a LOT of accuracy.
- Sometimes you can reorganize expressions to remove those types of operations. Other times you can't and those types of equations will become poorly behaved under certain situations.

Minute Essay

- What time Thursday do you want to have a review session? I'm free for most of the afternoon. It is likely that we will have to do it in either 340 or the Atlas lab because 228 will be in use most of the afternoon.
- There is a review sheet on the web. Note that the test is open book. If anything, that means the questions might be just a bit harder.