Subject: Re: How Computers Represent Floats
Posted by William Clodius on Thu, 30 Nov 2000 08:00:00 GMT
View Forum Message <> Reply to Message

"William B. Clodius" wrote:
> <snip> IEEE 754 requires that all intermediate calculations
> be performed a higher precision so
Ignore the above incomplete sentence. What I originally attempted to write was covered later.
>
> <snip>

Some other surprises.

The definition of the IEEE 754 mantisa, an integer with values from 2^n_mant to 2*2^n_mant-1, where n_mant is the number of bits available for the mantisa, is termed a normalized number. This is error prone for very small numbers. IEEE 754 mandates that there be available for such small numbers what are termed denorms where the mantissa is interpreted as an integer from 0 to 2^n_mant, so that accuracy degrades gradually for such values. However, this complicates the implementation of the floating point, so some processors, e.g., the DEC Alpha make this available only in software at a greatly reduced performance.

The special values for IEEE 754 were chosen with specific applications in mind and are not always the best for other applications. In particular some applications benefit from unsigned zeros and infinities others from signed NaNs none of which are provided by IEEE 754. Only sophisticated users tend to complain about this.

The default rounding behavior for IEEE 754 from an "extended" precision intermediate that is exactly halfway between values in the result representation, always to the same final bit (effectively alternately up and down), often surpises very observant users, although it is designed to reduce the propagation of systematic errors in most applications.