<https://patents.google.com/patent/US5892697A/en>

Method and apparatus for handling overflow and underflow in processing floating-point numbers

A method for processing floating-point numbers, each floating-point number having at least sign portion, an exponent portion and a mantissa portion, comprising the steps of converting a floating-point number memory register representation to a floating-point register representation; rounding the converted floating-point number; performing an arithmetic computation upon said rounded number resulting in a new floating-point value; and converting the resulting new floating-point register value to a floating-point memory register representation. An apparatus for processing floating-point numbers comprising means for converting a floating-point value from a random access memory register to a floating-point unit register; means for selecting the type of accuracy required of the converted value; means for selecting the type of rounding required of the converted floating-point value; means for performing an arithmetic operation upon at least one converted floating-point value based upon said selected accuracy and rounding; means for chopping the result of the arithmetic operation to fit a mantissa portion of a floating-point value; and means for converting the resulting floating-point value from the floating-point register representation to the random access memory representation.