Massachusetts Institute of Technology

Department of Electrical Engineering and Computer Science

6.313 Contemporary Computer Design

Problem Set #6 Solutions Fall, 1995

1. a)

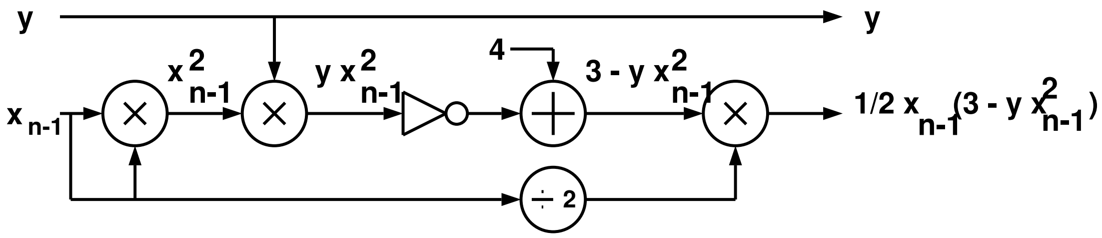
$$g(x) = \frac{1}{x^2} - y \tag{1}$$

$$g\prime(x) = -2x^{-3} \tag{2}$$

b)

$$\begin{aligned} x_n &= x_{n-1} - \frac{g(x_{n-1})}{g\prime(x_{n-1})} && (3)\\ &= x_{n-1} - \frac{x_{n-1}^{-2} - y}{-2x_{n-1}^{-3}} && (4)\\ &= x_{n-1} + \frac{1}{2}(x_{n-1} - yx_{n-1}^{3}) && (5)\\ &= \frac{1}{2}x_{n-1}(3 - yx_{n-1}^{2}) && (6)\end{aligned}$$

c)



d) Newton's method always converges quadratically (the number of bits of precision *approximately* doubles with each iteration).

e) The input number is normalized by bit pairs. When normalizing y, the number is shifted by M bits so we are looking up a value for $\frac{1}{\sqrt{\frac{y}{2^M}}} = \frac{2^{M/2}}{\sqrt{y}}$, meaning M must be even for us to be able to scale back for the normalization. So y is shifted an even number of times, and the leading two bits can be 01, 10, or 11. In other words, $1 \le y < 4$.

e) The number of iterations required depends on the number of bits in your initial estimate. For an 8 bit guess, only 3 iterations are required to reach a result accurate to 32 bits.

2. Along with the mantissa alignment, the two numbers must be checked for zeros

before computation. If the numbers are non-zero, after the mantissas are aligned, the signs of the numbers must be compared. This result is shared for addition and subtraction. Also, in the normalization stage, one of the results will need at most a single shift for normalization while the other requires checks for multiple shifts. This more difficult normalization only need be performed on one result.

3. a) The smallest single precision normalized number in the IEEE standard is 1.0×2^{-126}. The smallest denormalized number is $1 \times 2^{-23} \times 2^{-126}$ for an extra 23 bits of precision. And with double precision, 52 extra bit positions are available.

b) Adds and subtracts may require that the mantissa be shifted repeatedly to normalize the number. The MSB must be repeatedly checked and the exponent decremented, and if underflow occurs which requires gradual underflow, the exponent must be zeroed and the mantissa left in its denormalized position. Other operations require at most one shift. This extra compare and shift hardware could require 20-40% extra area.

c) Taken from Patterson & Hennessy's *Computer Architecture: A Quantitative Approach*. The IEEE standard specifies that if user trap handlers are enabled, the system must trap if the result is denormal. If trap handlers are disabled, the underflow flag is set only if the result must be rounded. There are a few practical codes that generate many underflows even when working properly, and these programs usually run slower on systems that require denormals to be processed by a trap handler.