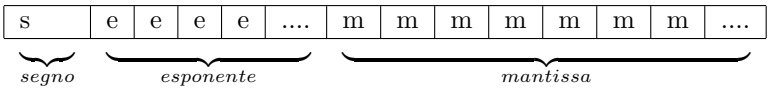
Calcolo Numerico: Q&A's

Giacomo Cusinato

November 7, 2014

1 Descrizione del sistema floating-point

Il sistema floating-point è un metodo di rappresentazione numerica utilizzato dai calcolatori per rappresentare numeri razionali ed approssimazioni dei numeri reali in modo tale da ottenere un compromesso tra range e precizione. La rappresentazione in floating-point normalizzata in base 2 è quella utilizzata dallo standard IEEE: i numeri sono scritti nrlla forma $x = f2^e$ dove $f = \pm 1.f_{-1}f_{-2}...f-n$ e $e = \pm e_{Ne-1}e_{Ne-2}...e0$ sono le cifre rispettivamente della mantissa e dell'esponente, in numero finito e valiri binari. La rappresentazione floating point può essere espressa così:



Nel sistema IEE, la rappresentazione in singola precisione è a 32bit con (1 bit per il segno, 8 per l'esponente e 23 bit per la mantissa) mentre quella in doppia precisione viene rappresentata in 32 bit (1 bit per il segno, 11 bit per l'esponente e 52 bit per la mantissa).

2 Precisione di macchina nel sistema floating-point