Arithmétique en virgule flottante

Pour l'addition et la soustraction, il faut d'abord ajuster l'exposant du plus petit des deux opérandes pour que les deux exposants soit identiques

Il suffit alors de décaler la mantisse du plus petit nombre à droite (right shift) (y compris le bit implicit) et d'incrémenter son exposant

Si les exposants sont très différents, on risque de perdre des bits significatifs du plus petit nombre

Pour la soustraction, si les deux nombres diffèrent de peu, il peut y avoir une perte considérable de bits significatifs du résultat

Pour la multiplication, il suffit de multiplier les mantisses et d'additionner les exposants. Si l'on utilise un additionneur normal, il faut soustraire 127 du résultat pour obtenir la bonne représentation de l'exposant. Il faut aussi éventuellement normaliser le résultat, car la mantisse du résultat peut être supérieure ou égale à 2

Pour la division, il faut diviser les mantisses et soustraire les exposants