

## How to Floating Point Quantize

To convert from a number into a scale-mantissa floating point code with  $R_s$  scale bits and  $R_m$  mantissa bits:

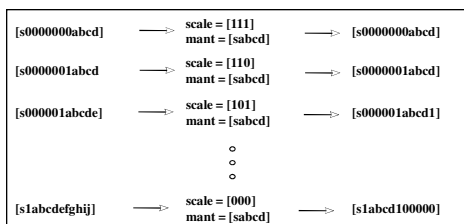
- I. Quantize the number as an  $R$  bit code where  $R=2^{R_s}-1+R_m$ .
- II. Count the number of leading zeros in  $|code|$ . If the number of leading zeros is less than  $2^{R_s}-1$  then set the scale equal to the number of leading zeros; otherwise set the scale equal to  $2^{R_s}-1$ .
- III. If scale equals  $2^{R_s}-1$ , then set the first mantissa bit equal to  $s$  and set the remaining  $R_m-1$  bits equal to the bits following the  $2^{R_s}-1$  leading zeros in  $|code|$ ; otherwise set the first mantissa bit equal to  $s$  and set the remaining  $R_m-1$  bits equal to the bits following the leading zeros, omitting the leading one.

## How to Floating Point De-Quantize

To convert from scale-mantissa floating point code with  $R_s$  scale bits and  $R_m$  mantissa bits into a number:

- I. Create an  $R$  bit code where  $R=2^{R_s}-1+R_m$  from the mantissa and scale factor where  $s$  is the first mantissa bit and  $|code|$ 
  - A. has scale leading zeros
  - B. followed by the remaining  $R_m-1$  mantissa bits if scale is  $2^{R_s}-1$ , otherwise followed by a one and then the remaining mantissa bits
  - C. followed by a one and as many trailing zeros as will fit if scale is less than  $2^{R_s}-1$ .
- II. Dequantize the  $R$  bit code into the number.

## Example: $R_s=3$ , $R_m=5$



## Block-Floating Point

- We often choose to do block-floating point quantization
  - We use only one scale factor for a group of adjacent samples (this step will become important when we quantize in the frequency domain and the grouping of adjacent frequency lines mirrors the critical bandwidth distribution)
  - We shift all samples in the group by the same value scale
- This implies that the number of leading zeros in scale is relative to the size of the largest sample in the group rather than relative to each samples
- When we quantize, we do not omit the one following the leading zeros when scale is less than  $2^{R_s}-1$
- When we de-quantize, if the mantissa is zero, we do not add trailing 1 when scale is less than  $2^{R_s}-1$