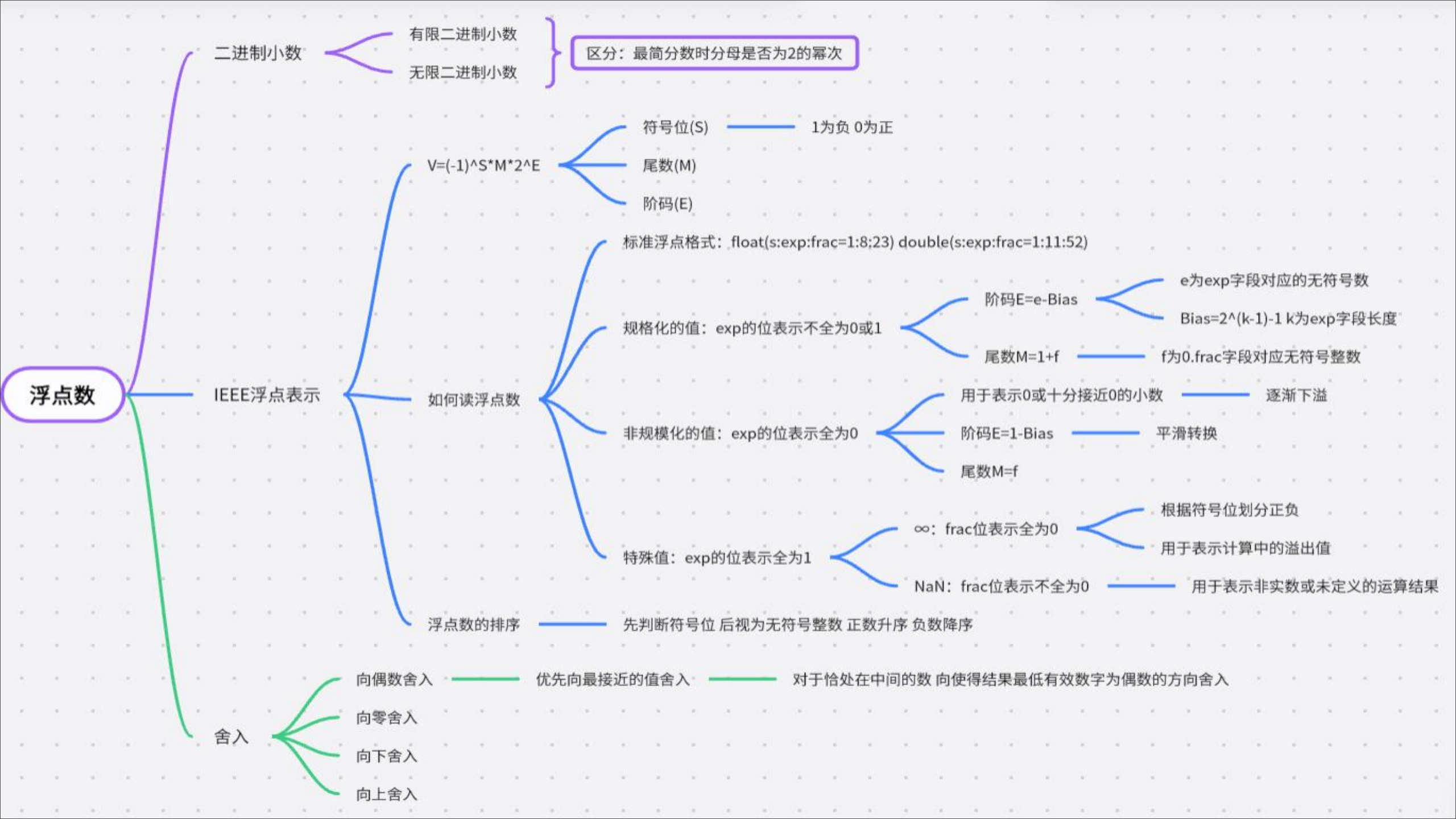
浮点数

连昊


浮点数
二进制小数
有限二进制小数
无限二进制小数
区分：最简分数时分母是否为2的幂次
IEEE浮点表示
V=(-1)^S*M*2^E
符号位(S)
1为负 0为正
尾数(M)
阶码(E)
如何读浮点数
标准浮点格式：float(s:exp:frac=1:8:23) double(s:exp:frac=1:11:52)
规格化的值：exp的位表示不全为0或1
阶码E=e-Bias
e为exp字段对应的无符号数
Bias=2^(k-1)-1 k为exp字段长度
尾数M=1+f
f为0.frac字段对应无符号整数
非规模化的值：exp的位表示全为0
用于表示0或十分接近0的小数
逐渐下溢
阶码E=1-Bias
平滑转换
尾数M=f
特殊值：exp的位表示全为1
∞：frac位表示全为0
根据符号位划分正负
用于表示计算中的溢出值
NaN：frac位表示不全为0
用于表示非实数或未定义的运算结果
浮点数的排序
先判断符号位 后视为无符号整数 正数升序 负数降序
舍入
向偶数舍入
优先向最接近的值舍入
对于恰处在中间的数 向使得结果最低有效数字为偶数的方向舍入
向零舍入
向下舍入
向上舍入


浮点数
浮点运算
浮点乘法
可交换 不可结合 不可分配
单调性(a、b、c不为NaN)
若a≥b,c≥0 则 a*c≥b*c
若a≥b,c≤0 则 a*c≤b*c
a*a≥0
运算方式
符号位 S=S1^S2
尾数 M=M1*M2
位码 E=E1+E2
调整M至[1,2)并对应调整E
若E过大 则溢出
否则进行舍入
浮点加法
可交换但不可结合
单调性：若 a≥b 则 x+a≥x+b （x不为NaN）
运算方式
符号位 S根据运算结果确定
尾数 M根据错位相加运算得到
位码 E=max{E1,E2}
调整M至[1,2)并对应调整E
若E过大 则溢出
否则进行舍入
C语言中的浮点数
表示范围（是否溢出）：double>float>int
表示精度（是否舍入）：double>int>float
float/double -> int 溢出：转化为Tmin

感谢聆听！