E指数设计文档

实现目标

基于FP16的基本性质，，y只使用规范数的情况下有：

因此有：

因此对 的指数部分，一种只有18中情况：00001（-14）~10010（3）

误差评估方式：

1. \*100% (表示计算值，表示标准值)
2. = （表fp16表示规格化数时的最小精度）

总体思路：查找表

算法实现：

根据FP16的基本表示,包括符号位，指数位，尾数位，并且有：

有：

**步骤一：指数查表**

这里定义为指数值，可以通过指数表获得e指数计算结果，表的深度为18，定义为指数表。根据实验结果，为整数，只有18种可能情况，其范围为：

**步骤二：尾数移位**

对于，在硬件上可以通过移位进行实现，因此有：

因此可以将表示为：

**步骤三：尾数查表**

进一步实验可知，当, 即达到FP16表示的最大范围时，对于x，有：

因此，对于中的值，有一下情况：

当时，的取值是：00,01,10,11

当时，的取值是：00,01,10,11

当时，的取值是：00,01

当为其他值时，的取值是：00

**步骤四：偏移查表**

因此，对于一共只有四种情况，也可以通过查表获得表的深度为4，查询结果为此表定义为偏移表。进一步的，因为与的表可以进行融合，融合后表的深度为4+4+2+15 = 25。

**步骤五：尾数查表**

对于，的位宽是10位，这里做了一个权衡，只选取b的前7位生成尾数表，权衡考虑的点为一下两点：

1. 对于左移的情况，已经基本上做到了全覆盖
2. 对于右移的情况，因为此时已经较小，因此误差所引起的整体值误差已经很小，另一方面是一个很小的值，因此b末3位带来的误差也会很小。

因此针对b前7位，生成一个深度为128的表，查询结果为定义为尾数表。

至此可以有：

**步骤六：结果补偿**

但是由于尾数段的省略，会引入一些误差，此时的一定小于真实值。因此这里引入补偿，基于一些实验，补偿方案选择如下：

对于的尾数段直接进行补偿，保留b的后三位，因为e指数和加法不是线性映射关系，因此，对于后三位进行非线性映射后。

当x为正数时，result的尾数段加上补偿值，这里需要考虑进位，又因为这种补偿只会在x的exp小于等于10001时发生，因此一定不会产生溢出。

当x为负数时，result的尾数段减去补偿值，这里需要考虑进位，当exp=00001时，会产生下溢出情况，但由于此时的数据已经非常小了，因此暂时考虑不做补偿。

对于后三位的非线性映射表：（待定）

尾数补偿示例：

其中为最终结果值

误差统计

正数部分

|  |  |  |  |
| --- | --- | --- | --- |
| **指数位** | **绝对误差** | **相对误差** | **尾数值** |
| **00001** | **0.0000000000** | **0.00000000%** | **1111111111** |
| **00010** | **0.0000000000** | **0.00000000%** | **1111111111** |
| **00011** | **0.0000000000** | **0.00000000%** | **1111111111** |
| **00100** | **0.0000000000** | **0.00000000%** | **1111111111** |
| **00101** | **0.0009765600** | **0.09746589%** | **1111111111** |
| **00110** | **0.0009765600** | **0.09737098%** | **1111111111** |
| **00111** | **0.0009765600** | **0.09718173%** | **1101111001** |
| **01000** | **0.0029296900** | **0.28873917%** | **1110110001** |
| **01001** | **0.0029296900** | **0.28653295%** | **1111000001** |
| **01010** | **0.0029296900** | **0.28195489%** | **0111101100** |
| **01011** | **0.0029296900** | **0.26501767%** | **1001110010** |
| **01100** | **0.0029296900** | **0.24213075%** | **1111101111** |
| **01101** | **0.0039062500** | **0.27322404%** | **1111110111** |
| **01110** | **0.0068359400** | **0.34146341%** | **0110001111** |
| **01111** | **0.0156250000** | **0.28985507%** | **1111011111** |
| **10000** | **0.0937500000** | **0.25759918%** | **1111111111** |
| **10001** | **8.0000000000** | **0.30565461%** | **1111110001** |
| **10010** | **96.0000000000** | **0.19286403%** | **0110001011** |

负数部分

|  |  |  |  |
| --- | --- | --- | --- |
| **指数位** | **绝对误差** | **相对误差** | **尾数值** |
| **00001** | **0.0000000000** | **0.00000000%** | **1111111111** |
| **00010** | **0.0000000000** | **0.00000000%** | **1111111111** |
| **00011** | **0.0004882800** | **0.04885198%** | **1111111111** |
| **00100** | **0.0004882800** | **0.04887586%** | **1111111111** |
| **00101** | **0.0009765600** | **0.09784736%** | **1111111111** |
| **00110** | **0.0014648400** | **0.14705882%** | **1111111111** |
| **00111** | **0.0024414100** | **0.24606299%** | **1111111111** |
| **01000** | **0.0034179700** | **0.34687810%** | **1101111111** |
| **01001** | **0.0029296900** | **0.30196276%** | **1110111111** |
| **01010** | **0.0029296900** | **0.31152648%** | **1111011111** |
| **01011** | **0.0029296900** | **0.31397174%** | **0001101111** |
| **01100** | **0.0024414100** | **0.28968714%** | **0101110111** |
| **01101** | **0.0014648400** | **0.20477816%** | **0101011011** |
| **01110** | **0.0019531200** | **0.38797284%** | **0101111110** |
| **01111** | **0.0007324200** | **0.29154519%** | **0110000111** |
| **10000** | **0.0003051800** | **0.26939655%** | **0001011011** |
| **10001** | **0.0000305200** | **0.29182879%** | **1110011111** |
| **10010** | **0.0000002400** | **0.17621145%** | **0011001101** |

是正数部分

时,数据表示的范围：

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

指数查表：

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

误差情况：

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 绝对误差 |  |  |
| 相对误差 |  |  |

最小相邻值误差情况：

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

时：

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

时,数据表示的范围：

|  |  |
| --- | --- |
|  |  |
| 65758.8801344867(65504) | 2980.9579870417 |
|  |  |