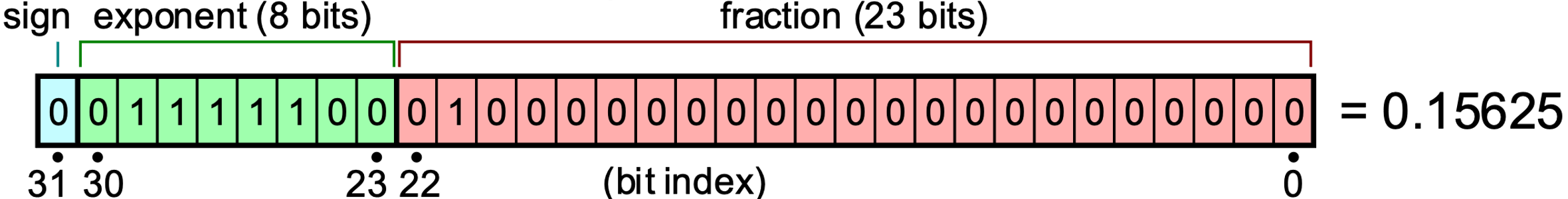
**Lecture 16: Scale for GPU Pool**

# IEEE754: Float Point System



可以理解为把 u8 转化为 i8

Mantissa 忽略先导1（考虑二进制，科学计数法总是 1Exxx，先导1总是出现，这也是为什么1+项对出现）。表示小数权重（）：



GPU Memory Footprint (llama 65B @ FP32)

|  |  |  |
| --- | --- | --- |
| Llama 1 65B | | |
|  | Model dimension | 8192 |
|  | Sequence Length | 2048 |
|  | Batch size | 1 |
|  | Number of layers | 80 |
|  | Number of attention heads | 64 |

Inference Stage:

Training Stage: Model: 260 GiB, Gradient: 260 GiB  
Optimiser: 520 GiB (Adam = Momentum + Variance)