Titan-TPU V2 数据格式与量化标准

1. 定点数表示 (Fixed Point Representation)

Titan-TPU V2 核心计算单元采用 **INT8** 和 **INT16** 混合精度计算，但在顶层接口表现为 **Q8.8 格式**。

1.1 Q8.8 格式定义

- **总位宽**: 16 bits
- **符号位**: 1 bit (MSB)
- **整数位**: 7 bits
- **小数位**: 8 bits
- **表示范围**: -128.00 到 +127.996
- **分辨率**: $2^{-8} = 0.00390625$

1.2 示例

- `0x0100` = $1 \times 2^0 + 0 = 1.0$
- `0x0080` = $1 \times 2^{-1} = 0.5$
- `0xFF00` = -1.0 (补码表示)

2. 累加器精度 (Accumulator Precision)

为了防止在长时间 MAC 运算中发生溢出，PE 内部累加器采用扩展精度。

- **输入位宽**: 16-bit
- **权重位宽**: 16-bit
- **部分和 (PSum) 位宽**: 32-bit
- **VPU 输出位宽**: 截断回 16-bit (带饱和截断 Saturation)

3. 饱和逻辑 (Saturation Logic)

当计算结果超出 16-bit 表示范围时，硬件自动将其钳位到最大值或最小值，而不是发生回绕 (Wrap-around)。

- `If Result > 32767, Output = 32767`
- `If Result < -32768, Output = -32768`