NPU接口驱动说明

1. 接口说明

|  |  |  |  |
| --- | --- | --- | --- |
| 接口信号名称 | I/O | 位宽 | 功能描述 |
| clk | I | 1 | 系统时钟 200MHz |
| rst\_n | I | 1 | 系统复位，低电平有效 |
| wr\_sop\_weight | I | 1 | 权重加载起始信号，1有效，持续一个周期 |
| wr\_eop\_weight | I | 1 | 权重加载结束信号，1有效，持续一个周期 |
| wr\_vld\_weight | I | 1 | 权重值有效信号，1有效 |
| wr\_data\_weight | I | 32 | 权重数据 |
| wr\_sop\_data | I | 1 | 特征图加载起始信号，1有效，持续一个周期 |
| wr\_eop\_data | I | 1 | 特征图加载结束信号，1有效，持续一个周期 |
| wr\_vld\_data | I | 1 | 特征图值有效信号，1有效 |
| wr\_data\_data | I | 32 | 特征图数据 |
| save\_finish | O | 1 | 卷积计算结束信号，1有效，持续到数据被全部读出为止 |
| rd\_sop\_x | I | 1 | 读取卷积计算结果起始信号，1有效，持续一个周期 |
| rd\_eop\_x | O | 1 | 卷积计算结果输出结束信号，1有效，持续一个周期 |
| rd\_vld\_x | O | 1 | 卷积计算结果值有效信号 |
| rd\_data\_x | O | 32 | 卷积计算结果 |
| 注：x取值范围为0~7，代表卷积计算结果输出端口共有8组。 | | | |

1. 接口时序
   1. 写权重/写数据

写数据和写权重端口的时序完全一致，且二者写入的时间没有限制，可同时写权重写数据。可先写权重，后写数据。可先写数据，后写权重。具体可参考图2.1。



图2.1 写权重/写数据接口时序

注意：

①wr\_data\_weight/wr\_data\_data位宽为32位，一个权重或一个特征的数据位宽是16位，因此一个wr\_data\_weight/wr\_data\_data能够表示2个权重或特征。

②图2.1中，wr\_data\_weight第一个数据是S，该数据目前没有用途，但必须要先写一个。

③数据在写入时，可以不连续，即wr\_vld\_weight可不一直保持高电平，但一次传输数据量要保证10个。

* 1. 读计算结果

读数据接口共有8组，每组接口时序完全一致。下图2.2展示了第一组接口的读取时序。



图2.2 读计算结果接口时序

注：

①rd\_sop\_0拉高后的一个周期，rd\_vld\_0必然为高电平，且输出是连续的。即当rd\_sop\_0拉高后，接下来8个时钟周期一定会将8个计算结果输出，并在第9个时钟周期拉高rd\_eop\_0信号。

②rd\_sop\_0可在save\_finish拉高后的任意时间间隔内拉高。当所有通道的数据全部读出后，save\_finish信号才会拉低。

1. 测试用例

NPU中加速单元采用脉动阵列，脉动阵列规格如图3.1所示，规格为8\*8，一次工作能够计算出8个特征图与8个通道之间的卷积运算。

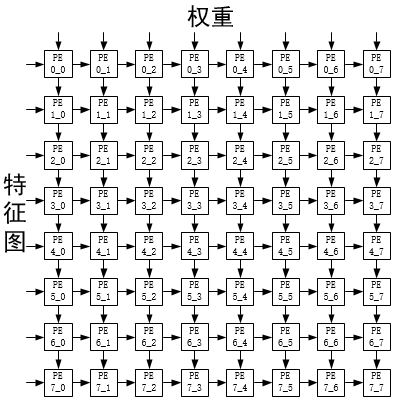


图3.1 脉动阵列结构

本测试用于kernel size为3，stride为2的运算，padding不在加速部分处理。该测试用例仅说明特征图数据导入方法，权重数据过程与特征图完全一致。

要测试的特征图可如图3.1所示。由于3\*3卷积步长为2，一次需要导入8个特征图。本项目导入过程分3次，每次导入8张3\*3特征图的第一行数据，例如图3.1中的绿色部分，一次脉动阵列计算开始前，需要导入3次。

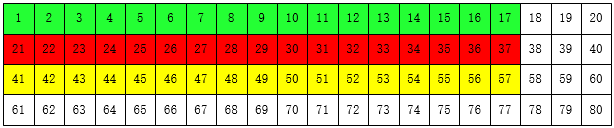


图3.1 特征图数据

导入数据需要3次，每次导入的方法可参考图3.2，图3.2说明了如何将图3.1中的第一行绿色数据通过写端口导入到NPU中。

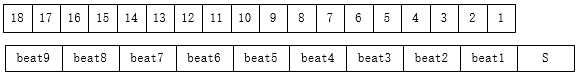


图3.2 数据导入过程