2024 春季学期-智能计算系统-作业三

(要求：独立完成，5 月 22 日前在 BB 平台提交)

1. (4 points)（课本 P225 页习题 7.2）本章所介绍的单核深度学习处理器的片内存储和传统 CPU 的片上缓存有什么区别？

> **参考答案:**
>
> 传统 CPU 只包含一块缓存供所有数据使用。而 DLP-S 中，由于输入数据和权重等的访存行为不同，因此分别设置了输入输出缓存 NRAM 和权重缓存 WRAM。

2. (4 points)（课本 P225 页习题 7.3）本章所介绍的单核深度学习处理器的访存行为与传统 CPU 有什么区别？

> **参考答案:**
>
> CPU 为通用处理器，很多计算任务中缺少数据复用等特性，虽然某些循环类型的计算存在访存局部性特征，但是为了通用性考虑，在设计时未考虑数据复用的特性。
>
> 在 DLP-S 等深度学习处理器中，充分考虑了数据和参数在计算过程中的复用特性，一次访存获得数据之后进行多次相关计算，使得访存次数减少。

3. (4 points)（课本 P225 页习题 7.4）本章所介绍的单核深度学习处理器的指令译码过程和传统 CPU 体系结构有什么区别？

> **参考答案:**
>
> 传统 CPU 中，指令大多按顺序执行。即使采用流水线架构，多条指令也存在先后的启动顺序。
>
> 在 DLP-S 中，为了提升整体性能，引入了多条指令发射队列。译码后根据指令的类型，发送到不同的队列。在执行过程中，只要不同队列的指令不存在依赖关系，不同队列中的多条指令可以并行地执行，实现指令级并行。

4. (20 points)（课本 P226 页习题 7.7、7.8）假设有一个单核的神经网络处理器，包含用于存放权重的片上存储 WRAM 共 256KB，用于存放输入/输出神经元数据的片上存储 NRAM 共 128KB，一个矩阵运算单元每个时钟周期内可完成 256 个 32 位浮点乘累加运算，该芯片运行频率为 1GHz，片外访存总带宽为 64GB/s。假设运算器利用率为 100% 且不考虑延迟，访存带宽利用率为 100% 且不考虑延迟。可以使用以下几种简化的指令：

   move ram_type1 ram_type2 size，用于从 ram_type1 向 ram_type2 传输 size 个字节的数据。其中，ram_type 可选 DRAM、NRAM 和 WRAM。

compute compute_type num，用于执行总运算量为 num 的 compute_type 类型的运算。其中，compute_type 可选 MAC_32、MAC_16、ADD_32、ADD_16、SUB_32、SUB_16、MUL_32、MUL_16、DIV_32、DIV_16 等。

loop loop_time...endloop，用于表示执行循环体 loop_time 次。

sync，同步指令，表示在此之前的指令必须都执行完成才能继续执行后续的指令。

请用上述指令完成以下任务，并估算执行时间：

（1）一个全连接层，其输入神经元的个数为 1×256，权重矩阵的大小为 256×1，所有数据均为 32 位宽的浮点数。

（2）一个全连接层，其输入神经元的个数为 32×256，权重矩阵的大小为 256×128，所有数据均为 32 位宽的浮点数。

参考答案：

（1）根据处理器参数，首先判断片上存储空间是否能够容纳所有数据：

该全连接层输入神经元数量为 1×256，每个数据为 32 位单精度浮点，因此，所需的存储空间大小为 $1 \times 256 \times 4Byte = 1KB$。

输出神经元只有一个，因此所需的存储空间大小为 $4B$。

因此片上 NRAM 足够容纳所有的输入输出神经元。

该全连接层权重数量为 256×1，每个数据为 32 位单精度浮点，因此，所需的存储空间大小为 $256 \times 1 \times 4Byte = 1KB$。因此，片上 WRAM 也能够容纳所有的权重数据。

其次，分析计算所需的时钟周期数：

该全连接层总的运算量为 256 次 MAC 运算。而 MFU 一个时钟周期能够计算 256 个 MAC 运算，因此，计算过程只需要一个周期即可。

完成该层计算的指令如下：

```
move DRAM WRAM 1KB  #载入权重
move DRAM NRAM 1KB  #载入数据
compute MAC_32 256  #执行MAC计算
sync  #等待计算完成
move NRAM DRAM 4B  #结果写回DRAM
```

从 DRAM 载入 1KB 数据所需的时间为 $1KB/(64GB/s) = 15.625ns$。

向 DRAM 写入 4B 数据的时间为 $4B/(64GB/s) = 0.0625ns$。

MFU 执行 256 个 MAC 计算的时间为一个时钟周期，即 $1ns$。

所以，该层计算的总时间为 $15.625 \times 2 + 1 + 0.0625 = 32.3125ns$。

（2）采用类似的分析思路。根据处理器参数，首先判断片上存储空间是否能够容纳所有数据：

该全连接层输入神经元数量为 32×256，每个数据为 32 位单精度浮点，因此，所需的存储空间大小为 $32 \times 256 \times 4Byte = 32KB$。根据权重矩阵大小判断，32 为批量的大小。

输出神经元数量为 128 个，批量为 32，每个数据为 32 位单精度浮点，因此，所需的存储空间大小为 $128 \times 32 \times 4Byte = 16KB$。

因此片上 NRAM 足够容纳所有的输入输出神经元。

该全连接层权重数量为 256×128，每个数据为 32 位单精度浮点，因此，所需的存储空间大小为 $256 \times 128 \times 4Byte = 128KB$。因此，片上 WRAM 也能够容纳所有的权重数据。

其次，分析计算所需的时钟周期数：

该全连接层每计算一个输出神经元所需的计算量为 256 次 MAC 运算，128 个输出神经元需要 256×128 次 MAC 运算，总共 32 个批量需要 $256 \times 128 \times 32$ 次 MAC 运算。MFU 一个时钟周期能够计算 256 个 MAC 运算，因此，计算过程需要 128×32 个周期。

完成该层计算的指令如下：

```
move DRAM WRAM 128KB  #载入权重
move DRAM NRAM 32KB  #载入数据
loop 4096  #循环次数
compute MAC_32 256  #执行MAC计算
end loop  #结束循环
sync  #等待计算完成
move NRAM DRAM 16KB  #结果写回DRAM
```

从 DRAM 载入 128KB 数据所需的时间为 $128KB/(64GB/s) = 2ms$。

从 DRAM 载入 32KB 数据所需的时间为 $32KB/(64GB/s) = 0.5ms$。

向 DRAM 写入 16KB 数据的时间为 $16KB/(64GB/s) = 0.25ms$。

MFU 执行 256×4096 个 MAC 计算的时间为 4096 个时钟周期，即 $4.096ms$。

所以，该层计算的总时间为 $2 + 0.5 + 4.096 + 0.25 = 6.846ms$。