智能计算系统第四次作业

朱浩泽 1911530 计算机科学与技术

6.3 假设设计一个深度学习处理器，它通过PCIe和DDR3的内存相连接。假设带宽为12.8Gb/s，那么和只有一个ALU的深度学习处理器相比，最理想情况下全连接层的加速比能有多少？

该深度学习处理器每个全连接层有12.8Gb的权重，12.8Gb的权重进行全连接层运算的时候一次性从DDR3中读入，然后通过并行方式运算

ALU只能一次读入一组数据进行运算，而一个ALU读入的数据量为一个字节是8b

所以加速比为 $\frac{12.8Gb}{8b} = 1.6G$

6.4 简述为什么指令级并行在深度学习处理器中一般情况下作用不大。什么情况下指令集并行也能在深度学习处理器中发挥作用？

深度学习主要是对tensor进行操作，这一点从tensorflow的名字中也可以看出。而tensor来说，其形状一般来说比较规整，而且在主要的卷积层和全连接层多为向量操作。而通过体系结构的课程我们可以知道，指令集并行性的主要优势在于灵活性高，但是流水线的控制逻辑复杂且功耗巨大，对于深度学习这种本身就功耗巨大且数据较为规整的任务不是很合适，所以只有在深度学习算法加速设计到复杂的控制逻辑时，指令集并行性才能够发挥作用，例如深度强化学习就非常需要指令集并行，用于加快速度，通常这些任务需要在某个环境中疯狂采集数据，于是就通过并行来加快采集速度；在pytorch中dataloader有一个参数叫做num_worker，就是用于开多少进程用于数据集加载，这样的好处就是可以用满机器的IO速度。这些都是指令集并行性在深度学习处理器中发挥的作用。

7.7 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，用执行运算总量为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。

请使用上述指令完成以下任务，并估计执行时间：一个全连接层，其输入的神经元个数为1X256、权重矩阵的大小为256X1，所有数据均为32位宽的浮点数。

Apache

```
compute MUL_32 256
    move WRAM DRAM 4
    move NRAM DRAM 4
    x[i,1] mul W[1,i]
sync
```

$$\text{时钟周期} = \frac{1}{1Ghz} = 1 \times 10^{-9}$$

$$\text{访存时间} = \frac{256B \times 4 \times 32 + 1B}{64GB} = 5.12 \times 10^{-7} s$$

$$\text{矩阵运算} = \frac{1}{1Ghz} = 1 \times 10^{-9}$$

$$\text{总时间} = \text{矩阵运算} + \text{访存时间} = 5.12 \times 10^{-7}$$

利用习题7.7所述的处理器和指令完成以下任务，并估算运行时间：一个全连接层，其输入神经元的个数为32x256，权重矩阵的大小为256x128，所有数据均为32位宽的浮点数。

Apache

```
loop 32
    compute MUL_32 256
    loop 128
        move WRAM DRAM 4
        move NRAM DRAM 4
        x[i,k] mul W[k,j]
    sync
    endloop
ednloop
```

$$\text{时钟周期} = \frac{1}{1Ghz}$$

$$\text{访存时间} = \frac{32 \times 32b \times 256}{64GB} = 5.12 \times 10^{-7}$$

$$\text{矩阵运算} = \frac{128 \times 32}{1Ghz} = 4.096 \times 10^{-6}$$

$$\text{执行时间} = \text{访存时间} + \text{执行时间} = 4.6 \times 10^{-6}$$