第三次作业

author: PB19061272_金桥

[2-47] 阐述 FPGA 计算相对于 CPU 计算的优势。

- 低延迟：FPGA 同时拥有流水线并行和数据并行，而 GPU 几乎只有数据并行，所以 FPGA 计算比 CPU 计算延迟低。
- 灵活：FPGA 各逻辑单元的连接方式可以由编程更改，具有更高的灵活性。

[2-49] 解释“Dark silicon”现象。

- 由于功耗的限制，一个很高端的处理器，比如多核的，其实同一时刻只能有很少的一部分门电路能够工作，其余的大部分处于不工作的状态，这部分不工作的门电路，就叫做“Dark silicon”。

[2-50] 名词解释：Amdahl’s Rule of Thumb。

- 系统中对某一部件采用更快执行方式所能获得的系统性能改进 程度，取决于这种执行方式被使用的频率，或所占总执行时间的比例。

[2-51] 名词解释：Memory Wall。

- 内存性能严重限制了处理器的性能发挥。

[2-53] 试举例说明“More Moore”、""“More than Moore”、“Beyond CMOS”。

- More Moore：继续改进芯片制作工艺，以提高单位面积晶体管数目为目标，例如性能参数从 22nm 逐步进步到 7nm。
- More than Moore：侧重于功能的多样化，包括：
  - 不再单纯靠增加晶体 管数目提升性能，而是更多地靠电路设计以及系统算法优化；
  - 不再单一追求把更多功能模块放到同一块芯片，可以靠封装技术来实现集成；
  - 芯片不仅仅是努力实现更高的性能，也可以是增加一些有用的新功能。
- Beyond CMOS：使用 CMOS 以外的新器件提升集成电路性能，例如未来可期的量子计算机。

[2-55] 高通公司 Snapdragon 888+ 芯片中 Hexagon™ 780 Processor 适用于什么类别的计算?

- 人工智能算法加速
- 数字图像处理

[2-57] 试分析以 TPU 和 Cambricon 为代表的人工智能算法加速计算芯片在设计思路上有哪些共性。

- 由于人工智能计算经常出现数据的重复利用，所以二者都设计了容量较大的片上内存。
- 都为加矩阵计算与卷积计算而设计了针对性的硬件逻辑。

[2-58] 试述 Roofline 模型中的计算强度 (Operational Intensity)。

- Roofline 模型中的计算强度为计算量和访存量的比值，即 $\mathrm{AI} = \frac{\mathrm{FLOPS}}{\mathrm{Bytes}}$
- 计算强度表示此模型在计算过程中，每 `Byte` 内存交换到底用于进行多少次浮点运算。

[2-59] 假设矩阵 A、B 维度是 1920 × 1080，估算完成矩阵加法 C=A+B 的计算量，并估算该运算的计算强度 AI (Arithemtic Intensity)。

- C 中每个元素需要进行一次加法运算，所以计算量为

$$1920 \times 1080 = 2073600\ \mathrm{FLOPS}$$

- 计算强度 AI 为

$$\mathrm{AI} = \frac{\mathrm{FLOPS}}{\mathrm{Bytes}} = \frac{1920 \times 1080}{3 \times 1920 \times 1080} = \frac{1}{3}$$

[2-62] 试述对缓存一致性问题的理解。

- 在多核且每个核有独立缓存的情况下，如果某个处理器更改了内存中的值，则需要告知所有的 Cache，即保持 Cache 副本的一致性。
- 为了向所有其他缓存副本拥有者广播，需要缓存一致性协议与专用的控制器。可以通过缓存一致性协议 (Coherency protocols) 有多种，多数属于“窥探 (snooping）”协议。
- “窥探”的基本思想是，所有数据传输都发生在一条共享的总线上，而所有的处理器都能看到这条总线；缓存本身是独立的，但是内存是共享资源，所有的内存访问都要经过仲裁 (arbitrate）；同一个指令周期中，只有一个缓存可以读写内存。

[2-66] 举例说明 HBM 潜在的应用场景。

- HBM 优势在于高速、高带宽、高位宽，缺陷在于高延迟、容量小、扩展性差。所以 HBM 非常适合对带宽要求高、对延迟要求低的领域，例如作为 GPU 的显存。
- 在计算强度不变的情况下，HBM 的高带宽提供了更高的算力，所以也能用于对算力需求大的人工智能计算加速。

[2-67] SPM (ScratchPad Memory) 和 Cache 都是片内集成 SRAM 存储单元，为何不能用 SPM 代替 Cache?

- Cache 的硬件结构更复杂，访存由硬件控制，能够完成比较复杂的访存的行为。
- 而 SPM 的存储位置要由程序员指定，无法完成较复杂的访存操作。

[2-68] 脉动阵列 (Systolic Array) 适用于哪些计算场景?

- 脉动阵列适用于处理重复计算，这种计算通常都需要庞大的计算能力，但一 般都是高度规则和可并行化的，而脉动阵列充分利用了这种规律性和并行性。
- 例如数字图像处理、神经网络卷积加速。

[2-70] 举例说明“算力”和“算法”的匹配。

- Roofline 模型中的计算强度 (Operational Intensity) 取决于算法，由计算强度与带宽共同决定了需要的算力，即带宽不变的情况下，所需的算力依赖于选择的算法。