# 第 2 章 作业

PB19061221 王城冰

**PART-III**

[2-47] 阐述 FPGA 计算相对于 CPU 计算的优势。
答：既能管理又能运算；处理速度快，流水线并行和数据并行（延迟低，流处理），可编程逻辑器件，功能可以随时改变（这一特性使得其在深度学习中占据不可或缺的位置）；功耗相比之下较低。
[2-49] 解释“Dark silicon”现象。
答：在电子行业，暗硅是指在给定的热设计功率（TDP）约束下，集成电路在额定工作电压下不能通电的电路数量。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“暗硅”。

[2-50] 名词解释：Amdahl’s Rule of Thumb。
答：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。
答：内存墙是指在处理大量数据时，处理器经常需要等待来自相当慢的 RAM 内存的数据的现象。

[2-53] 试举例说明“More Moore”、”“More than Moore”、“Beyond CMOS”。
答：More Moore：Fin Field-Effect Transistor(FinFET), Fully Depleted Silicon On Insulator(FD-SOI)
More than Moore: Analog/Mixed-signal, RF, MEMS, Sensors
Beyond CMOS: Tunneling FET, Impact Ionization MOS, Rapid single flux quantum, SpinFET

[2-55] 高通公司 Snapdragon 888+芯片中 Hexagon™ 780 Processor 适用于什么类别的计算？
答：Artificial Intelligence (AI) Engine

[2-57] 试分析以 TPU 和 Cambricon 为代表的人工智能算法加速算芯片在设计思路上有哪些共性。
答：都有大容量的存储电路设计，且针对在矩阵乘法和卷积运算中，最大化数据复用，并减小内存访问次数，在降低内存带宽压力的同时也减小了内存访问的能量消耗。都有为向量和矩阵的操作加速而专门设计的模块。

[2-58] 试述 Roofline 模型中的计算强度 (Operational Intensity)。
答：指每字节 DRAM 流量的操作，将访问的总字节定义为那些经过缓存层次过滤后进入主内存的字节。测量缓存和内存之间的流量，而不是处理器和缓存之间的流量。因此，操作强度预测了特定计算机上的内核所需的 DRAM 带宽

[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)。
答：1920*1080/(1920*1080*4Bytes)=0.25ops/Byte

[2-62] 试述对缓存一致性问题的理解。
答：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。缓存控制器不停地在窥探总线上发生的数据交换，跟踪其他缓存在做什么。当一个缓存代表它所属的处理器去读写后端存储时，其他处理器都会得到通知，它们以此来使自己的缓存保持同步。只要某个处理器一写后端存储，其他处理器马上就知道这块后端存储数据在它们自己的缓存中对应的 Cache line 已经失效。

[2-66] 举例说明 HBM 潜在的应用场景。
答：在学术领域，尤其在 DSA（Domain Specific Acceleration）场景，如果能将 embedding 数据合理分配到 32 路 HBM 当中，同时做到系统级 pipeline 数据访存和计算。在 AI 计算、区块链和数字货币挖矿等对大数据处理访存需求极高等领域都有广泛运用。

[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？
答：SPM 和 Cache 都是由 SRAM 组成，都能提高系统的平均性能，但它们在结构、访问延时和操作能耗特性等方面存在明显的差别。Cache 额外还有 TagRAM 部件和地址比较逻辑电路器件，比较逻辑电路器件用来判定访问 Cache 的操作是否命中。

[2-68] 脉动阵列(Systolic Array)适用于哪些计算场景？
答：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。脉动阵列的优势在于让数据在基本处理单元的阵列中进行传递和计算，降低了数据访存次数，采用了模块化和流水线的设计，使其结构更加整齐，布线更加一致，从而极大的提高了频率。

[2-70] 举例说明“算力”和“算法”的匹配。
答：在深度学习中的 transformer 很火，但 transformer 模型的参数量极为庞大，如果要让模型能够训练，必须要合理优化算法。transformer 之所以这么有效，关键还在于它里面设计了很多可以并行计算的结构，这些结构，可以实现数据并行和模型并行，可以很好的适配 GPU，可以说是很充分的利用了算力。