北京大学
PEKING UNIVERSITY

人工智能的硬件基石：从物理器件到计算架构

2026年春季 - 作业3

一、 单项选择题（每题 4 分，共 12 分）

1. 神经网络的“量化（Quantization）”技术将浮点模型转换为定点模型，这主要改善了 AI 芯片设计中的哪项指标？

   A. 算术逻辑单元（ALU）的峰值算力

   B. 存储容量需求与片上访存带宽（Memory Bandwidth）

   C. 芯片内部的布线拥塞

   D. 外部 DDR 的传输频率

2. 典型的卷积神经网络（CNN）加速器中，哪种数据复用（Data Reuse）策略主要用于减少权重的片外数据搬运？

   A. 卷积核复用（Convolution Reuse）

   B. 输入特征图复用（Input Feature Map Reuse）

   C. 输出通道复用（Output Channel Reuse）

   D. 空间复用（Spatial Reuse）

3. 随着存算一体（Processing-in-Memory, PIM）架构的发展，用于实现神经网络乘加运算（MAC）的核心物理器件通常是（ ）

   A. 静态随机存取存储器（SRAM）或阻变存储器（RRAM）

   B. 现场可编程门阵列（FPGA）

   C. 锁相环（PLL）

   D. 逻辑与门（AND Gate）

4. 对于云端推理芯片，通常更关注的性能指标是（ ）

   A. 峰值浮点算力（Peak TFLOPS）

   B. 能量效率（TOPS/W）与特定批次大小（Batch Size）下的延迟（Latency）

   C. 支持的最大模型参数量（Parameter Capacity）

   D. 芯片面积（Die Size）与静态功耗

二、 多项选择题（每题 5 分，共 25 分）

1. 与传统 CPU 相比，AI 加速芯片（如 TPU、NPU）在架构设计上有何显著特

点？

A. 包含大规模并行计算阵列（如脉动阵列）

B. 针对特定操作（如矩阵乘法、激活函数）进行了硬化处理

C. 对标量运算（Scalar Operations）和控制流代码进行了深度优化

D. 高度重视片上缓存（SRAM）的层次结构与数据局部性

2. 脉动阵列（Systolic Array）是目前 AI 芯片中最常用的加速器架构，其设计优势包括哪些？

   A. 极高的数据吞吐量

   B. 高度规则的布局布线（Regular Layout）

   C. 能够灵活处理任意复杂的控制流分支

   D. 减少了全局数据总线的读写频率

3. 在 AI 编译器的设计中，针对 AI 加速芯片的优化挑战主要涵盖哪几个维度？

   A. 算子融合（Operator Fusion）以减少访存

   B. 循环展开（Loop Unrolling）以增加并行度

   C. 存储层次结构映射（Memory Hierarchy Mapping）

   D. 量子力学算法推演（Quantum Mechanics Deduction）

4. 针对 Transformer 架构中大语言模型（LLM）的生成阶段，AI 芯片设计在面临“内存墙”挑战时，可以采用的硬件优化技术包括哪些？

   A. 增大片上 SRAM 容量以缓存 Key-Value Cache

   B. 采用存算一体（PIM）技术

   C. 引入稀疏化（Sparse）加速引擎

   D. 支持多核间的超高速片间互连（Interconnect）

5. 人工智能芯片按照应用场景可分为端侧与云端，以下关于两者的设计侧重点描述正确的是哪些？

   A. 云端 AI 芯片侧重于超高并发训练和推理吞吐量

   B. 端侧 AI 芯片对功耗和发热（Thermal Dissipation）有极严格的限制

   C. 端侧 AI 芯片通常集成了专用的视觉或语音处理 DSP

   D. 云端 AI 芯片必须支持片间多卡互连通信（如 NVLink、PCIe Gen5）

三、 简答题（每题 13 分，共 13 分）

请简要调研 Neural ODE Transformer 模型，并阐述实现该模型加速芯片架构与常规 Transformer 加速芯片架构的不同，以及可能需要重点考虑的设计难点。