

第 1 题

比较 ARM 体系中 A32、A16、T16、T32 指令集的特点，并说明为什么 T16 代码并不总比 A32 代码优化，请尝试举出代码实例和应用场景进行佐证说明。

答：

A32 指令集，有固定的 32 位指令长度，并在 4 字节边界上对齐，采用 32 位长度，能够携带更多的操作信息和寻址模式，适合高性能计算任务。A32 指令集实际上是在 ARMv6 和 ARMv7 架构中我们常说的 ARM 指令集，ARMv8 及之后改名 A32（为了与 A64 进行区分）。

T16 是 Thumb 指令集的早期版本，仅支持 16 位指令。它以牺牲一些性能为代价提供了更小的程序体积。T16 指令集适合内存受限的环境，但在处理复杂操作时可能需要更多的指令来实现相同的功能，这可能导致执行效率降低。

T32，即 Thumb-2，是对 T16 的扩展，支持 16 位和 32 位指令的混合，提供了与 ARM 相似的性能，同时保留了缩小的代码体积。T32 指令集在 AArch32 状态下执行，由于其尺寸和性能优势，编译或组合所有 32 位代码以利用 Thumb-2 技术越来越普遍。

用了几种搜索方式，没有找到 A16 指令集的资料，这里是不是想给出的是 A64 与 A32？

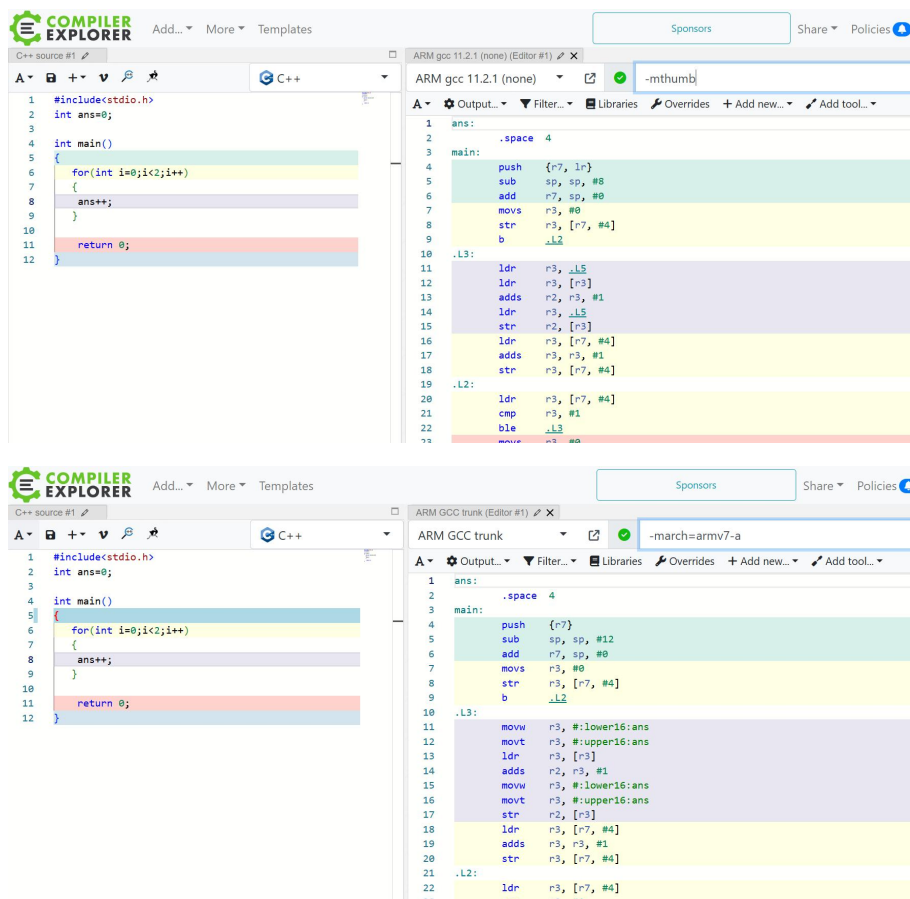
A64 指令集，是 ARMv8 架构中的 64 位指令集，提供对 64 位宽整数寄存器和数据操作的访问，以及使用 64 位大小的内存指针的能力。A64 指令集在 AArch64 状态下执行，与旧的 A32 指令集相比，它增加了一些附加功能，同时删除了有可能限制性能速度或能耗的其他功能

T16 代码并不总比 A32 代码优化，原因在于 T16 的 16 位指令虽然在某些情况下可以减少代码体积，提高内存利用率，但在处理复杂操作时可能需要更多的指令来实现相同的功能，这可能导致执行效率降低。

例如，在进行复杂的数学运算或数据处理时，A32 指令集的 32 位指令可以提供更多的操作数和更复杂的寻址模式，从而实现更高效的代码执行。此外，现代处理器通常对 32 位指令有更优化的处理，因此在性能关键的应用中，A32 指令集可能会提供更好的性能。

在实际应用中，如果一个程序主要执行简单的控制流和数据传输，T16 指令集可能会因为代码密度高而表现出优势。然而，对于需要大量计算和复杂数据处理的应用，如图形处理或科学计算，A32 指令集可能会因为其能够执行更复杂的操作而提供更好的性能。

下图是同一段代码使用不同指令集的区别的比较，T16 代码并不总比 A32 代码优化。



第2题

简述“超大核+大核+小核”多核处理器架构的优势和特点, 并尝试结合实例解读该架构可以同时大幅提升处理器的性能与能耗比。

答:

“超大核+大核+小核”多核处理器架构是一种异构多核架构。在传统的单核处理器中, 所有的计算任务都由一个核心执行, 这可能导致性能瓶颈, 特别是在处理大量数据或多任务处理时。多核处理器通过在同一芯片上集成多个核心, 使得多个任务可以同时进行, 从而提高整体性能和效率。异构多核架构中, 处理核心可以是不同的, 它们可能具有不同的架构、时钟频率和功耗特征。这种设计的目标是通过将不同类型的核心结合在一起, 使得处理器可以更好地适应不同种类的任务。例如, 一个异构多核处理器可能包含高性能核心和低功耗核心, 以在需要时提供更好的性能, 而在轻负载时降低功耗。异构多核则更具灵活性, 可以更好地平衡性能和功耗。

超大核 (Super Core): 高性能核心, 用于处理最复杂的任务和提供最高的单线程性能。它们通常用于需要大量计算能力的应用, 如高端游戏、科学计算和图形处理。

大核 (Large Core): 提供中等水平的性能, 适用于多线程应用和日常计算任务。它们在性能和功耗之间提供了良好的平衡。

小核 (Small Core): 低功耗核心, 用于处理轻量级任务和和设备空闲时减少能耗。它们在需要最小能耗时运行, 如在待机模式或执行后台任务时。

优势和性能：

性能优化：超大核提供了高性能，确保了处理器在需要大量计算能力的应用是表现良好。

能效提升：小核在低负载时使用，大幅降低了能耗，延长电池寿命。

灵活性：可以根据不同的工作负载动态调整核心的使用，实现性能和功耗的最佳平衡。

响应速度：大核提供了快速的响应速度，适合日常使用和多任务处理。

以高通骁龙 8 Gen 3 处理器为例，它提供了：

1 个基于 Arm Cortex-X4 技术的主处理器核心，主频最高可达 3.3 GHz

5 个最高 3.2GHz 的性能核心

2 个最高 2.3GHz 的效率核心

据高通称，新款 8 Gen 3 的性能比前代产品提高了 30%，能效提高了 20%。它还提供了 25% 的 GPU 性能提升和 20% 的能效提升。在高性能需求的游戏或应用中，超大核会被激活以提供必要的计算能力。在日常使用中，大核会处理大多数任务，而小核则在后台运行，如同步数据或执行低功耗的传感器处理任务。当用户不活跃时，大部分核心可以关闭或进入低功耗状态，只有小核保持运行以监听通知和执行基本任务，从而显著降低整体能耗。

这种架构使得处理器在提供高性能的同时，也优化了能耗，特别是在移动设备和需要长时间运行的嵌入式系统中，这种设计尤为重要。通过智能地调度任务到最合适的核心上，异构多核架构能够显著提升处理器的性能与能耗比。

THE TITAN of on-device intelligence

10 billion+ parameters
20+ tokens/sec
Meta Llama 2/BaiChuan
On-device Personalization
Qualcomm® Sensing Hub
First to support multi-modality gen AI models

Qualcomm® AI Stack
PyTorch ExecuTorch delegate and fully optimized models
Qualcomm® Hexagon® NPU
98% faster and 40% more efficient
Fastest in the world
Stable diffusion & ControlNet < 1 second

Qualcomm® Kryo® CPU
Up to 3.3GHz 15/2 Configuration
30% Faster **20% More Efficient**

4nm
Processing technology

Video Object Eraser
For video capture

Cognitive ISP
12 Layer real-time Semantic Segmentation

Generative AI Backgrounds
Video capture with stable diffusion

Dolby
HDR photo capture

Global Illumination with Ray Tracing
Next-Gen Light Reflection System

Unreal Engine 5 with Lumen

Snapdragon® Game Super Resolution
Single pass spatial aware upscaling for gaming up to 8K

Photo Expansion
Filled by AI

Night Vision for video capture
Enhanced with Frame Rate Conversion

New Computer Vision Engine
Support for 3D Time-of-Flight sensors for higher resolution depth mapping

4th Gen Computational HDR
Support for DCG Image Sensor

240 FPS Gaming
On 240Hz Displays

Adreno Frame Motion Engine 2.0

Qualcomm® Adreno® GPU
25% faster **25% more efficient**
240Hz to 1Hz QSYNC
Variable refresh rate for extreme power savings

5G Modem-RF
5G Advanced-ready
AI hardware acceleration
10 Gbps down / 3.5 Gbps up

Qualcomm® FastConnect® System
Fastest Wi-Fi 7 (5.8Gbps)
10x Multi-Link
Dual Bluetooth®

Truepic
C2PA compliant photo capture

Dual Always-Sensing ISPs

24-bit 96kHz lossless
Qualcomm® Expanded Personal Area Network Technology (xPAN)
Whole home coverage

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

第3题

我们已经在课堂/课程网站上学习了 SRAM 存储单元的组成与工作原理、工作过程，那么请尝试阐释：

- 1) DRAM 的存储单元如何构成？其工作过程是什么？
- 2) Nor Flash 和 NAND Flash 各自的架构特点和工作过程是什么？为什么通常认为, Nor Flash 适合存放代码, NAND Flash 适合存放大块数据？
- 3) 你使用过的嵌入式开发设备或者手机、平板，是否符合前一个问题中的 Flash 功能设定？为什么？

答：

(1) SRAM 和 DRAM 同属于随机访问存储器，可被随机读写。它的存储单元可以有单管、三管、四管等形式。

单管 DRAM 由相连接的电容和晶体管构成。电容代表数据位（0 或 1），晶体管连接到行选择线和位选择线，作为开关，控制电容的充放电。

充电阶段，电容被充电到一定的电压水平以代表数据位。由于电容会自然放电，因此需要定期刷新以保持数据。

读取数据时，通过激活行和列的交叉点来访问特定的存储单元，电容的电荷状态通过放大器读取并转换为数据位。

写数据时，两个 MOS 管被选中，数据从 IO 引脚输入，写入 1 时，MOS 管对电容充电，写入 0 时，I/O 为低电平，电容放电。

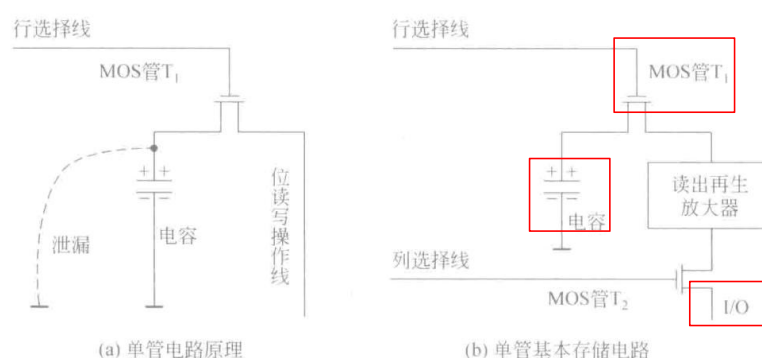


图 4.11 单 MOS 管与电容构成的存储元结构

(2)

Nor Flash:

架构特点：Nor Flash 的存储单元以阵列形式排列，每个存储单元都直接连接到一个唯一的地址和数据线，就像是并联一样。它的地址线和数据线分开，可随机任意读取，时序相对简单，读取、传输速度快，擦除、写入较慢。

工作过程：NOR Flash 的每个存储单元都连接到输出端，构成 NOR 电路结构，通过 MOS 电路进行数据的存储和读取，可以长时间保存数据，不需要维持持续

的电流输入。存储单元由一对 P 型和 N 型 MOS 电晶体组成，形成双稳态电路。通过输入脉冲的高低电平来控制电荷的存储和释放。

Nor Flash 的读取、传输速度快，擦除、写入较慢，代码通常需要频繁地读取，而写入操作相对较少，故 Nor Flash 常用于代码存储。

NAND Flash:

架构特点：NAND Flash 的存储单元以串行方式连接，使用复杂 IO 接口穿行读取数据，共用一套总线（数据总线和地址总线），高密度，多用于 data 大量存储

工作过程：NAND Flash 使用电荷量来存储数据。每个存储单元中的栅极上存储了一定数量的电子，表示为 1 或 0。当需要读取或写入数据时，通过对栅极施加适当的电压来控制电荷量。具体来说，Flash 的基本组成单元是浮栅晶体管，其状态可以用来指示二进制的 0 或 1。写操作就是往晶体管中注入电子，使之充电；擦除操作则是把晶体管中的电子排出，使之放电。

NAND Flash 通常以块为单位进行擦除，然后以页面为单位进行写入。由于其结构，NAND Flash 适合存放大块数据，如文件系统或媒体文件，这些数据通常以较大的块进行读写。

（3）符合。

嵌入式开发设备、手机和平板通常使用 NAND Flash 作为主要的非易失性存储，可能还会包含一些 Nor Flash，用于存放启动代码或固件。

例如，手机可能使用 NAND Flash 来存储操作系统、应用程序和用户数据，而 Nor Flash 可能用于存储引导加载程序，这些程序在设备启动时需要快速加载。

NOR Flash 的读取和我们常见的 SDRAM 的读取是一样，用户可以直接运行装载在 NOR FLASH 里面的代码，这样可以减少 SRAM 的容量从而节约了成本。

NAND Flash 没有采取 **内存 Q** 的随机读取技术，它的读取是以一次读取一块的形式来进行的，通常是一次读取 512 个字节，采用这种技术的 Flash 比较廉价。用户不能直接运行 NAND Flash 上的代码，因此好多使用 NAND Flash 的开发板除了使用 NAND Flash 以外，还作上了一块小的 NOR Flash 来运行启动代码。

一般小容量的用 NOR Flash，因为其读取速度快，多用来存储操作系统等重要信息，而大容量的用 NAND FLASH，最常见的 NAND FLASH 应用是 **嵌入式系统** 采用的 DOC (Disk On Chip) 和我们通常用的“闪存”，可以在线擦除。目前市面上的 FLASH 主要来自 Intel, AMD, Fujitsu 和 Toshiba，而生产 NAND Flash 的主要厂家有 Samsung 和 Toshiba。