
大模型-AI集群 (存)
数据 存储
硬件 & 连接
ZOMI

大模型业务全流程

数据 & 模型算法

2. 数据处理

- 开源数据
- 数据预处理
- 向量数据库

3. 模型算法

- LLM模型架构
- 多模态一切皆Tokens

模型训练 & 微调

4. 模型训练

- 混合精度
- 梯度检查
- 梯度累积
- …

5. 分布式并行

训练集群稳定性

6. 模型微调

- 全参微调
- 低参微调
- 指令微调

模型验证 & 推理部署

7. 模型验证

- 下游任务
- 测评标准
- bench mark

8. 推理与智能体

- 量化压缩
- 推理加速
- 9. Agent 智能体

集群算力准备

1. AI 集群建设

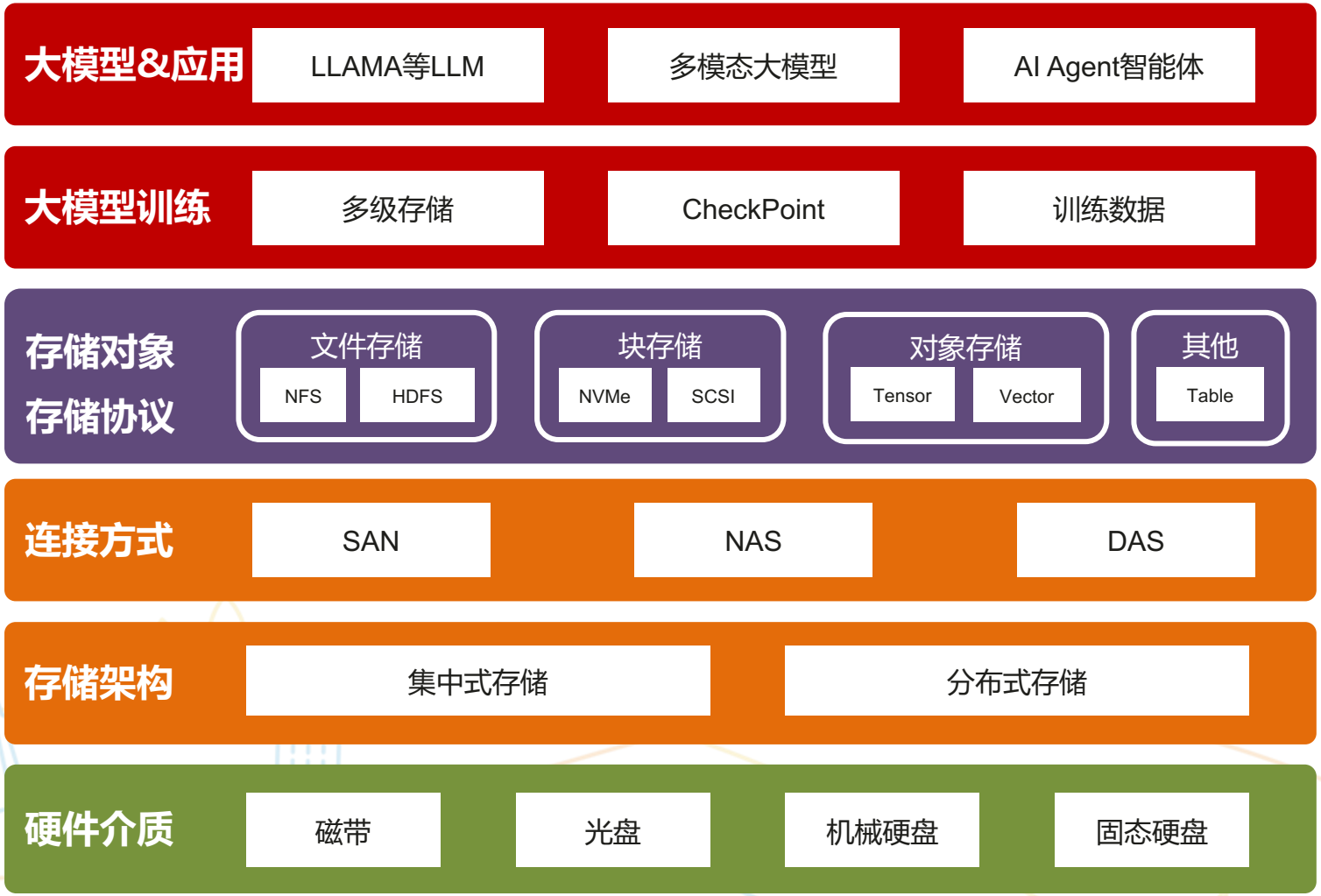
- 计算、存储、网络
- AI 集群机房建设
- AI 集群上线与运维

关于本内容

- **内容背景**
  - AI 集群 + 大模型
- **具体内容**
  - **数据存储现状和场景：**存储软件类型、存储硬件类型的发展
  - **大模型对存储的挑战：**存储性能指标、存储遇到大模型挑战与新机会点
  - **大模型训练CKPT优化：**大模型训练过程、CKPT过程分解、CKPT优化
  - **大模型时代对存储的思考：**什么样的存储架构才是AI大模型时代的选择？

增长的存储主要靠数据中心和存储服务器拉动

存储技术架构

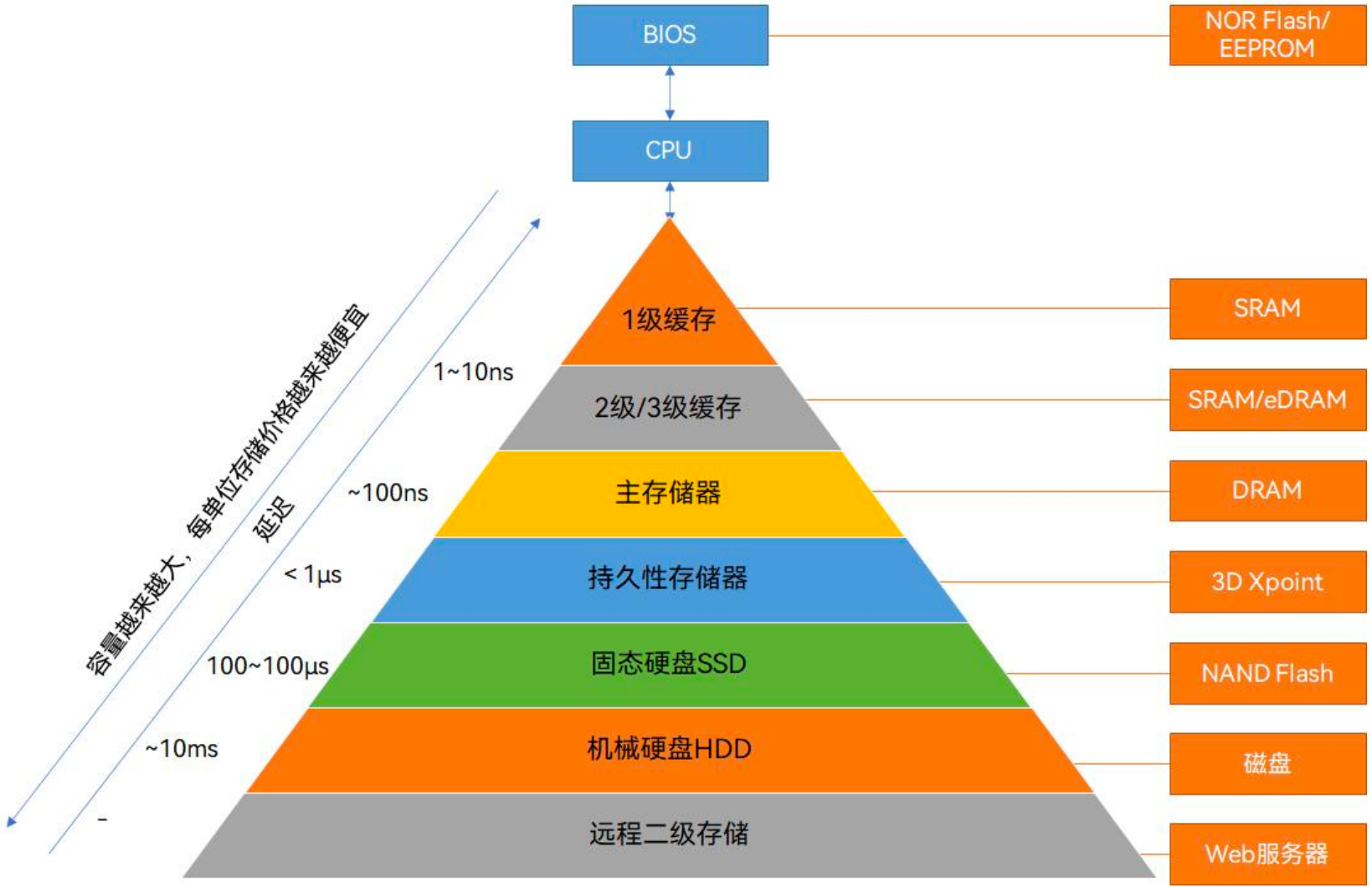


1. 数据存储的硬件介质

存储介质的组织层次

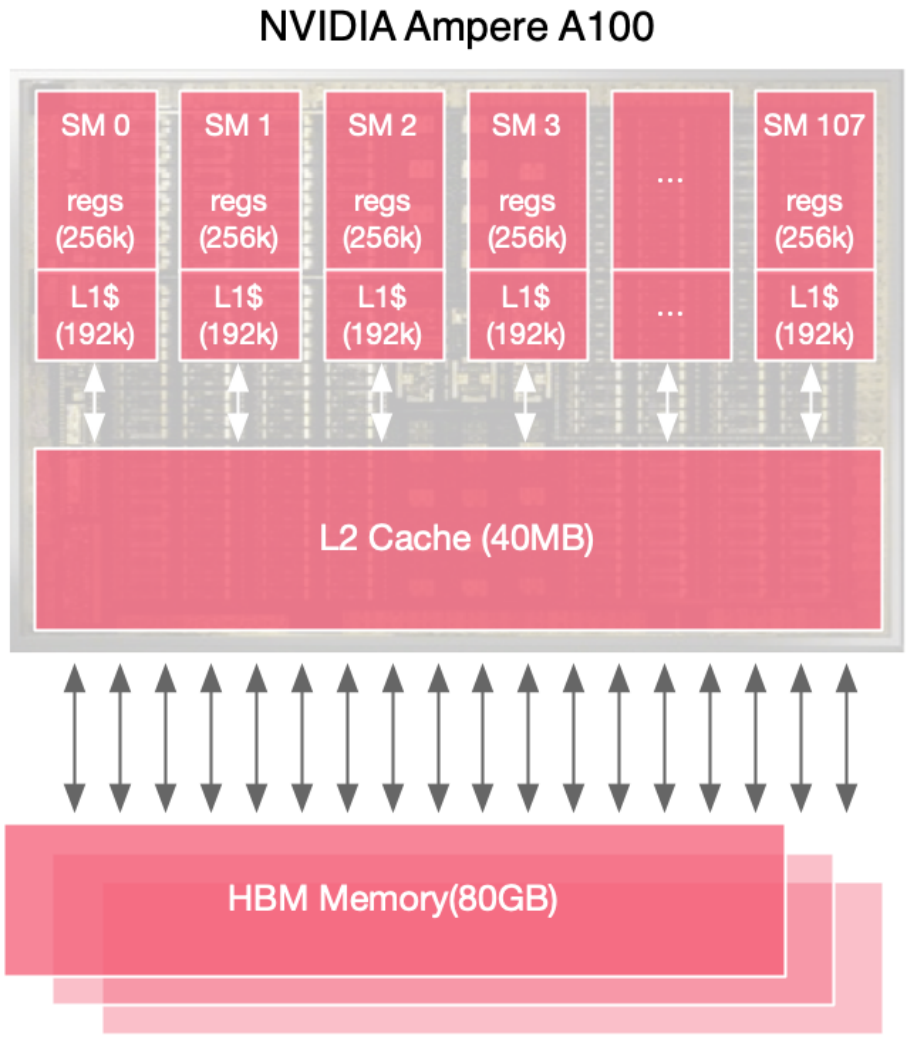
- 容量更小
- 速度更快
- 价格更高

- 容量更大
- 速度更慢
- 价格更低



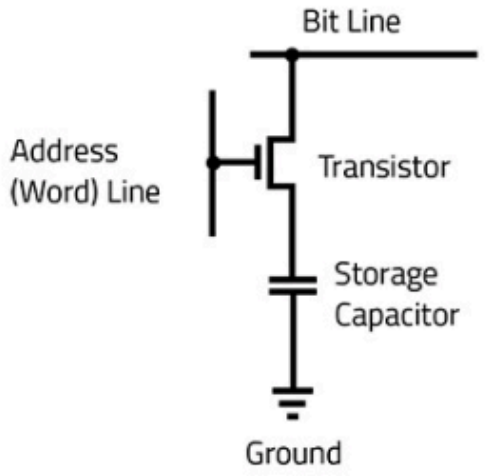
Cache 高速缓存存储器

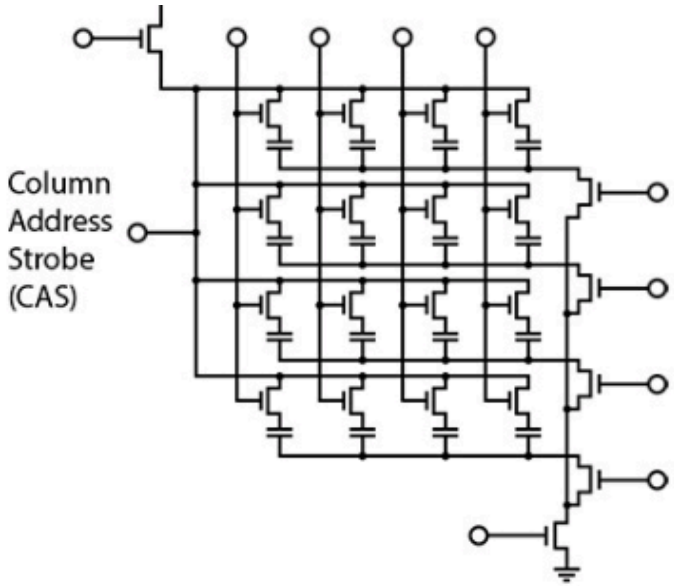
- Cache 介于 CPU 和主存间的高速小容量存储器，由 SRAM 组成（可以分为L1/L2/L3多层），容量小但比主存 DRAM 技术更快速。
- CPU 往往需要重复读取同样的数据块，Cache 的引入与缓存容量增大，可以大幅提升 CPU 内部读取数据的命中率，从而提高系统性能。



DRAM 动态随机存储器

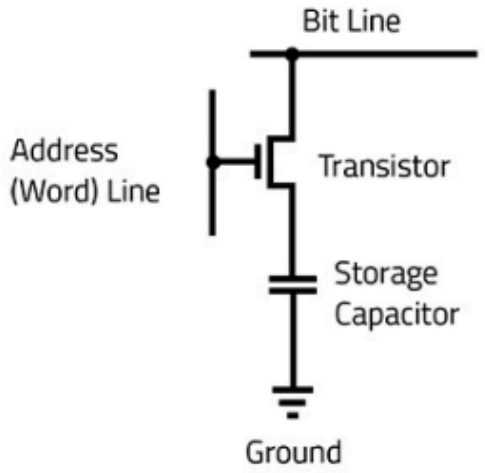
- **组成**：由许多重复的位元格（Bit Cell）组成，每一个基本单元由一个电容和一个晶体管构成：
  1. 电容中存储电荷量，用于表示“0”和“1”
  2. 晶体管用来控制电容的充放电
- **方式**：由于电容会存在漏电现象，为了避免丢失数据。必须在数据改变或断电前，周期性动态充电，保持电势。因此，DRAM被称为“动态”随机存储器。



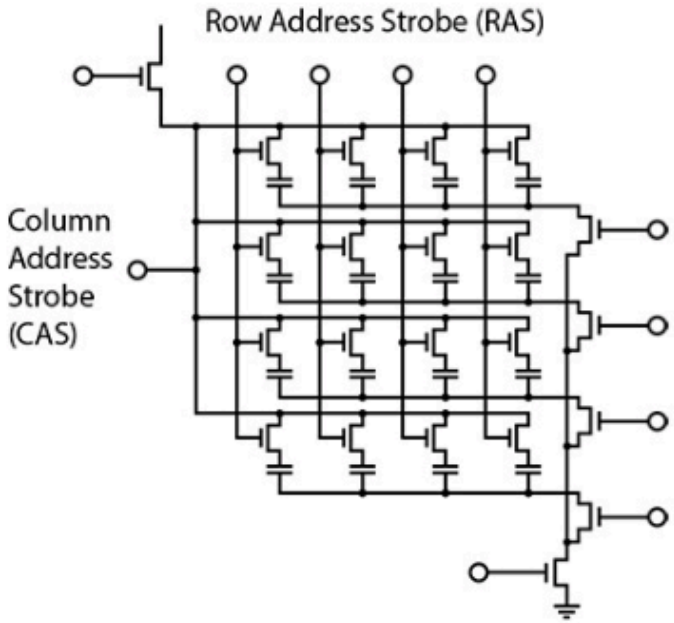


DRAM 动态随机存储器

- **应用场景：**DRAM 作为 PC、移动手机的内存主流方案。录入 PC 内存条（DDR）、早期显卡显存（GDDR）、手机运行内存（LPDDR），都是DRAM 一种具体协议实现。
  - DDR（DDR SDRAM ），双倍速率同步动态随机存储器。



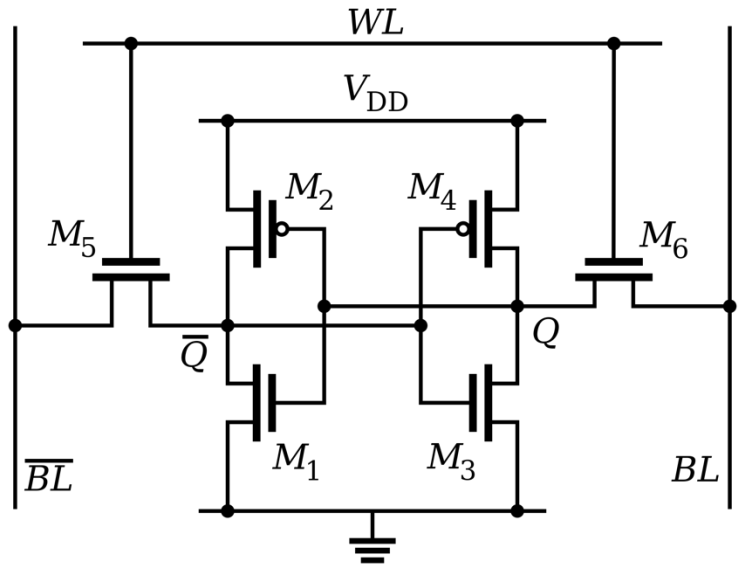
Single Memory Cell



Memory Cell Array

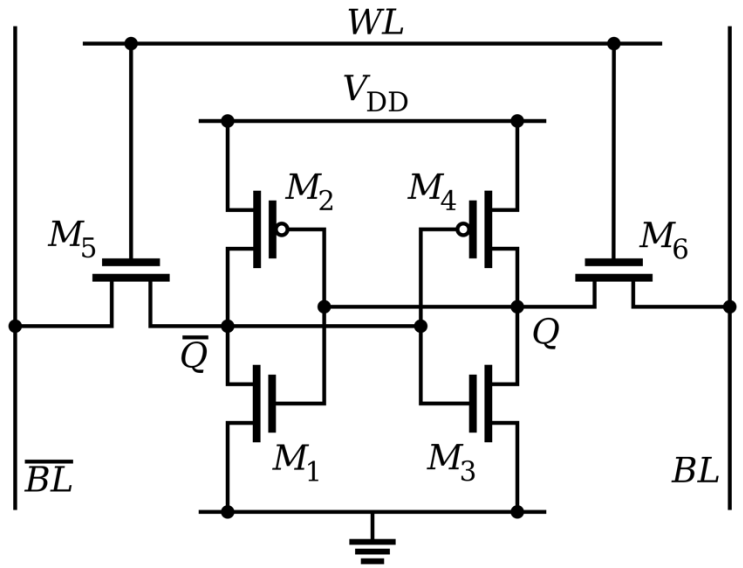
SRAM 静态随机存储器

- SRAM 基本单元最少由 6 个晶体管组成
  1. 4个场效应管（M1, M2, M3, M4）构成两个交叉耦合的反相器
  2. 2个场效应管（M5, M6）用于读写的位线（Bit Line）控制开关，
  3. 场效应管构成一个锁存器（触发器），通电时锁住二进制数 0 和 1，因此被称为静态随机存储器



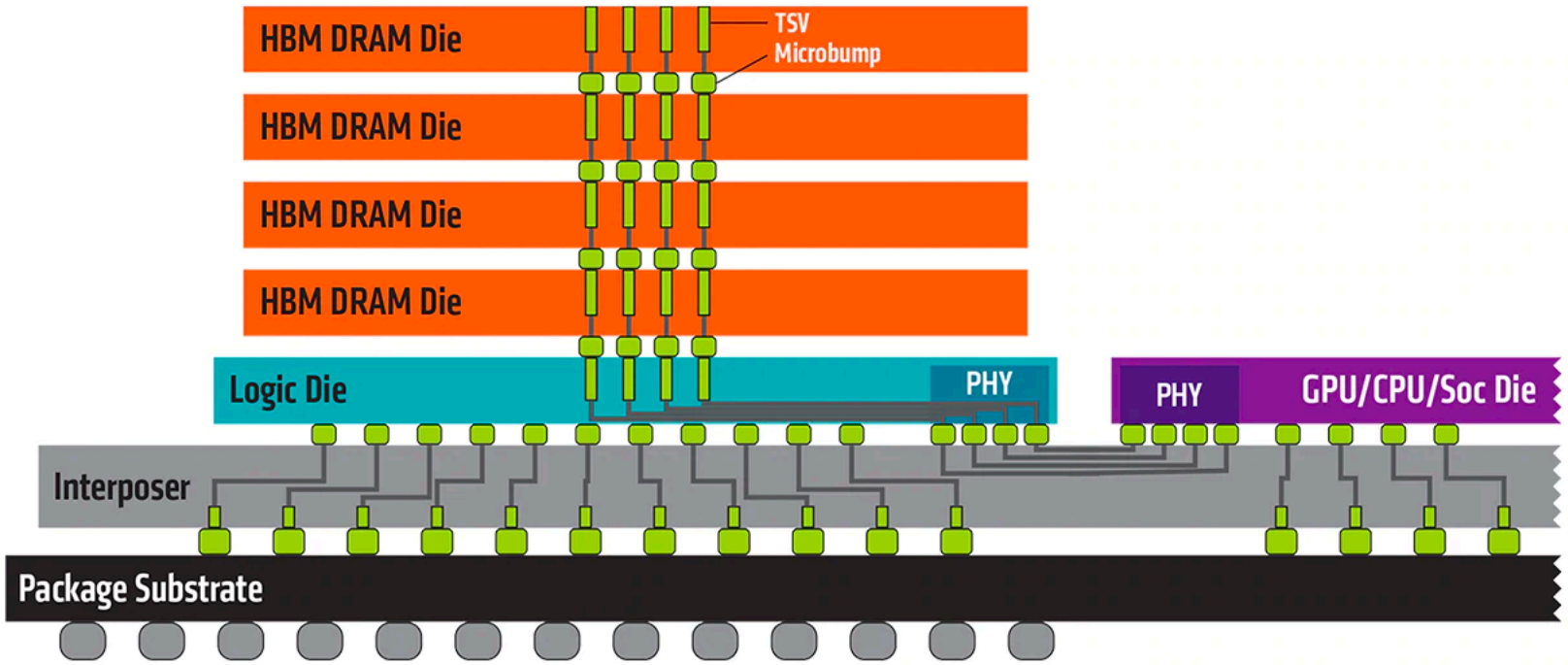
SRAM 静态随机存储器

- SRAM 不需要定期刷新，响应速度快，但功耗大、集成度低、价格昂贵。
- **应用场景**：主要用于 CPU/GPU 片内的主缓存 and/or 辅助缓存。



HBM High Bandwidth Memory，高带宽存储器

- HBM 由 DRAM 颗粒进行堆叠的架构，硬件上大幅提升了每个 DRAM 存储堆栈的容量和位宽，从而能提供超强的传输速度。具体将多层 DRAM 芯片通过硅通孔（TSV）和微型凸点（uBump）连接在一起，形成一个存储堆栈（stack），多个堆栈与逻辑芯片（GPU/CPU）通过硅中介层（interposer）封装在一起的技术。



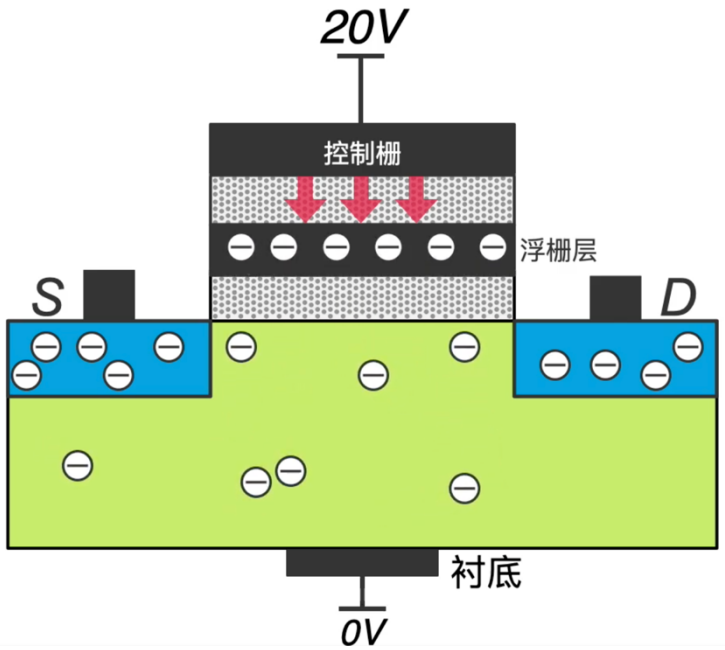
FLASH 类型

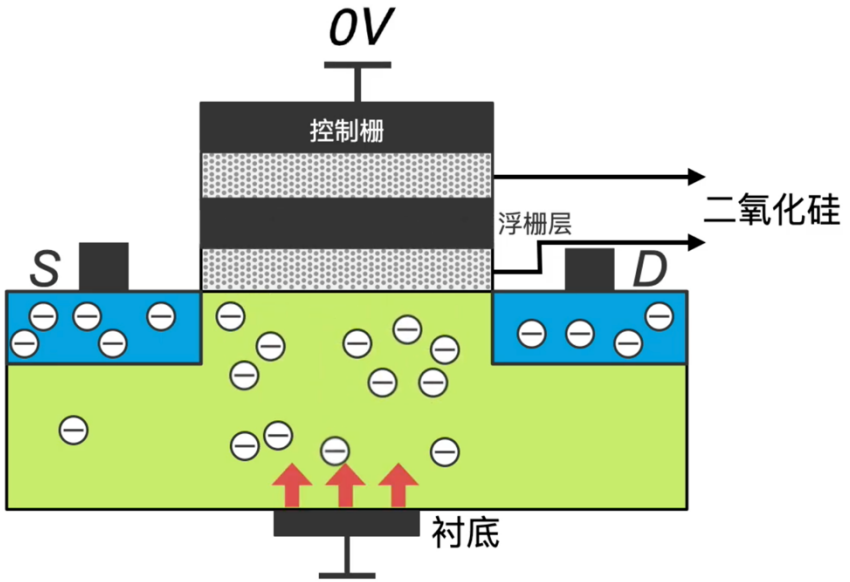
FLASH 有 NAND and Nor 两大类

- **Nor Flash：主要用来执行片上程序**
  - 优点：具有很好的读写性能和随机访问性能，先得到广泛的应用；
  - 缺点：单片容量较小且写入速度较慢，决定了其应用范围较窄。
- **NAND Flash：主要用在大容量存储场合**
  - 优点：优秀的读写性能、较大的存储容量和性价比，因此在大容量存储领域得到了广泛的应用；
  - 缺点：不具备随机访问性能。

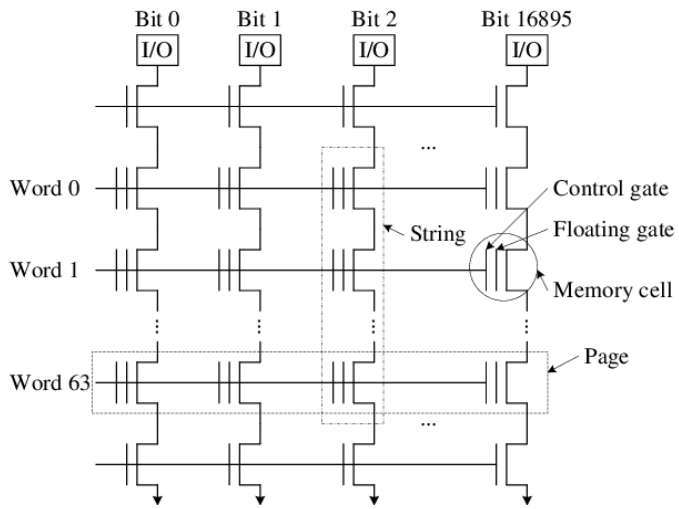
NAND Flash 工作原理

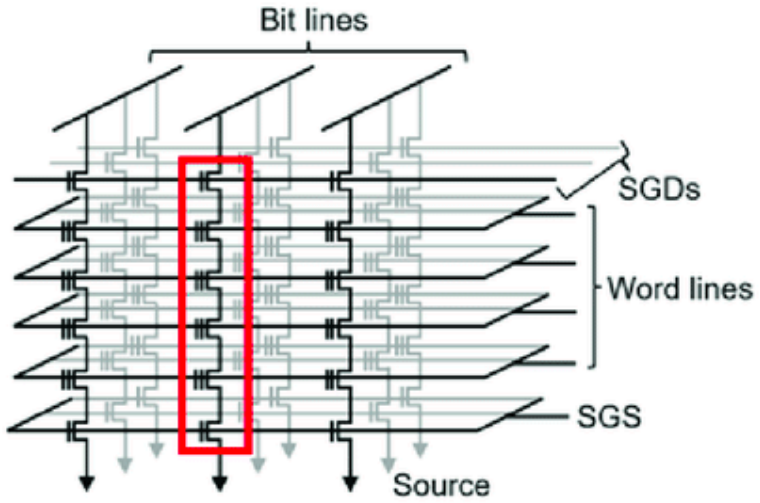
- NAND Flash 最核心的是浮栅（Floating Gate）晶体管，数据在Flash内存单元中是以电荷(electrical charge)形式存储。存储电荷多少，取决于控制门（Control gate）被施加电压，其控制是向存储单元中冲入电荷还是使其释放电荷。而数据的表示，以所存储的电荷的电压是否超过一个特定的阈值Vth 来表示。
- 发展到 CT（Charge Trap）晶体管之后，电荷被存储在CT的绝缘材料上，电荷更不易移动，相比浮栅 CT 更是将 NAND Flash 非易失性发挥到极致。

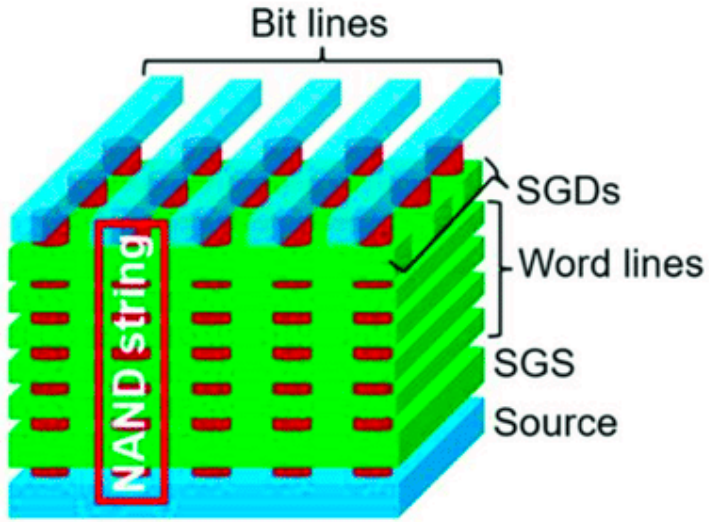




- 3D NAND Flash 仍然采用与平面存储器相同SLC、MLC和TLC技术，以进一步提高可用的存储密度。 3D 堆叠主要优点是降低每字节成本，在芯片每单位面积上封装更多的 bit。缺点是更高存储密度带来了成本增加：制造过程的额外复杂性。
- 值得注意的是，有效利用 3D NAND Flash 很大程度上取决于闪存控制器。







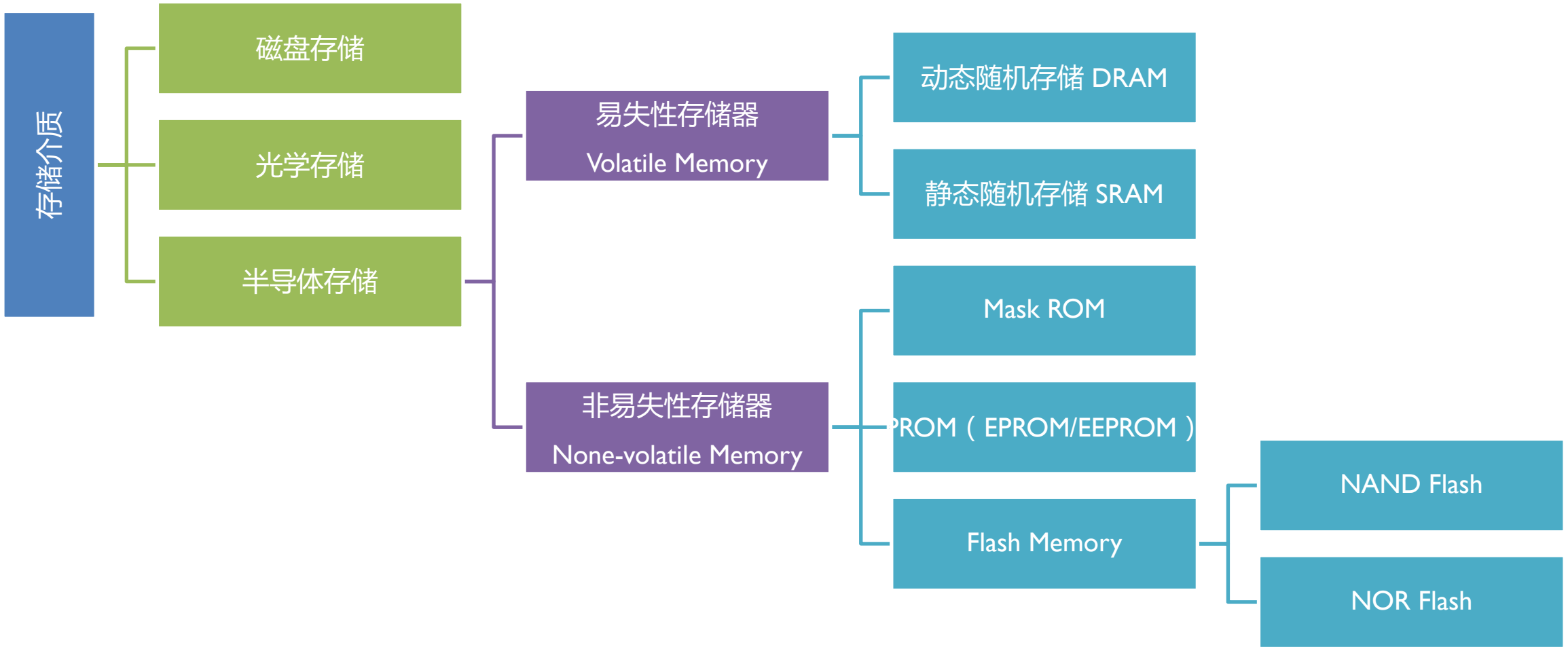
FLASH 应用场景

- NAND Flash 占绝对主导（如手机128G容量、1T固态硬盘），现在大部分 SSD 都用于存储不易丢失资料，所以 SSD 存储单元会选择 NAND Flash 芯片。
- Nor 容量不大，速度也不快，但是可以支持芯片内执行。Nor 上可以写入系统、程序和算法，而且这些程序在直接在 Nor 内部就能运行，不需要拷贝到 RAM（内存）再进行处理，因此省下 RAM 器件，所以 Nor 在嵌入式领域用途十分广泛，Nor 和其他处理器芯片（如 MCU）结合就能执行应用程序，且无需 RAM，因此在工控领域使用广泛（如 TWS 蓝牙耳机）。

SSD vs HDD

- 固态硬盘（Solid State Disk，SSD）是以NAND闪存介质为主的一种存储产品，现已广泛应用于笔记本电脑、台式电脑、移动终端、服务器和数据中心等场合，在很多应用场合可以直接替换机械硬盘。
- 固态硬盘没有机械硬盘（Hard Disk Drive，HDD）的马达+磁盘+磁头等结构，而是由纯电子电路设计，包括 NAND 闪存控制器和 NAND 闪存颗粒等，在传输等性能上比机械硬盘具有较大优势，例如读写速度远超机械硬盘的读写速度，没有机械硬盘的机械性延迟，还抗震防摔，功耗低，没有噪音。机械硬盘存在毫秒级的存取延迟，而固态硬盘的存储延迟为数百微秒，这差别的倍数还是很大的。

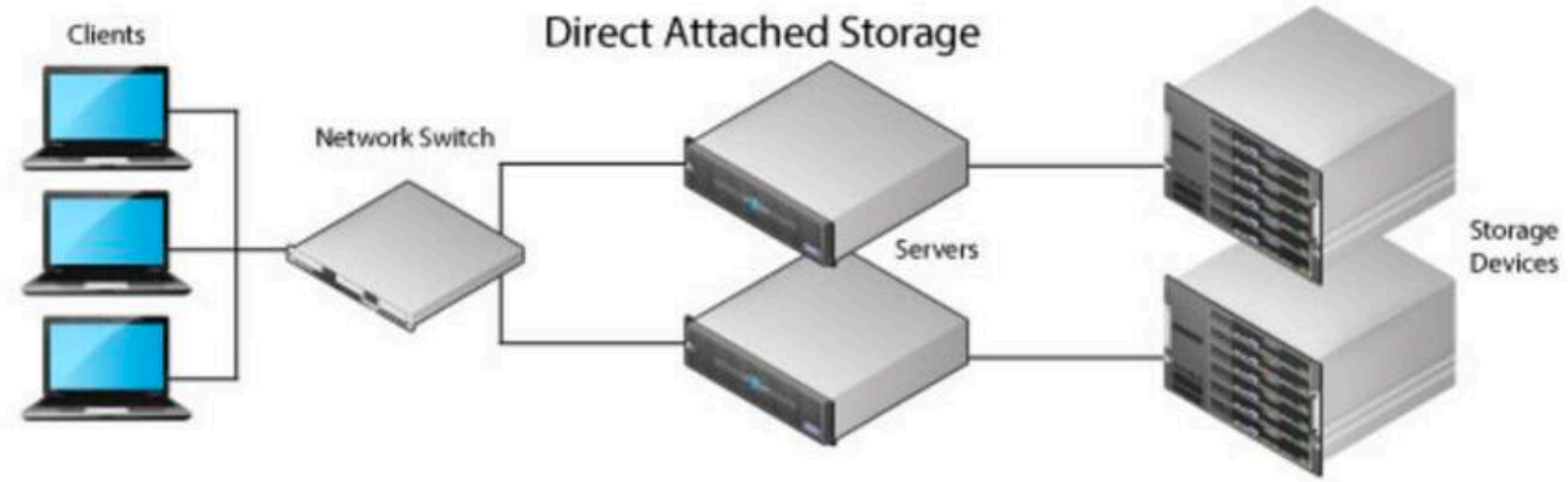
存储介质分类


存储介质
磁盘存储
光学存储
半导体存储
易失性存储器
Volatile Memory
动态随机存储 DRAM
静态随机存储 SRAM
非易失性存储器
None-volatile Memory
Mask ROM
ROM（EPROM/EEPROM）
Flash Memory
NAND Flash
NOR Flash

2. 数据存储の連接方式

DAS Direct Attached Storage，直连存储

- 比较熟悉的两种主流存储类型是块存储和文件存储，块存储中 DAS（Direct-Attached Storage 直连存储）是主流的企业存储方案。直连存储就是存储设备直接与主机服务器连接，其他主机不能使用这个存储设备（广义上，日常生活使用的 HDD 硬盘可以理解为最简单 DAS 存储）

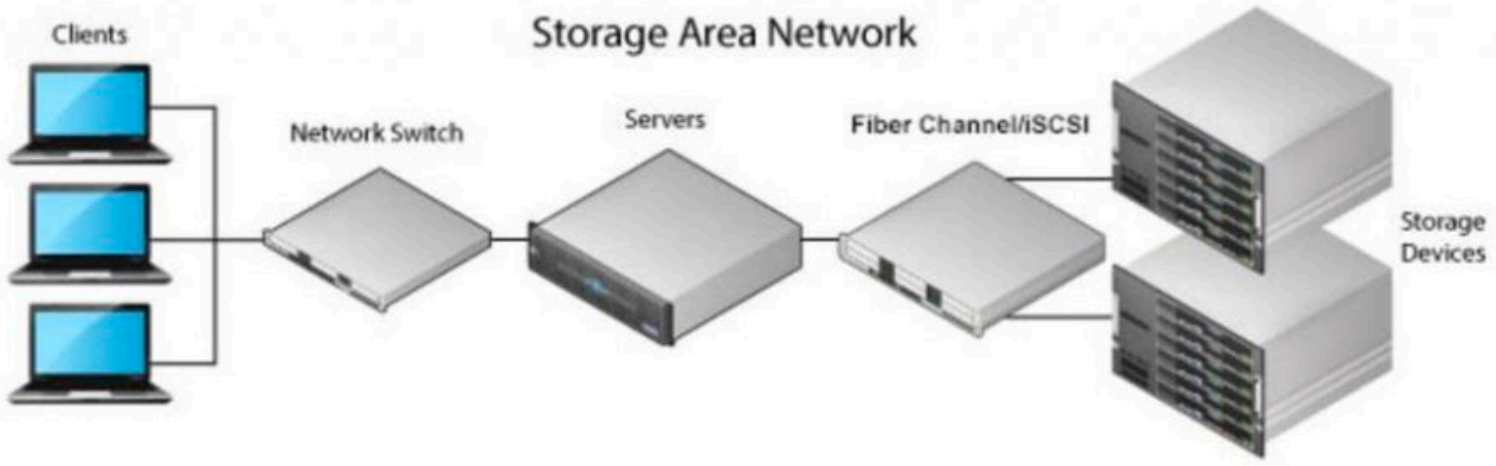


DAS Pons and Cons

- **DAS 优点**
  1. **实施简单**：无须专业人员操作和维护，节省用户投资。
  2. **大容量存储**：内置较多磁盘和较大存储空间，后端可以挂接磁盘扩展柜，满足海量存储需求。
  3. **提高存取性能**：操作单个应用数据，同时有多个物理磁盘并行工作，运行速度比单个磁盘运行速度高。
  4. **数据和 OS 分离**：OS 一般存放在服务器硬盘中，应用数据放置于存储阵列中，实现数据和 OS 分离，操作系统升级和应用数据运行相互没有直接影响。
- **DAS 缺点**
  1. **故障恢复难**：服务器本身容易成为系统瓶颈，服务器发生故障时，整个存储数据将不可访问。
  2. **管理复杂**：对于存在多服务器的系统来说，设备分散，不便于管理；数据备份时的操作会比较复杂。
  3. **动态分配**：同时多台服务器使用 DAS，存储空间不能在服务器之间动态分配，可能造成相当的资源浪费。

SAN Storage Area Network，存储区域网络

- 通过交换机将磁盘阵列等存储设备与相关服务器连接起来的高速专用存储网络。1）FC-SAN，即基于 FC 设备及通信协议的存储区域网络。2）光纤网络成本高，又发展出IP SAN，即基于以太网 SAN 存储形式。 IP SAN 利用 TCP/IP 协议实现了对 SCSI 协议封装，主机端相当于 CS 架构客户端，存储端则是服务端。

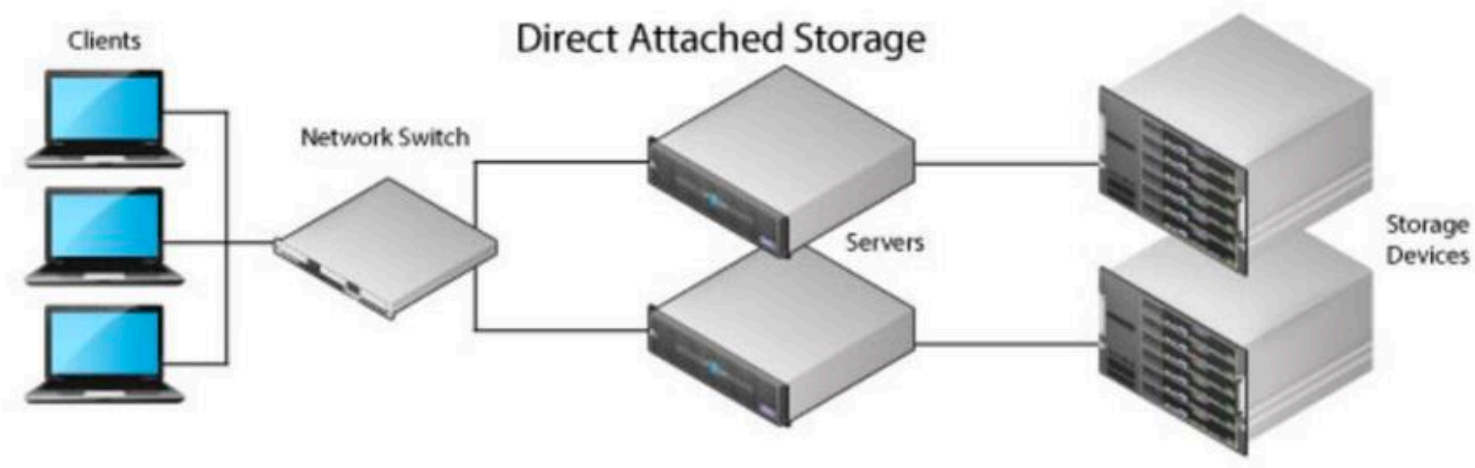


SAN 优缺点

- **SAN 优点**
  1. **易于扩容**：SAN 提供一个共享存储池；存储空间不足时，可以进行空间扩容，SAN 可以随时增减服务器。
  2. **集中管理**：SAN 把前端服务器数据集中到一起，使得数据管理和系统运维都得到极大的简化。
  3. **优异存取性能**：SAN 使用光纤传输数据，数据读取性能非常高效非常高速。
  4. **方便灵活备份**：重要数据集约化，更有效备份来实施数据保护，运行备份操作时无需考虑对网络总体性能影响。
- **SAN 缺点**
  - **价格昂贵**：FC SAN 应用光纤作为通信信道，使得 SAN 整体构架成本居高不下，但是因为SAN高性能、易管理等诸多优点，使得SAN仍然是目前应用最普遍的存储解决方案。

NAS Network Attached Storage，网络接入存储

- 存储系统直接接入网络，通过网络交换机，将服务器与存储连接在一起，用户通过 TCP/IP 协议访问数据，并通过标准文件共享协议（CIFS、NFS等）实现目录级共享。
- NAS 和 SAN 都构成一个存储网络，可以实现设备集中管理。NAS 访问需要经过文件系统格式转换，所以是文件级访问。SAN 提供是裸设备，适合块存储相关的应用（如数据库）。

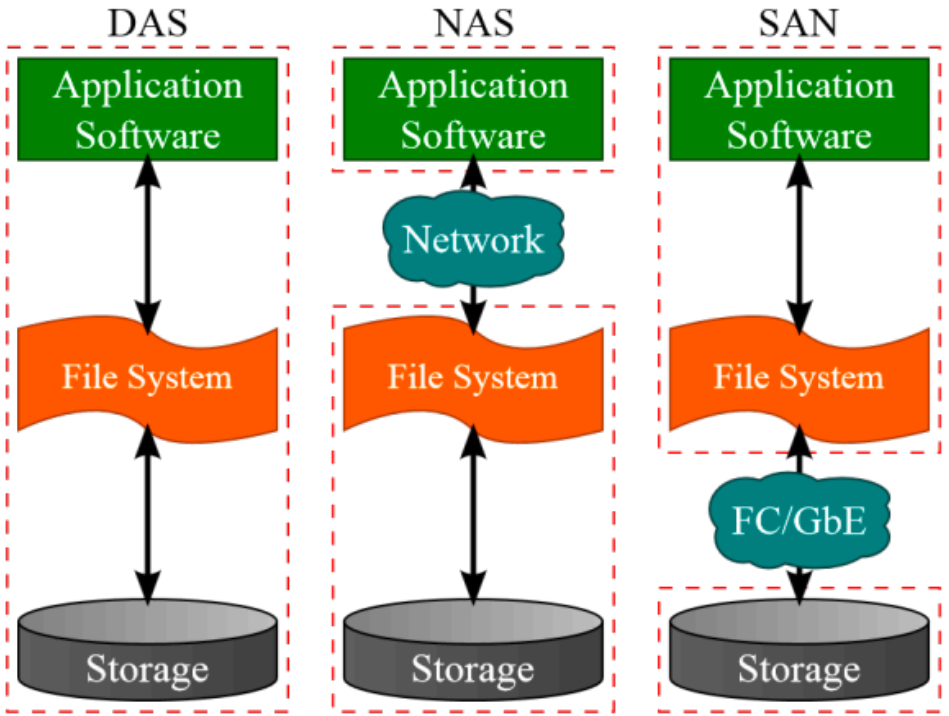


NAS 优缺点

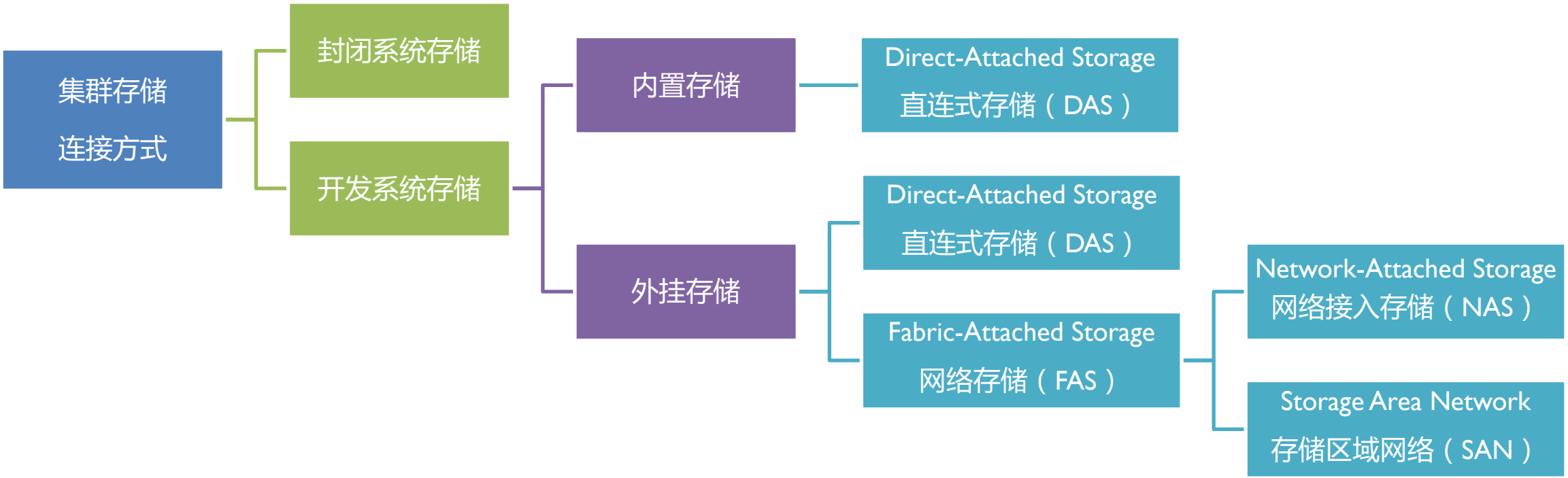
- **NAS 优点**
  1. **可访问**：所有联网设备都可以访问 NAS。
  2. **高性能**：NAS 专用于提供文件服务，其他联网设备无需再提供文件服务。
  3. **容错性**：可以对 NAS 进行格式化，以支持复制磁盘、独立磁盘冗余阵列或纠删码，进而确保数据的完整性。
  4. **易于设置**：NAS 架构可通过简化脚本，或以简化 OS 预装设备形式交付，大幅缩短存储设置和系统管理的时间。
  5. **横向扩展功能**： NAS 增加存储容量简单。不必升级或更换现有服务器，在不中断网络下启用新存储。
- **NAS 缺点**
  - 比如前期安装和设备成本比较高，可扩展性受到设备大小限制，存储性能有局限性
  - NAS受限企业网络带宽，可能出现当多台客户端访问 NAS 导致性能下降，最终不能满足用户需求等。

DAS、SAN、NAS 关系

- DAS 存储一般应用在中小企业，与计算机采用直连方式，NAS 存储则通过以太网添加到计算机上，SAN 存储则使用 FC 接口，提供性能更加的存储。NAS 与 NAS 主要区别体现在操作系统所在=位置。



连接方式分类


集群存储
连接方式
封闭系统存储
开发系统存储
内置存储
外挂存储
Direct-Attached Storage
直连式存储（DAS）
Direct-Attached Storage
直连式存储（DAS）
Fabric-Attached Storage
网络存储（FAS）
Network-Attached Storage
网络接入存储（NAS）
Storage Area Network
存储区域网络（SAN）

AI 集群的服务器存储方式 I

I. 实现统一存储系统？

- 前端主机接口可支持FC 8Gb、iSCSI 1Gb和iSCSI 10Gb，后端具备SAS 6Gb硬盘扩展接口，可支持SAS、SATA硬盘及SSD固态硬盘具备极佳的扩展能力。实现FC SAN与IP SAN、各类存储介质的完美融合。
- 有效整合用户现有存储网络架构，实现高性能SAN网络统一部署和集中管理，以适应业务和应用变化动态需求。主机接口及硬盘接口均采用模块化设计，更换主机接口或硬盘扩展接口，无须更换固件，简化升级维护的难度和工作量。

AI 集群的服务器存储方式 II

1. AI 业务会产生 and/or 利用更多非结构化数据，这对 NAS 和 SAN 都是挑战，NAS+SAN 或许将是未来主要的存储解决方案？
   - 作为目前比较热门的统一存储方案。其既然提供集中化磁盘阵列，那么就支持主机系统通过 IP 网络进行文件级数据访问，或光纤协议在 SAN 网络进行块级别数据访问。
   - iSCSI 作为通用 IP 协议，只提供块级别数据访问。磁盘阵列配置多端口的存储控制器和一个管理接口，允许存储管理员按需创建存储池或空间，并将其提供给不同访问类型的主机系统。

3. 小结&思考

小结

1. 了解数据存储的硬件介质和对应的分类级别（RAM、FLASH）
2. 了解数据存储的集群主要提供的连接方式（DAS、NAS、SAN）

Reference 引用&参考

1. https://www.linkedin.com/pulse/dram-scaling-challenges-grow-aken-cheung
2. https://nn.wikipedia.org/wiki/Random_Access_Memory
3. https://www.youtube.com/watch?v=OwtV0NrTLHA
4. https://www.agrade.com.cn/IndustryNews/683.html
5. https://www.cnblogs.com/FireLife-Cheng/p/17183491.html
6. https://www.wpgdadatong.com/blog/detail/43099
7. https://www.sohu.com/a/476456430_121149658
8. https://www.programmersought.com/article/66421832717/
9. https://blog.router-switch.com/2021/03/storage-architecture-nas-vs-san-vs-das/
10. https://www.linkedin.com/pulse/dram-scaling-challenges-grow-aken-cheung
11. https://36kr.com/p/1933583910734464
12. https://lvxixiao.gitbook.io/blog/os/store/hardware
13. https://www.redhat.com/zh/topics/data-storage/network-attached-storage
14. https://community.fs.com/cn/case-study/the-common-types-of-storage.html
15. https://help.aliyun.com/zh/ecs/user-guide/nvme-protocol
16. https://blog.csdn.net/fgf00/article/details/52592651

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem