信息存储理论与技术

主讲：冯　丹
陈俭喜
(chenjx@hust.edu.cn)

参考书目

1.《海量数据存储》 方粮 机械工业出版社

2.《海量信息存储》张江陵 冯丹 科学出版社

3.《大话存储【终极版】——存储系统底层架构原理极限剖析》张冬 清华大学出版社

4.《信息存储技术原理》 张江陵 金海 华中理工大学出版社

5.《数据存储架构与技术》（第2版） 舒继武 人民邮电出版社

课程安排及要求

授课方式：课堂讲授与研讨结合

研讨要求：读书报告 课堂讨论

考核方式：开卷考试(70%) + 平时成绩(30%)

报告选题参考

1. 新型存储技术及应用：

   3D Flash、NVM、DNA、量子 ……

2. 存储设备:

   SSD，RAID，可计算存储、存算融合……

3. 存储系统:

   对象存储系统，云存储架构及关键技术，新型分布式并行文件系统，分离内存……

4. 存储通道技术:

   FC，SAS，IB， NVMe，CXL ……

报告选题参考

5. 存储系统关键技术：

   存储安全问题及解决方案

   存储系统可靠性问题及解决方法

   存储系统性能分析、评价及提高性能的手段

   Storage for AI，AI + Storage

6. 其他与存储密切相关的有价值的论题

得分不高的情况：
1、单篇论文简单翻译
2、网上搜索资料拼接
不及格的情况：
1、抄袭，2、态度问题

建议：调研阅读近几年顶级存储论文，选定某个主题做一个综述性的报告

（会议：FAST，ISCA，HPCA，ASPLOS，MICRO，OSDI，SOSP，ATC，EuroSys，…，
期刊：IEEE/ACM TC， ToS， TPDS…）

报告/考试时间安排

1. 提交题目、内容提要、参考文献列表及小组成员，每小组不超过4人：

   9月30日之前邮件提交至 chenjx@hust.edu.cn

   邮件标题：组长学号-姓名-报告主题词

2. 公布课堂报告人选:

   提交题目后一周内，邮件通知确认

3. 课堂报告、交流讨论:

   第8、9周

4. 开卷考试:

   第11周周末（待定）

数据信息爆炸增长

大数据需求——以腾讯为例

月活跃用户 >8亿
同时在线用户 >2亿

月活跃用户 >13亿
(2024)

QQ空间相册

图片 4000亿张
存储量 200PB

微信朋友圈

日上传 10亿张
日下载 1200亿张

- 社交、游戏，访问密度高达100万次/秒/100GB量级的数据读写；
- 在线业务，应保证良好的用户体验，不论数据访问密集程度如何，均要求延迟在100毫秒以内；
- 服务器数量10万级别。

急剧增长的问题——仅靠扩大规模难以为继

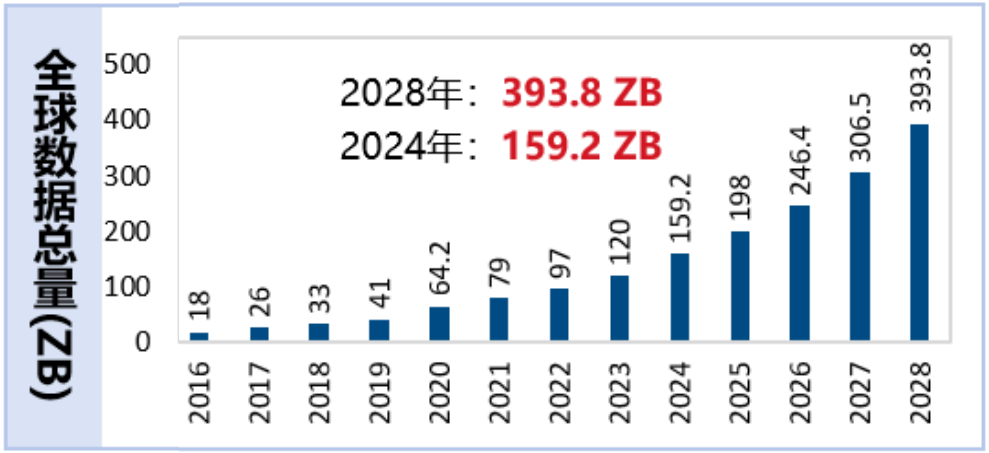
高性能、低延迟、大容量、可扩展 | 负载复杂、能耗高 | 长期保存困难

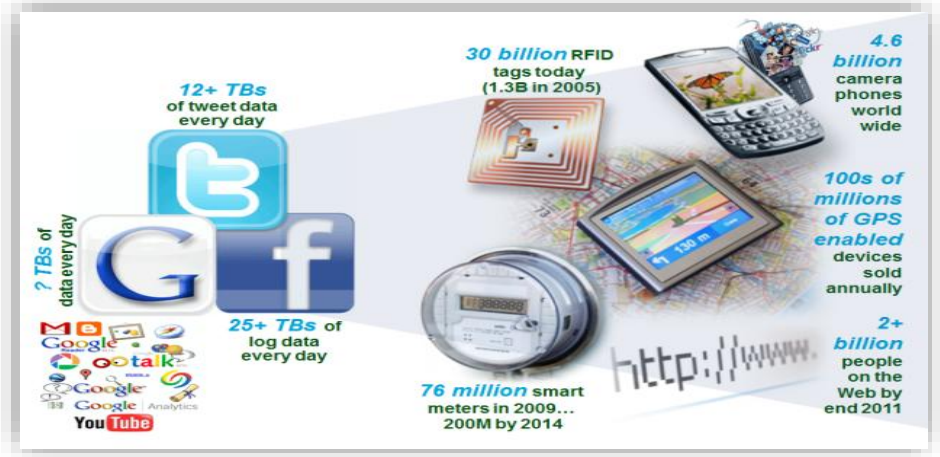
数据信息爆炸增长

- **大数据时代**：人类每18个月产生的信息量，是人类之前全部信息量的总和！

—— 新摩尔定律，Jim Gray

2024年已达159.2ZB，2028年将达到393ZB

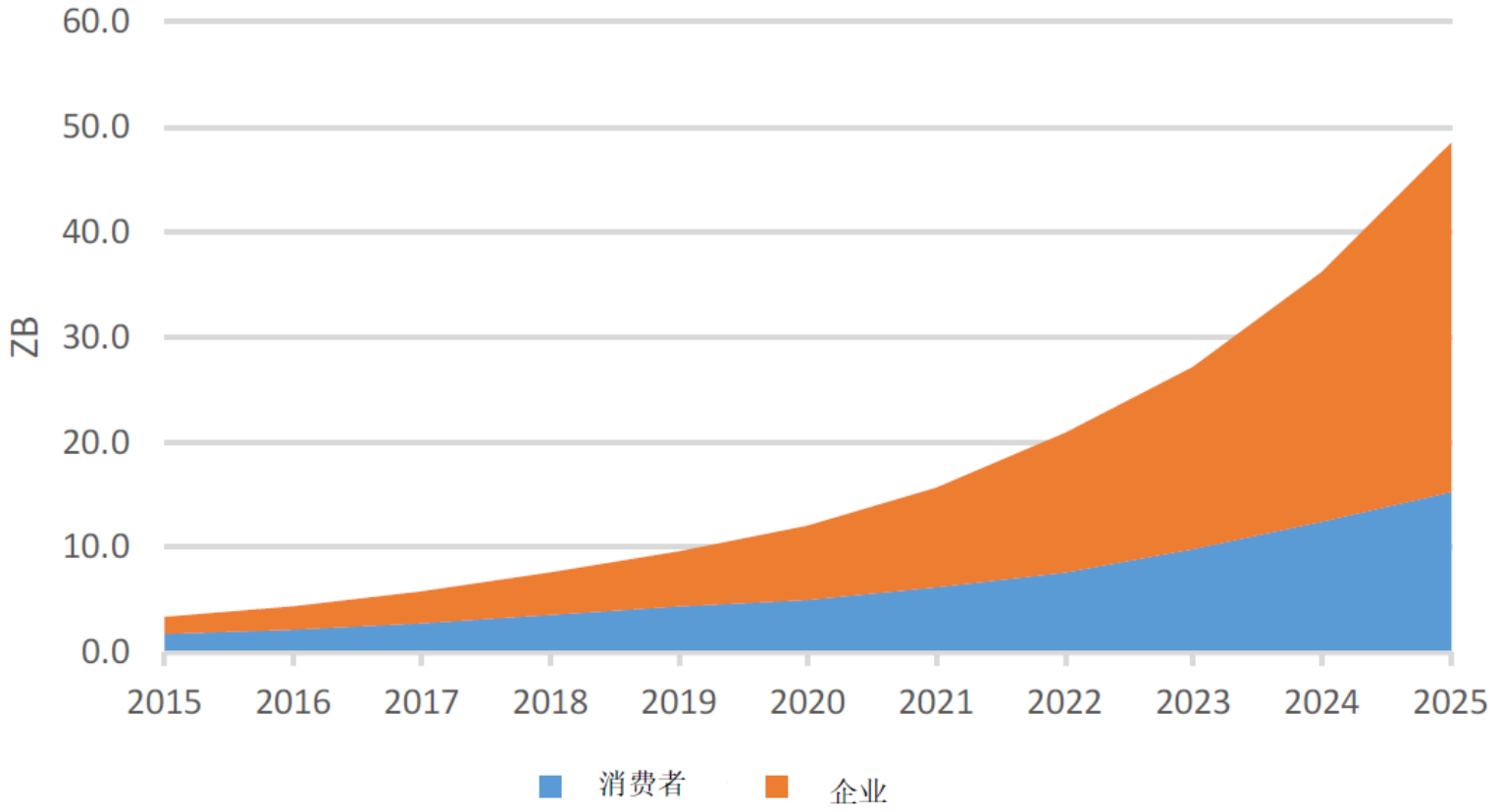




数据爆炸式增长
令存储系统面临严峻挑战

数据信息爆炸增长

- **IDC：我国拥有全球最大的数据量（48.6ZB，2025年）**

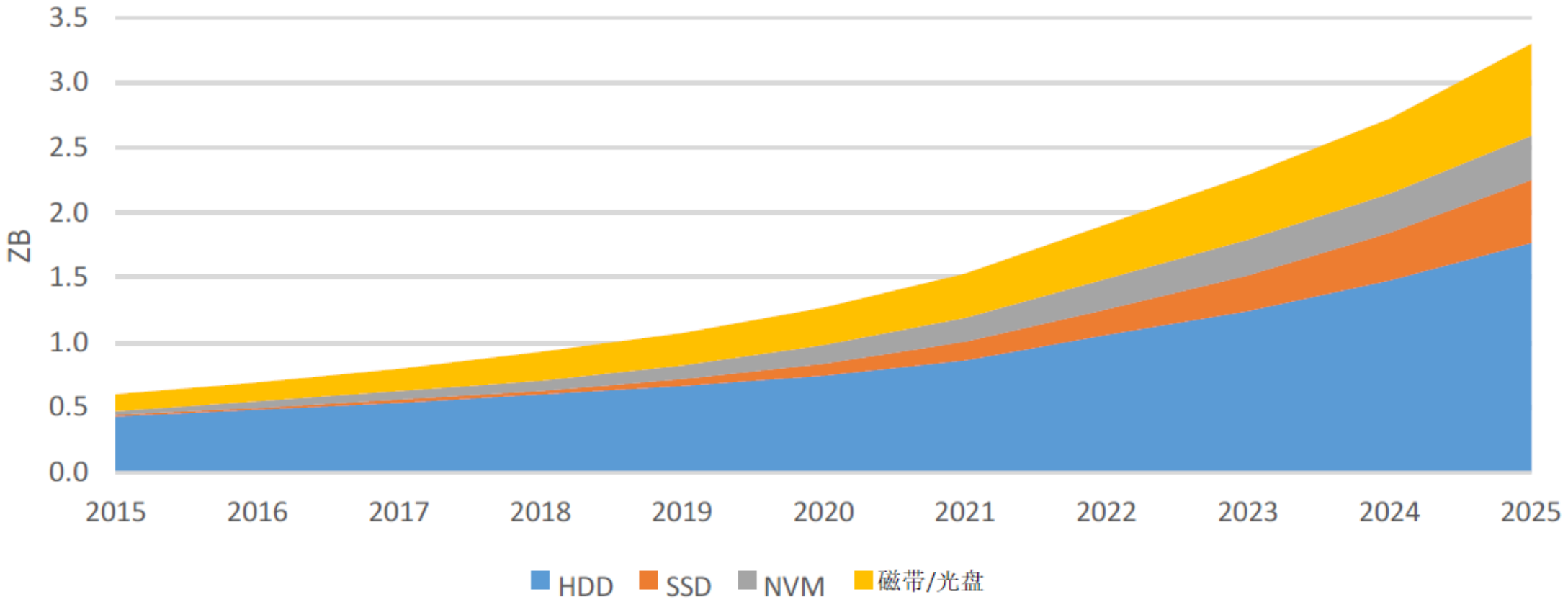


- 2018到2023年，企业级存储装机容量年增长率将达到**25.1%**
- 个人存储容量年复合增长率为**5.9%**

数据时代 2025，2018 年 11 月

数据信息爆炸增长

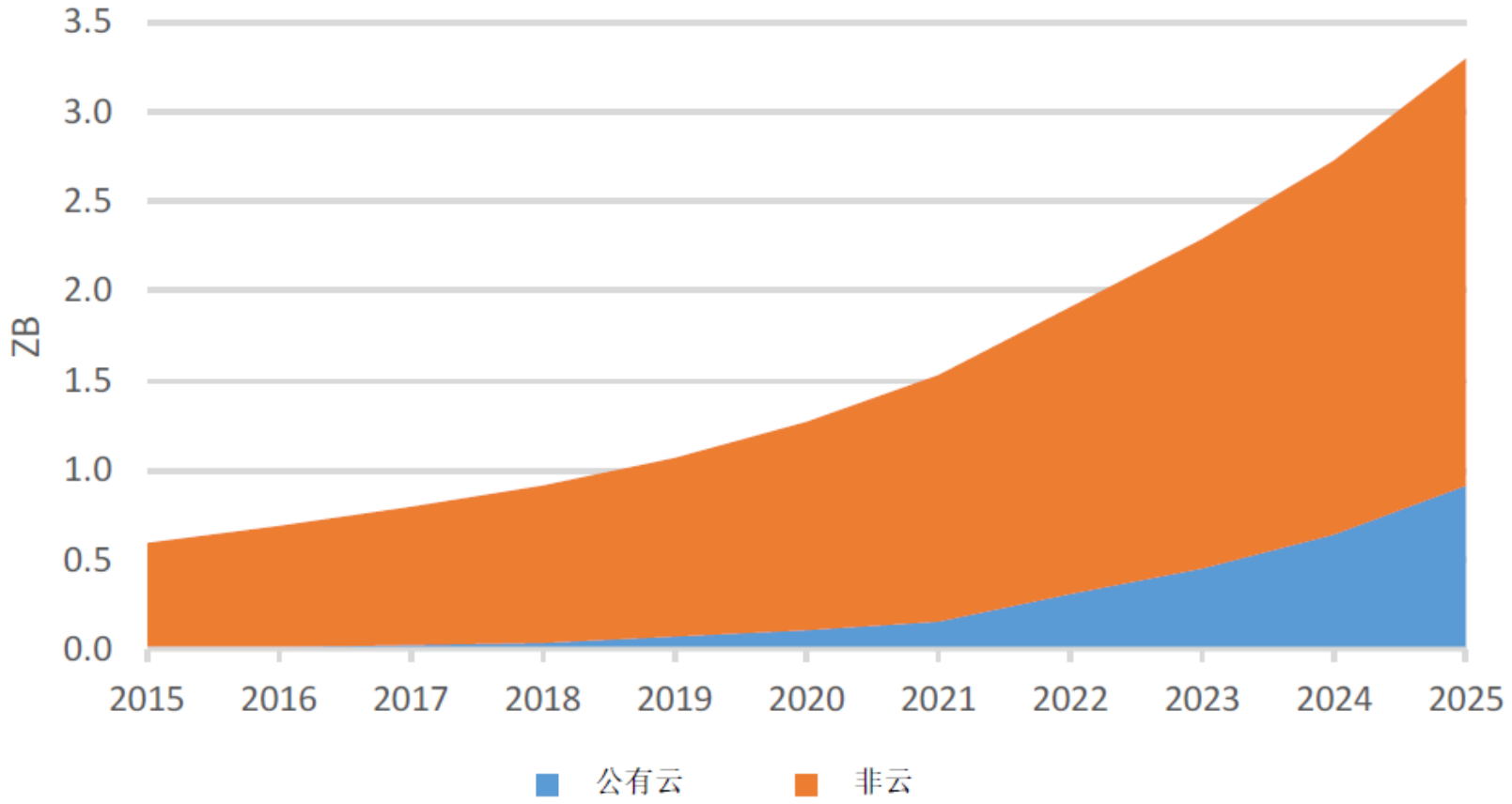
- 不同介质类型的存储容量变化



整体存储装机容量年复合增长率为18.4%，未来五年**SSD**的年复合增长率将达到**44%**

数据信息爆炸增长

- 数据存储云化



国家需求

- 数据是国家基础性战略资源，存储是保障信息安全和发展数字经济的基石
- 硬盘：Seagate, Western Digital；存储芯片：Samsung，Micron；
- 国家存储器基地落户武汉，1600亿元，2016.12

长江存储 YANGTZE MEMORY

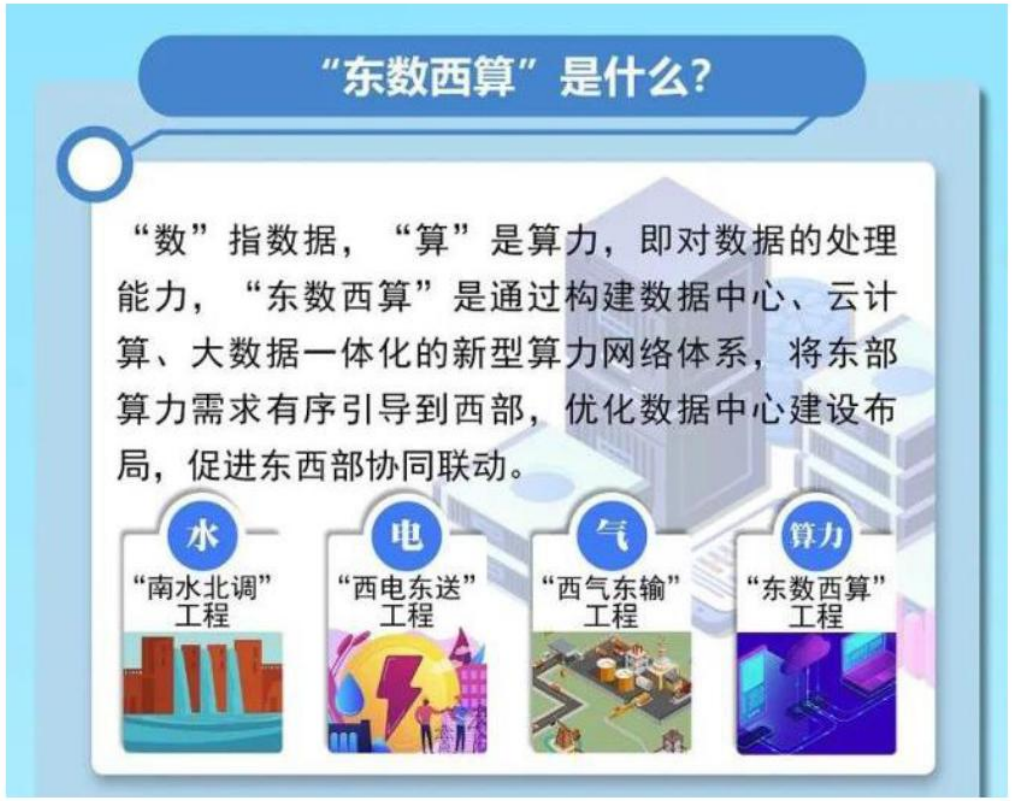
信息存储技术需要自主创新

国家需求

“东数西算”工程

新基建：大数据中心

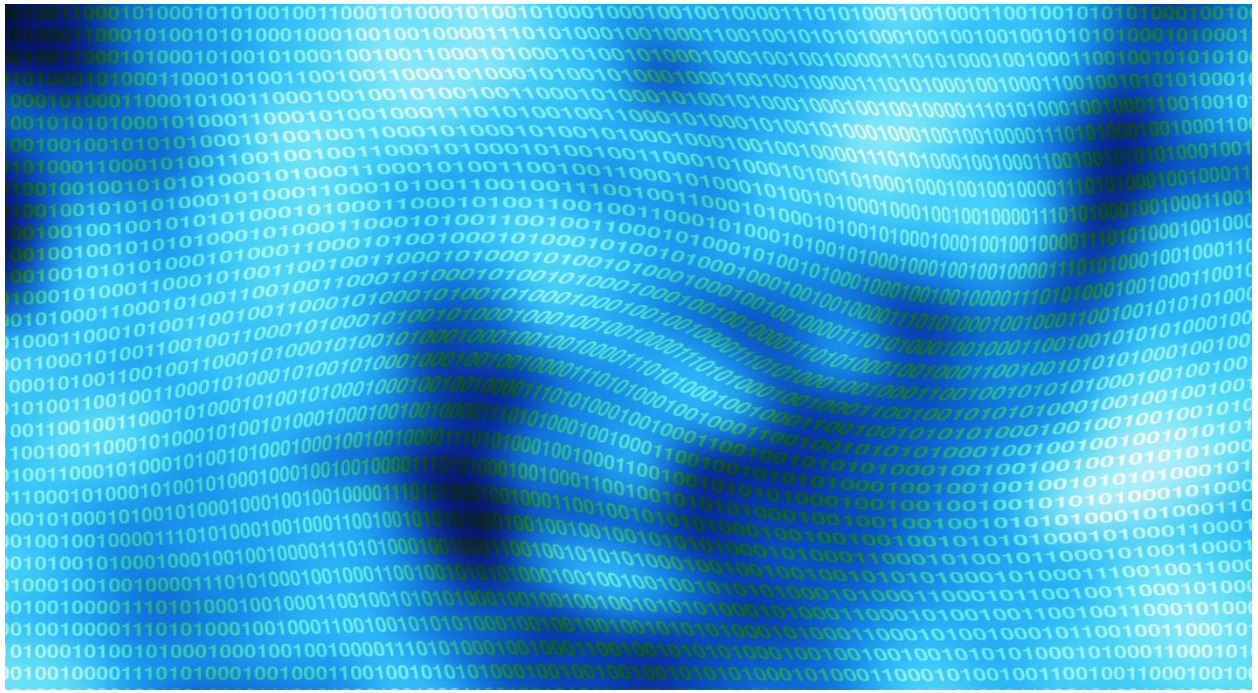
西部：
温、冷数据，时间响应要求低的计算



东部：
热数据，对时间响应要求高的计算

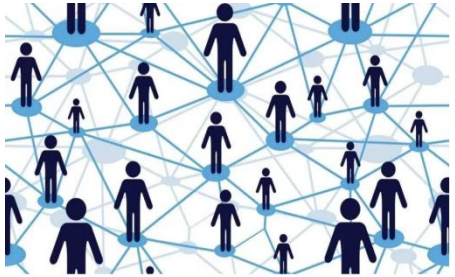
算力、**存力**、运输力

什么是信息存储？



信息

- 是指通过传递和交流的方式，向接收者传达一定的内容、知识或者数据的过程
- 具有传递、共享和存储的功能
- 信息的传递和交流是人类社会活动的基础，它能够促进人们之间的理解、沟通和合作

信息的传递

- 信息是需要传递
- 传递：**空间**、**时间**
- 跨越空间的传递：称之为**通讯**、**传输**
  例：打电话，信息跨越空间传递
- 跨越**时间**的传递：称之为**记忆**、**存储**
  例：读李白的诗，信息跨越时间传递

孤帆远影碧空尽，唯见长江天际流

信息存储的本质

跨越时间进行信息传递的过程

记录当下，相约未来

存储技术的前世

-- 跨越时间的信息传递

岩画

时间跨度6000年

石刻楔形文字(古巴比伦)

汉穆拉比法典

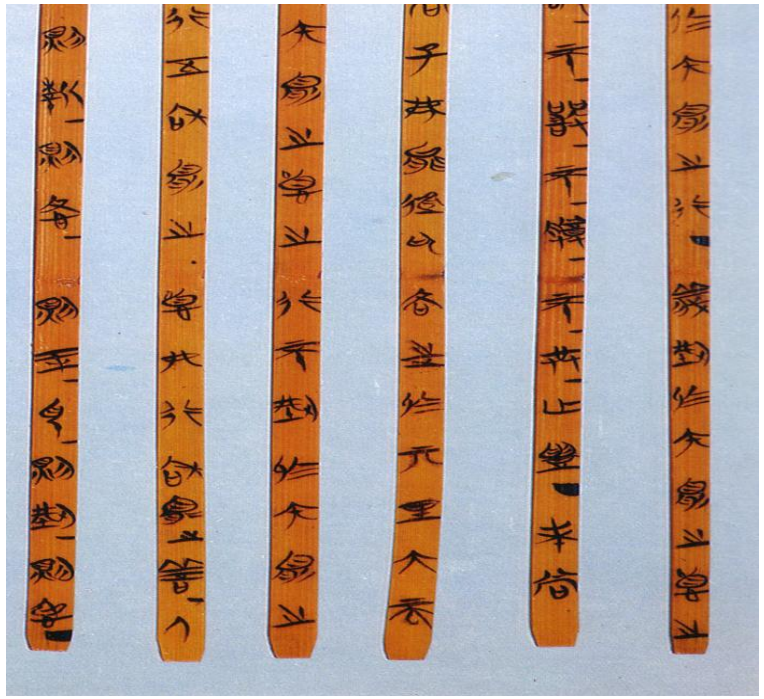
时间跨度 3800年

前人传给我们的信息

甲骨文(商代)

时间跨度：

3000年



竹简 (汉代)

时间跨度：

2000年

韦编三绝

伟大的存储发明：造纸和印刷

蔡伦造纸（东汉）

时间跨度：1900年

木活字

活字印刷（宋）

时间跨度：950年

(农书)

伟大的存储发明：照相术和留声机

(法国)路易斯·达盖尔(LOUIS DAGUERR)

发明照相术

时间跨度:
183年

爱迪生发明的留声机

时间跨度:145年

存储技术留下历史时刻

伟大的存储发明：电影和录像

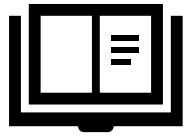
爱迪生发明电影机
时间跨度：135年

录像机发明
时间跨度：95年

信息存储的重要性

- 通讯是传播知识，存储是积累知识，它们是促进人类文明发展的左膀右臂
- 每当存储技术有一个划时代的发明和应用，社会进步明显加快

什么是信息存储技术？



- 利用存储工具加强和扩展人的存储(记忆)能力
- 信息存储技术的本质是利用具有时间稳态的物理原理和现象，实现跨越时间的信息传递
- 什么物理现象具有时间稳态：

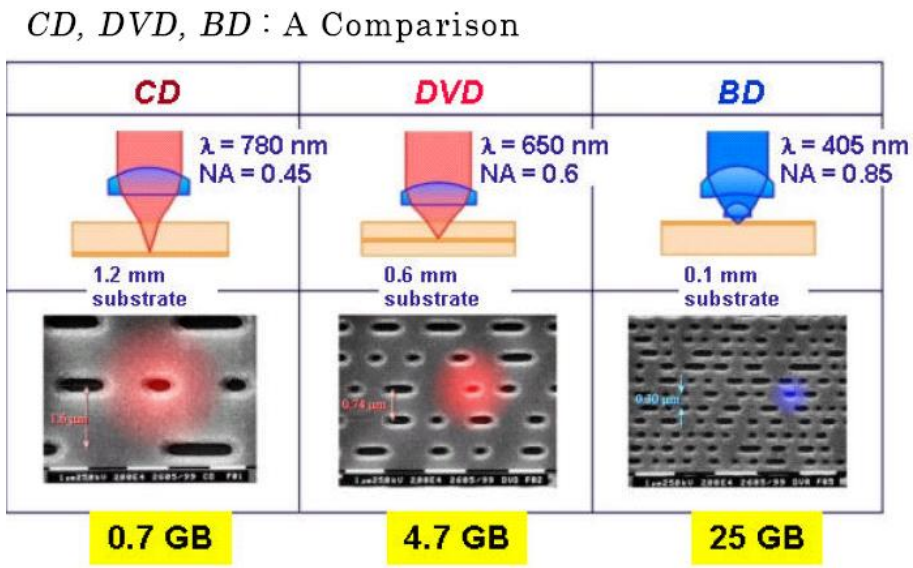
  形变/色变对光的反射和透射，磁稳态，半导体稳态，量子稳态，生物稳态
- 满足三个条件：

  具有2个或以上状态，状态可识别，状态可改变

磁　　光　　电　　量子　　DNA

时间稳态的物理现象

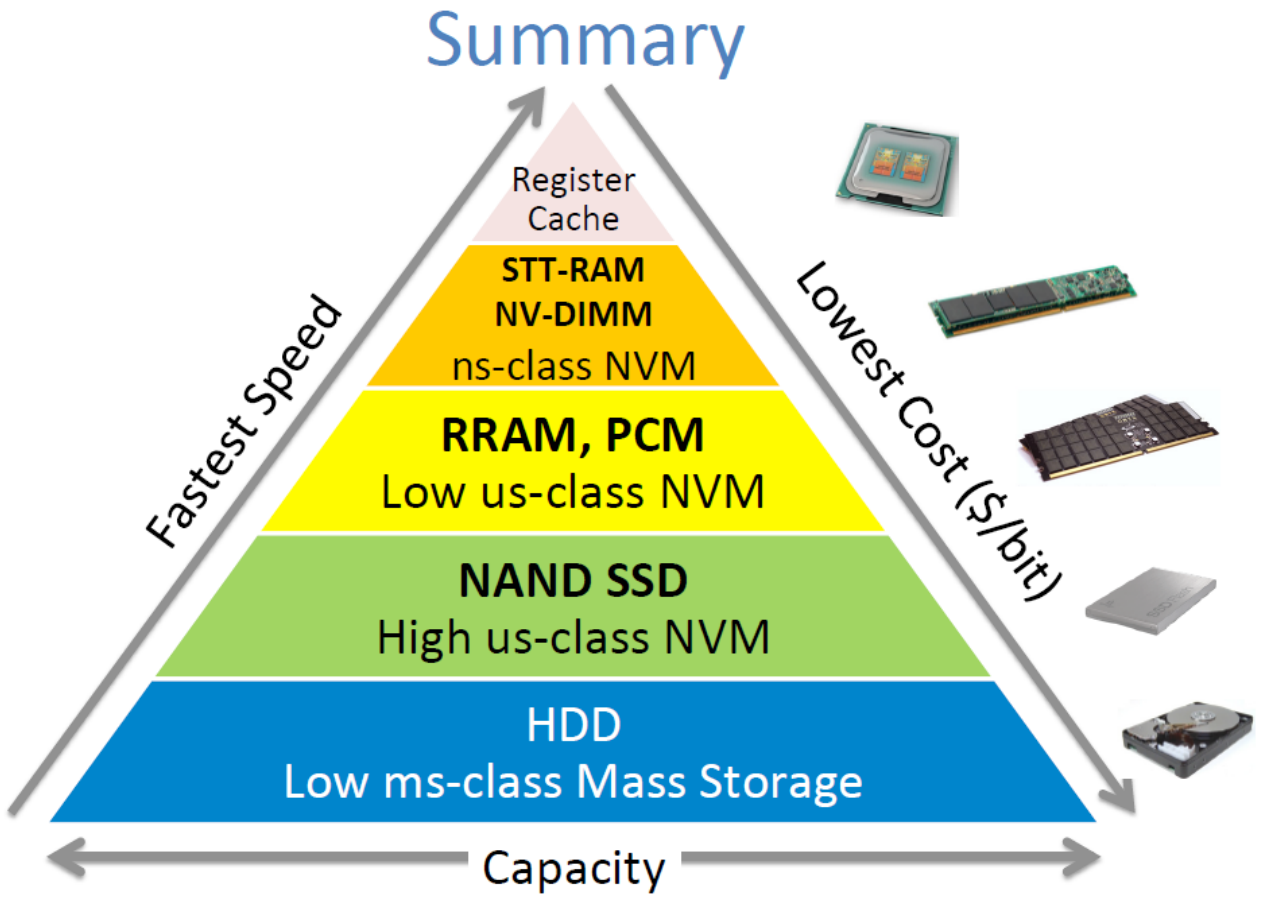

泉州萬安渡石橋始造於皇祐五
年四月庚寅以嘉祐四年十二月
辛未訖功衆趾于渕釀水爲四十
七道梁空以行其長三千六百尺
廣丈有五尺翼以扶欄如其長之
數而兩之靡金錢一千四百萬求



玻璃硬盘

古代的石存储器和现代的光存储器利用同一物理现象

目前所用的主要存储器


Summary
Register
Cache
STT-RAM
NV-DIMM
ns-class NVM
RRAM, PCM
Low us-class NVM
NAND SSD
High us-class NVM
HDD
Low ms-class Mass Storage
Fastest Speed
Lowest Cost ($/bit)
Capacity

最古老而最先进的存储器

- 上帝（大自然）创造了最先进的存储器---大脑
- 其智慧程度与人类当前所能创造的存储器相差n个数量级
- 但是，这种存储器会产生遗忘，不精确，难以复现和表达，只能靠语言一代一代的传承知识（大脑转存）
- 所以，人们开始发明各种存储技术。正是借助这些存储技术，我们才能了解远古时代人类的生活状况，更进一步说，我们现今的一切知识和文明，都是通过这些存储技术才得以积累和发展的

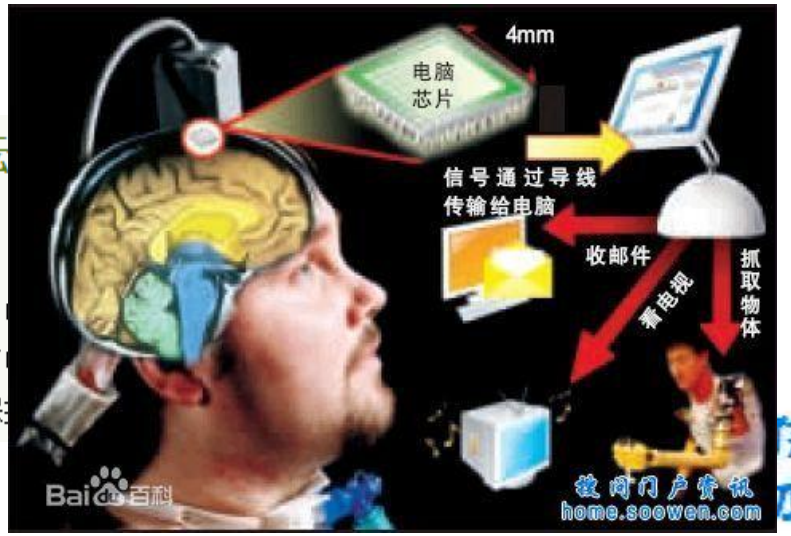
- 人脑：巨大数量的神经元 (10^{11}) 和突触 (10^{15})
- 神经元之间复杂的互联和通信

美国脑计划经费突破45亿美元

2014年6月，美国总统巴拉克·奥巴马的
目--该项目旨在绘制活体人脑图谱，并命令
为该计划拨款1亿美元，覆盖国立卫生研究

“中国脑计划”酝

随着欧、美、日相继启动各种人脑计划，
在复旦大学举办的东方科技论坛上了解到，“
发展的重大科技项目”之一，将从认识脑、保



当代信息存储技术的特征

- **数字化：**

  一切形式的信息统一转换为0，1的集合而存储

- **网络化：**

  所有的存储器都可通过网络（广义的网络：有线、无线、互联总线）互连起来，首次实现了时间和空间的二维任意传送

网络存储系统的两种资源

存储资源（容器）：

半导体存储（RAM、NVRAM、Flash)、磁存储（硬盘、磁带）、光存储（CD-ROM、DVD、BD、MO)

传输资源（管道）：

总线、网络和交换、路由设备

SCSI, FC, IP, iSCSI, IB, SATA, SAS, NVMe，CXL…

存储系统发展趋势

电信行业	社交网络	物联网	金融行业	高性能计算	卫星数据中心	国家安全
业务数据繁多：联通每天入库超过42亿条用户上网数据记录，日均1.2TB以上。	微信社交网络：点：6.97亿以上 边：千亿以上	信息源复杂多样，数据量PB、EB以上	银行：7PB的磁盘存储和超过20PB的磁带存储	NERSC：数据传输速率可高达80GB/s	资源卫星中心：每日生成数据大于12TB	AFCCC：数据达到数PB，采用4 级数据保护方式

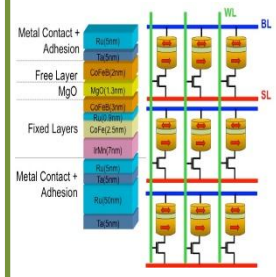
各个行业大数据应用对存储提出更高要求

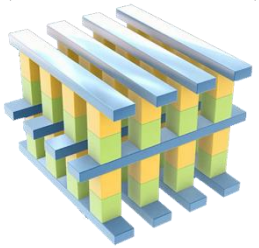
高并发、高带宽、低延迟、可扩展、低能耗、长程保存

面向大数据的新一代存储技术？？？

需要在存储器件、设备、系统三个层面实现理论和技术突破

新器件 (机理、结构)





新设备 （方法、技术）

新系统

课程内容

需求 → 多层次研究：器件、设备、系统

1. 磁存储器基本原理
2. 光存储器基本原理
3. 固态存储技术
4. 存储系统接口与互联
5. 磁盘阵列
6. 网络存储系统
7. 并行与分布式存储系统
8. 存储虚拟化与云存储
9. 数据保护与存储安全

1. 存储器类型

2. 存储器分层结构

1.存储器类型

a. 存取方法：

顺序存取：访问时间与存储单元的物理位置密切相关，磁带

随机存取：每一位置有唯一的寻址机制直接达到，RAM

直接存取：块间直接到达、块内顺序存取，磁盘

关联存取：一个字通过其部分内容而不是地址进行访问，Cache、TLB

存取方法

存取方式	特点	访问时间与位置关系	典型例子	主要用途
随机存取	可直接访问任一单元	**无关**	内存条、CPU缓存	**主存**、缓存、外部存储
顺序存取	必须按顺序访问	**密切相关**（线性顺序）	磁带	海量数据备份
直接存取	先直接定位到区域，再在区域内顺序查找	**部分相关**	**机械硬盘**、光盘	**辅助存储器**（传统磁盘）
关联存取	按内容并行查找，而非按地址	**无关**（但原理完全不同）	TLB Cache	专用高速缓存

固态盘(SSD)属于哪一类?

1.存储器类型

b. 功能：

只读存储器：ROM,CD-ROM

可重写的存储器：Disk

可擦除的存储器：EPROM，Flash

c. 物理类型：

半导体、磁表面、光

d. 物理特性：

易丢失/不易丢失、可擦除/不可擦除

1.1. 半导体主存储器

（1）分类

a. 随机存储器RAM： 数据易失

静态RAM：触发器中逻辑门，快；

动态RAM：电容充电存储数据，需刷新

b. 只读存储器ROM： 数据永久保存

微程序设计、常用函数库、系统程序、功能表

⟶ 批量生产

c. 可编程ROM（PROM）：

写一次，数据不丢失 ⟶ 少量生产

写一次读多次：

PROM（EPROM）：

紫外线擦除、电写、数据不易失

闪存（FLASH Memory）：

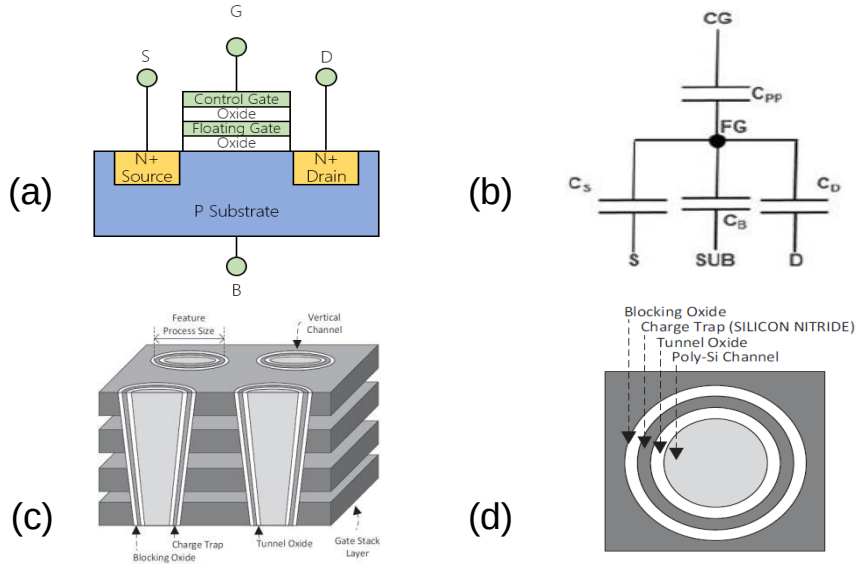
一至几秒内被擦除，数据不易失

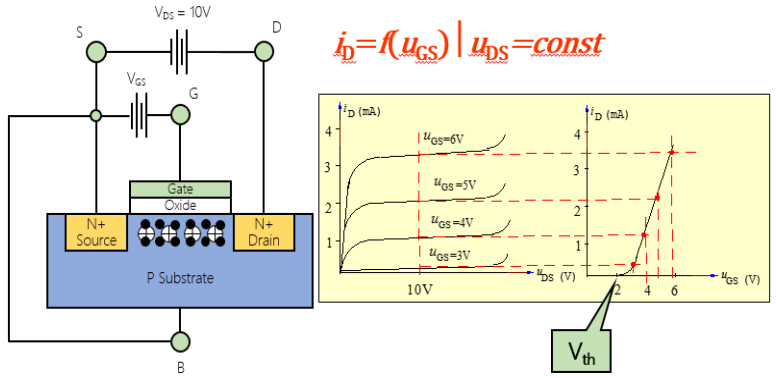
电可擦PROM（EEPROM）

存放配置信息

FLASH MEMORY 闪存单元结构

- 目前主流闪存单元分为浮栅单元和电荷俘获单元；
- 通过一定外加电场作用，使闪存单元俘获/排出电荷；改变存储单元阈值电压的高低，表示逻辑0、1；
  - 写：字线与位线间施加较大电压(20V)
  - 读：字线与位线间施加判断电压(≈2V)
  - 擦除：字线与位线间施加反向电压(20V)





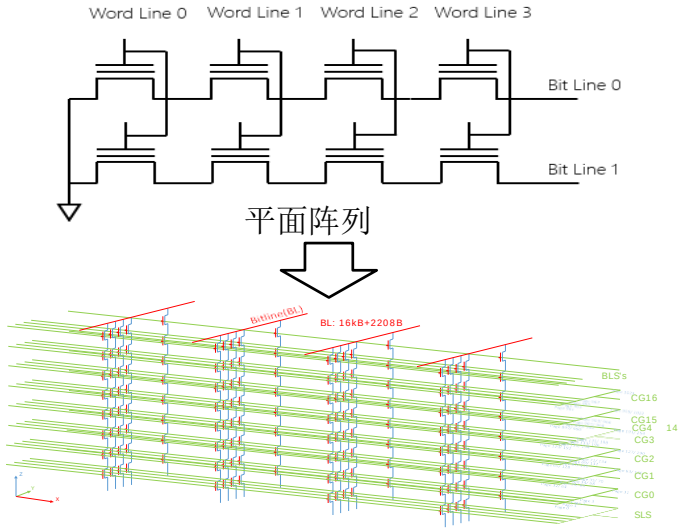
阈值电压转移特性曲线

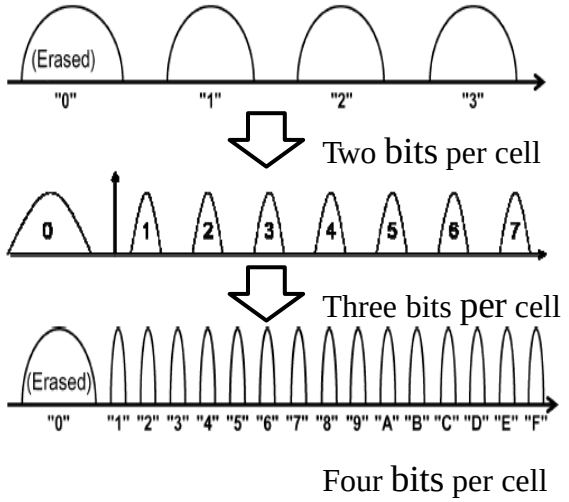
（a）浮栅单元结构
（b）浮栅单元电容模型
（c）电荷俘获单元闪存垂直通道
（d）电荷俘获单元结构

闪存阵列结构

- 从**2D**阵列转向**3D**堆叠阵列
  - 工艺尺寸缩小到达物理极限
  - 垂直堆叠进一步增加存储密度，堆叠层数逐年增加
- 单元存储更多逻辑比特
  - 存储密度进一步提高（**MLC->TLC->QLC**）
  - 阈值电压区间被压缩，可靠性和编程速度需要进一步研究

容量更大！





1.2 磁存储器

借助磁性材料的两种剩余磁化状态，或磁化与非磁化的两种材料状态，或有磁化翻转和无磁化翻转的两种状态记录二进制数据信息。

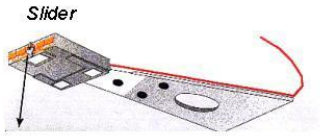
种类：

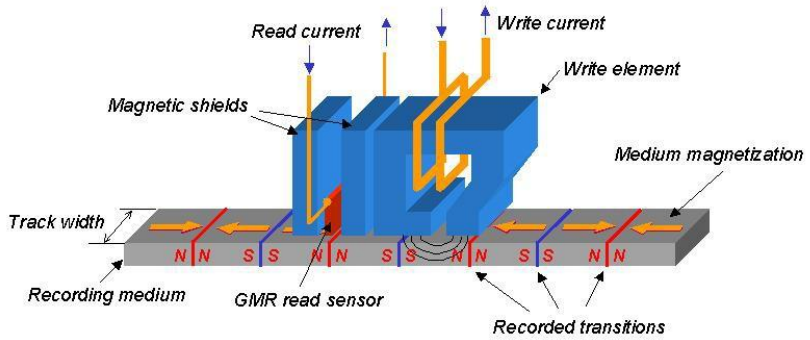
磁芯存储器 存储单元是铁氧体圆环

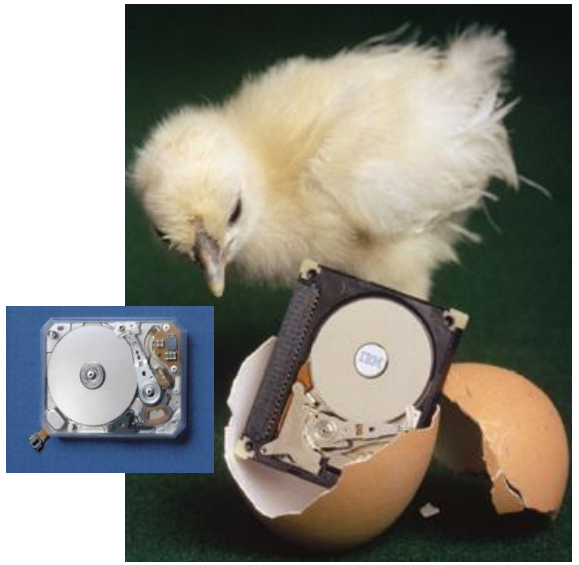
磁泡存储器 存储单元是圆柱形的磁畴，称为磁泡

磁面存储器 磁层上记录的是正、负磁化状态或磁化状态的变化，即磁化翻转。 按媒体基底材料的不同，分为两类。使用柔性基底材料的设备有磁带机，软磁盘机。使用刚性基底材料的是硬磁盘驱动器。

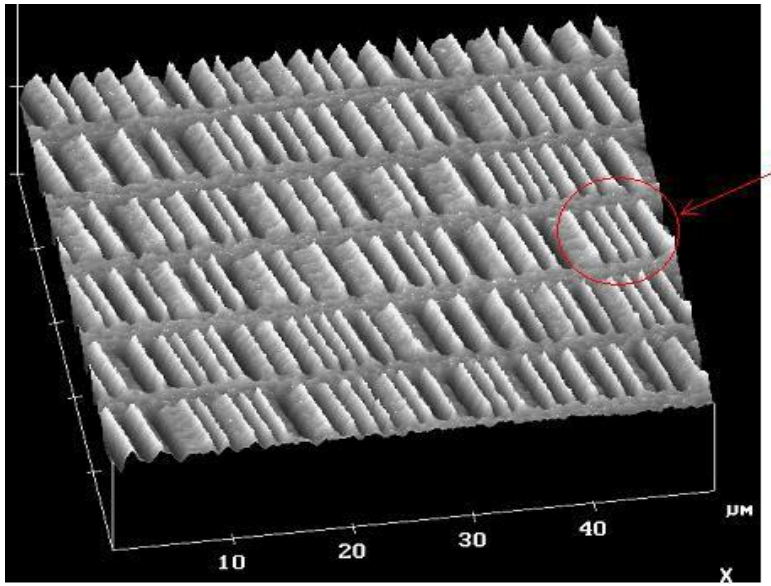
Hard Disk Drives

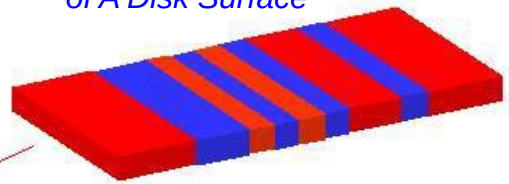


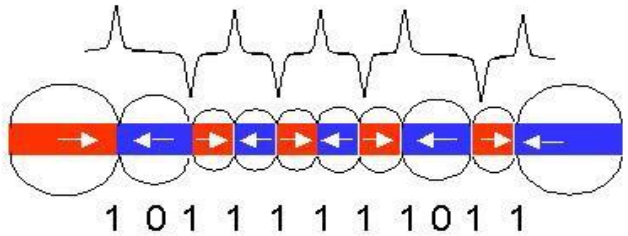




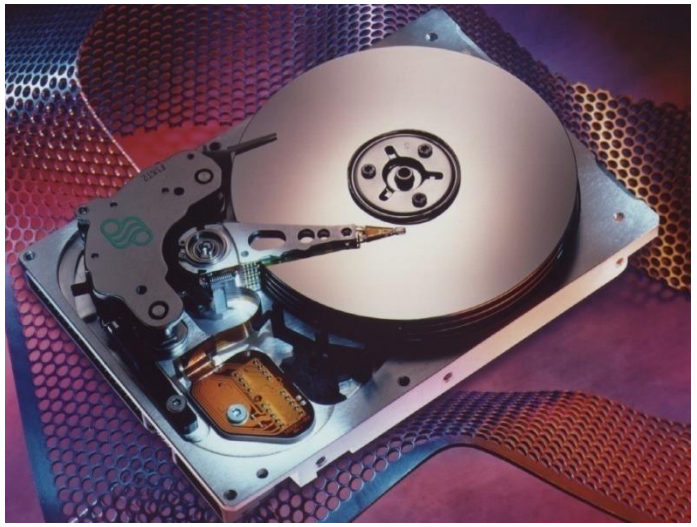
Magnetic Force Microscopy Image of A Disk Surface







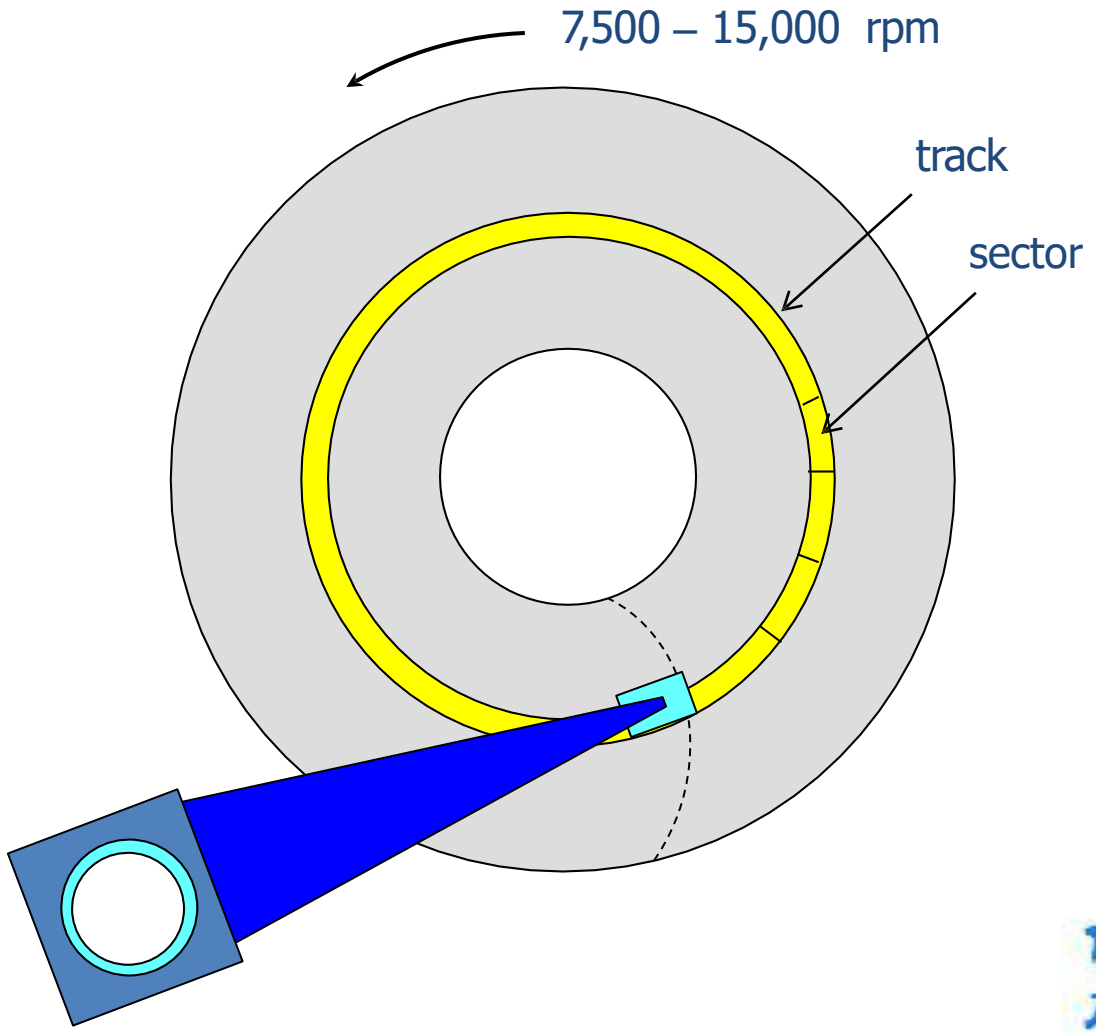
Rotational Latency



Inexpensive: $0.001/1MByte

Rotational Latency

- **Average latency: 3 – 6 ms**
- **Wait until desired sector passes under head**
- **Worst case: a complete rotation 7,500 rpm = 8 ms 15,000 rpm = 4 ms**



Disk Technologies

Disk access time = seek time (disk arm) ~ 8ms

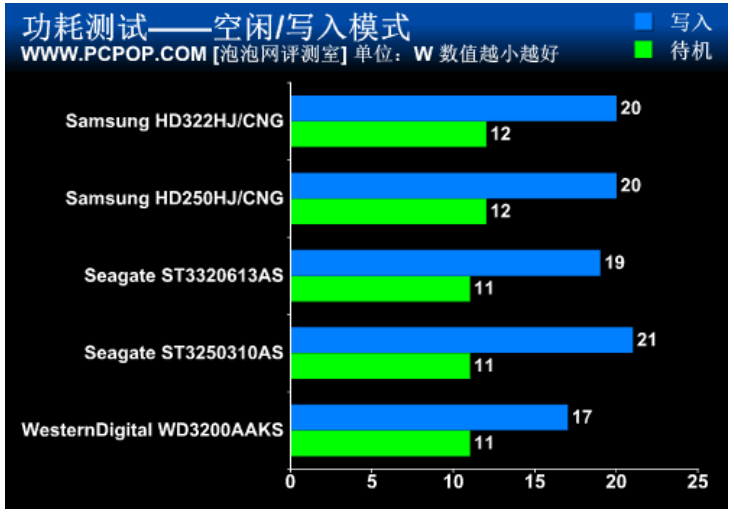
+ rotational delay (disk platter) ~ 3 ms @ 10000 rpm

+ transfer time (media transfer rate) ~ 60 MB/s

- **Today: Processor performance increases roughly 60% per year.**
- **Today: Disk capacity increases 100% per year.**
- **Today: *Disk performance increases only 20% per year!***

- 硬盘：在大规模存储系统中仍占重要地位
  - 耗能在不同状态下有很大差别：
    休眠（0.8W)、空闲(5-11w)、读写(10-21w)
    随机读写，顺序读写，内圈读写，外圈读写：转速不同

- 闪存和固态盘（Flash/SSD)
  - 用于移动和高端存储，能耗为硬盘的十分之一
  - ZNS SSD、FDP、MultiStream…
- 冷存储系统？
  - 硬盘/固态盘/光盘？



1.3 光存储器

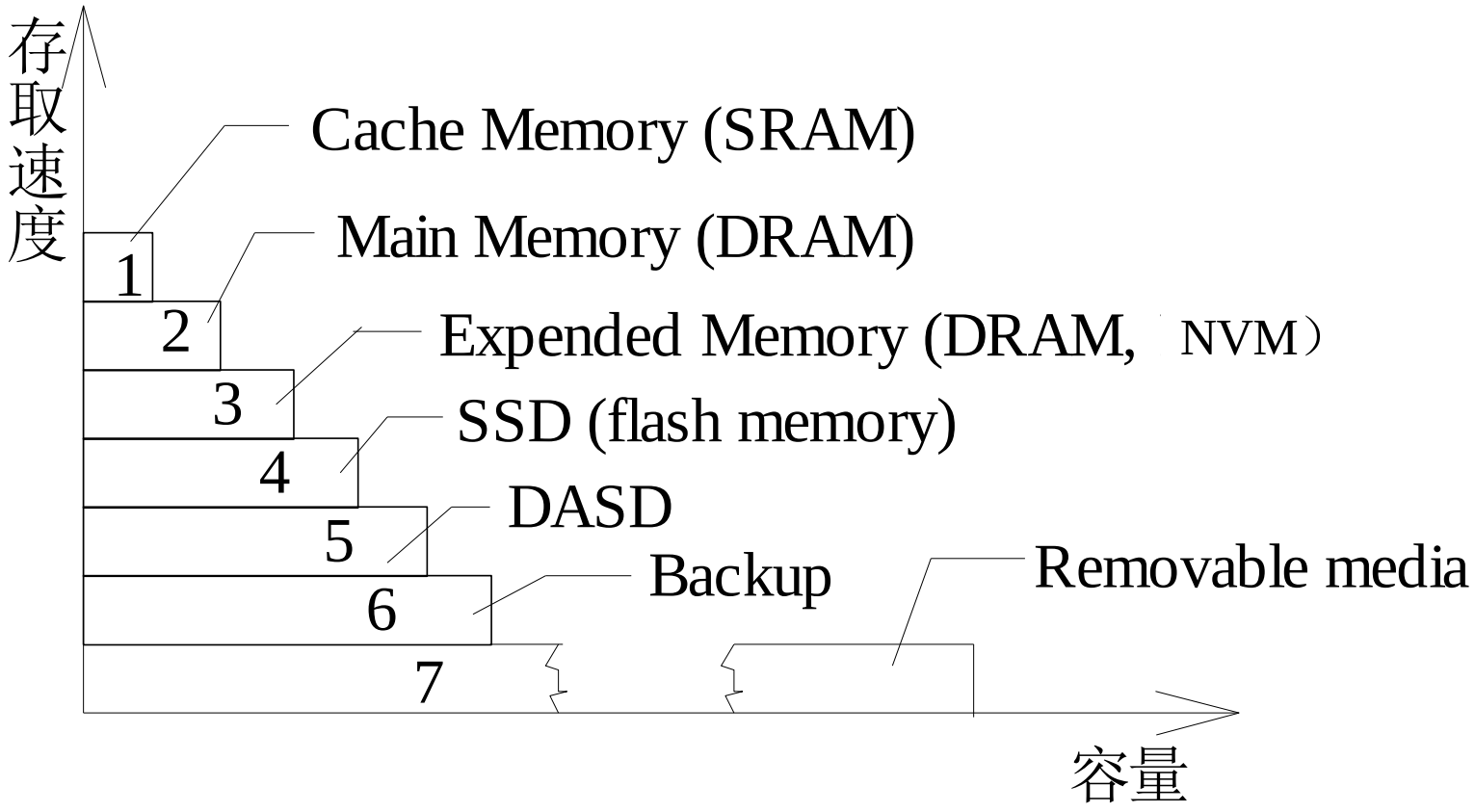
利用微小的激光束照射光记录媒体上，使被照射部位发生热效应或光效应，从而改变媒体的光学（或光磁）性质以记录信息的一类存储设备。读出时，媒体表面的状态转变为反射光强或偏振光的偏转角旋转，还原出记录的信息。

种类：

只读光盘存储器
只写一次读多次光盘存储器
可擦光盘存储器

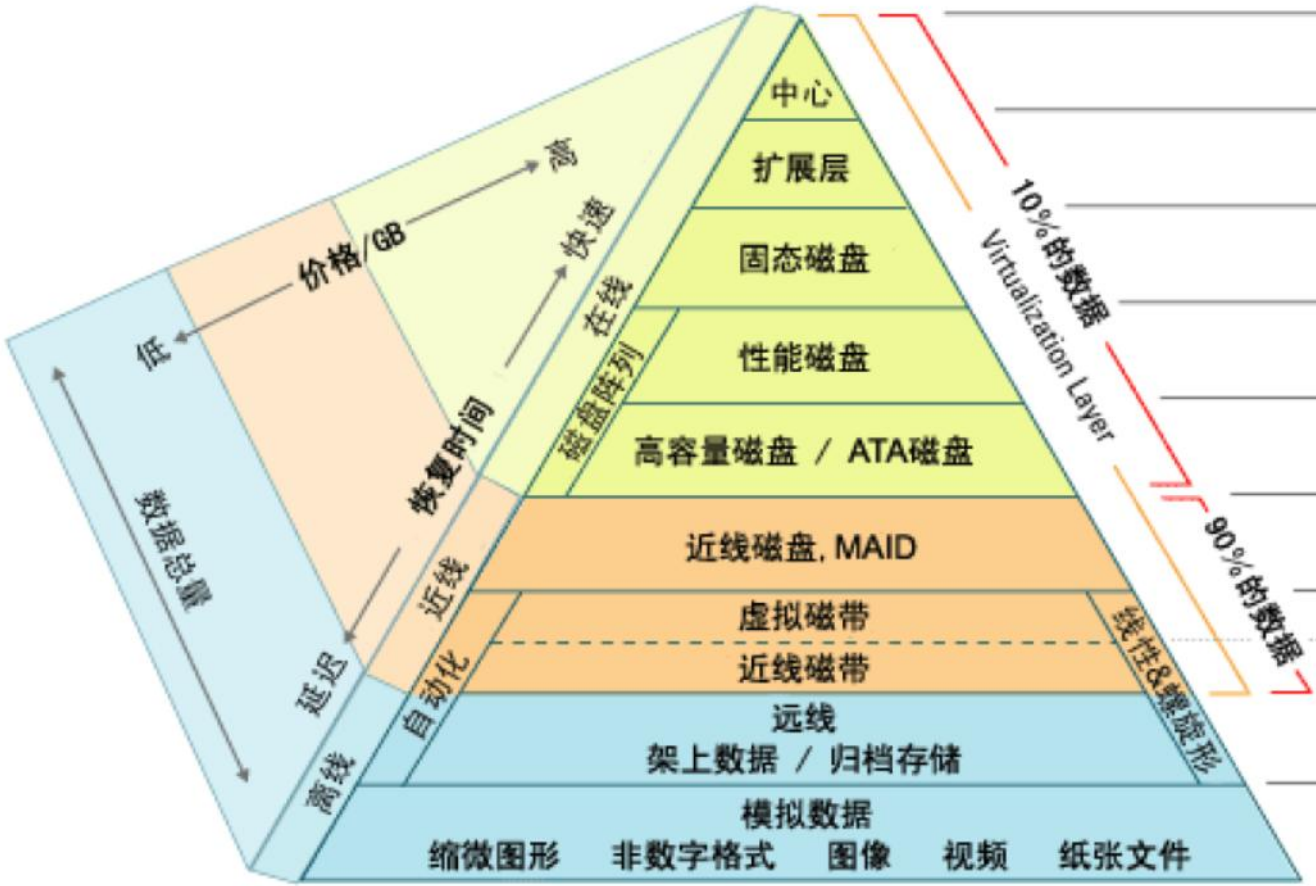
2. 存储器分层结构

存储器设计的三个问题：容量、速度、价格



存储级别

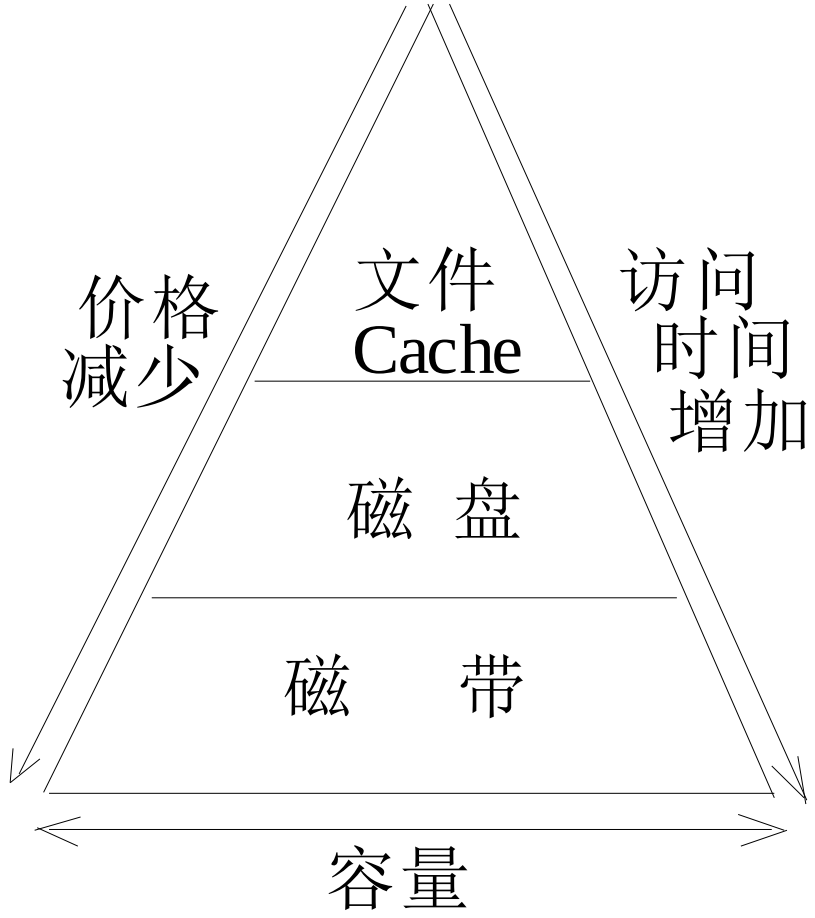
存储的级别



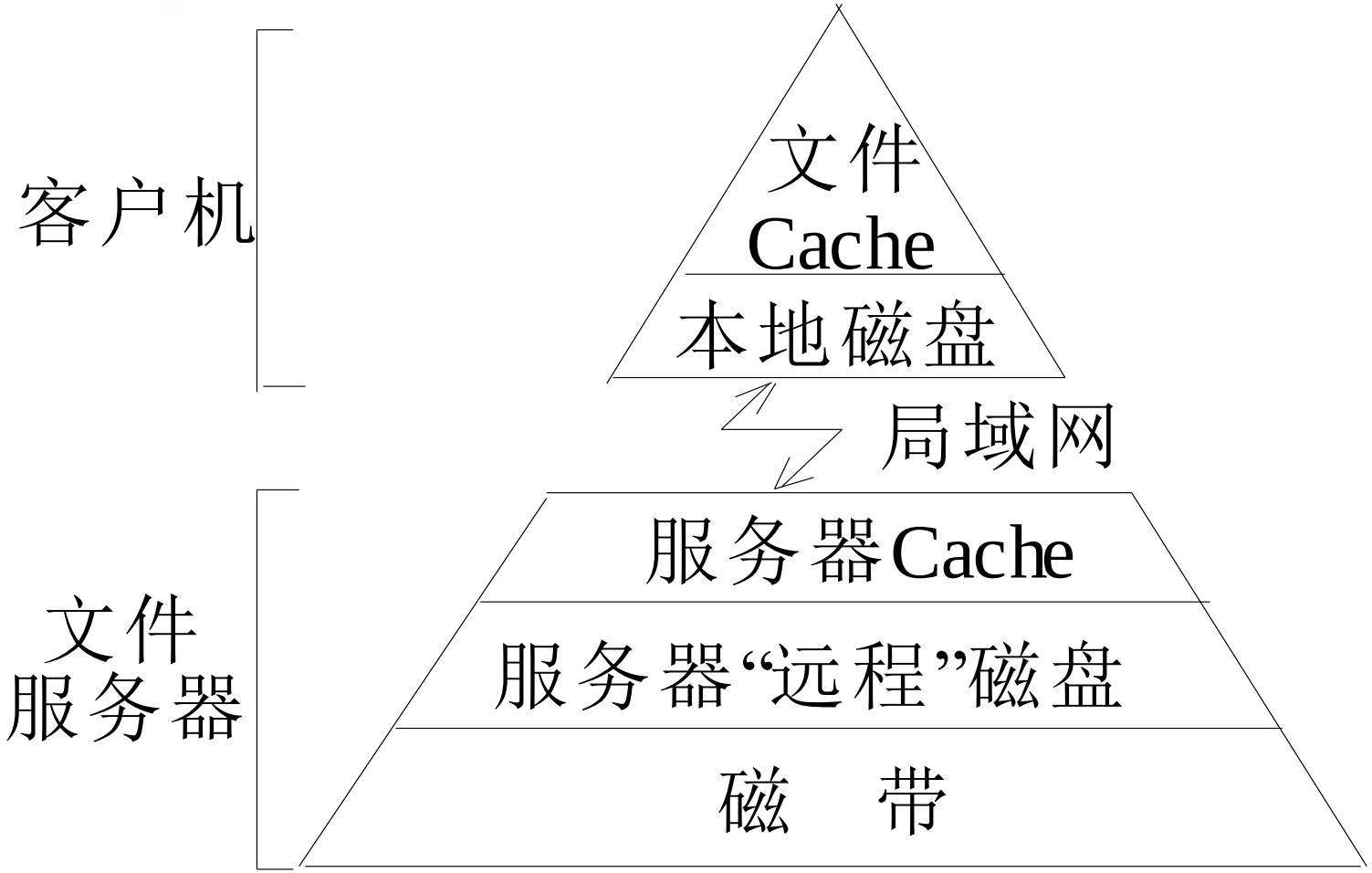
分级参数

	价格/GB	最大容量	访问时间/性能
中心	$100K–300K	256 GB	±200 GB/sec
扩展层	< $75K	nx32 GB	20–40 ns
固态磁盘	$5K–10K	50 GB	< 0.1 ms
性能磁盘	$40–70	GB–TB	4–10 ms
高容量磁盘 / ATA磁盘	$30–40	GB–100 TB	10–25 ms
近线磁盘, MAID	$2–20	TB–PB	< 250 ms
虚拟磁带	$2–10	TB–PB	5–10 sec
近线磁带	$.20–$4*	TB–PB	5–10 sec
远线 架上数据 / 归档存储	< $.001*	TB–PB	min, hr, day
模拟数据 缩微图形 非数字格式 图像 视频 纸张文件	+$20	PB–EB	min, hr, day

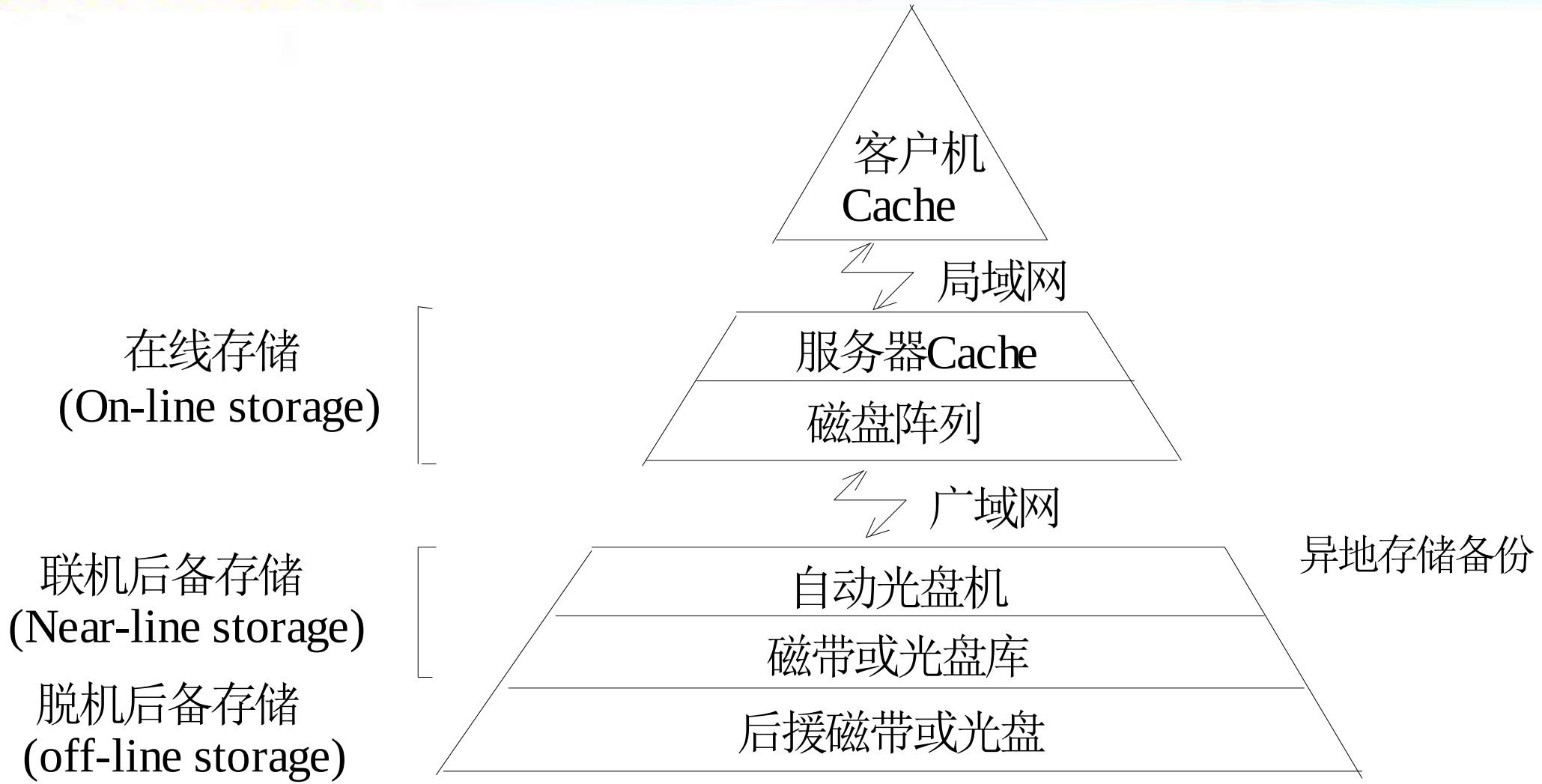
* 基于录制技术



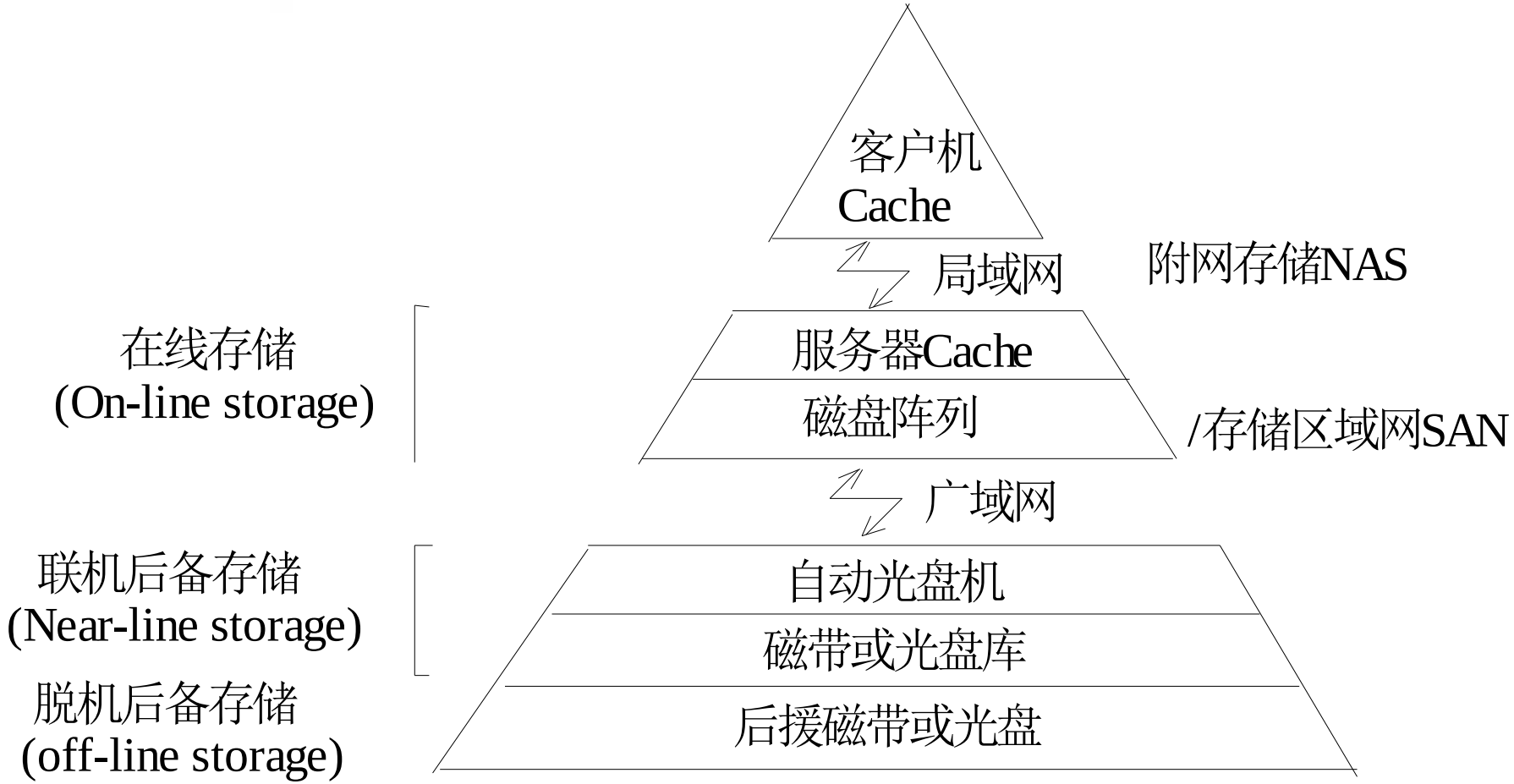
(a) 网络存储系统层次结构(1980年)



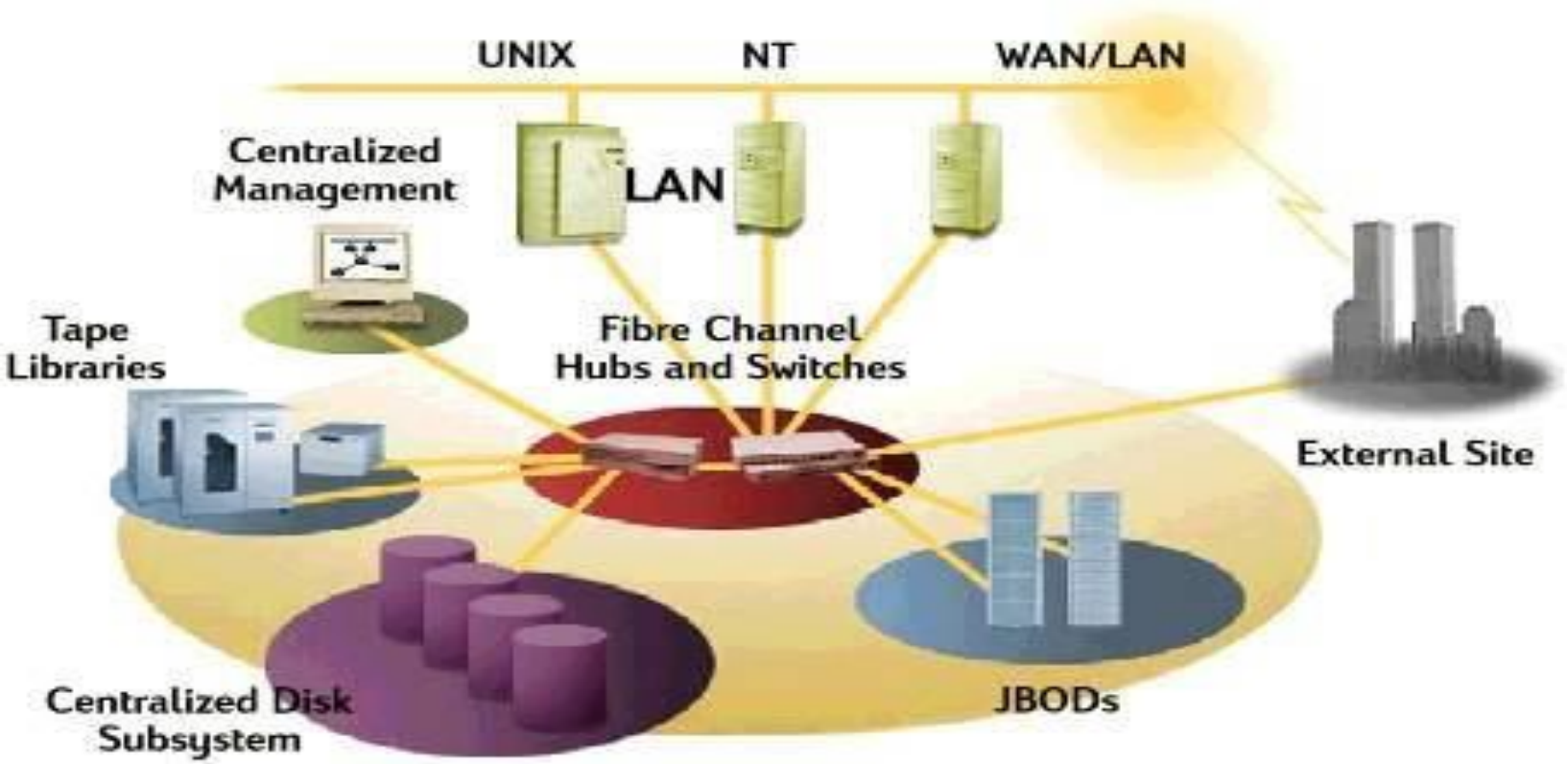
(b)　网络存储系统层次结构(1990年)



(c) 网络存储系统层次结构(1995年)



(d) 网络存储系统层次结构(2000年)


UNIX
NT
WAN/LAN
Centralized Management
LAN
Tape Libraries
Fibre Channel Hubs and Switches
External Site
Centralized Disk Subsystem
JBODs

网络存储