

持久化：磁盘驱动器

邵颖

南京大学

智能科学与技术学院





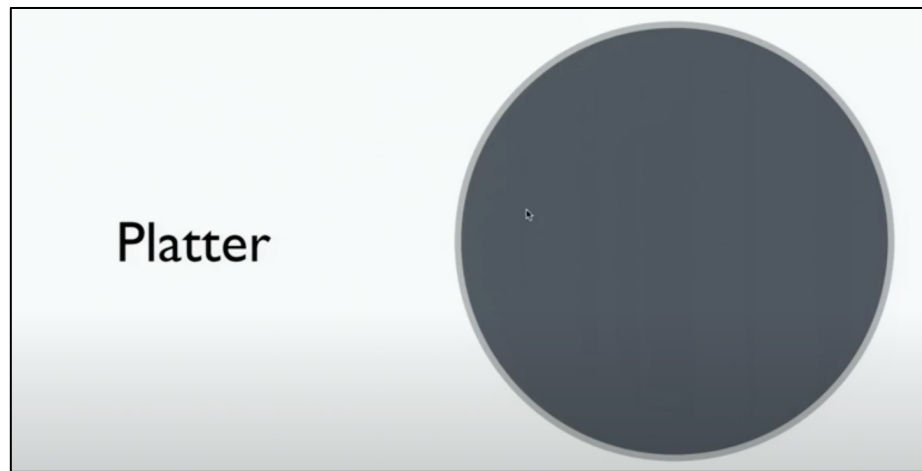
接口

- 磁盘拥有一个以扇区为单位的可寻址空间：
 - 表现为一组扇区（sector）数组
 - 扇区大小通常为 **512 字节**
- 主要操作包括：
 - 对扇区的读写操作

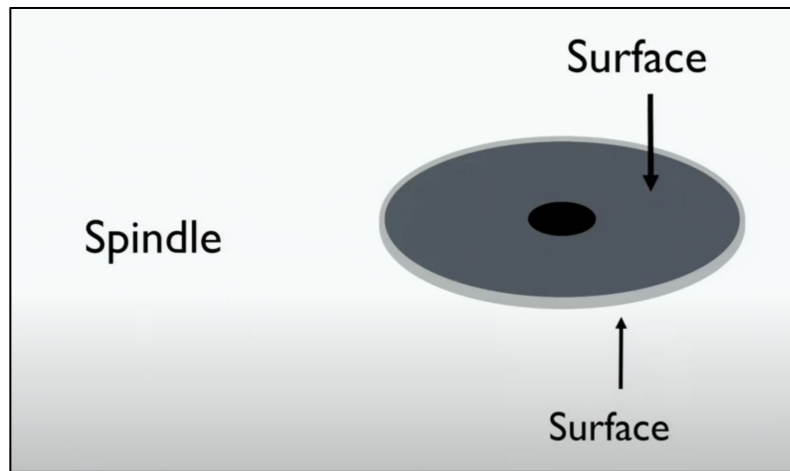




磁盘组件



盘片 (Platter)



盘片表面 (Surface)

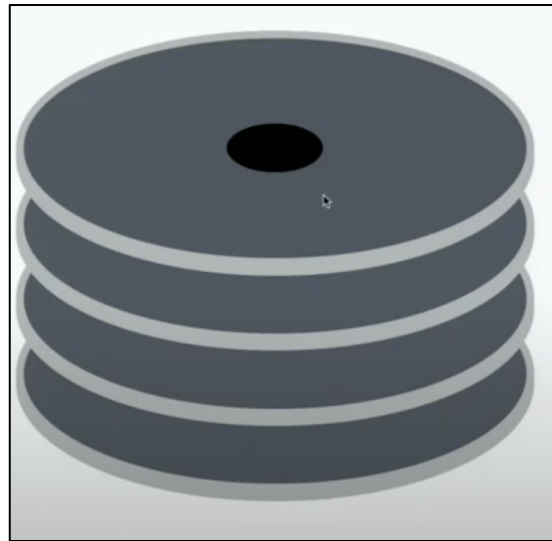
主轴 (Spindle)





磁盘组件

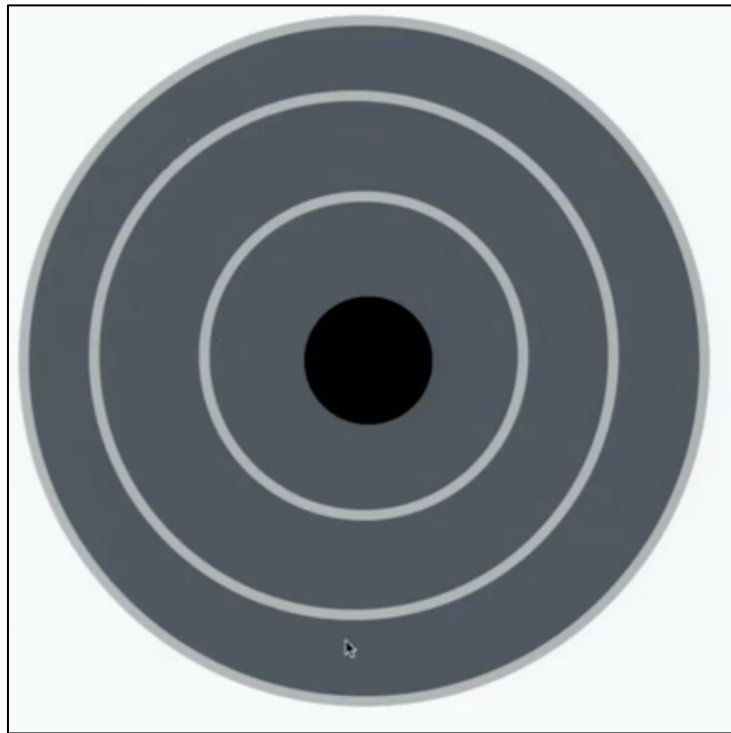
- 多个盘片通常连接在同一主轴上。
- 电机连接主轴，使盘片旋转（spins）。
- 旋转速率以每分钟转数（ Rotations Per Minute， RPM ）表示。
- 例如：转速为 10000 RPM 的磁盘，旋转一圈的时间约为 6 ms。





磁盘组件

- 磁盘表面被划分为多个环形区域，称为磁道（**track**）。
- 多个盘面上半径相同的磁道组合成一个柱面（**cylinder**）。





磁盘组件

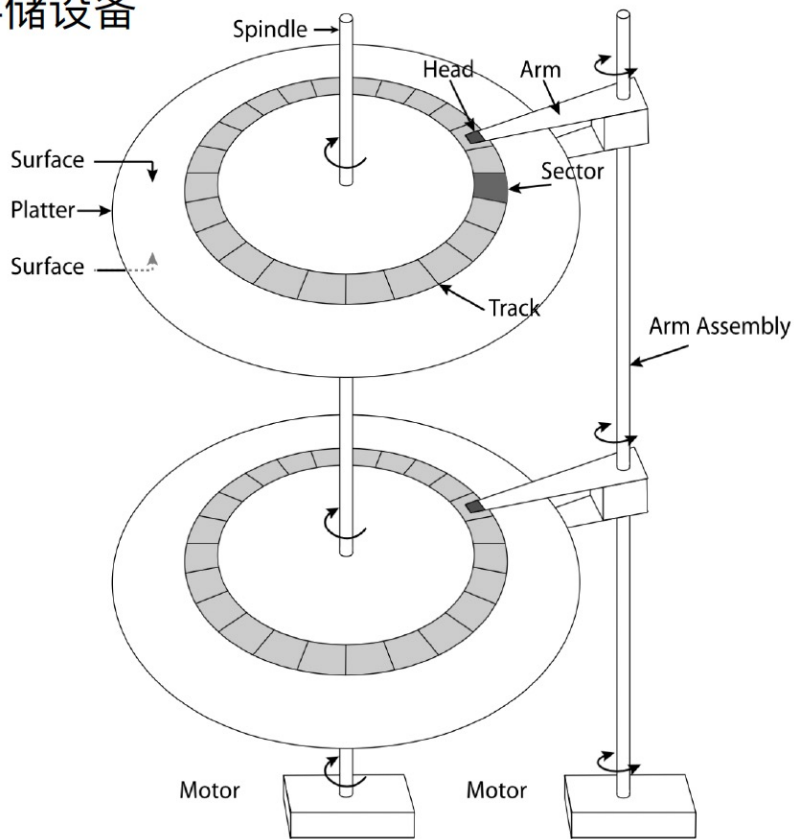
- 磁道（tracks）进一步划分成带编号的扇区（sectors）
 - 操作系统将扇区视作一个线性数组。
 - 扇区的实际物理映射方式取决于具体磁盘硬件细节，操作系统无需了解这些物理细节。
- 移动臂上的磁头可以从每个表面读取数据
 - 这指的是硬盘驱动器，其中读取/写入磁头安装在一个可移动的臂上，能够访问磁盘盘片的每个表面以读取或写入数据。





磁盘组件

一种典型的大容量数据持久化存储设备



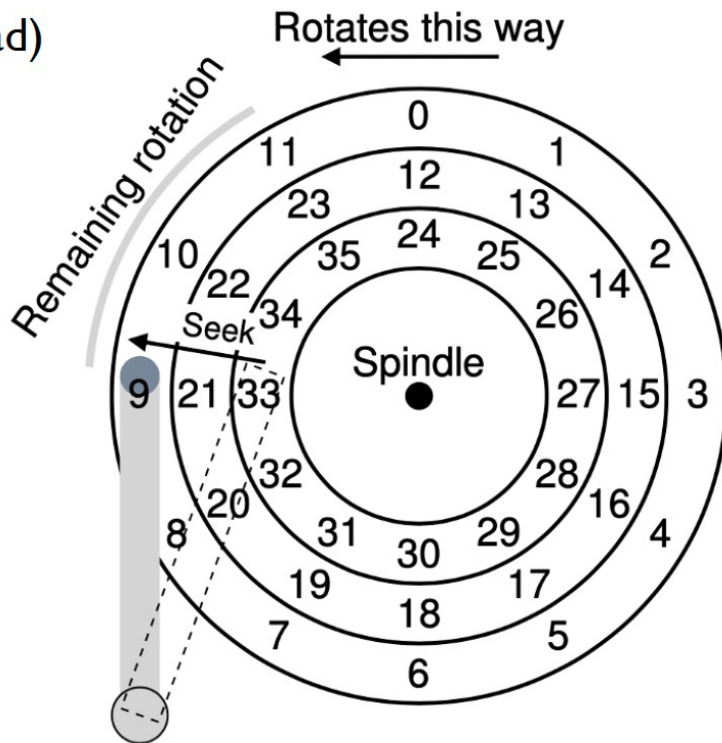


寻道、旋转、传输

数据读写开销 $T_{I/O} = T_{seek} + T_{rotation} + T_{transfer}$

- 寻道时间 (seek time): 将磁头 (head) 移动到指定磁道 (track)
- 旋转延迟 (rotational delay): 等待扇区 (sector) 旋转到磁头下
- 数据传输 (data transfer): 向扇区读写数据

I/O 速率: $R_{I/O} = \frac{\text{大小}_{\text{传输}}}{T_{I/O}}$





磁盘规格：顺序与随机吞吐量对比

表 37.1

磁盘驱动器规格：SCSI 与 SATA

	Cheetah 15K.5	Barracuda
容量	300GB	1TB
RPM	15000	7200
平均寻道时间	4ms	9ms
最大传输速度	125MB/s	105MB/s
磁盘	4	4
缓存	16MB	16/32MB
连接方式	SCSI	SATA

顺序工作负载：每个磁盘的吞吐量是多少？

- Cheetah: 125 MB/s
- Barracuda: 105 MB/s





磁盘规格：顺序与随机吞吐量对比

表 37.1

磁盘驱动器规格：SCSI 与 SATA

	Cheetah 15K.5	Barracuda
容量	300GB	1TB
RPM	15000	7200
平均寻道时间	4ms	9ms
最大传输速度	125MB/s	105MB/s
磁盘	4	4
缓存	16MB	16/32MB
连接方式	SCSI	SATA

- 随机工作负载：每个磁盘的吞吐量是多少？
- 假设每次随机读取的数据大小为16 KB。



磁盘规格（续）

- Cheetah硬盘随机16KB读取吞吐量分析

表 37.1

磁盘驱动器规格：SCSI 与 SATA

	Cheetah 15K.5	Barracuda
容量	300GB	1TB
RPM	15000	7200
平均寻道时间	4ms	9ms
最大传输速度	125MB/s	105MB/s
磁盘	4	4
缓存	16MB	16/32MB
连接方式	SCSI	SATA

- 总耗时 = 寻道时间 + 旋转延迟 + 数据传输时间

- 平均寻道时间 = 4毫秒

- 平均旋转延迟计算

$$\text{平均旋转延迟} = \frac{1}{2} \times \frac{1 \text{ 分钟}}{15000 \text{ 转}} \times \frac{60 \text{ 秒}}{1 \text{ 分钟}} \times \frac{1000 \text{ 毫秒}}{1 \text{ 秒}} = 2 \text{ 毫秒}$$

- 传输16 KB数据所需的时间计算：

$$\text{传输时间} = \frac{16 \text{ KB}}{125 \text{ MB/s}} \times \frac{1,000,000 \text{ 微秒}}{1 \text{ 秒}} = 125 \text{ 微秒}$$





磁盘规格（续）

表 37.1

磁盘驱动器规格：SCSI 与 SATA

	Cheetah 15K.5	Barracuda
容量	300GB	1TB
RPM	15000	7200
平均寻道时间	4ms	9ms
最大传输速度	125MB/s	105MB/s
磁盘	4	4
缓存	16MB	16/32MB
连接方式	SCSI	SATA

• Cheetah磁盘随机16KB读取的吞吐量分析

• 总耗时 = 寻道时间 + 旋转延迟 + 数据传输时间

• Cheetah磁盘的总耗时 = 4 ms（寻道）+ 2 ms（旋转延迟）+ 125 μ s（传输）= **6.1 ms**

• 随机吞吐量计算（单位：MB/s）：

$$\text{吞吐量} = \frac{16 \text{ KB}}{6.1 \text{ ms}} \times \frac{1 \text{ MB}}{1024 \text{ KB}} \times \frac{1000 \text{ ms}}{1 \text{ s}} = 2.5 \text{ MB/s}$$





磁盘规格（续）

- **Barracuda**磁盘随机**16KB**读取的吞吐量分析

表 37.1 磁盘驱动器规格：SCSI 与 SATA		
	Cheetah 15K.5	Barracuda
容量	300GB	1TB
RPM	15000	7200
平均寻道时间	4ms	9ms
最大传输速度	125MB/s	105MB/s
磁盘	4	4
缓存	16MB	16/32MB
连接方式	SCSI	SATA

- 总耗时 = 寻道时间 + 旋转延迟 + 数据传输时间

- 平均寻道时间 = **9毫秒**

- 平均旋转延迟计算：
$$\text{平均旋转延迟} = \frac{1}{2} \times \frac{1 \text{ 分钟}}{7200 \text{ 转}} \times \frac{60 \text{ 秒}}{1 \text{ 分钟}} \times \frac{1000 \text{ 毫秒}}{1 \text{ 秒}} = 4.1 \text{ 毫秒}$$

- 传输**16 KB**数据所需的时间计算：

$$\text{传输时间} = \frac{1 \text{ 秒}}{105 \text{ MB}} \times 16 \text{ KB} \times \frac{1,000,000 \text{ 微秒}}{1 \text{ 秒}} = 149 \text{ 微秒}$$





磁盘规格（续）

表 37.1

磁盘驱动器规格：SCSI 与 SATA

	Cheetah 15K.5	Barracuda
容量	300GB	1TB
RPM	15000	7200
平均寻道时间	4ms	9ms
最大传输速度	125MB/s	105MB/s
磁盘	4	4
缓存	16MB	16/32MB
连接方式	SCSI	SATA

工作负载类型	Cheetah	Barracuda
顺序访问	125 MB/s	105 MB/s
随机访问	2.5 MB/s	1.2 MB/s





磁盘调度

给定一组磁盘访问请求，决定以什么样的顺序进行响应 (disk schedule)

- 由于磁盘 I/O 成本很高，操作系统曾经在磁盘调度上发挥过重要作用
- 通过优化请求调度顺序来最小化磁头移动 (最大化磁盘 I/O 吞吐量)
 - 先到先服务 (First Come First Service, FCFS)
 - 最短寻道时间优先 (Shortest-Seek-Time-First, SSTF)
 - 电梯算法 (SCAN)
 - 最短定位时间优先 (Shortest Positioning Time First, SPTF)





先到先服务：FCFS

按磁盘访问请求的到达顺序进行调度

- 假设磁头 (arm head) 当前位于编号为 11 的磁道 (track)，后续对磁道访问请求的到达顺序为 1, 36, 16, 34, 9, 12
- 在 FCFS 下磁头总共需要移动 $10 + 35 + 20 + 18 + 25 + 3 = 111$ 个磁道

简单易实现，但性能不理想

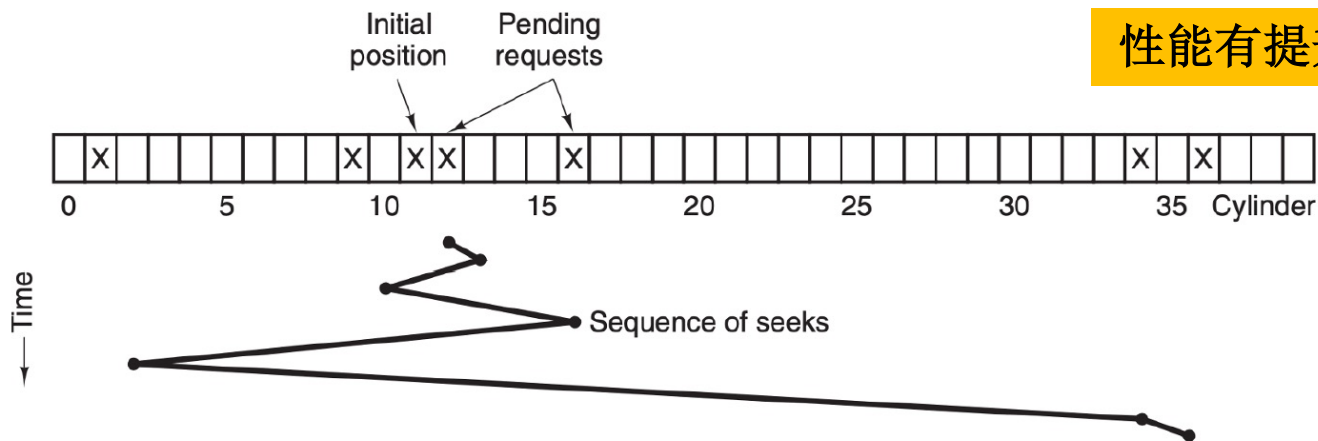




最短寻道时间优先：SSTF

优先调度和当前磁道距离最近的访问请求 (最小化寻道时间)

- 其实就是 Shortest Job First 的思想
- 在 SSTF 下磁头总共需要移动 $1 + 3 + 7 + 15 + 33 + 2 = 61$ 个磁道



性能有提升，但产生饥饿问题

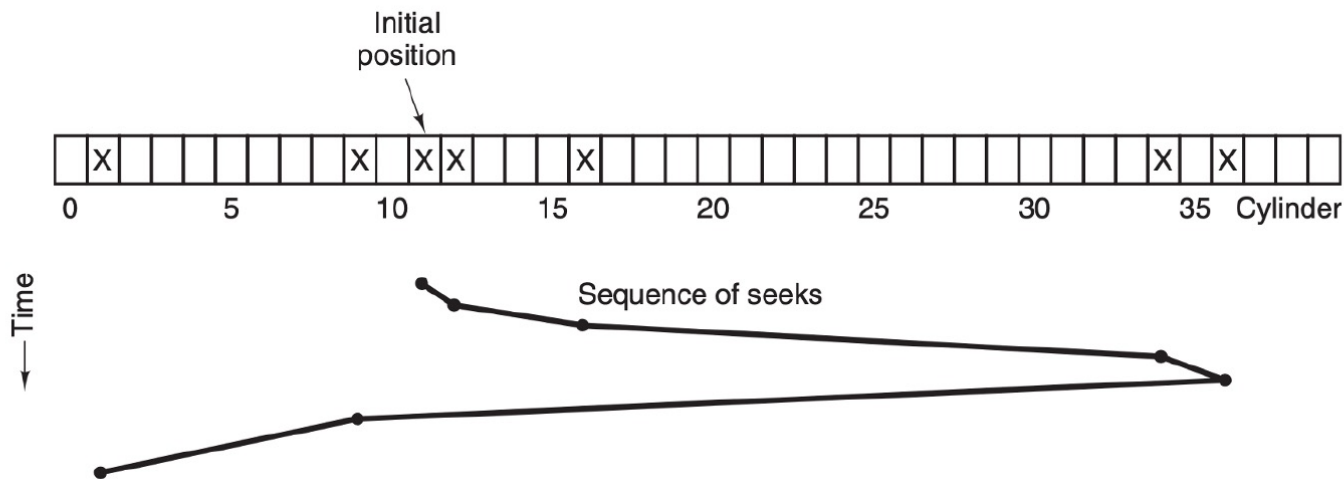




电梯算法：SCAN，解决饥饿问题

让磁头在磁道上来回移动，并依次处理遇到的访问请求

- 在 SCAN 下磁头总共需要移动 $1 + 4 + 18 + 2 + 27 + 8 = 60$ 个磁道
- 可以在此基础上实现很多变种方法 (例如，在每次扫描时是否冻结待处理请求、是否只从一个方向扫描)

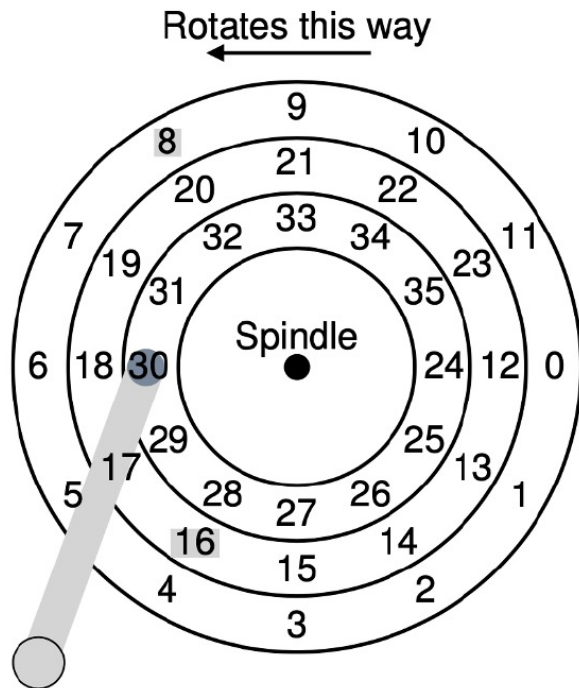




最短定位时间优先：SPTF

同时考虑寻道时间 (seek time) 和旋转延迟 (rotational delay) 的开销

- 假设当前磁头位于 30 扇区 (sector), 后续请求为 16 和 8 扇区
 - 如果 seek time 开销明显高于 rotational delay → sector 16
 - 如果 seek time 开销仅略高于 rotational delay → sector 8
- 需要精确的知道磁盘的底层结构信息 (扇区布局、磁头移动速度等)





磁盘调度总结

- 现代操作系统通常把一系列磁盘访问请求直接发送给磁盘，由磁盘控制器负责调度
- 现代磁盘同时还拥有较大的缓存
 - 在读取一个扇区 (sector) 时将整个磁道 (track) 都缓存起来
 - 和缓存内存中数据不同，这里缓存的是没有被显示读取的数据
- 在写一个扇区时，在数据写入缓冲区后就响应 (write back) 或等到数据确实写入扇区后才响应 (write through)



小结

- 介绍了磁盘组件：盘片、主轴、磁道、柱面、扇区、磁头等
- 介绍了数据读写开销的三个关键部分：寻道时间、旋转延迟、数据传输
- 介绍了顺序和随机两种工作负载
- 介绍了4种磁盘调度方法：
 - 先到先服务（First Come First Service, FCFS）
 - 最短寻道时间优先（Shortest-Seek-Time-First, SSTF）
 - 电梯算法（SCAN）
 - 最短定位时间优先（Shortest Positioning Time First, SPTF）

