Storage Systems

Disk, RAID, Dependability

1960s – 1980s
Computing Revolution

1990 – Information Age

Communication

Computation 

Storage

Computation

Communication

Storage

Storage

Storage

requires **higher standard of dependability**
than the rest of the computer

Storage ?

requires **higher standard of dependability**
than the rest of the computer

program crash

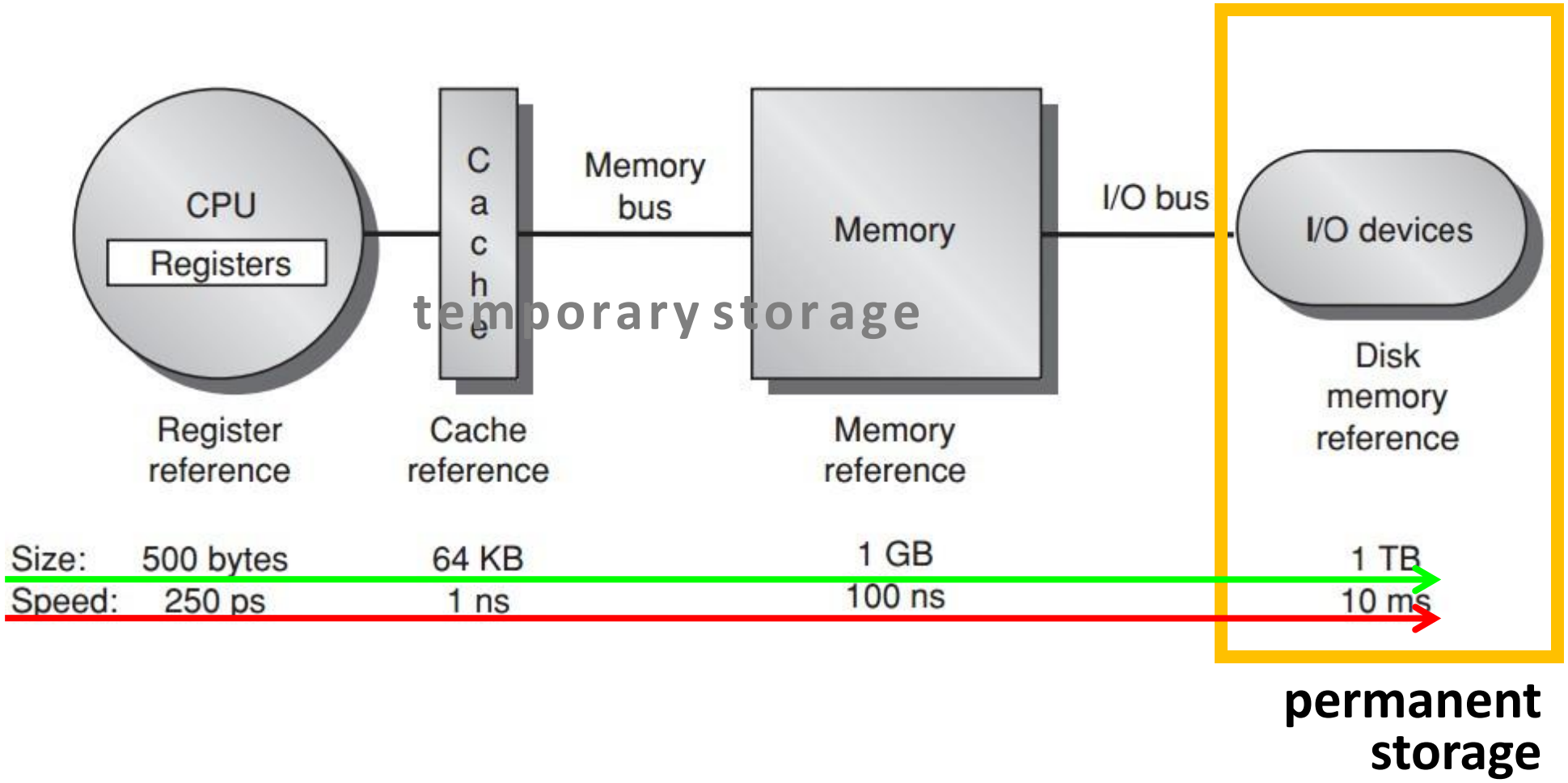
Storage

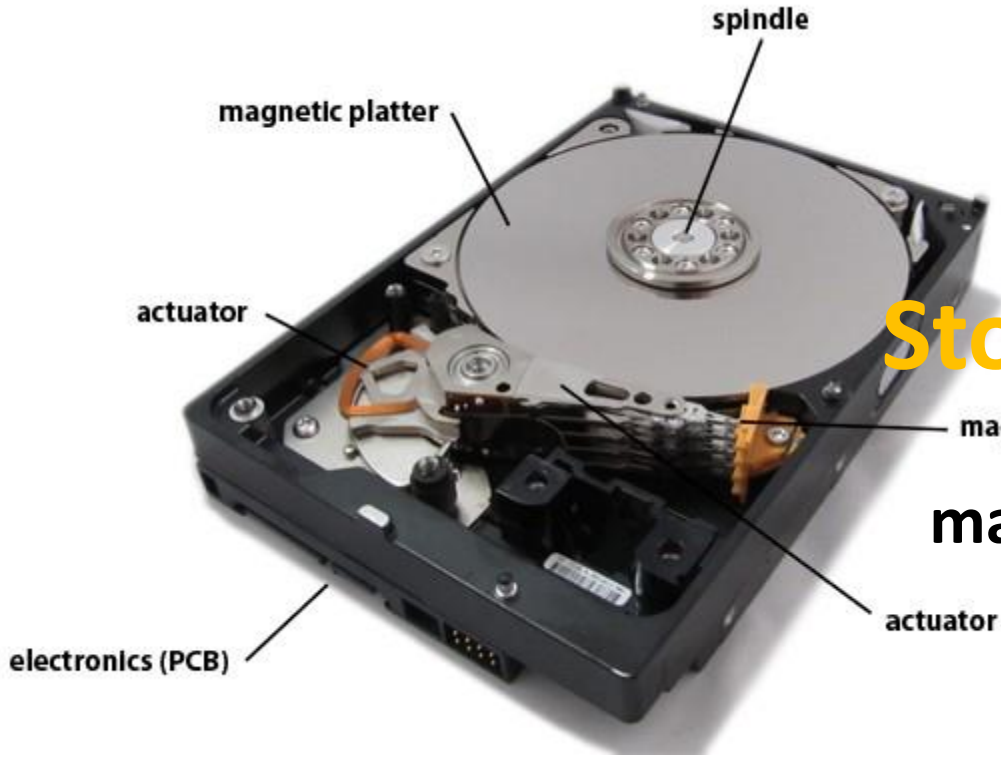
requires **higher standard of dependability**
than the rest of the computer

data loss ← **Storage**

requires **higher standard of dependability**
than the rest of the computer

Memory Hierarchy





Storage

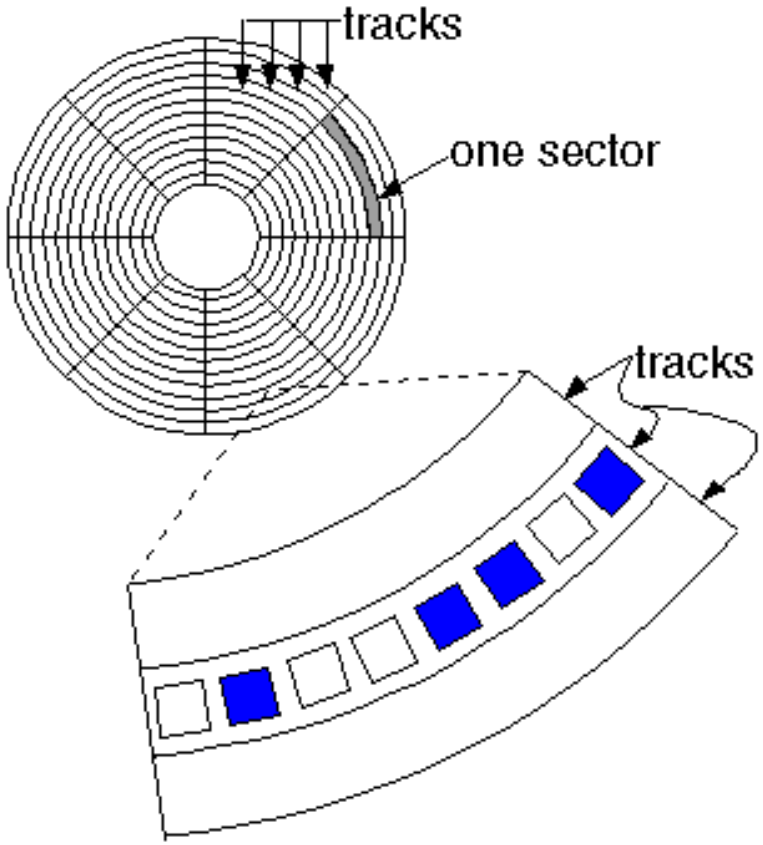
magnetic disks dominate

Preview

- Disk
- Disk Array: RAID
- Dependability: Fault, Error, Failure

let’s start from a single disk

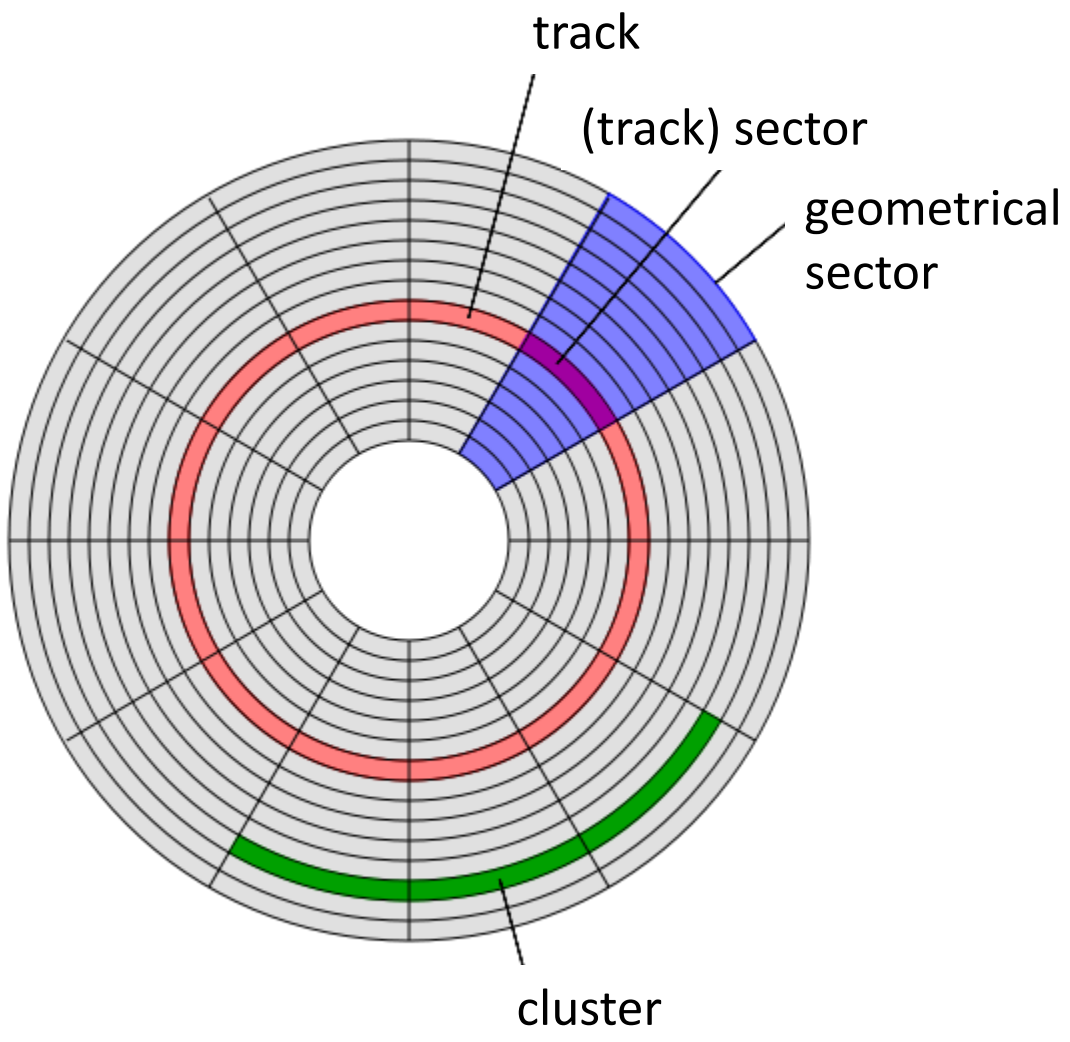
Disk



http://cf.ydcdn.net/1.0.1.19/images/computer/MAGDISK.GIF

Disk

: wiki



https://en.wikipedia.org/wiki/Data_cluster

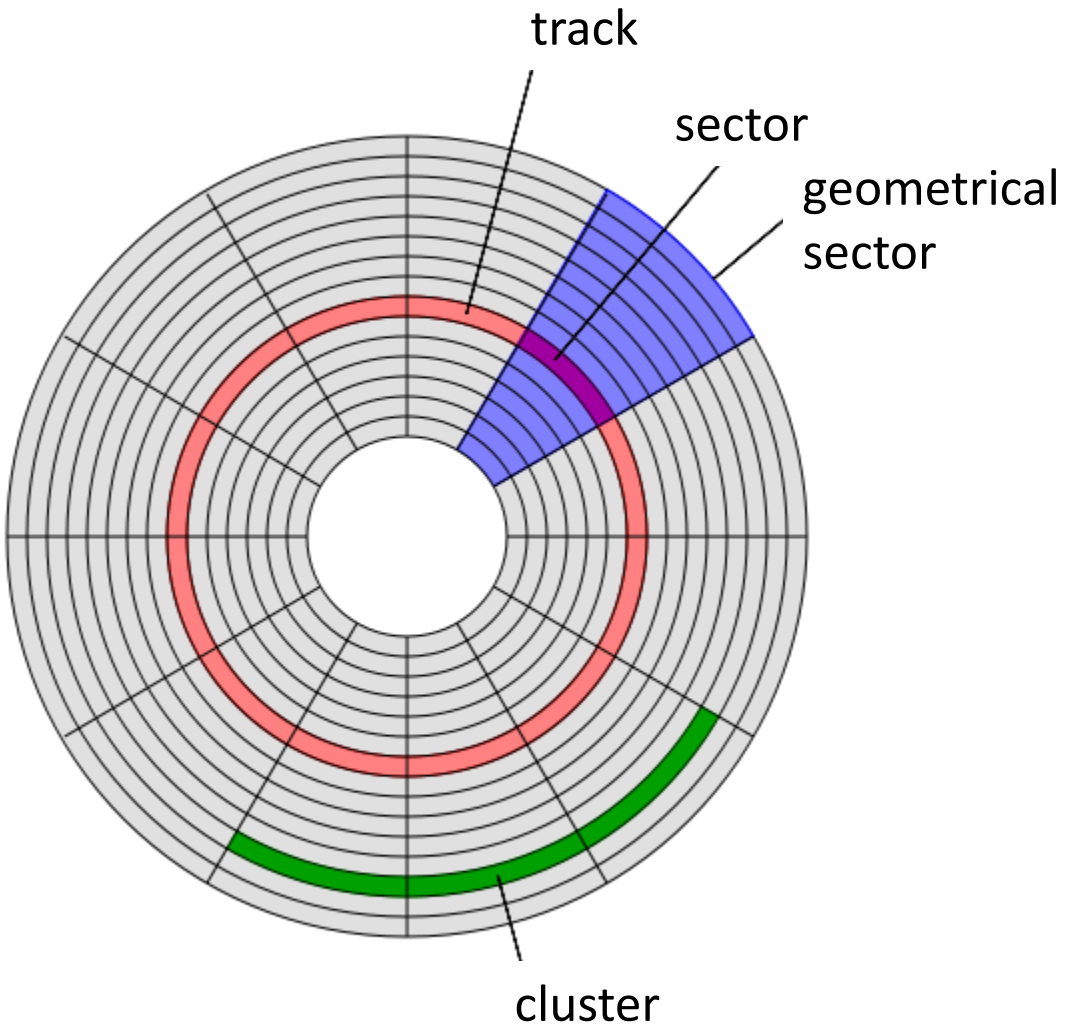
Disk

- Sector

  minimum storage unit

  a block may span multiple sectors

: wiki



https://en.wikipedia.org/wiki/Data_cluster

Disk

: wiki

- Sector

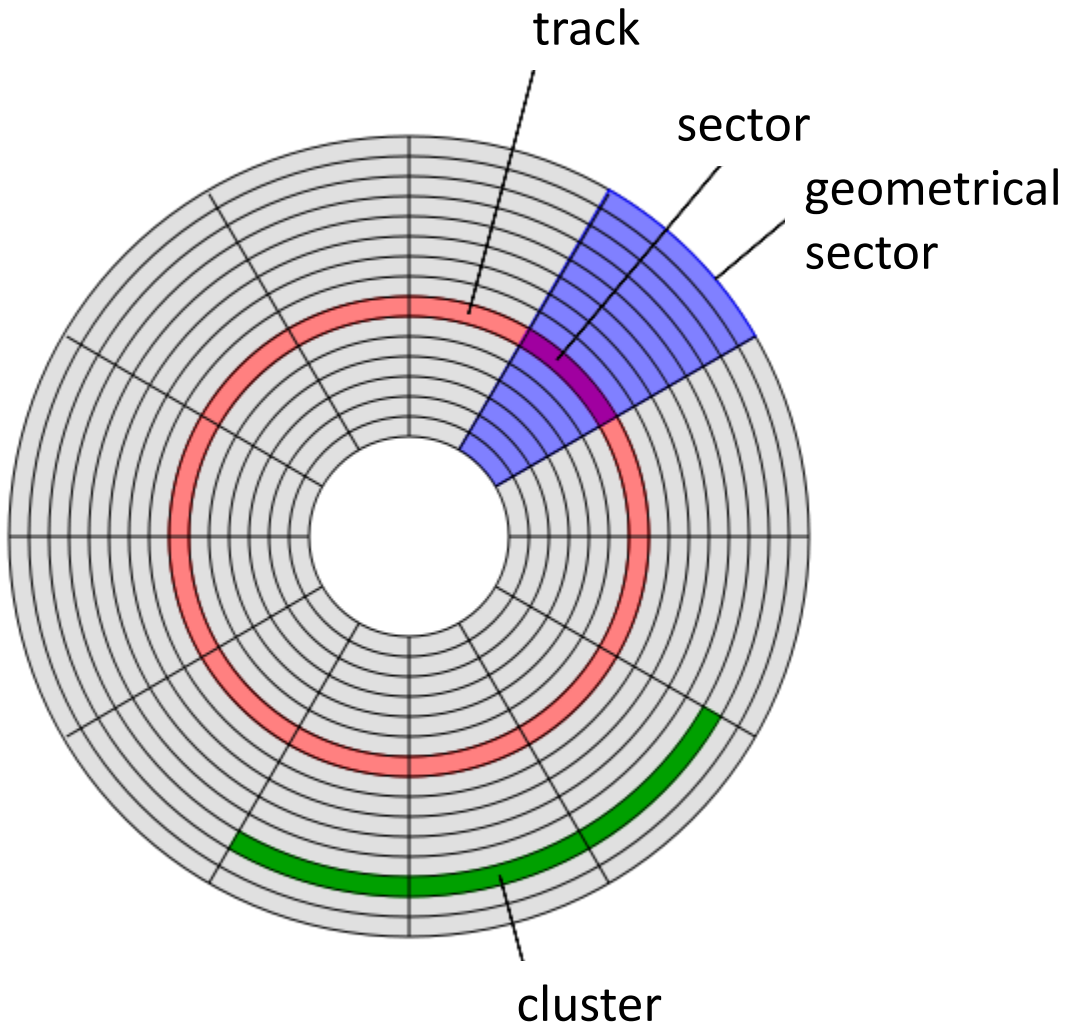
  minimum storage unit

  a block may span multiple sectors

- Cluster

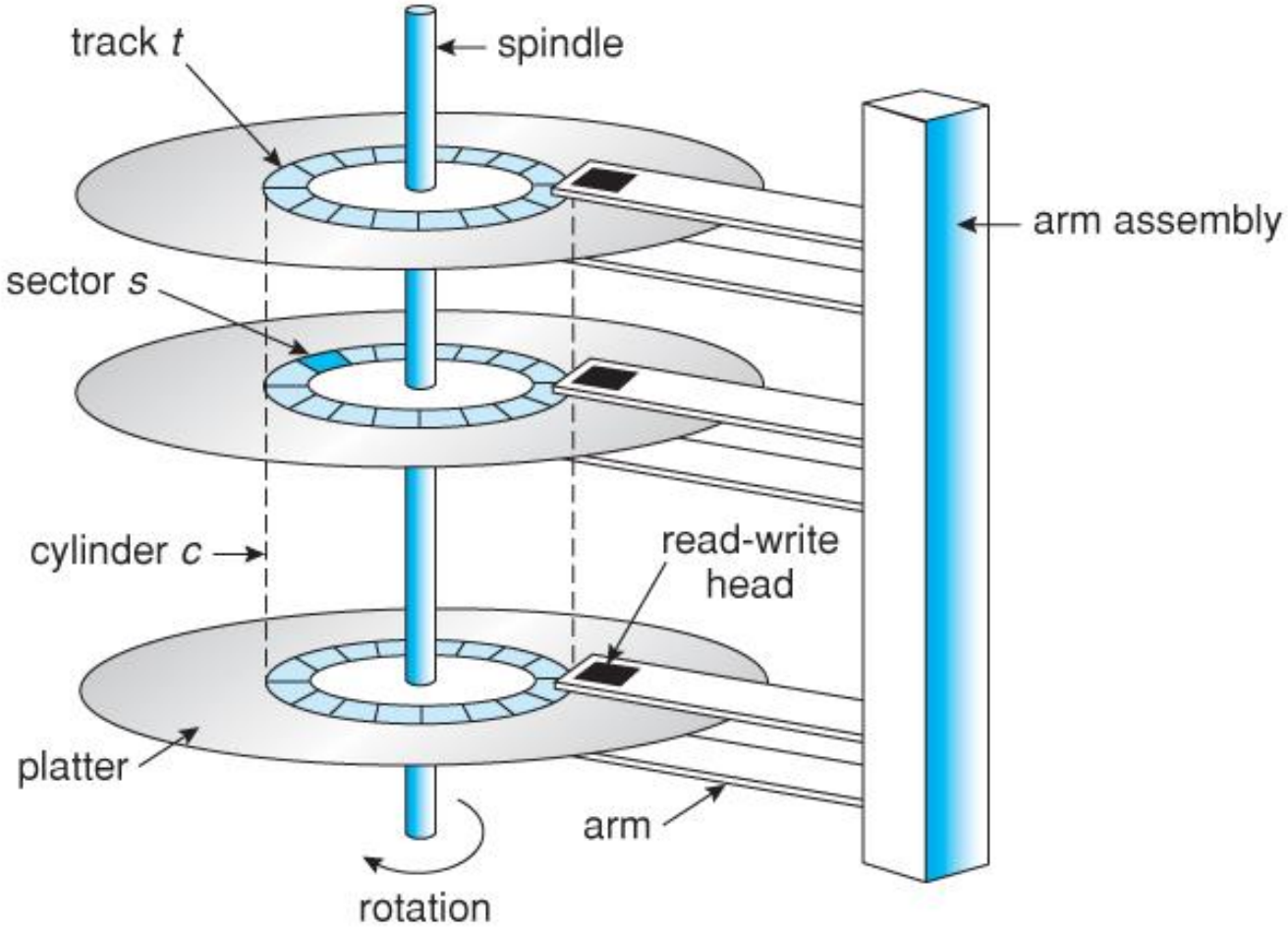
  (dis)contiguous groups of sectors to reduce the overhead of managing on-disk data structures;

  may span more than one track



https://en.wikipedia.org/wiki/Data_cluster

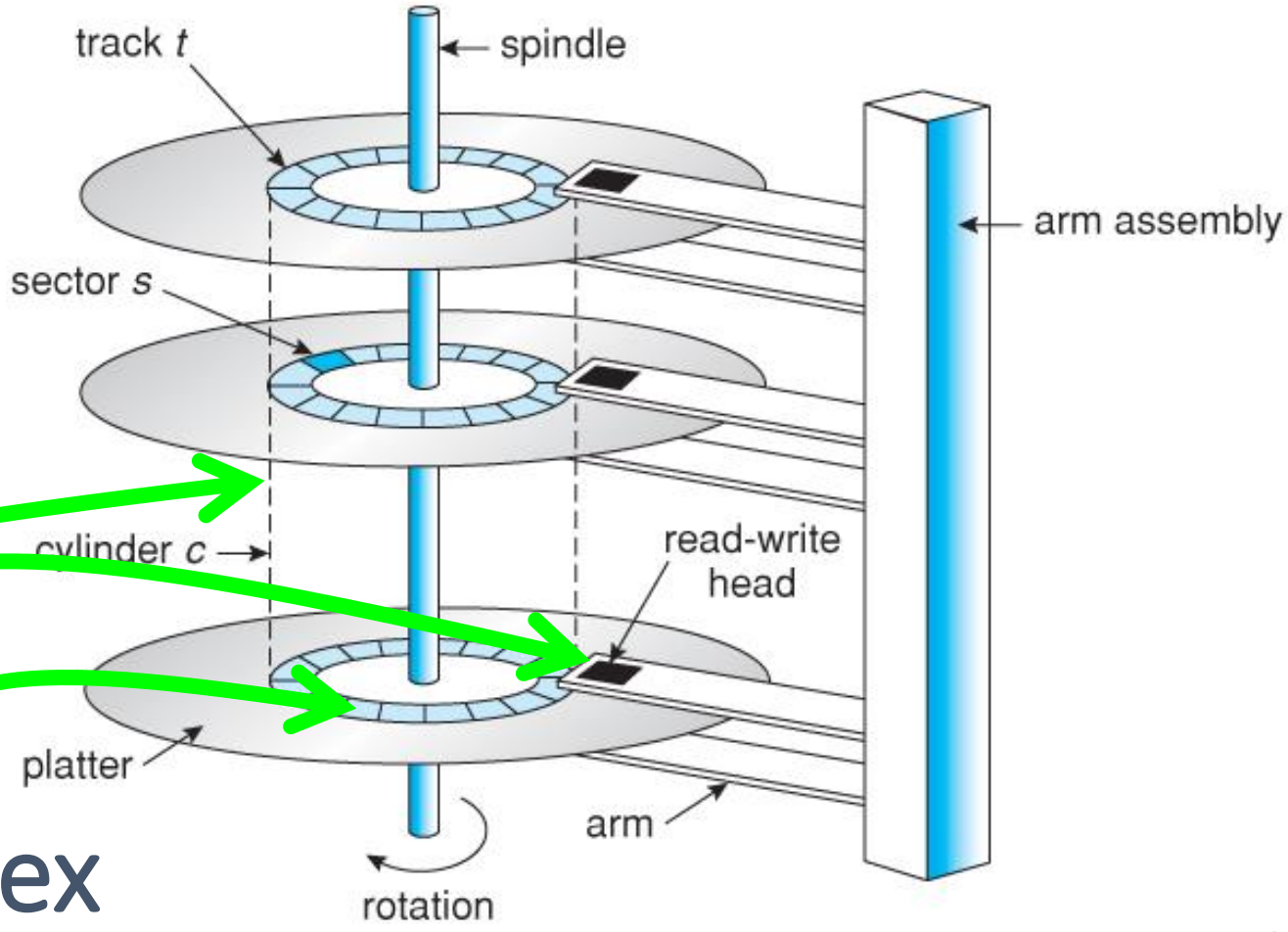
Disk



http://www.cs.uic.edu/~jbell/CourseNotes/OperatingSystems/images/Chapter10/10_01_DiskMechanism.jpg

Disk

: locate data



CHS index

Seek time

Latency

Transmission time

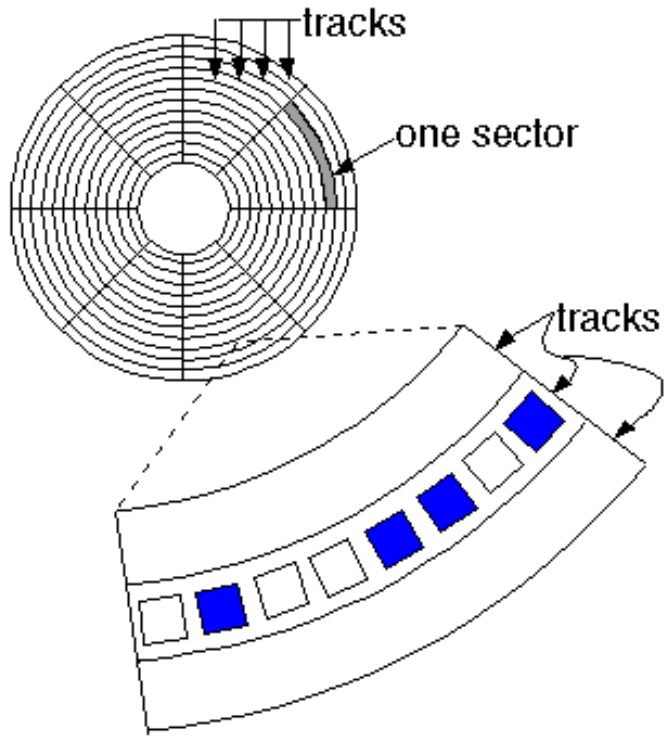
http://www.cs.uic.edu/~jbell/CourseNotes/OperatingSystems/images/Chapter10/10_01_DiskMechanism.jpg

Disk Capacity

- **Areal Density**

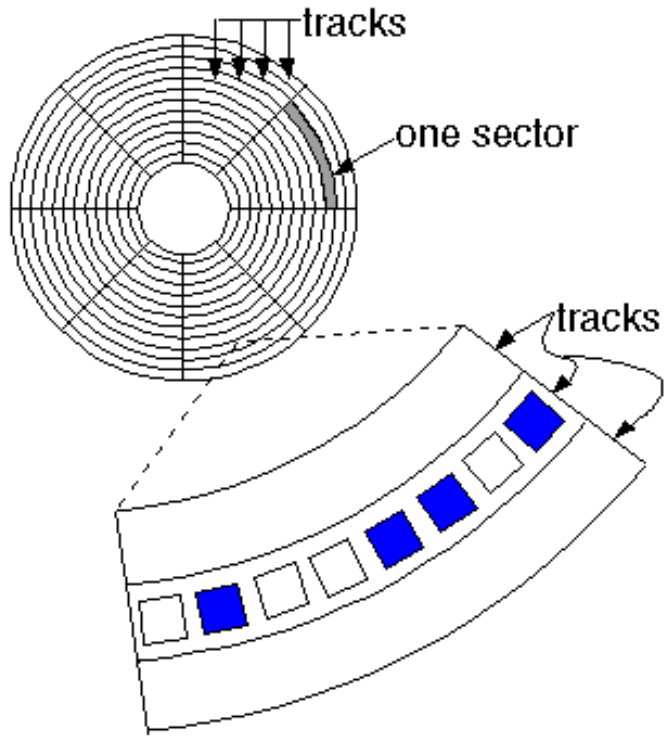
=bits/inch^2

=(tracks/inch) x (bits-per-track/inch)

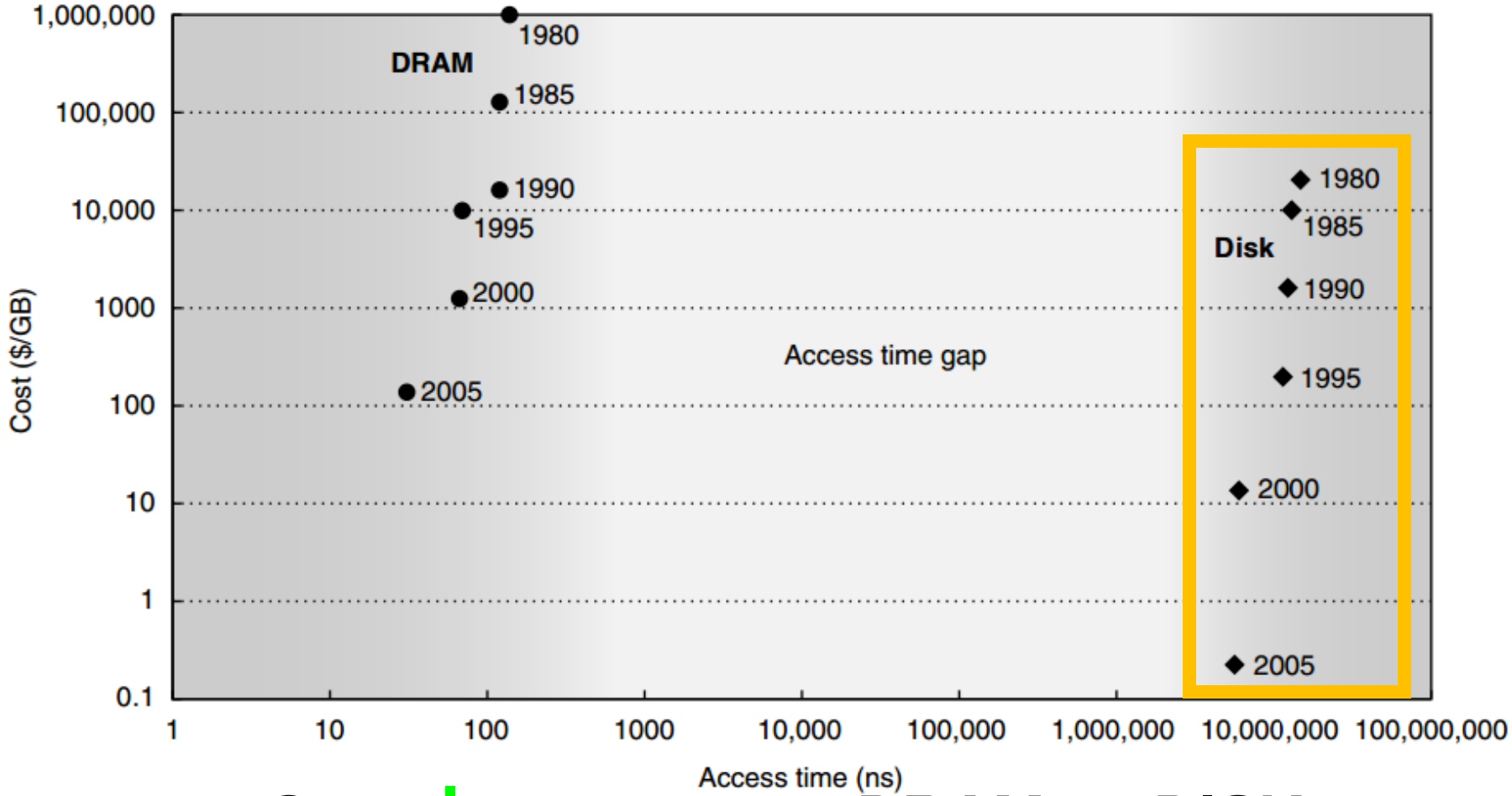


Disk Capacity

- **Areal Density**
  in 2011, the highest density
  400 billion bits/inch^2
- **Costs per gigabyte**
  between 1983 and 2011,
  improved by almost
  a factor of 1,000,000



Disk vs DRAM



Cost ↓ **DRAM >> DISK**

Access time ↓ **DRAM << DISK**

Disk's Competitor

- **Flash Memory**
  non-volatile semiconductor memory;
  same bandwidth as disks;
  100 to 1000 times faster;
  15 to 25 times higher cost/gigabyte;
- **Wear out**
  limited to 1 million writes
- **Popular in cell phones,**
  but not in desktop and server

Disk Power

- **Power** by disk motor

≈Diameter$^{4.6}$ x RPM$^{2.8}$ x No. of platters

RPM: Revolutions Per Minute *rotation speed*

Disk Power

- **Power** by disk motor

≈Diameter$^{4.6}$ x RPM$^{2.8}$ x No. of platters

RPM: Revolutions Per Minute *rotation speed*

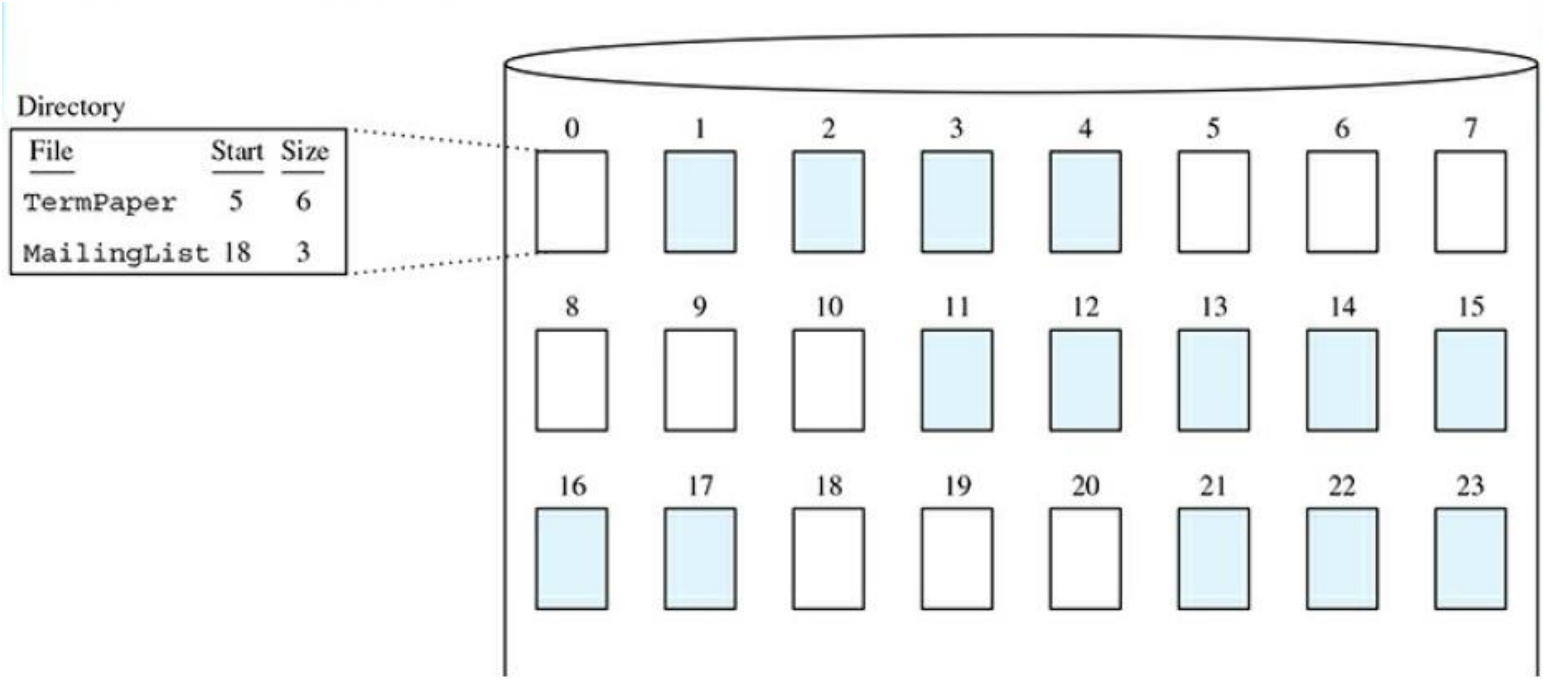
- Smaller patters, slower rotation, and fewer platters reduce disk motor power

Disk Power

disk	Capacity (GB)	Price	Platters	RPM	Diameter (inches)	Average seek (ms)	Power (watts)	I/O/sec	Disk BW (MB/sec)	Buffer BW (MB/sec)	Buffer size (MB)	MTTF (hrs)
SATA	2000	$85	4	5900	3.7	16	12	47	45–95	300	32	0.6M
SAS	600	$400	4	15,000	2.6	3–4	16	285	122–204	750	16	1.6M

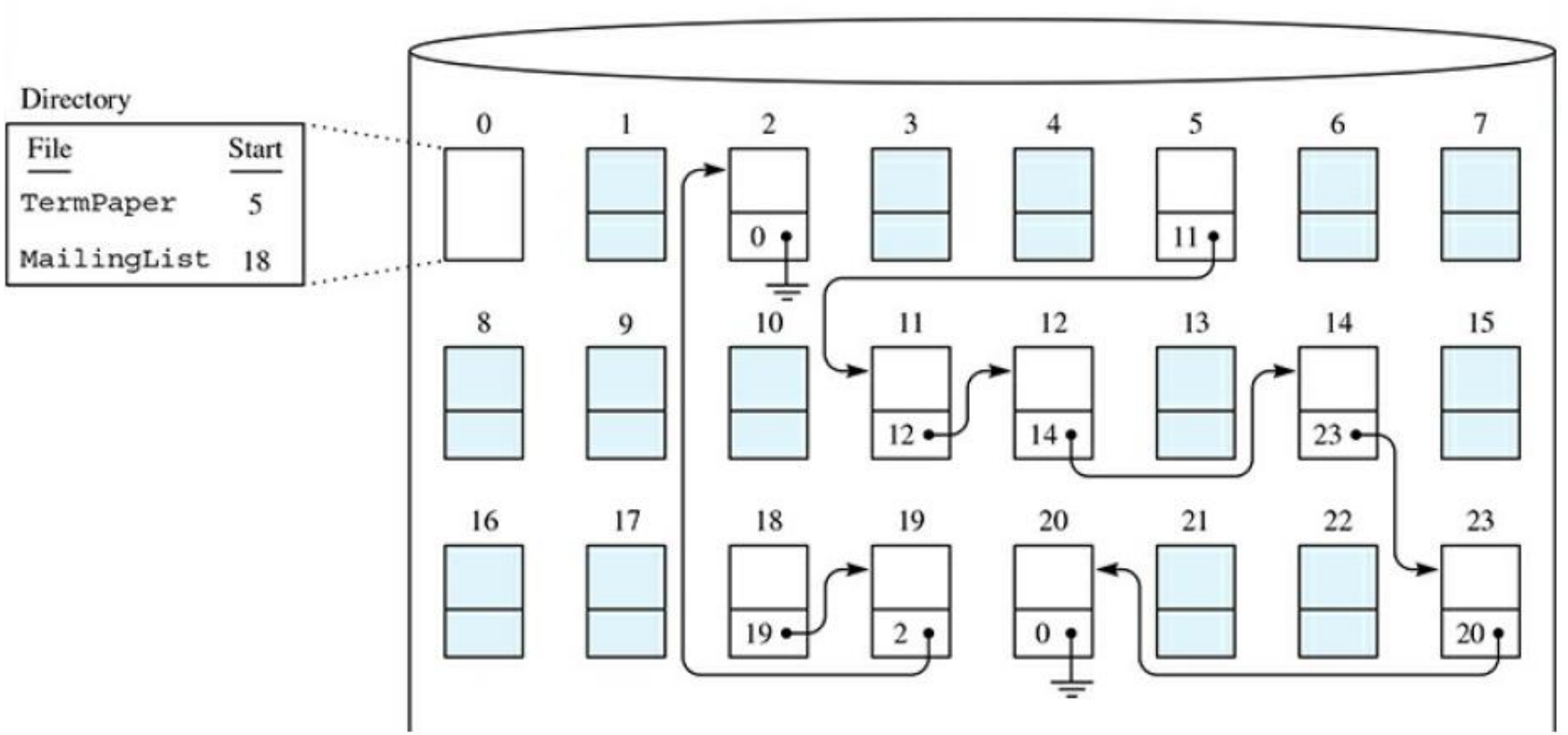
Disk & File Abstract

fscanf (fp, “%d”, &mydata)



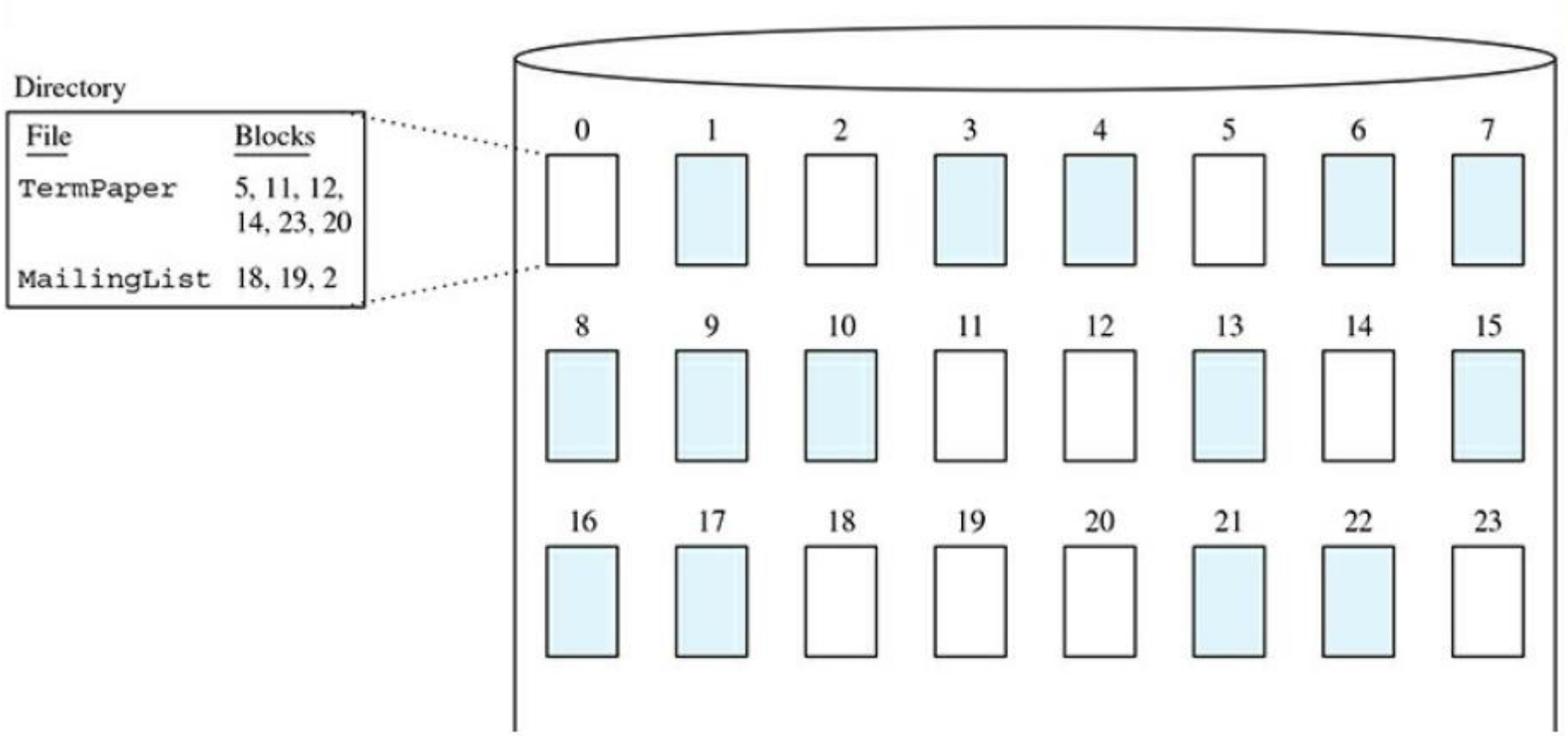
Contiguous allocation on a disk

Disk & File Abstract



Linked allocation on a disk

Disk & File Abstract



Indexed allocation on a disk

what if one is not enough...

what if one is not enough…
disk failure

what if one is not enough…
disk failure
all or nothing 

Disk Arrays

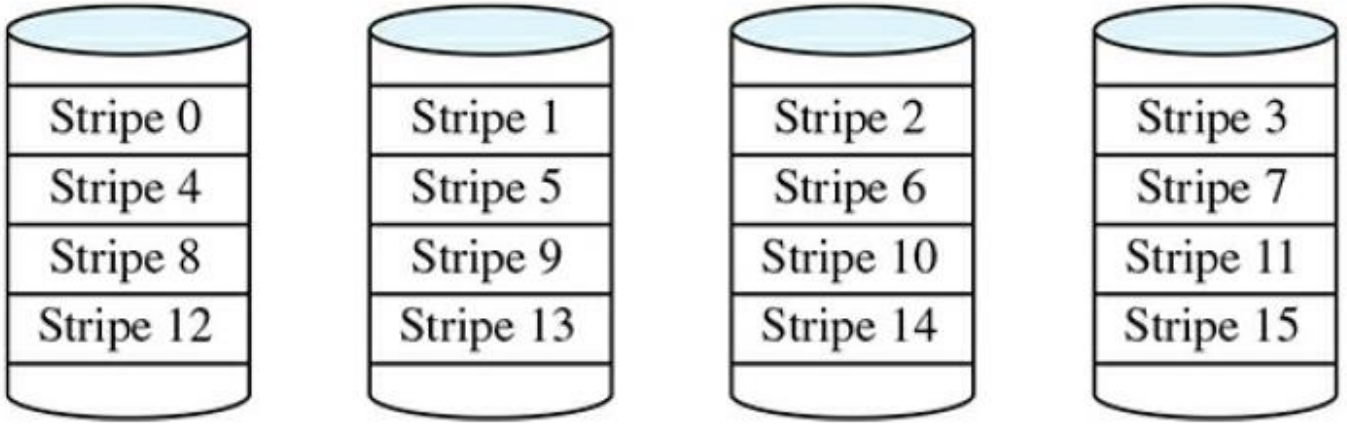
- Disk arrays with **redundant disks** to tolerate faults
- If a single disk fails, the lost information is reconstructed from redundant information
- **Striping**: simply spreading data over multiple disks
- **RAID:** redundant array of inexpensive/independent disks

RAID
Redundant Array of Indepent Disks

RAID level		Disk failures tolerated, check space overhead for 8 data disks	Pros	Cons	Company products
0	Nonredundant striped	0 failures, 0 check disks	No space overhead	No protection	Widely used
1	Mirrored	1 failure, 8 check disks	No parity calculation; fast recovery; small writes faster than higher RAIDs; fast reads	Highest check storage overhead	EMC, HP (Tandem), IBM
2	Memory-style ECC	1 failure, 4 check disks	Doesn't rely on failed disk to self-diagnose	~ Log 2 check storage overhead	Not used
3	Bit-interleaved parity	1 failure, 1 check disk	Low check overhead; high bandwidth for large reads or writes	No support for small, random reads or writes	Storage Concepts
4	Block-interleaved parity	1 failure, 1 check disk	Low check overhead; more bandwidth for small reads	Parity disk is small write bottleneck	Network Appliance
5	Block-interleaved distributed parity	1 failure, 1 check disk	Low check overhead; more bandwidth for small reads and writes	Small writes → 4 disk accesses	Widely used
6	Row-diagonal parity, EVEN-ODD	2 failures, 2 check disks	Protects against 2 disk failures	Small writes → 6 disk accesses; 2× check overhead	Network Appliance

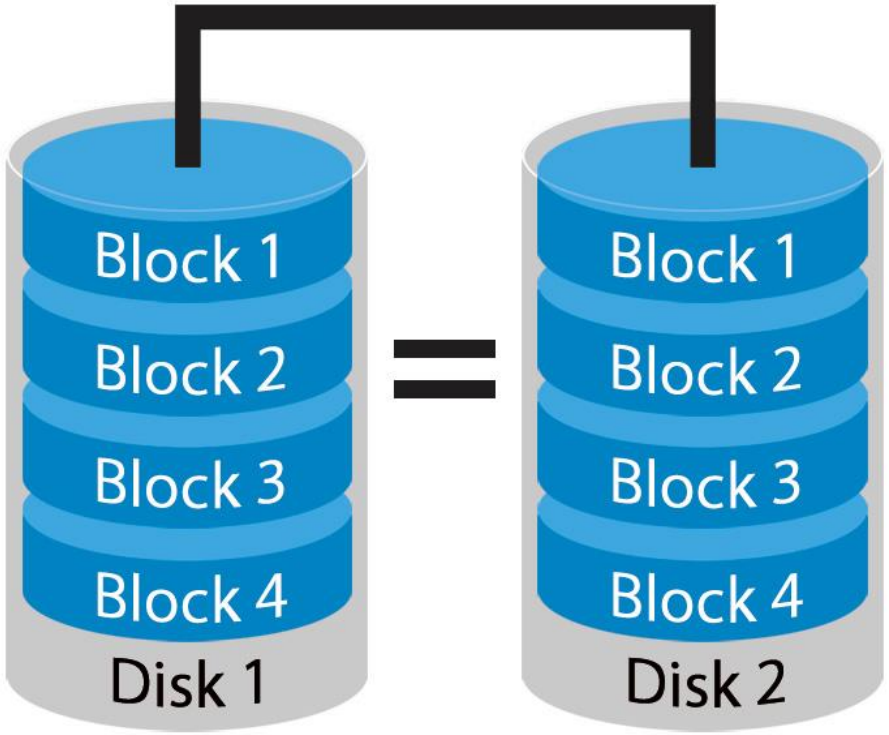
RAID 0: Nonredundant striped

- **JBOD:** just a bunch of disks
- No redundancy
- No failure tolerated
- Measuring stick for other RAID levels: cost, performance, and dependability



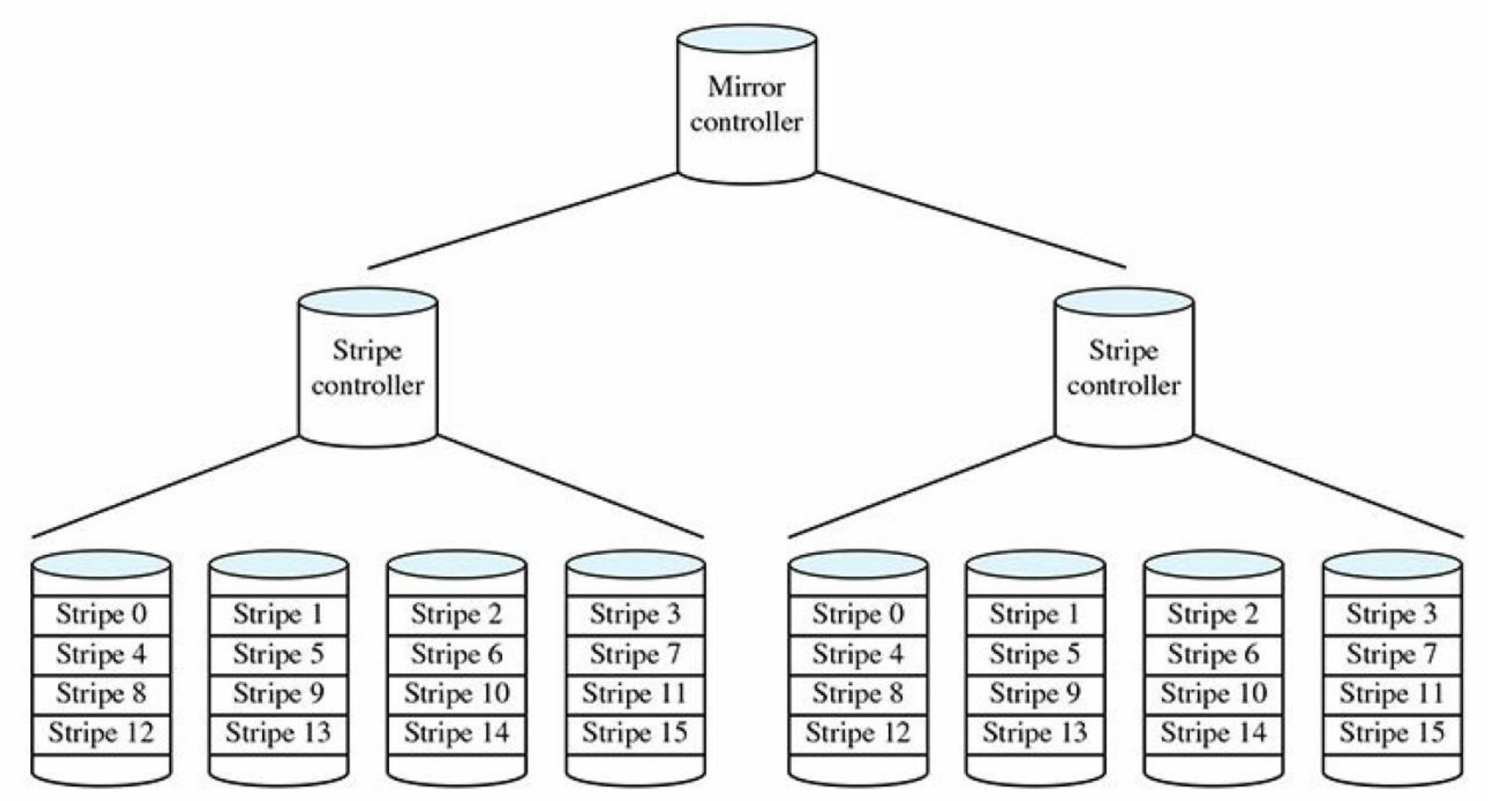
RAID 1

- **Mirroring** or **Shadowing**
- Two copies for every piece of data
- one logical write = two physical writes
- 100% capacity/space overhead

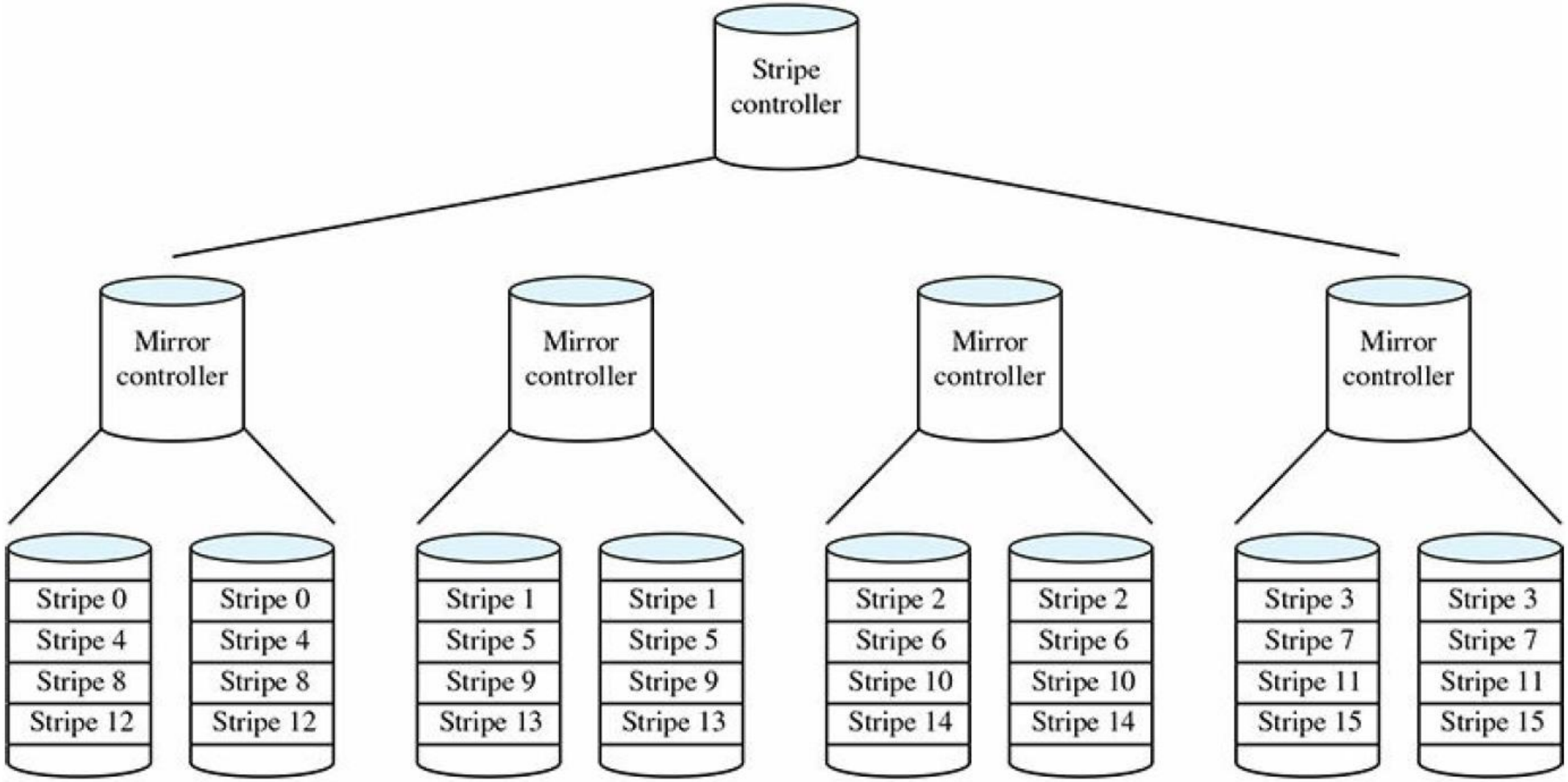


http://www.petemarovichimages.com/wp-content/uploads/2013/11/RAID1.jpg

RAID 01

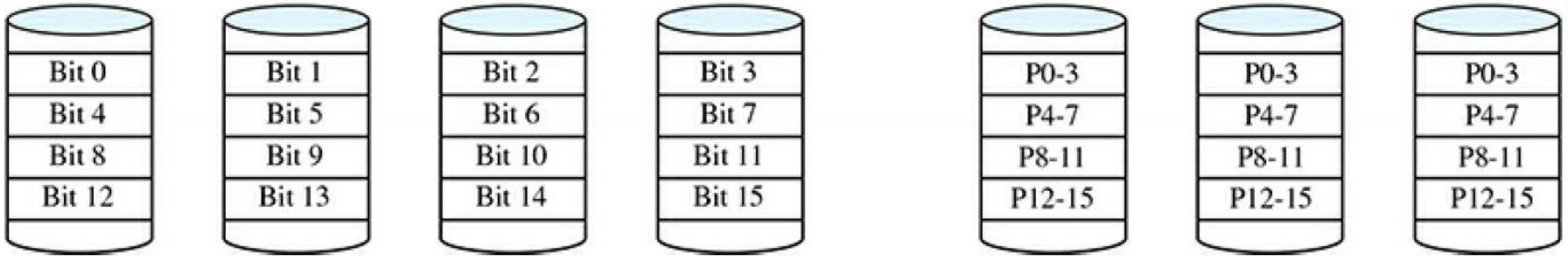

Mirror controller
Stripe controller
Stripe controller
Stripe 0
Stripe 4
Stripe 8
Stripe 12
Stripe 1
Stripe 5
Stripe 9
Stripe 13
Stripe 2
Stripe 6
Stripe 10
Stripe 14
Stripe 3
Stripe 7
Stripe 11
Stripe 15
Stripe 0
Stripe 4
Stripe 8
Stripe 12
Stripe 1
Stripe 5
Stripe 9
Stripe 13
Stripe 2
Stripe 6
Stripe 10
Stripe 14
Stripe 3
Stripe 7
Stripe 11
Stripe 15

RAID 10


Stripe controller
Mirror controller
Mirror controller
Mirror controller
Mirror controller
Stripe 0
Stripe 4
Stripe 8
Stripe 12
Stripe 0
Stripe 4
Stripe 8
Stripe 12
Stripe 1
Stripe 5
Stripe 9
Stripe 13
Stripe 1
Stripe 5
Stripe 9
Stripe 13
Stripe 2
Stripe 6
Stripe 10
Stripe 14
Stripe 2
Stripe 6
Stripe 10
Stripe 14
Stripe 3
Stripe 7
Stripe 11
Stripe 15
Stripe 3
Stripe 7
Stripe 11
Stripe 15

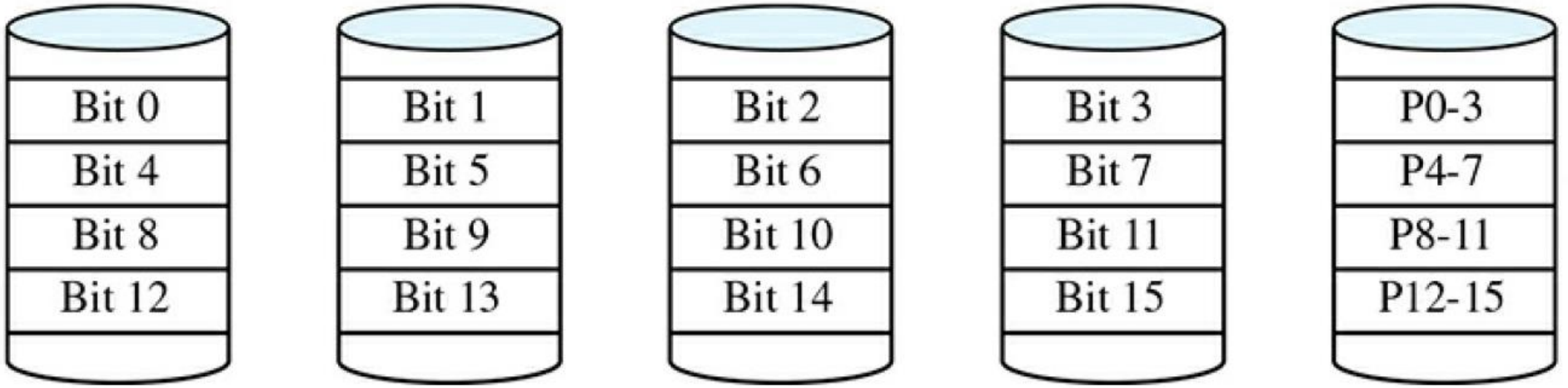
RAID 2

- Each bit of data word is written to a data disk drive
- Each data word has its (Hamming Code) ECC word recorded on the ECC disks
- On read, the ECC code verifies correct data or corrects single disks errors



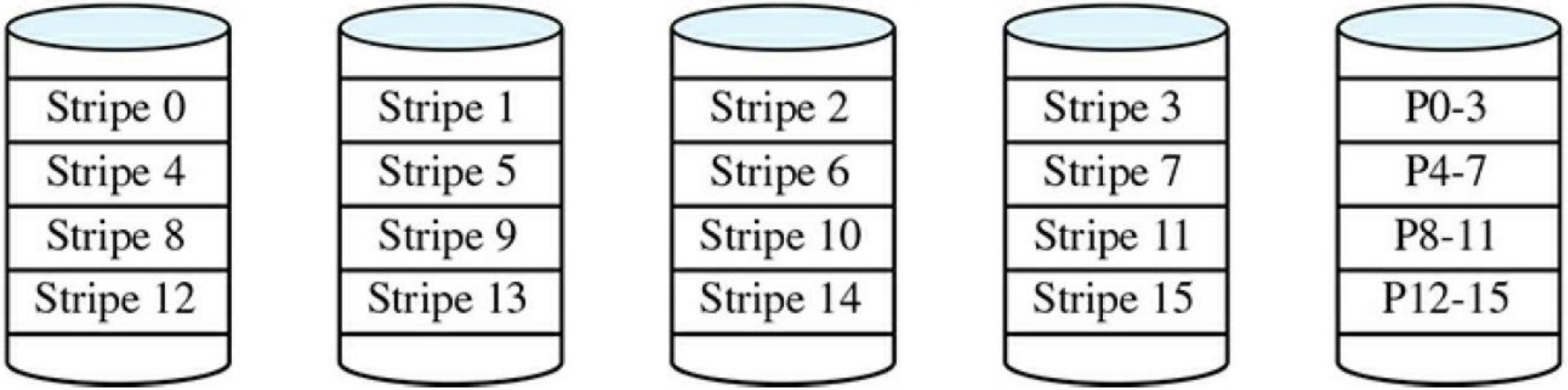
RAID 3

- Data striped over all data disks
- Parity of a stripe to parity disk
- Require at least 3 disks to implement

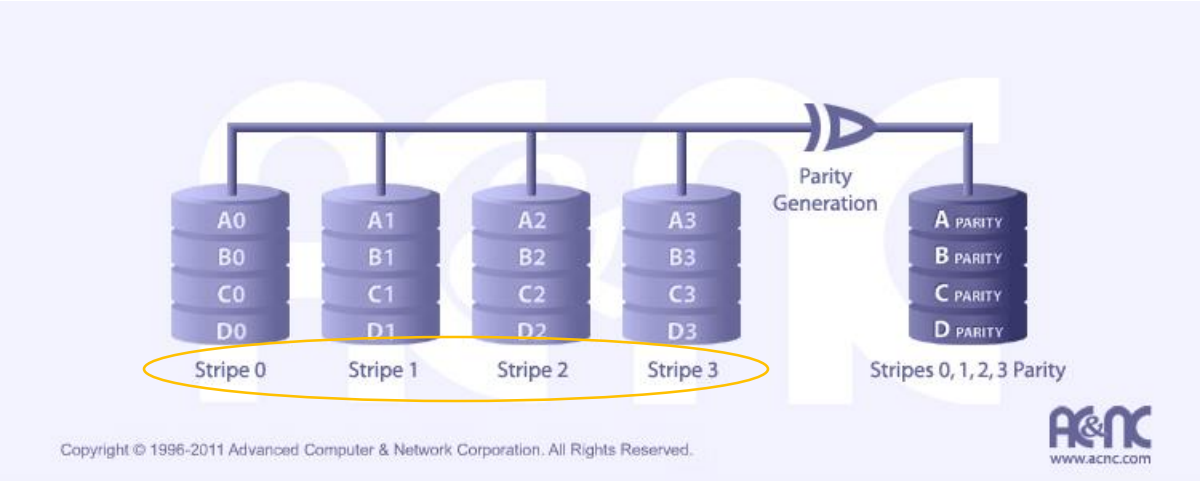


RAID 4

- Favor small accesses
- Allows each disk to perform independent reads, using sectors' own error checking



RAID 3 & RAID 4

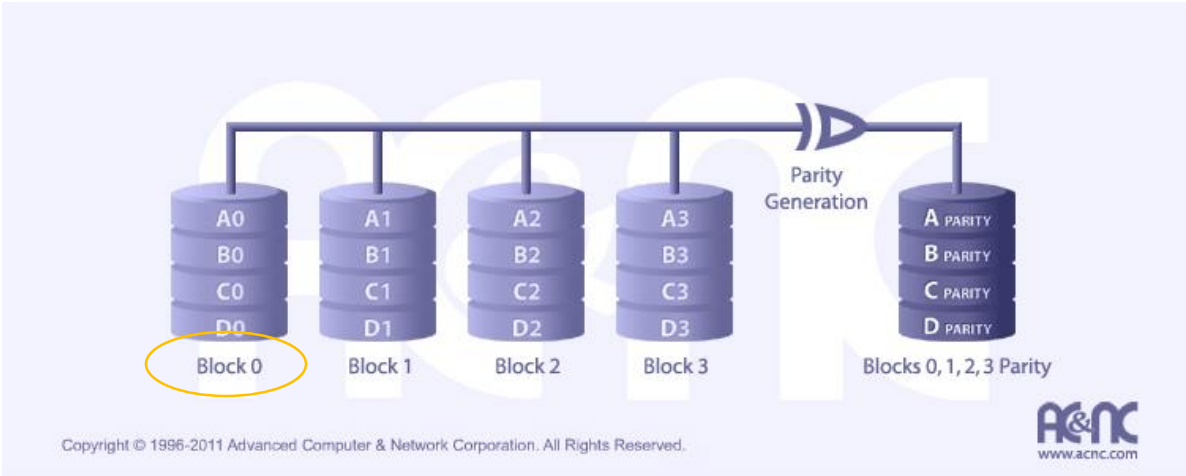


bottleneck:

single parity disk

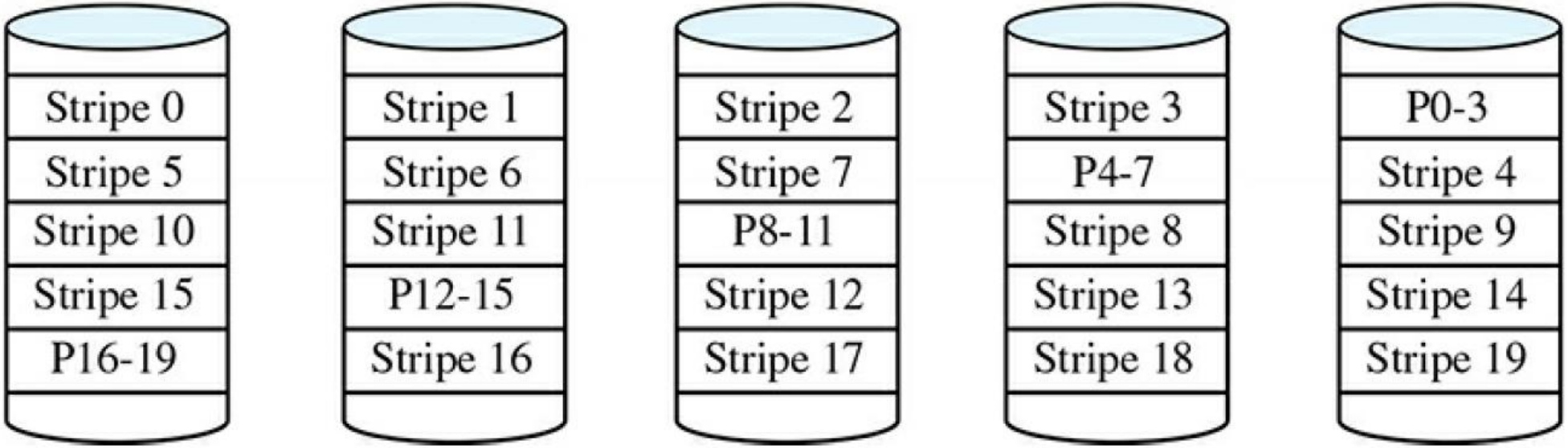
access:

parallel vs independent



RAID 5

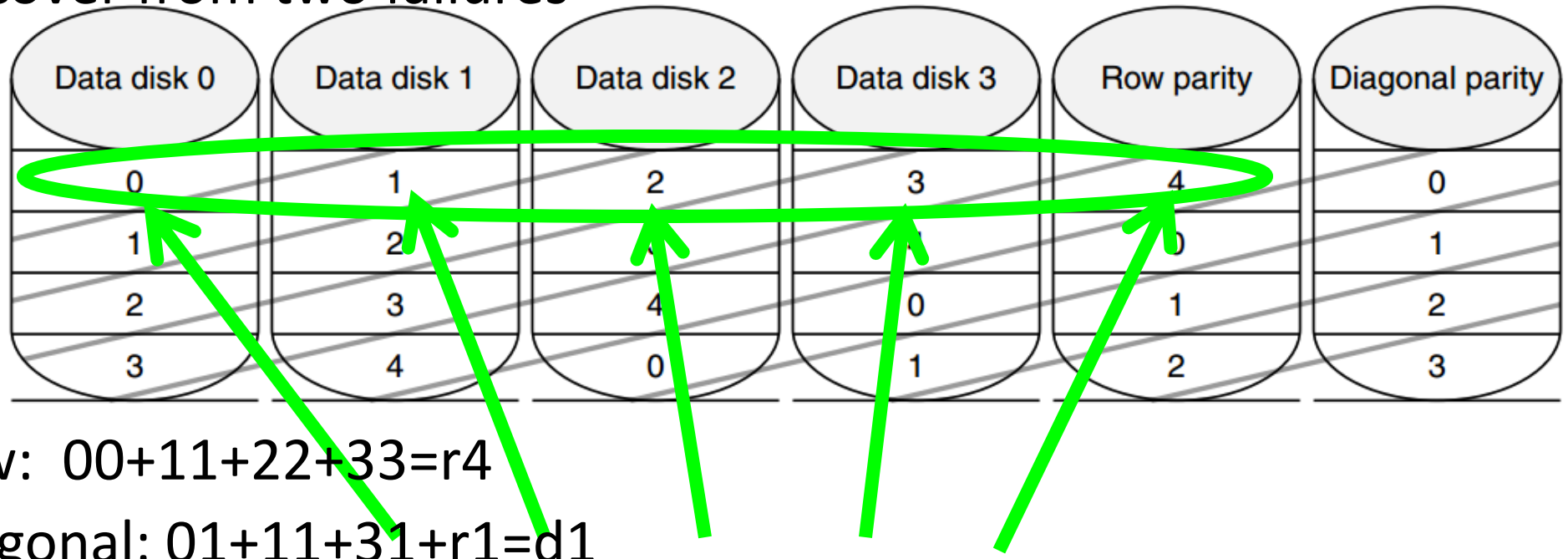
- Distributes the parity info across all disks in the array
- Removes the bottleneck of a single parity disk as RAID 3 and RAID 4



RAID 6: Row-diagonal Parity

- **RAID-DP**

  Recover from two failures



  xor

  row: 00+11+22+33=r4

  diagonal: 01+11+31+r1=d1

RAID: Further Readings

- Raid Types – Classifications

  BytePile.com

  https://www.icc-usa.com/content/raid-calculator/raid-0-1.png
- RAID

  JetStor

  http://www.acnc.com/raidedu/0

- More error detection and recovery schemes
- Reed Solomon coding: https://www.youtube.com/watch?v=jgO09opx56o
- Erasure coding, etc.

When are disks dependable and when are they not?

Dependability

- Computer system **dependability** is the quality of delivered service such that reliance can justifiably be placed on this service.

- The **service** delivered by a system is its observed **actual behavior** as perceived by other system(s) interacting with this system’s users.

- Each module also has an ideal **specified behavior**, where a **service specification** is an agreed description of the expected behavior.

Failure

- A system **failure** occurs when the actual behavior deviates from the specified behavior.

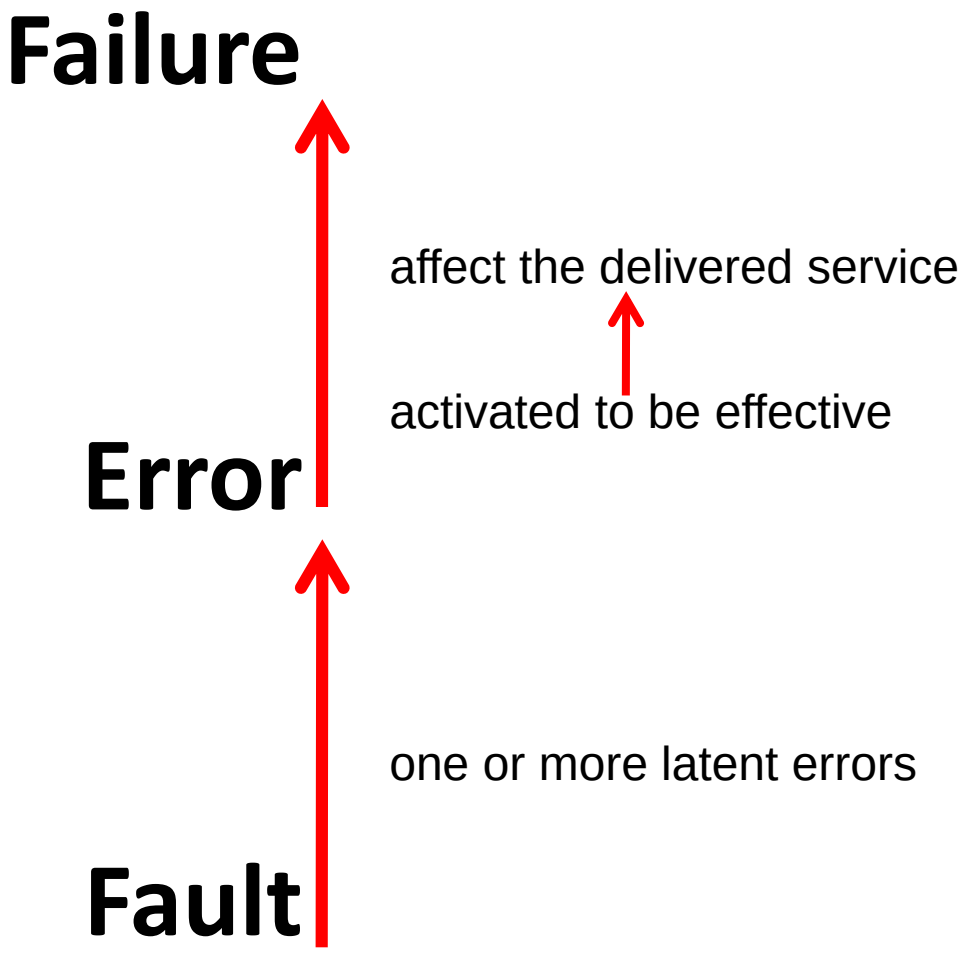
Error, Fault

- The failure occurred because of an **error**, a defect in that module.
- The cause of an error is a **fault**.

- When a fault occurs, it creates a **latent error**, which becomes **effective** when it is activated;
- When the error actually affects the delivered service, a **failure** occurs.

Fault, Error, Failure

- A fault creates one or more latent errors
- Either an effective error is a formerly latent error in that component or it has propagated from another error in that component or from else where
- A component failure occurs when the error affects the delivered service


Failure
affect the delivered service
activated to be effective
Error
one or more latent errors
Fault

Categories of Faults by Cause

- **Hardware faults**
  failed devices
- **Design faults**
  usually in software design;
  occasionally in hardware design;
- **Operation faults**
  mistakes by operations and maintenance personnel;
- **Environmental faults**
  fire, flood, earthquake, power failure, sabotage;

Categories of Faults by Duration

- **Transient faults**

  exist for a limited time and are not recurring;
- **Intermittent faults**

  cause a system to oscillate between faulty and fault-free operation
- **Permanent faults**

  do not correct themselves with the passing of time;

Threats/Vulnerabilities/Attacks

Threats

- A threat is a specific means by which a risk can be realized by an adversary
- Context specific (a fact of the environment)

Vulnerabilities

- A Vulnerability is a flaw that is accessible (threat) to an adversary who has the capability to exploit that flaw

Attacks

- An attack occurs when someone attempts to exploit a vulnerability
- Kinds of attacks: Passive/Active/Denial of Service

Attacks

- 缓冲区溢出攻击：Buffer overflow attack (stack/heap)
- 返回编程攻击：ROP attack
- 密码攻击：Password attack
- 钓鱼攻击：Phishing attack
- 跨站脚本攻击：XSS attack
- SQL注入攻击：SQL injection attack
- 冷启动攻击：Cold-boot attack
- 整数溢出攻击：Integer overflow attack
- 竞争条件攻击：Race condition attack
- 回放攻击：Replay attack
- 返回用户攻击：Return-to-user attack
- 拒绝服务攻击：Deny-of-service attack
- 暴力穷举攻击：Brute-force attack
- 回滚攻击：Rollback attack
- 社会工程攻击：Social engineering attack
- 侧信道攻击：Side-channel attack
- 隐蔽信道攻击：Covert-channel attack
- 物理攻击：Physical attack
- ...

Safety/Security/Trust

NTNU definition (Skavland Idsø and Mejdell Jakobsen, 2000):

Safety is protection against random incidents. Random incidents are unwanted incidents that happen as a result of one or more coincidences.

Security is protection against intended incidents. Wanted incidents happen due to a result of deliberate and planned act.

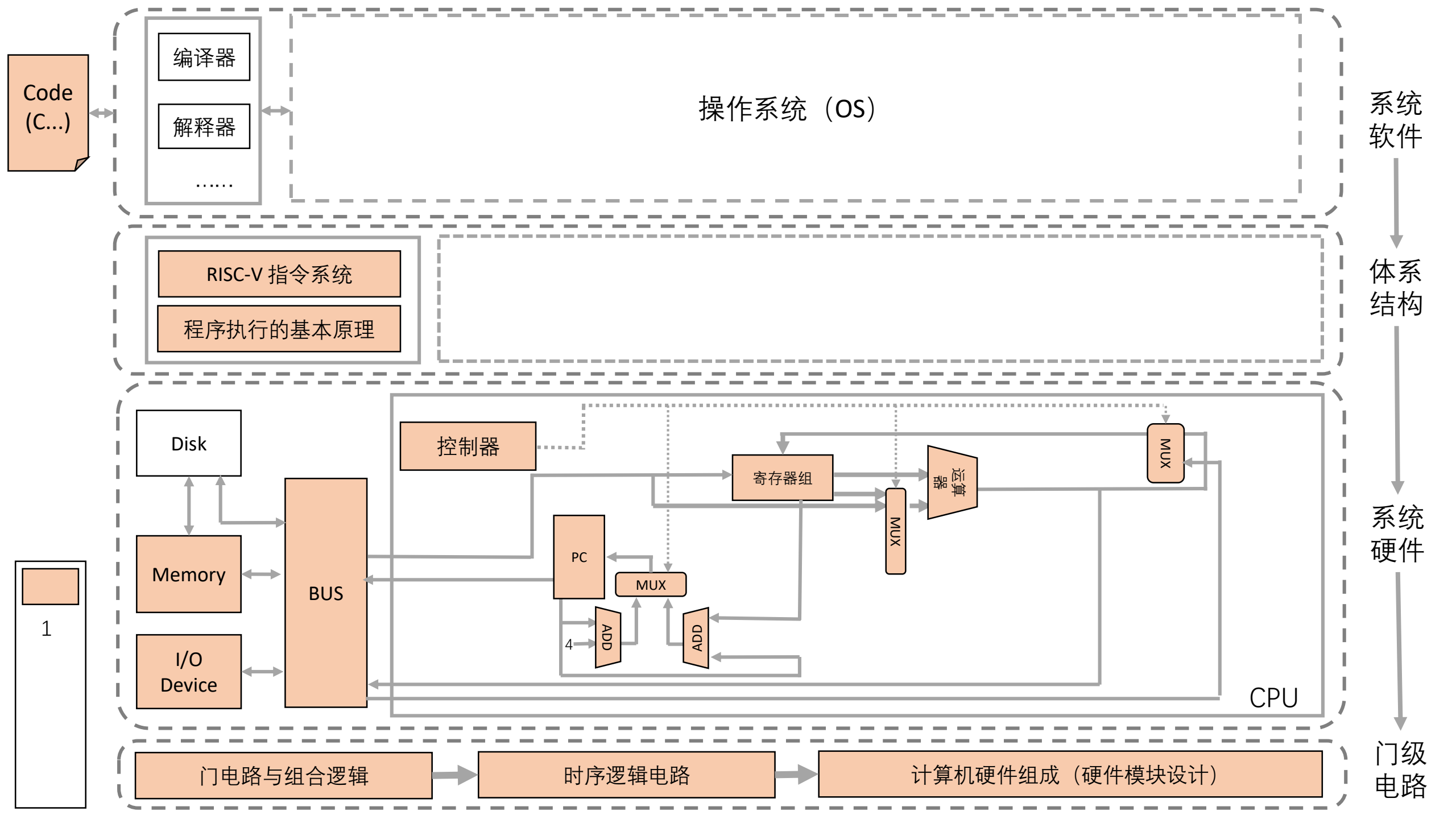
Security

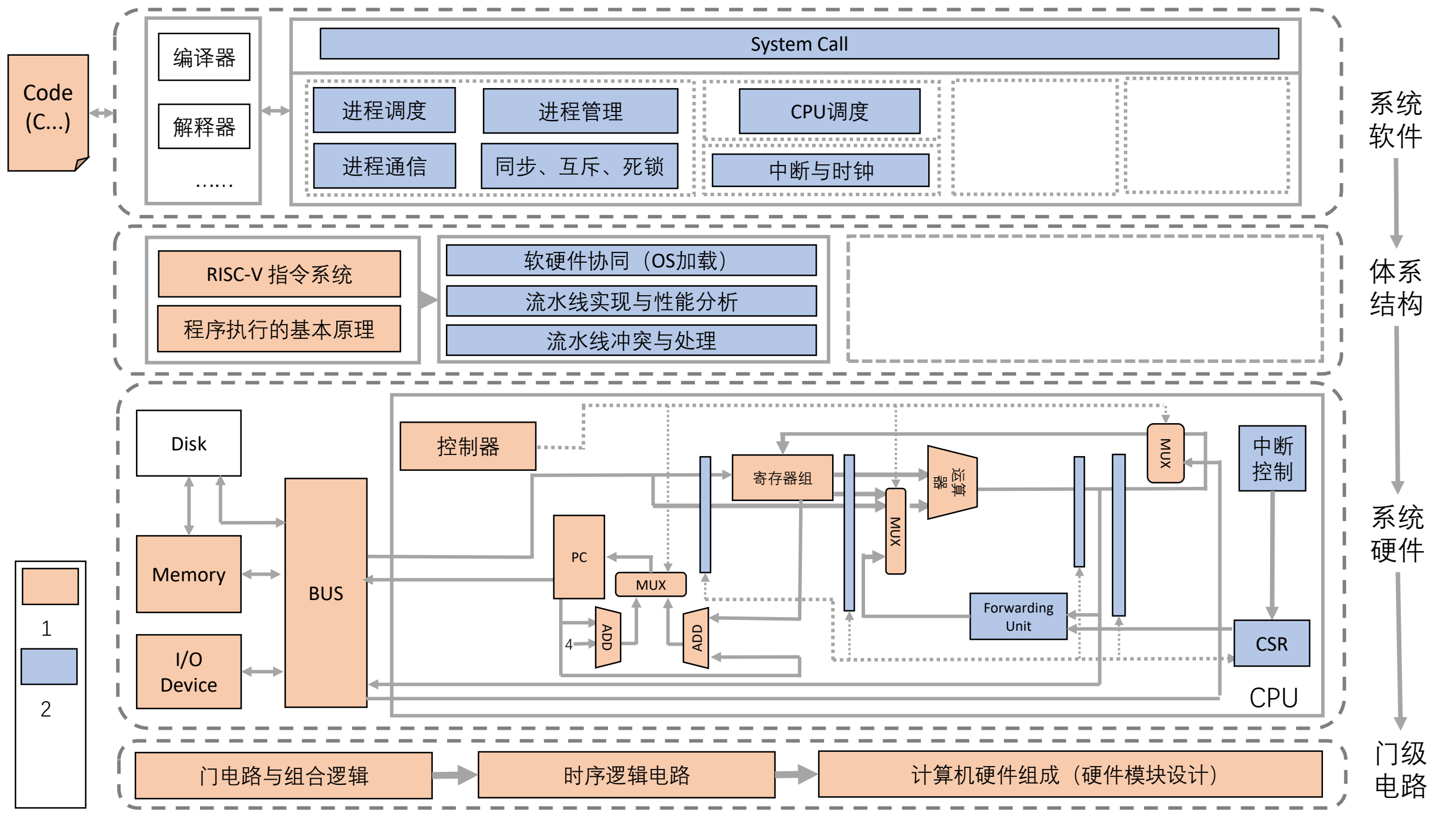
- Confidentiality
- Integrity
- Availability

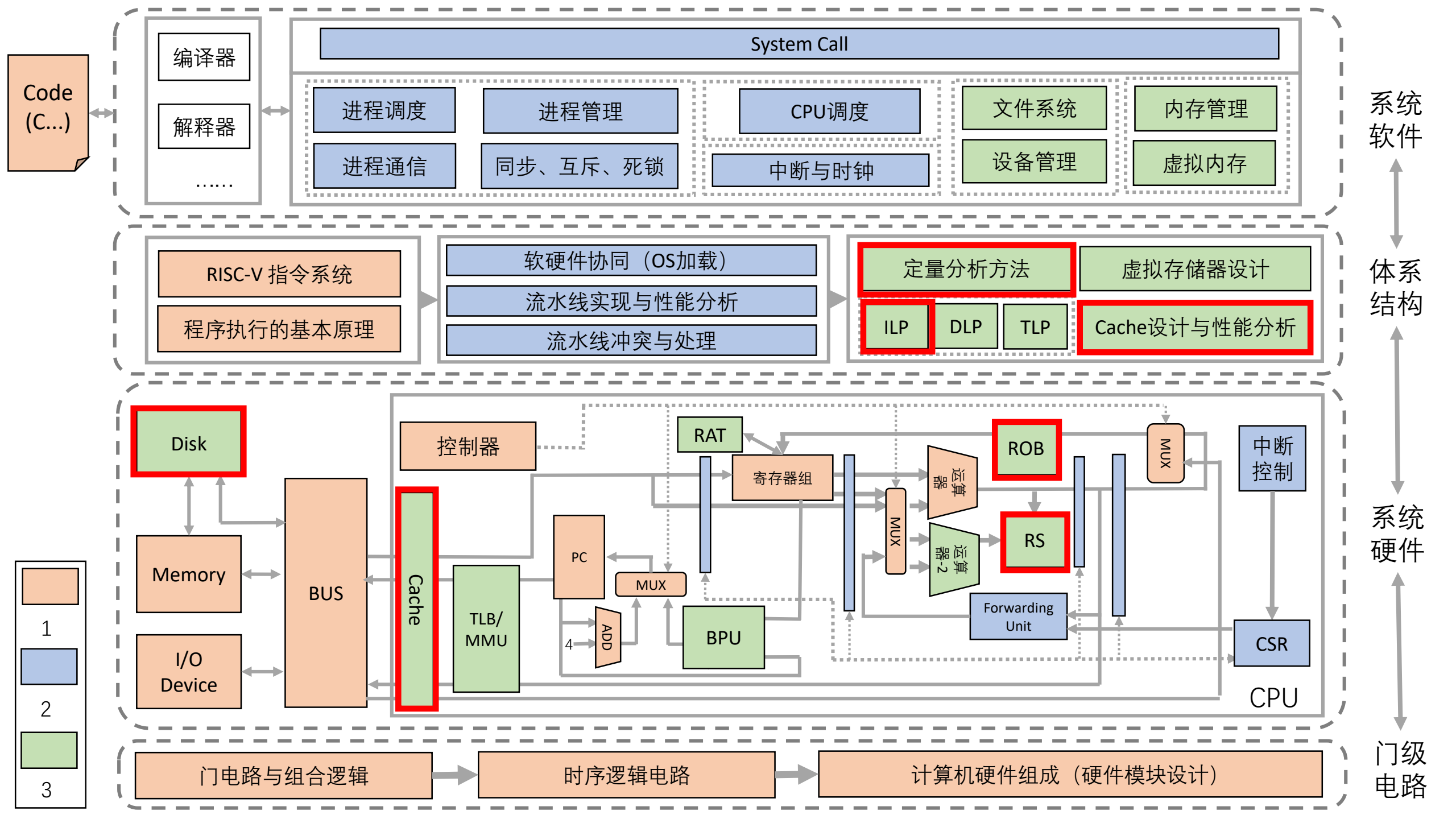
Example: Berkeley's Tertiary Disk

Component	Total in system	Total failed	Percentage failed
SCSI controller	44	1	2.3%
SCSI cable	39	1	2.6%
SCSI disk	368	7	1.9%
IDE/ATA disk	24	6	25.0%
Disk enclosure—backplane	46	13	28.3%
Disk enclosure—power supply	92	3	3.3%
Ethernet controller	20	1	5.0%
Ethernet switch	2	1	50.0%
Ethernet cable	42	1	2.3%
CPU/motherboard	20	0	0%

knowledge map


Code
(C...)
编译器
解释器
……
操作系统（OS）
系统
软件
RISC-V 指令系统
程序执行的基本原理
体系
结构
Disk
Memory
I/O
Device
BUS
控制器
寄存器组
运算
器
MUX
MUX
PC
MUX
4
ADD
ADD
CPU
系统
硬件
1
门电路与组合逻辑
时序逻辑电路
计算机硬件组成（硬件模块设计）
门级
电路


Code
(C...)
编译器
解释器
……
System Call
进程调度
进程管理
CPU调度
进程通信
同步、互斥、死锁
中断与时钟
RISC-V 指令系统
程序执行的基本原理
软硬件协同（OS加载）
流水线实现与性能分析
流水线冲突与处理
Disk
Memory
I/O
Device
BUS
控制器
寄存器组
运算器
MUX
PC
MUX
4
ADD
ADD
MUX
中断
控制
Forwarding
Unit
CSR
CPU
1
2
门电路与组合逻辑
时序逻辑电路
计算机硬件组成（硬件模块设计）
系统
软件
体系
结构
系统
硬件
门级
电路


Code
(C...)
编译器
解释器
……
System Call
进程调度
进程管理
CPU调度
文件系统
内存管理
进程通信
同步、互斥、死锁
中断与时钟
设备管理
虚拟内存
系统
软件
RISC-V 指令系统
程序执行的基本原理
软硬件协同（OS加载）
流水线实现与性能分析
流水线冲突与处理
定量分析方法
虚拟存储器设计
ILP
DLP
TLP
Cache设计与性能分析
体系
结构
Disk
Memory
I/O
Device
BUS
Cache
控制器
TLB/
MMU
PC
4
ADD
MUX
BPU
RAT
寄存器组
MUX
运算
器
运算
器-2
ROB
RS
Forwarding
Unit
MUX
中断
控制
CSR
CPU
系统
硬件
1
2
3
门电路与组合逻辑
时序逻辑电路
计算机硬件组成（硬件模块设计）
门级
电路