Storage and reliability

Computer Architecture

J. Daniel García Sánchez (coordinator)
David Expósito Singh
Javier García Blas

ARCOS Group
Computer Science and Engineering Department
University Carlos III of Madrid

Magnetic disks

- High storage capacity (hundreds of GBs).
- Spin at constant angular velocity.
- Access time for data stream:
  - T = track seek + rotation latency.
  - Depends on the stream access sequence.

Density

- Bits stored along track (BPI).
- Number of tracks per surface (TPI).
- Disks design trend to increasing density of bits stored per area unit (Areal Density).
- Areal Density = BPI $\times$ TPI

Year	Density
1973	2
1979	8
1989	63
1997	3,090
2000	17,100
2006	130,000

History perspective

- 1956 IBM Ramac $\rightarrow$ Early 70s Winchester.
  - Developed for mainframes.
  - Proprietary interfaces.
  - Constant reduction of size: from 27 to 14 inches.
- 1970s.
  - 5.25 inches.
  - Industry of standard interfaces for storage emerge.
- Early 1980s: Personal Computers (PCs) and first generations of desktop computers.

History perspective

- Mid 1980s: Client/server computing.
  - Centralized storage in file servers.
  - Miniaturization increases: 8 inches to 5.25.
  - Mass production of disk units in the market.
  - Standards: SCSI, IPI, IDE.
  - 5.25 inches to 3.5 inches for PCs.
- 1900s: Laptops => 2.5 inches.
- 2000s: New devices leading to new units:
  - 1.8 inches: iPods, MP3 players.
  - 1 inch IBMs microdrive.
  - 0.85 inches (Toshiba) mobile phones.

Illiac IV

- University of Illinois (1974)
  - 30,000,000$.
  - Solid state memory.
  - Laser memory.
  - Fastest in the world until 1981.
  - Numeric computing for NASA.

Disk capacity and performance

- Continuous increase in capacity (60%/year) and bandwidth (40%/year).
- Slow increase of disk rotation (8%/year).
- Time to read the whole disk.

Year	Sequentially	Randomly (1 sector/seek)
1990	4 min.	6 hours
2000	12 min.	1 week
2006 (SCSI)	56 min.	3 weeks
2006 (SATA)	171 min.	7 weeks

2 Reliability and availability

- Reliability
- Availability

Reliability

- The lifetime of a system represented as a random variable X.
- System reliability defined as function $R(t)$

$$R(t) = P(X > t) : R(0) = 1 \quad y \quad R(\inf) = 0 \tag{1}$$

Reliability and failures

- From study of components failures we obtain reliability

- **Reliability**: Probability that a device works properly during a given period of time under specific operating conditions.

Reliability distributions

- Examples of distributions used for reliability:

  - Exponential:
    - If error rate is constant (generally true for electronic components), reliability follows an exponential distribution.

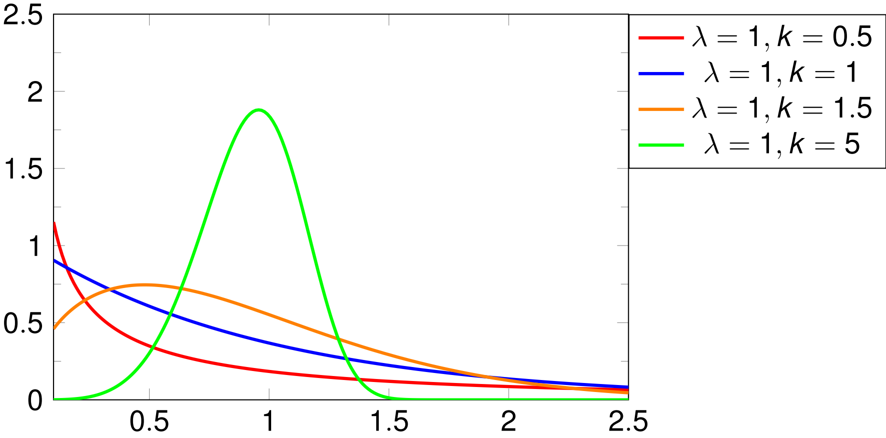
Reliability distributions

- Weibull:
  - Density function:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda} \cdot \left(\frac{x}{\lambda}\right)^{k-1} \cdot e^{-\left(\frac{x}{\lambda}\right)^k} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (2)$$

  - Parameter k also called shape factor:
    - $k < 1 \rightarrow$ failure rate decreases over time.
    - $k = 1 \rightarrow$ failure rate is constant over time.
    - $k > 1 \rightarrow$ failure rate increases over timem.
  - Models failure distribution whe failure rate is proportional to a power of time.

Weibull

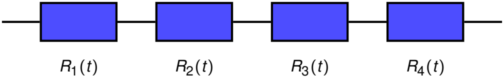

2.5
2
1.5
1
0.5
0
0.5
1
1.5
2
2.5
$\lambda = 1, k = 0.5$
$\lambda = 1, k = 1$
$\lambda = 1, k = 1.5$
$\lambda = 1, k = 5$

Reliability distributions

- Weibull:
  - Characteristic life η (time in which 63.2% of population fails) and form factor β
    - Associated to error rate, with $\beta = 1 \rightarrow$ constant error rate.

Serial systems

- Let $R_i(t)$ reliability for component **i**.
- System fails when some component fails.



- If failures are independent then:

$$R(t) = \prod_{i=1}^{N} R_i(t)$$

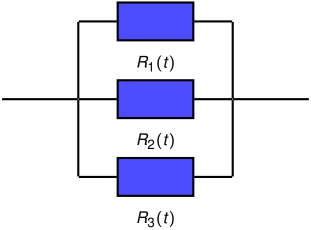
- System reliability is lower:

$$R(t) < R_i(t) \forall i$$

Paralel system

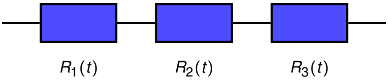
- System fails when all components fail.

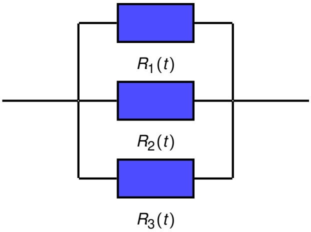
$$R(t) = 1 - \prod_{i=1}^{N} Q_i(t) : Q_i(t) = 1 - R_i(t)$$



Example

Para $t = 100$ $R_i(t) = 0.9$





$$R(t) = 0.9 \cdot 0.9 \cdot 0.9 = 0.729$$

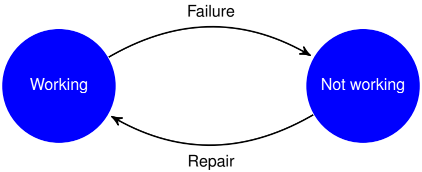
$$R(t) = 1 - (1 - 0.9)^3 = 0.999$$

2 Reliability and availability

- Reliability
- Availability

Availability

- In many cases, it is more interesting to know availability.
- Availability of a system $A(t)$ defined as the probability that the system is working correctly at instant t.
  - Reliability considers interval $[0, t]$.
  - Availability considers a concrete instant in time.
- A system modelled as following state diagram.



Availability measurement

- Let MTTF the average time to failure.
- Let MTTR the average time to repair.
- System availability A is defined as:

$$A = \frac{MTTF}{MTTF + MTTR}$$

- What does a reliability of 99% mean?
  - In 365 days, it works correctly $\frac{99 \cdot 365}{100} = 361.35$ days.
  - Out of service 3.65 days.

Annual time without service

Availability (%)	Days without service in a year
98%	7.3 days
99%	3.65 days
99.8%	17 hours y 30 minutes
99.9%	8 hours y 45 minutes
99.99%	52 minutes y 30 seconds
99.999%	5 minutes y 15 seconds
99.9999%	31.5 seconds

Computing availability

- Elements availability
  - HW: 99.99%
  - Disk: 99.9%
  - SO: 99.99%
  - Application: 99.9%
  - Communications: 99.9%

- System availability:
  - Product of elements availability.

$$A(t) = \prod_{i=1}^{N} A_i(t) = 99.6804 \Rightarrow 1.17 \text{days without service}$$

Sectors with most service interruptions

Sector	Percentage
Bank and finance	26%
Government, public administrations and institutions	19.1%
Education	11.3%
Industry	10.9%
Services	9.5%
Communications	8.2%

Cost of stopping one hour

Cost	Percentage
Up to 50,000$	46%
50,000$ – 100,000$	15%
100,000$ – 250,000$	13%
250,000$ – 500,000$	9%
500,000$ – 1,000,000$	9%
1,000,000$ – 5,000,000$	4%
More than 5,000,000$	4%

What to do with failures?

- Problems in disks:
  - Failure in the disk itself.
  - Failure in the disk controller.
  - Failure in block (damaged sectors).
  - Transient failures.

- Using a redundant storage system:
  - **R**edundant **A**rray of **I**nexpensive/Independent **D**isks.
  - Proposed for the first time in 1998 by David A. Patterson, Garth A. Gibson and Randy H. Katz.
  - *"A case for inexpensive arrays of redundant disks (RAID)"*

RAID Disks

- Several types of RAID:
- Basic levels:
  - **RAID 0**: block distribution (striping) without fault tolerance.
  - **RAID 1**: disk mirroring.
  - **RAID 2**: bit level interleaving with Hamming.
  - **RAID 3**: bit level interleaving with redundant information (parity)
  - **RAID 4**: block distribution with parity disk.
  - **RAID 5**: block distribution with distributed parity.
- Combinations:
  - **RAID 10**: Stripping and mirroring (RAID 0 and 1).
  - **RAID 51**: Combination of RAID 5 and RAID 1.
  - . . .

RAID 0 (striping)

- Fault tolerance:
  - Does not offer fault tolerance.
- Performance:
  - Higher throughput in read/write operations.
- Capacity:
  - Addition.

RAID 1 (mirroring)

- Fault tolerance:
  - 1 failure.
- Performance:
  - Higher throughput in read operations.
- Capacity:
  - 50% of total.

RAID 2

- Failure detection.
- Use Hamming code.
- Bit level *Striping*.
- Very costly implementation.
- Not used.

RAID 3

- RAID 3 (*striping with dedicated parity, bit level*.
- Byte level stripping.
- Parity of written bytes.
- Tolerance to 1 failure.
- Use byte level redundancy.
- Improve throughput:
  Parallel access to block.
- Parity disk is a bottleneck.

RAID 4

- RAID 4 (*striping with dedicated parity*.
- Block level striping.
- Fault tolerance: 1 failure.
- Performance:
  - Costly writes (parity).
  - Parity disk is a bottleneck.
- Capacity: $\frac{100 \cdot (n-1)}{n}\%$

RAID 3 versus RAID 4

- **RAID 3**: Each byte in a disk.
- **RAID 4**: Each block in a disk.

RAID 5

- RAID 5 (*striping with distributed parity*.
- Block level striping.
- Parity striping.
- Parity is not in the same disk as associated blocks.
- Fault tolerance: 1 failure.
- There is no bottleneck in access to parity.
- Capacity: $\frac{100\cdot(n-1)}{n}\%$

RAID 6

- RAID 6 (*striping with distributed redundant parity*.
- Block level striping.
- Parity striping.
- Parity is replicated twice.
- Parity is not in the same disk than the associated blocks.
- Fault tolerance: 2 failures.
- There is no bottleneck in access to parity.

Reads in RAID 4-5

- If disk works:
  - Corresponding disk is read.
- If disk does not work:
  - Blocks in other disks and parity disk are read to compute new block.

Writes in RAID 4-5

- If disk works:
  - Write a block and the new parity, by:
    1. Read the old block OB and the parity block OP.
    2. New parity will be: $NP = (OB \oplus NB) \oplus OP$.
    3. Write the new block NB and the parity block NP.
- If disk does not work:
  - Update block and parity in working disk.

- Whe disk fails is substituted and its information is reconstructed.

Summary

- Reliability models system life time.
- Parallel systems allow improving system reliability while serial systems worsen system reliability.
- Availability models the probability of failures at instant in time.
- RAID systems improve both performance and reliability of storage systems.

References

- **Computer Architecture. A Quantitative Approach**
  5th Ed.
  Hennessy and Patterson.
  Sections D.1, D.2, D.3, D.4.

Storage and reliability

Computer Architecture

J. Daniel García Sánchez (coordinator)
David Expósito Singh
Javier García Blas

ARCOS Group
Computer Science and Engineering Department
University Carlos III of Madrid