Almacenamiento y fiabilidad

Arquitectura de Computadores

J. Daniel García Sánchez (coordinador)
David Expósito Singh
Javier García Blas
Óscar Pérez Alonso
J. Manuel Pérez Lobato

Grupo ARCOS
Departamento de Informática
Universidad Carlos III de Madrid

1 Almacenamiento

2 Fiabilidad y disponibilidad

3 RAID

4 Conclusión

Discos magnéticos

- Elevada capacidad de almacenamiento (cientos de GB).
- Giran a velocidad angular constante.
- Tiempo de acceso a un flujo de datos:
  - T = posicionamiento en pista + latencia de rotación.
  - Depende de la secuencia de acceso a los flujos.

Densidad

- Bits almacenados a lo largo de la pista (BPI)
- Número de pistas por superficie (TPI)
- El diseño de los discos apunta a incrementar la densidad de bits almacenados por area (Areal Density)
- Areal Density = BPI x TPI

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

Perspectiva histórica

- 1956 IBM Ramac — primeros de los ‘70 Winchester.
  - Desarrollado para computadores centrales.
  - Interfaces propietarias.
  - Constante reducción de tamaño: 27 pulgadas a 14 pulgadas.
- 1970s.
  - 5.25 pulgadas.
  - Emergen la industria de interfaces estándard de almacenamiento.
- Primeros 1980s: PCS y primeras generaciones de computadores domésticos.

Perspectiva histórica

- Mediados 1980s: computación cliente/servidor.
  - Almacenamiento centralizado en servidores de ficheros.
  - Aceleración de la miniaturización: 8 pulgadas a 5.25.
  - Producción en masa de unidades de disco en el mercado.
  - Estandards: SCSI, IPI, IDE.
  - 5.25 pulgadas a 3.5 pulgadas para PCs.
- 1900s: Ordenadores portátiles => 2.5 pulgadas.
- 2000s: ¿Qué nuevos dispositivos que conducen a nuevas unidades?
  - 1.8 pulgadas: iPods, reproductores MP3.
  - 1 pulgada IBM's microdrive.
  - 0.85 pulgada (Toshiba) teléfonos móviles.

Illiac IV

- Universidad de Illinois (1974)
  - 30,000,000$.
  - Memoria estado sólido.
  - Memoria con láser.
  - Más rápido del mundo hasta 1981.
  - Cálculos numéricos para la NASA.

Capacidad del disco y rendimiento

- Aumento continuo en capacidad (60%/año) y ancho de banda (40%/año).
- Lenta mejora de la rotación del disco (8%/año).
- Tiempo para leer todo el disco.

Año	Secuencialmente	Aleatoriamente (1 sector/búsqueda)
1990	4 min.	6 horas
2000	12 min.	1 semana
2006 (SCSI)	56 min.	3 semanas
2006 (SATA)	171 min.	7 semanas

1 Almacenamiento

2 Fiabilidad y disponibilidad

3 RAID

4 Conclusión

2 Fiabilidad y disponibilidad
- Fiabilidad
- Disponibilidad

Fiabilidad

- El tiempo de vida de un sistema se representa mediante una variable aleatoria X.
- Se define la fiabilidad del sistema como una función $R(t)$

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

Fiabilidad y fallos

- A partir del estudio de los fallos de los componentes se obtiene la fiabilidad.
  - http://www.jmcprl.net/ntps/@datos/ntp_418.htm.

Distribuciones de fiabilidad

- Ejemplos de distribuciones utilizadas para fiabilidad:
  - `http://www.relexsoftware.com/resources/art/art_distrib.asp`.

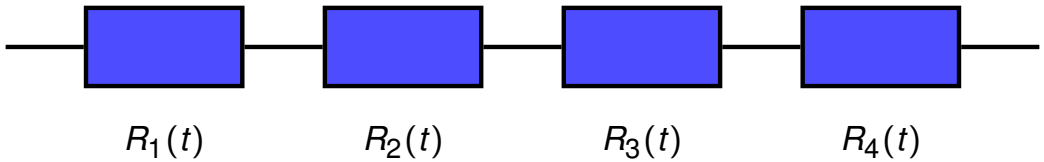
- Exponencial:
  - Si la tasa de errores es constante (generalmente verdadero para componentes electrónicos), la fiabilidad sigue una exponencial.

Distribuciones de fiabilidad

- Weibull:
  - Vida característica η (tiempo en el que el 63.2% de población falla) y factor de forma β
    - Asociado a la tasa de error, siendo b=1 $\rightarrow$ tasa de error constante.

Sistemas serie

- Sea $R_i(t)$ la fiabilidad del componente **i**.
- El sistema falla cuando algún componente falla.



- Si los fallos son independientes entonces:

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

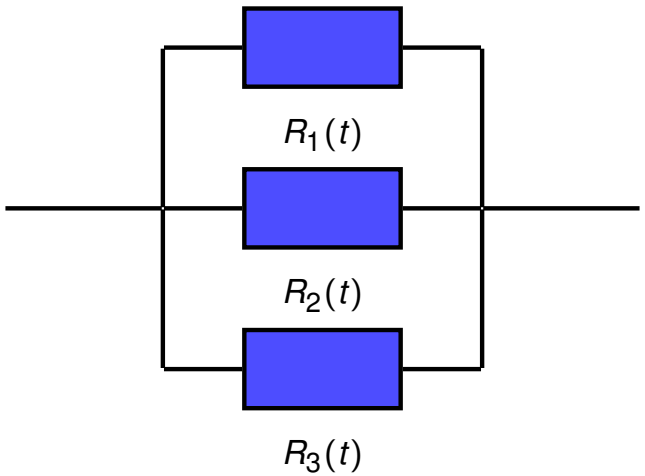
- La fiabilidad del sistema es menor.

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

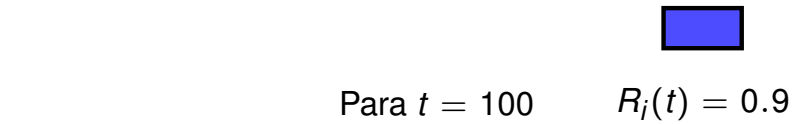
Sistema paralelo

- El sistema falla cuando todos los componentes fallan.

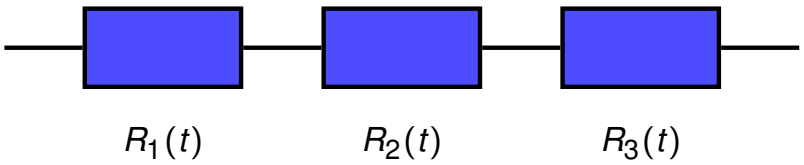
$$R(t) = 1 - \prod_{i=1}^{N} Q_i(t) : Q_i(t) = 1 - R_i(t)$$

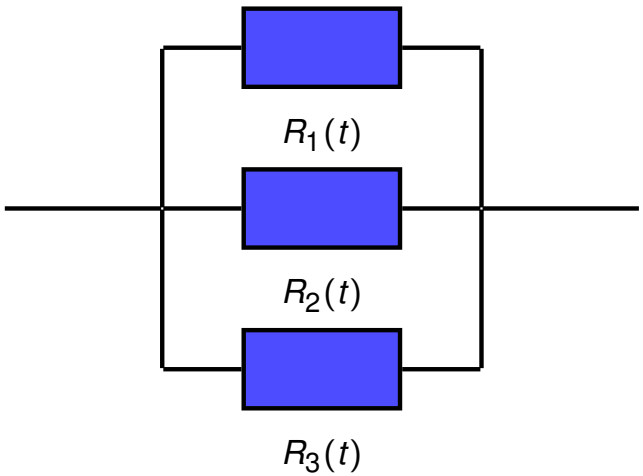


Ejemplo



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





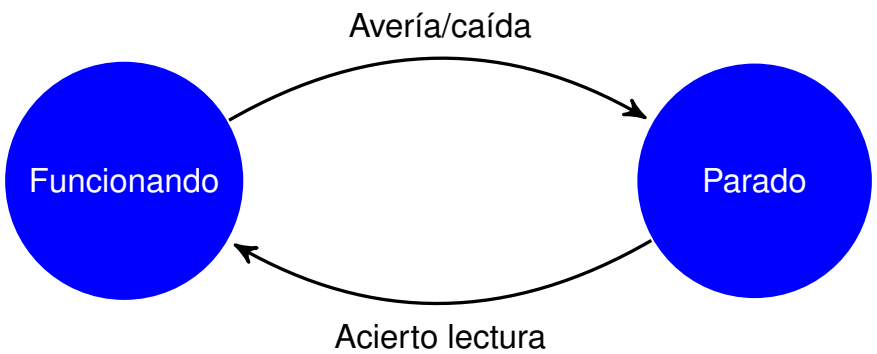
$$R(t) = 0.9 \cdot 0.9 \cdot 0.9 = 0.729$$

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

2 Fiabilidad y disponibilidad
- Fiabilidad
- Disponibilidad

Disponibilidad

- En muchos casos es más interesantes conocer la disponibilidad.
- Se define la disponibilidad de un sistema $A(t)$ como la probabilidad de que el sistema esté funcionando correctamente en el instante t.
  - La fiabilidad considera el intervalo $[0, t]$.
  - La disponibilidad considera un instante concreto de tiempo.
- Un sistema se modela según el siguiente diagrama de estados.



Medida de la disponibilidad

- Sea TMF el tiempo medio hasta el fallo.
- Sea TMR el tiempo medio de reparación.
- Se define la disponibilidad *A* de un sistema como:

$$A = \frac{TMF}{TMF + TMR}$$

- ¿Qué significa una disponibilidad del 99%?
  - En 365 días funciona correctamente $\frac{99 \cdot 365}{100} = 361.35$ días.
  - Está sin servicio 3.65 días.

Tiempo anual sin servicio

Disponibilidad (%)	Dias sin servicio al año
98%	7.3 días
99%	3.65 días
99.8%	17 horas y 30 minutos
99.9%	8 horas y 45 minutos
99.99%	52 minutos y 30 segundos
99.999%	5 minutos y 15 segundos
99.9999%	31.5 segundos

Cálculo de la disponibilidad

- Disponibilidad de los elementos:
  - HW: 99.99%
  - Disco: 99.9%
  - SO: 99.99%
  - Aplicación: 99.9%
  - Comunicación 99.9%

- Disponibilidad del sistema:
  - Producto de las disponibilidades de los elementos.

$$A(t) = \prod_{i=1}^{N} A_i(t) = 99.6804 \Rightarrow 1.17\text{días sin servicio}$$

Sectores que sufren más interrupciones

Sector	Procentaje
Banca y finanzas	26%
Gobierno, administraciones públicas e instituciones	19.1%
Educación	11.3%
Industria	10.9%
Servicios	9.5%
Comunicaciones	8.2%

Coste de la hora de parada

Coste	Porcentaje
Hasta 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%
Más de 5,000,000$	4%

1 Almacenamiento

2 Fiabilidad y disponibilidad

3 RAID

4 Conclusión

¿Cómo hacer frente a los fallos?

- Problemas en los discos:
  - Fallo en el propio disco.
  - Fallo en el controlador del disco.
  - Fallo en un bloque (sectores dañados).
  - Fallos transitorios.

- Uso de un sistema redundante de almacenamiento:
  - **R**edundant **A**rray of **I**nexpensive/Independent **D**isks.
  - Propuesto por primera vez en 1998 por David A. Patterson, Garth A. Gibson y Randy H. Katz.
  - *"Un Caso para Conjuntos de Discos Redundantes Económicos (RAID)"*

Discos RAID

- Varios tipos de RAID:
- Niveles básicos:
  - **RAID 0**: distribución de bloques (stripping) sin tolerancia a fallos
  - **RAID 1**: discos espejos (mirroring)
  - **RAID 2**: entrelazados a nivel de bit con Hamming
  - **RAID 3**: entrelazados de bits con información redundante (paridad)
  - **RAID 4**: distribución de bloques con disco de paridad
  - **RAID 5**: distribución de bloques con paridad distribuida
- Combinaciones:
  - **RAID 10**: Stripping y mirroring (RAID 0 y 1)
  - **RAID 51**: Combinación de RAID 5 y RAID 1
  - ...

RAID 0 (striping)

- Tolerancia a fallos:
  - No ofrece tolerancia a fallos.
- Rendimiento:
  - Mayor ancho de banda en operaciones de lectura/escritura.
- Capacidad:
  - La suma.

RAID 1 (mirroring)

- Tolerancia a fallos:
  - 1 fallo.
- Rendimiento:
  - Mayor ancho de banda en operaciones de lectura.
- Capacidad:
  - 50% del total.

RAID 2

- Detección de fallo.
- Uso de código Hamming.
- *Stripping* a nivel de bit.
- Implementación muy costosa.
- No se suele usar.

RAID 3

- RAID 3 (*striping with dedicated parity, bit level*.
- Stripping a nivel de byte.
- Paridad de bytes escritos
- Tolerancia a 1 fallo.
- Uso de redundancia a nivel de byte.
- Mejora ancho de banda: Acceso paralelo a un bloque.
- Disco de paridad es cuello de botella.

RAID 4

- RAID 4 (*striping with dedicated parity*.
- Stripping a nivel de bloque.
- Tolerancia a fallos: 1 fallo
- Rendimiento:
  - Costoso en escrituras (paridad).
  - Disco de paridad es cuello de botella.
- Capacidad: $\frac{100 \cdot (n-1)}{n}\%$

RAID 3 frente RAID 4

- **RAID 3**: Cada byte en un disco.
- **RAID 4**: Cada bloque en un disco.

RAID 5

- RAID 5 (*striping with distributed parity*.
- Stripping a nivel de bloque.
- Stripping de la paridad.
- La paridad no está en el disco que tiene los bloques asociados.
- Tolerancia a fallos: 1 fallo
- No existe cuello de botella en acceso a paridad.
- Capacidad: $\frac{100 \cdot (n-1)}{n}\%$

RAID 6

- RAID 6 (*striping with distributed redundant parity*.
- Stripping a nivel de bloque
- Stripping de la paridad
- La paridad está replicada por partida doble
- La paridad no está en el disco que tiene los bloques asociados.
- Tolerancia a fallos: 2 fallos
- No existe cuello de botella en acceso a paridad.

Lecturas en RAID 4-5

- Si el disco funciona:
  - Se lee del disco correspondiente.
- Si el disco no funciona:
  - Leer los bloques de los otros discos y el de paridad y calcular el nuevo bloque.

Escrituras en RAID 4-5

- Si el disco funciona:
  - Escribir el bloque y la nueva paridad. Para ello:
    1. Se lee el bloque antiguo BA y el de paridad PA
    2. La nueva paridad será: $PN = (BA \oplus BN) \oplus PA$
    3. Escribir el nuevo bloque BN y el de paridad PN
- Si el disco no funciona:
  - Se actualiza el bloque o paridad en disco que funcione

- Cuando un disco falla se sustituye y se reconstruye la información del mismo.

1 Almacenamiento

2 Fiabilidad y disponibilidad

3 RAID

4 Conclusión

Resumen

- La fiabilidad modela el tiempo de vida del sistema.
- Los sistemas paralelos permiten mejorar la fiabilidad del sistema mientras que los sistemas en serie empeoran la fiabilidad del sistema.
- La disponibilidad modela la probabilidad de fallo en un instante.
- Los sistemas RAID permiten mejorar el rendimiento y la fiabilidad de los sistemas de almacenamiento.

Referencias

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

Almacenamiento y fiabilidad

Arquitectura de Computadores

J. Daniel García Sánchez (coordinador)
David Expósito Singh
Javier García Blas
Óscar Pérez Alonso
J. Manuel Pérez Lobato

Grupo ARCOS
Departamento de Informática
Universidad Carlos III de Madrid