SESIÓN DE LABORATORIO 9
Matrices redundantes de discos (RAID)

Objetivos

- Entender el funcionamiento y las características básicas de las matrices redundantes de discos (RAID, Redundant Array of Independent Disks).
- Comprender las diferentes organizaciones o niveles de RAID y las implicaciones que tienen en la capacidad, el rendimiento y la confiabilidad del sistema de almacenamiento.
- Interpretar adecuadamente las especificaciones de un disco magnético dadas por el fabricante.
- Manejar y comprender el significado de parámetros de confiabilidad de los discos magnéticos tales como AFR y MTBF.

Desarrollo

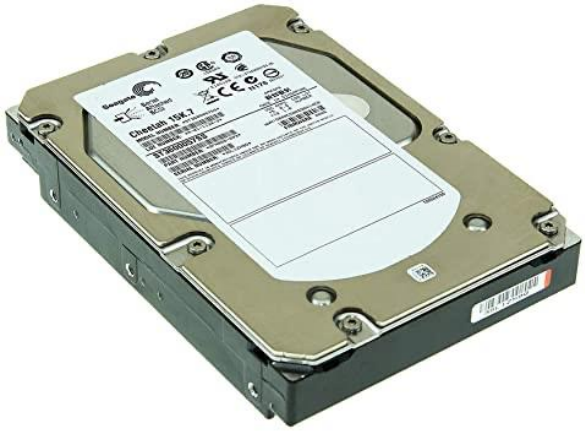
En esta sesión de laboratorio se van a resolver una serie de cuestiones y problemas sencillos relacionados con el diseño e implementación de las matrices redundantes de discos. Para ello el alumno deberá aplicar los conocimientos teóricos sobre discos magnéticos y su combinación en una matriz a una serie de supuestos prácticos elementales. En particular, en esta sesión se trabajarán cuestiones relativas a la capacidad efectiva de las matrices de discos, el rendimiento y la confiabilidad.

Mientras no se diga lo contrario, se trabajará con las especificaciones de los discos magnéticos Seagate Cheetah 15K.7. Se trata de una serie de discos de alta gama que engloba diversos modelos diseñados para implementar sistemas de almacenamiento con altos requerimientos de rendimiento, capacidad y confiabilidad. Estos discos se pueden encontrar con interfaces SAS (Serial Attached SCSI) y FC (Fibre Channel).

Para un mayor detalle, consúltese el documento del Manual del Producto Cheetah 15K.7 SAS Product manual. Un resumen de este documento puede verse en las Hojas de Especificaciones técnicas Cheetah 15K.7 SAS Data Sheet.

Dimensionamiento de la capacidad efectiva

En esta sección se supondrá que se quiere diseñar un sistema de almacenamiento con una capacidad total efectiva de 4 TB (recuérdese que la constante T = 10^{12} es diferente de Ti = 2^{40}). Para ello se van a utilizar discos del modelo ST3600057SS de la serie Cheetah 15K.7.



Las siguientes cuestiones hacen referencia al cómputo del número de discos necesarios para conseguir esta **capacidad total efectiva de 4 TB** teniendo en cuenta tanto la capacidad de este modelo en particular como las características propias de las diversas configuraciones de RAID.

RAID 0

Calcule el número de discos necesarios si el sistema de almacenamiento se configura como RAID 0. ¿Cuánta capacidad se usa para almacenar información redundante?

```
4TB / 600 GB = 4000GB / 600GB = 6,67 = 7 discos para tener una capacidad de 4TB
```

RAID 5, 5E y 5EE

Repita el cálculo anterior considerando que los discos se configuran para formar un RAID 5, 5E y 5EE. De toda la capacidad de la matriz de discos, ¿qué porcentaje se dedica al almacenamiento de la información redundante?

```
RAID 5, deben ser 7 discos de capacidad y un disco de paridad. 1 / 8 = 12.5
RAID 5E, deben ser 7 discos de capacidad y el disco de repuesto ahora trabaja.
RAID 5EE, 7 deben ser 7 discos de capacidad y el disco de repuesto ahora trabaja
1 / 9 = 11,11%
```

RAID 6

¿Y si los discos se configuran como RAID 6? ¿Qué porcentaje se dedica ahora al almacenamiento de la información redundante?

```
7 discos de capacidad y 2 de paridad.
2 / 9 = 22,22%
```

RAID 1+0

Calcule de nuevo el número de discos necesarios si el sistema se configura como RAID 1+0. Dibuje un pequeño esquema de los discos implicados en esta configuración.

Si se necesitan 7 discos para el almacenamiento total, pues para el RAID 1+0 se necesita el doble para el respaldo completo, o sea 14 DISCOS

RAID 5+0

Por último, repita de nuevo el cálculo para una configuración RAID 5+0, donde cada RAID 5 está formado por un grupo de 3 discos. Dibuje un esquema de los discos en esta nueva configuración del sistema de almacenamiento.

7 discos / 2 = 4 niveles totales, 3 niveles completos y 1 un nivel con un solo disco y otro de paridad.
3 x 3 + 2 = 11 DISCOS

Estimación del rendimiento

En esta sección se incidirá en aquellos aspectos relacionados con las prestaciones del sistema de almacenamiento según su configuración en RAID.

Un disco

¿Cuánto tiempo tarda un disco de la serie Cheetah 15K.7 en leer un fichero de 8 GiB? Suponga que la velocidad de transferencia sostenida del disco es el promedio de los dos valores extremos publicados por el fabricante. Vale la pena remarcar que las velocidades de transferencia sostenidas de todos los discos de la serie son idénticas, independientemente de la capacidad y de la interfaz de salida (SAS y FC). Algo parecido ocurre con un gran número de parámetros del disco, como los relativos a la confiabilidad.

$8 \times 2^{30} / 163$ MB/s = 52,7 segundos

RAID 0

¿Cuánto tiempo tardaría un RAID 0 constituido por seis discos como el anterior en leer el mismo fichero de 8 GiB?

```
8 x 2^30 / 6 x 163 MB/s = 8,78 segundos
```

RAID 5

Repita el cálculo anterior suponiendo ahora que los seis discos se configuran como un RAID 5.

```
8 x 2^30 / 5 x 163 MB/s = 10,54 segundos
```

RAID 5E

Repita el cálculo anterior suponiendo ahora que los seis discos se configuran como un RAID 5E.

```
8 x 2^30 / 5 x 163 MB/s = 10,54 segundos
```

Según los resultados obtenidos en los apartados anteriores, ¿qué nivel resulta más rápido, RAID 0, RAID 5 o RAID 5E? ¿A qué se debe esta diferencia? Cuantifique la aceleración obtenida por la opción más rápida respecto de las otras dos.

```
El RAID 0 es más rápido que el RAID 5 o RAID 5E, esta diferencia se debe a que en
el 5E los 6 discos trabajan en la lectura mientras que en el resto trabaja 1 disco
menos.
```

Aspectos de confiabilidad

Las especificaciones sobre confiabilidad de los discos dadas por el fabricante son variadas y atienden a diversos parámetros. Por ejemplo, uno de ellos cuantifica los errores no recuperables de lectura (UBER, Unrecoverable Bit Error Rate). También se suele proporcionar la tasa anual de fallos (AFR, Annualized Failure Rate) y el tiempo medio entre

fallos (MTBF, Mean Time Between Failures); en ocasiones, este último parámetro es reemplazado por el tiempo medio para el fallo (MTTF, Mean Time To Failure).

Ahora bien, es muy importante tener en cuenta que, en cualquier caso, se trata de valores estadísticos estimados haciendo pruebas con una población concreta de discos (por ejemplo, varios centenares). Por esta razón, el uso de estos indicadores debe tratarse con cautela cuando se intenta predecir el comportamiento de un disco individual.

Finalmente, no debe obviarse que el número de años de garantía también puede ser un valor muy importante a tener en cuenta a la hora de valorar la confiabilidad del disco porque obliga legalmente al fabricante a hacerse cargo de cualquier problema en su funcionamiento.

AFR, Annualized Failure Rate

Consulte las especificaciones de los discos Cheetah 15K.7 publicadas por el fabricante en el manual del producto e indique el valor del AFR y las condiciones en las que se ha estimado.

```
AFR = 0,55%
• 8,760 power-on hours per year.
• 250 average on/off cycles per year.
• Operations at nominal voltages.
• Systems will provide adequate cooling to ensure the case temperatures specified
in Section 6.4.1 are not exceeded.
```

En la práctica, ¿qué parece sugerir el sentido común sobre la influencia de la temperatura de operación del disco en su AFR?

```
Que el disco cumplirá con su AFR en las temperaturas óptimas, y que si estas no se
cumplen, aumenta el AFR.
```

Sin embargo, ¿a qué conclusión se llega en el estudio de Pinheiro, Weber y Barroso, Failure trends in a large disk drive population, llevado a cabo en Google en el año 2007?

```
Se llegó a la conclusión de que no hay una relación consistente entre el aumento de
fallos cuando aumentamos la temperatura de trabajo.
```

Relación entre AFR y MTBF

En los modelos de confiabilidad más sencillos se asume que el tiempo para el fallo sigue una distribución exponencial, lo que permite estimar la probabilidad de fallo de un disco magnético con la fórmula:

$$P[t \leq T] = 1 - e^{-\frac{1}{MTBF}}$$

A partir de la fórmula anterior, ¿qué relación hay entre los valores de AFR y MTBF?

```
La relación es que el AFR es el valor de la fórmula anterior.
```

Compruebe numéricamente que el valor del AFR publicado por el fabricante se puede obtener sustituyendo en la fórmula anterior el parámetro MTBF por el valor particular para este disco (1.600.000 horas).

```
AFR = 8760 horas / 1.600.000 horas
AFR = 0,005475 x 100
AFR = 0,55%
```

En vez utilizar la fórmula anterior, que incluye una operación de exponenciación, en algunas ocasiones se suele utilizar una relación matemática mucho más sencilla entre AFR y MTBF:

$$AFR \approx \frac{8760}{MTBF}$$

Utilizando la fórmula anterior, ¿cuál sería ahora el valor estimado del AFR? ¿Qué diferencia existe entre los dos valores calculados del AFR?

```
AFR = 8760 horas / 1.600.000 horas
AFR = 0,005475 x 100
AFR = 0,55%
```

Reemplazamiento de discos

Considere un sistema de almacenamiento de grandes dimensiones implementado con un total de 500 discos Cheetah 15K.7. ¿Cuántos discos habrán tenido que ser reemplazados, por término medio, al cabo de un año de funcionamiento?

```
500 x AFR = 500 x 0,55% = 2,75 discos
3 DISCOS DAÑADOS
```

Suponiendo que el AFR se mantiene constante durante los primeros cinco años, ¿cuántos discos se habrán reemplazado al final de este periodo de tiempo en el sistema anterior?

```
500 x AFR x 5 años = 500 x 0,55% x 5 = 13,75 discos
14 DISCOS DAÑADOS
```

Suponga ahora que, en el sistema anterior, el AFR evoluciona de la siguiente manera:

- El AFR es tres veces mayor en el primer mes de funcionamiento,
- A partir del quinto año se duplica cada año.

¿Cuántos discos se habrán reemplazado, por término medio, después de 7 años de funcionamiento?

```
500 x 0,55% x ( (3x1/12) + 11/12 + 1 + 1 + 1 + 1 + 2 + 4) ) = 30,71 discos
31 DISCOS DAÑADOS
```

Errores no recuperables

¿Cuántas veces se puede leer completamente un disco magnético modelo ST3600057SS antes de que experimentemos un error de lectura no recuperable?

```
1 error / 10^16 bits leidos
600 GB x 8 bits / 10^16 bits leidos = 0,48x10^-3 errores
1 / 0,48x10^-3 errores = 2083,33 veces
2083 veces
```

Proceso de reconstrucción

Considere un RAID 5 compuesto por 8 discos como el indicado en el apartado anterior (modelo ST3600057SS). En un momento determinado un disco de la matriz falla e, inmediatamente, es reemplazado por un disco de repuesto (on line spare).

En esta situación, el sistema inicia el proceso de reconstrucción de la información del disco estropeado en el disco de repuesto. Durante este proceso suponemos que el sistema solo se dedica a la reconstrucción y no admite peticiones externas hasta su finalización. Estime una cota inferior del tiempo que se tarda en reconstruir la información del disco perdido.

```
600 GB / 122 MB/s = 4918 segundos
```

¿Cuántos errores no recuperables de lectura se estima que pueden acontecer, por término medio, durante el proceso anterior de reconstrucción?

600 GB x 8 bits / 10^{16} bits = 0,48 x 10^{-12} errores