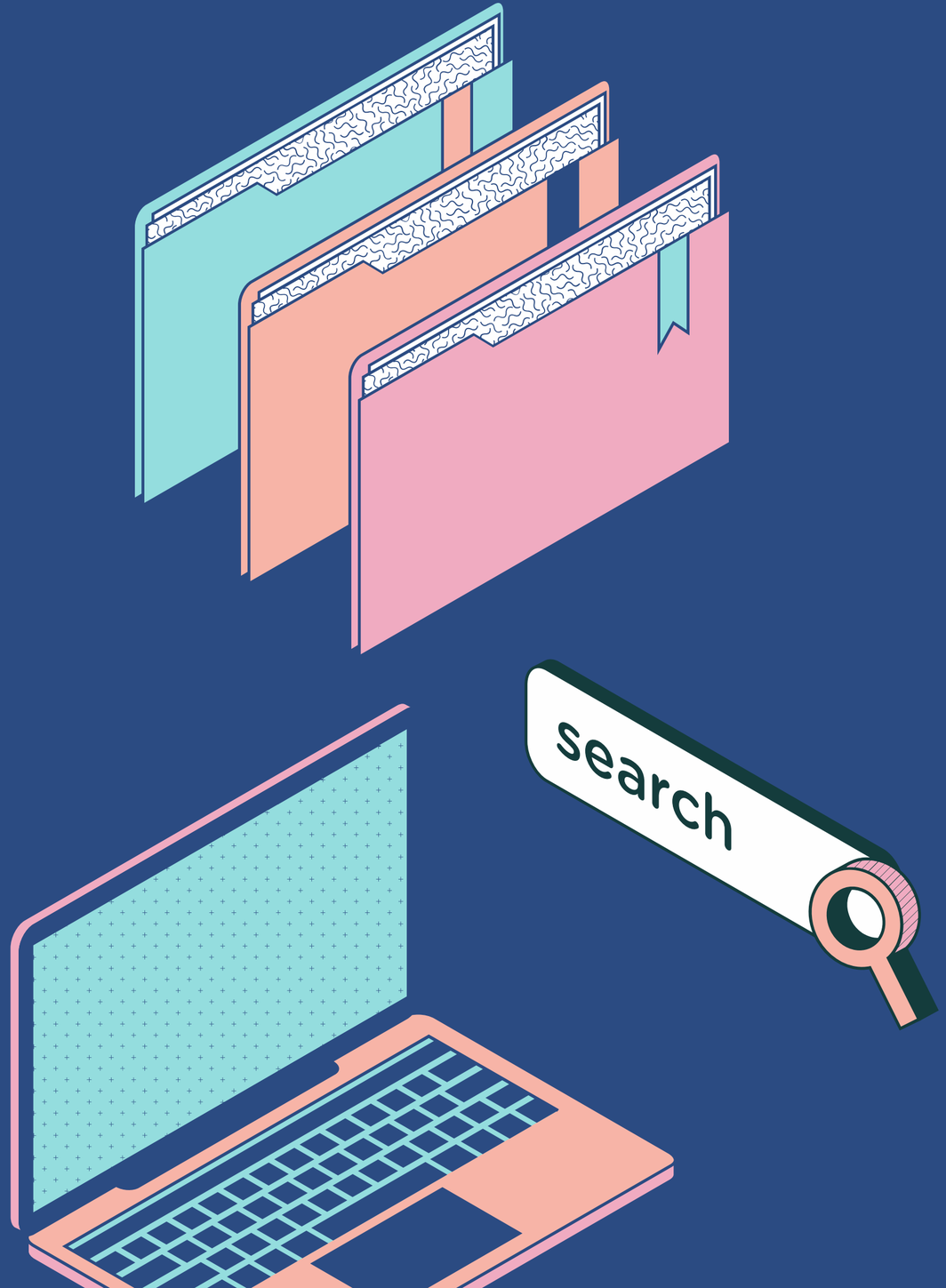


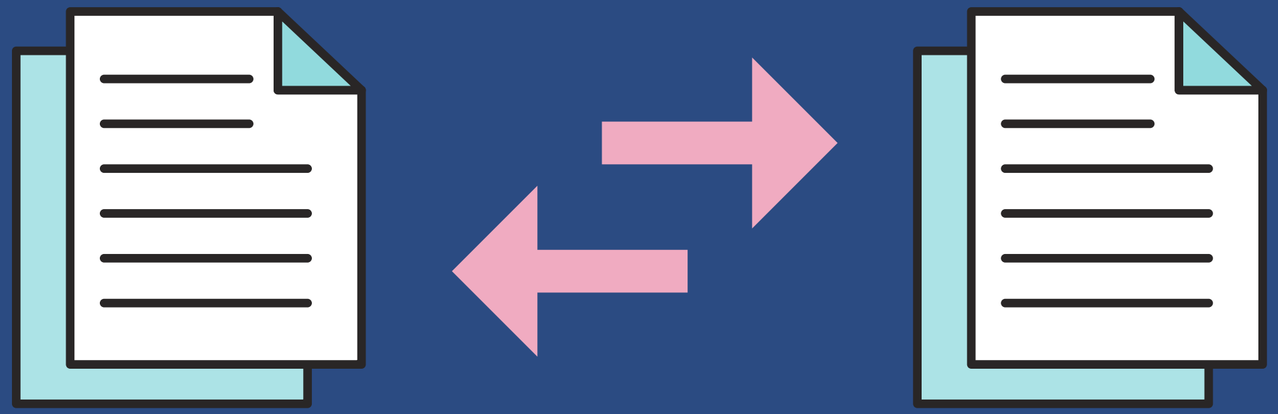
SISTEMAS OPERATIVOS

Deduplicación de Datos

Jessica Zepeda Baeza

Puntos a desarrollar

- ¿Qué es?
- Clasificaciones de la deduplicación
- ¿Cómo funciona?
- Implementación
- Compresión de Datos y SIS
- Puntos en Contra



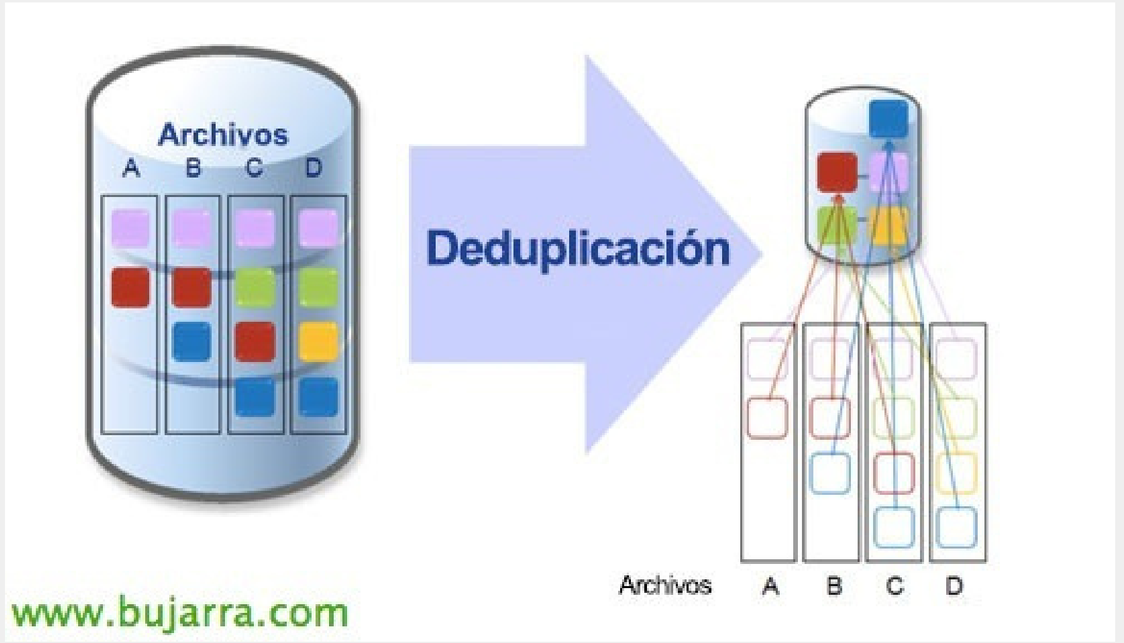
¿Qué es?

Proceso que reconoce la repetición de datos y la elimina o sustituye por una referencia a los datos.

Objetivo:
Optimizar espacio eliminando redundancia de datos.

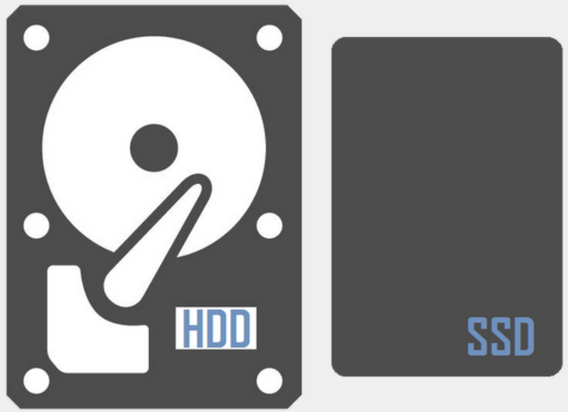
Consiste en:

- Almacenamiento
- Índice



¿Qué es?

Realizado en almacenamiento a largo plazo:
discos, unidades SSD



Chunking: segmentación de datos en bloques

menor tamaño de bloque
mayor deduplicaión
menor desempeño

mayor tamaño de bloque
menor deduplicaión
mayor desempeño

¿Qué es?

Ventajas

- Aumenta tiempo de vida del disco
- Mayor eficiencia de espacio
- Facilita el proceso del recolector de basura

Desventajas

- Rendimiento
- Confiabilidad

Importancia

Debido a la genración de datos y almacenamiento: **Backups**

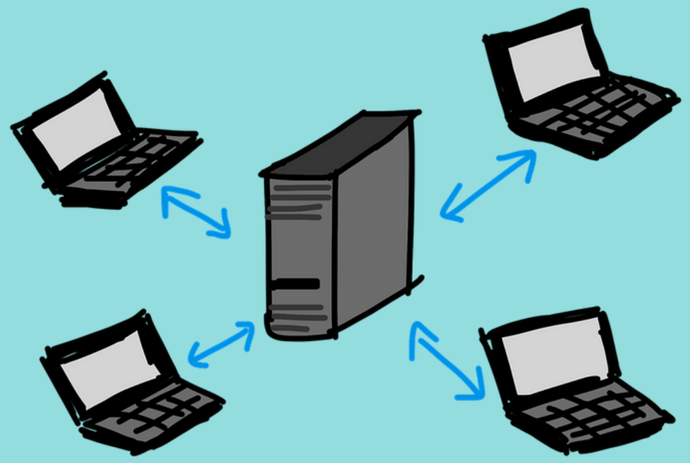
Clasificaciones de la deduplicación

- **En origen / En destino**

En origen:
Se realiza en el sistema de origen, se transmite al sistema de destino
Se evita el paso de datos duplicados
Ahorro de volumen enviado

En destino:
Aquel con almacenamiento en disco

- En vivo / Inline / Online
- Después del hecho / Offline / Post process



Clasificaciones de la deduplicación

- **En vivo / Después del hecho**

Inline / Online:
Se realiza al recibir los datos, antes de escribirlos en el disco
Previene escritura en disco de datos duplicados
Aumenta tiempo de escritura

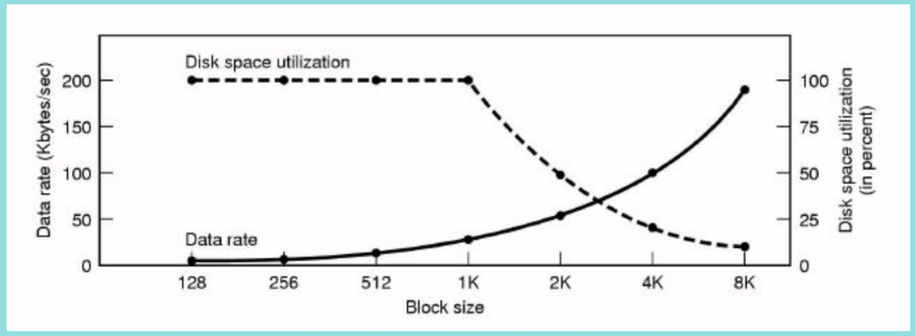
Offline:
Se copian todos los datos en disco y se hace el proceso en tiempo de inactividad

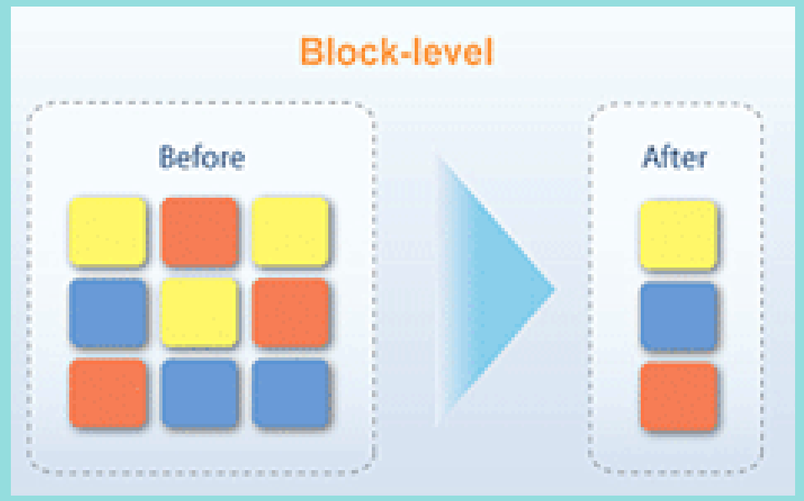
Clasificaciones de la deduplicación

- **Por archivo / por bloque**

Por archivo (File):
Se hace el análisis a nivel de archivo

Por bloque (Subfile):
Tamaño fijo y predecible
Aprovechamiento vs Transferencia





Clasificaciones de la deduplicación

- **Con / sin conocimiento de contenido**

Con conocimiento (Content-Aware):
Se tiene acceso a los metadatos (estructura de datos)
Se identifican datos duplicados más rápido

Sin conocimiento:
Mayor posibilidad de datos de entrada

Clasificaciones de la deduplicación

- **Por algoritmo**

Hashing

SHA-1/2
MDA5

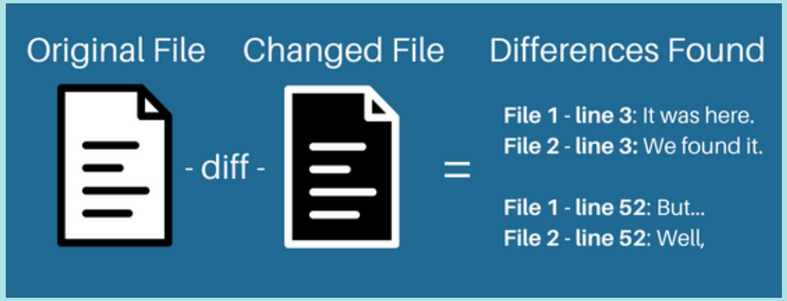
Whole File Hashing

Fixed Block Hashing
Variable Block Hashing

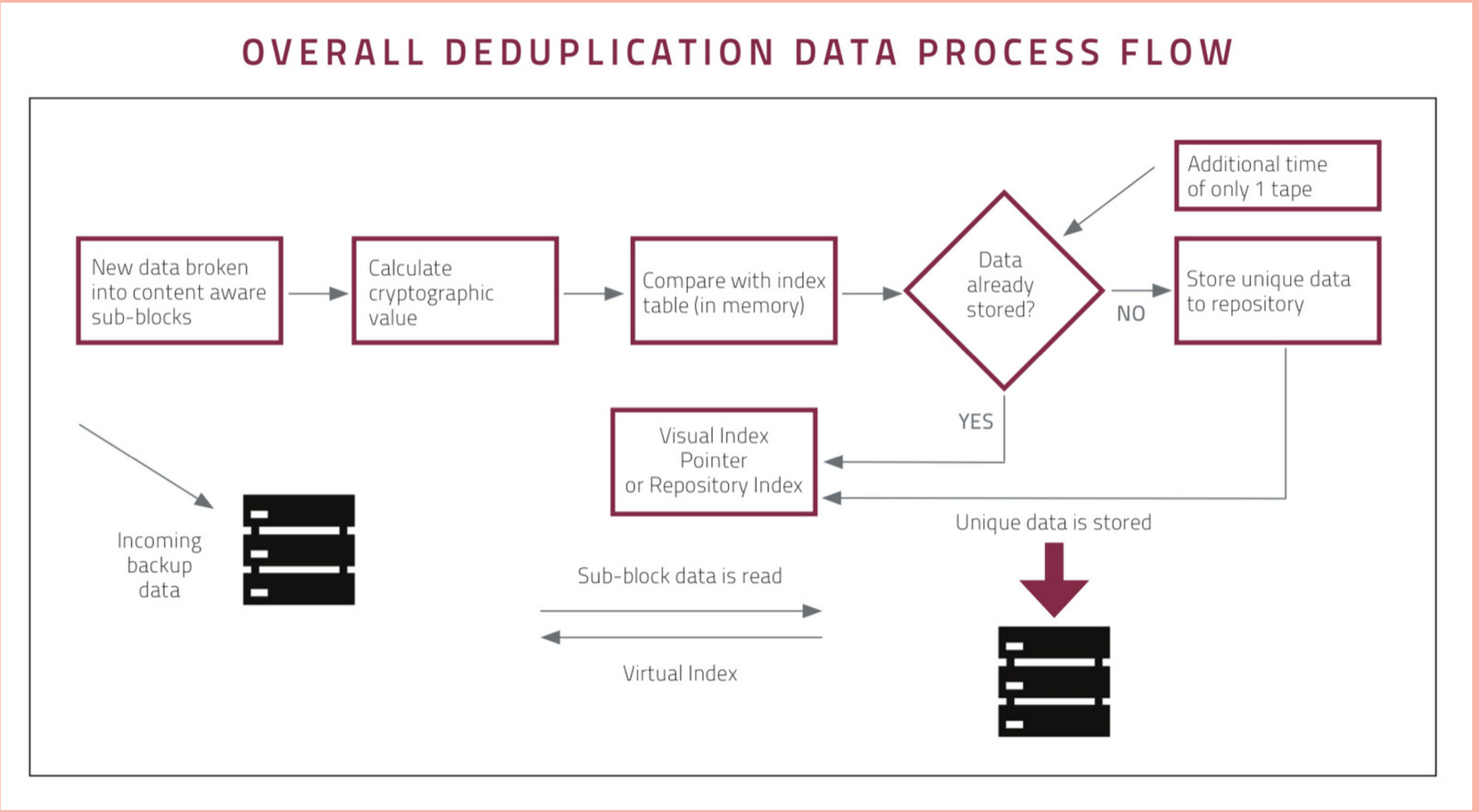
Se aplica el hash y si se encuentra duplicado se compara byte a byte para evitar **colisiones**

Delta Encoding

Se genera una delta o "patch" con las diferencias entre dos archivos



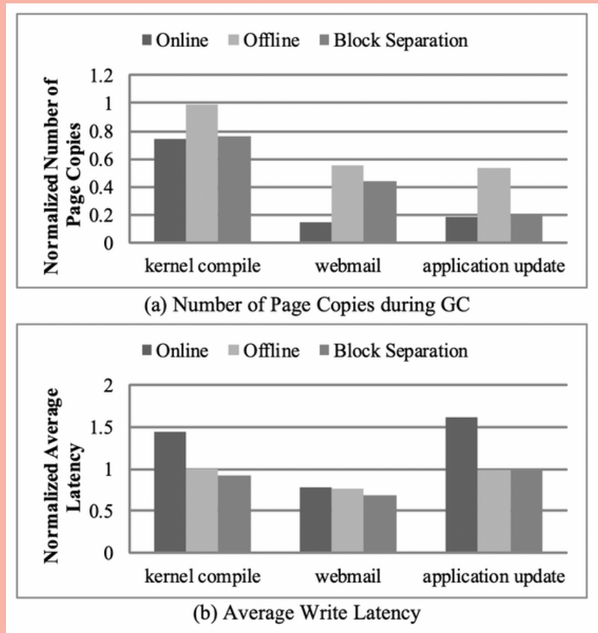
¿Cómo funciona la deduplicación?

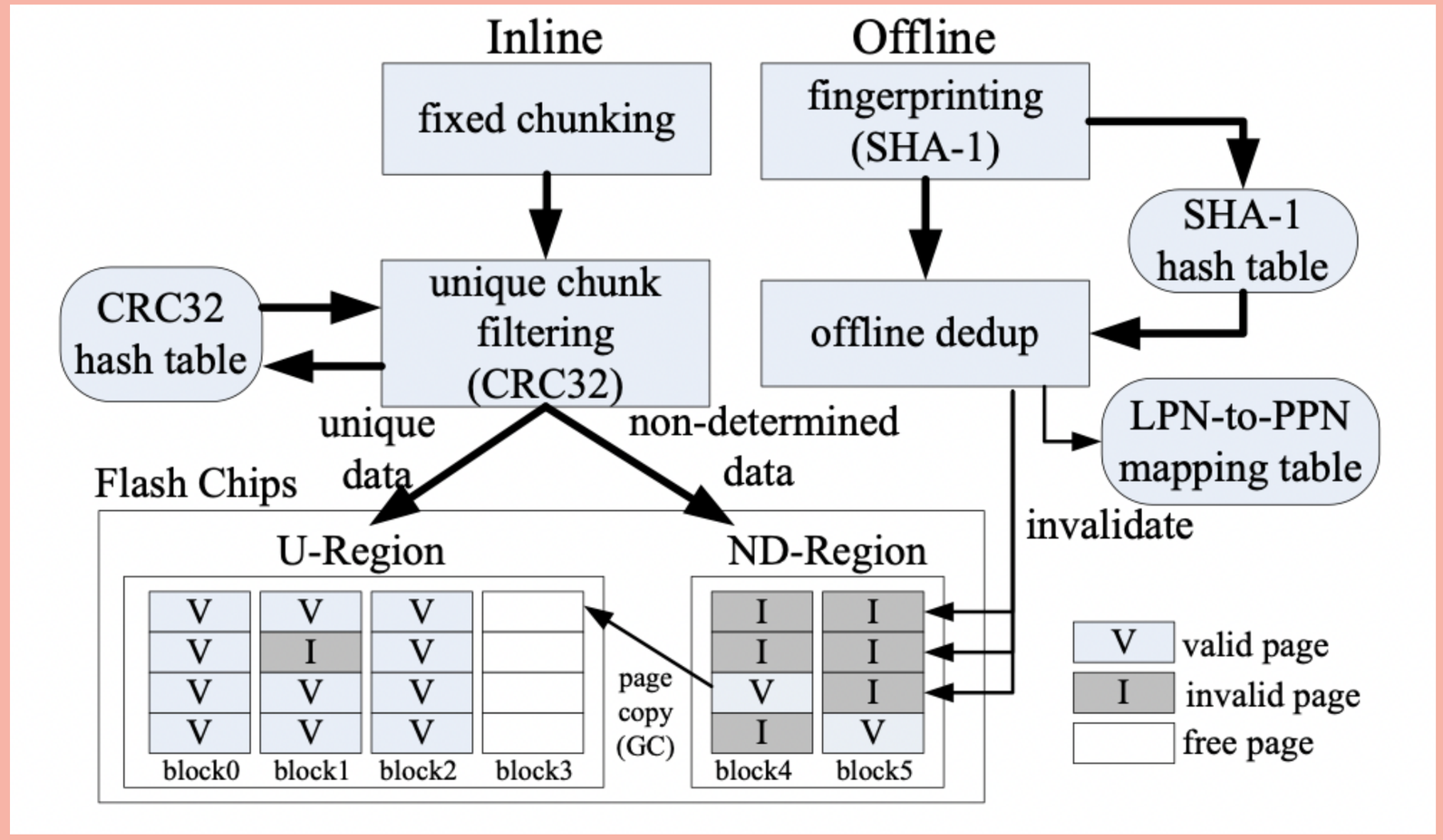


¿Cómo funciona la deduplicación?

Block Segmentation

Mayor eficiencia de tiempo y espacio





Implementación

Sistemas de Backup

Debido a la duplicación de datos y uso del disco.

Máquinas Virtuales

Se comparten bloques de datos grandes con el Sistema Operativo
Mayor beneficio si las máquinas virtuales emplean el mismo SO

Sobrecompromiso (overcommitment)

Asignar más almacenamiento para una máquina virtual del que se tiene.

Compresión de datos y

Objetivo:
Decodificar la información de ciertos archivos a menos bits.
Optimiza espacio y duplicación de patrones dentro de un archivo y entre archivos.

Diferencia:
Descompresión

Single Instance Store

Almacenamiento de instancia única

Objetivo:
Identifica archivos idénticos y los reemplaza por una referencia.
Predecesor de la deduplicación
Deduplicación por archivo

Microsoft Exchange Server Enterprise Edition

1997 → 2007

Puntos en contra

1 Confiabilidad

Daño en un bloque altamente refrenciado implica daños en varios archivos

2 Rendimiento

En archivos con bloques duplicados
Bloques no contiguos
Recuperación de archivo más tardado
Más desplazamiento de la cabeza lectora

Referencias

- An, J., y Shin, D. (2013). Offline deduplication-aware block separation for solid state disk. Descargado de https://www.usenix.org/system/files/fastpw13-paper6_0.pdf
- IONOS. (2021). Reducción de datos por deduplicación o por compresión. IONOS Digital Guide. Descargado de https://www.ionos.mx/digitalguide/servidores/know-how/ reduccion-de-datos-deduplicacion-o-compresion/
- Mandagere, N., Zhou, P., Smith, M. A., y Uttamchandani, S. (2008). Demystifying data deduplication. Descargado de http://cs.brown.edu/courses/cs227/archives/2016/ papers/p12-mandagere.pdf#page9
- Microsoft. (2022). Understanding data deduplication. Descargado de https://docs.microsoft.com/en-us/windows-server/storage/data-deduplication/ understand
- Patricio, F. J. J. (2009). Técnicas de deduplicación de datos y aplicación en librerías virtuales de cintas. Lenguajes y Sistemas Informáticos e Ingeniería del Software. Descargado de https://oa.upm.es/1803/1/PFC FRANCISCO_JAVIER_JIMENEZ_PATRICIO.pdf
- Shin, D. (2015). Offline deduplication with lightweight hash for solid state disk. The 22nd Korean Conference on Semiconductors. Descargado de http://nyx.skku.ac.kr/wp-content/uploads/2014/07/offline_deduplication_with_lightweight_hash_for_ssd.pdf
- Solís, D. G. M. (2015). Análisis de métodos de deduplicación de datos aplicados en repositorios linux para la facultad de ingeniería en sistemas electrónica e industrial. UNIVERSIDAD TÉCNICA DE AMBATO. Descargado de http://repositorio.uta.edu.ec/bitstream/123456789/19367/1/Tesis_t1081si.pdf