Стабилни складишта и SSD

Оперативни системи 2024

проф. д-р Димитар Трајанов,
проф. д-р Невена Ацковска,
проф. д-р Боро Јакимовски,
проф д-р Весна Димитрова,
проф. д-р Игор Мишковски,
проф. д-р Сашо Граматиков,
вонр. проф. д-р Милош Јовановиќ,
вонр. проф. д-р Ристе Стојанов,
доц. д-р Костадин Мишев

"Ss. Cyril and Methodius" University - Skopje
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

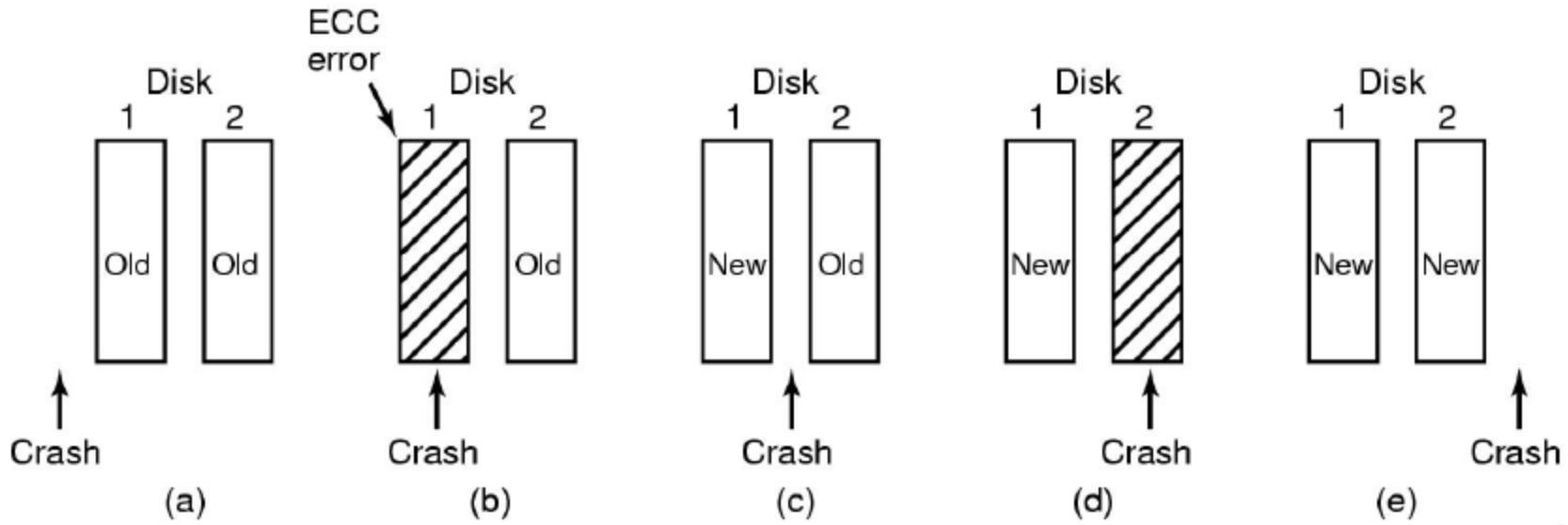
Стабилни складишта

- Кога дисковиот систем треба да запише или точно ги запишува податоците или не запишува ништо
- Причини:
  - ECC не е доволно
  - Самиот сектор може со тек на време да се расипе
  - Може да падне процесорот
- Се користи пар на идентични дискови
- Софтверски имплементиран

Стабилни складишта

- Стабилни запишувања
  - Се запишуваат и проверуваат податоците на двата диска.
  - Најпрво на дискот 1, а потоа на дискот 2.
  - Се прави тоа n пати, при неуспех се ремапира секторот на помошен сектор и се повторува операцијата
- Стабилни читања
  - Се чита првин блокот од диск 1 и се проверува ECC
  - Доколку по n пати се воспостави дека има грешка, се чита од вториот диск
- Оporавување по пад
  - Двата блока се во ред и се исти
  - Едниот има грешка – читај од другиот и запиши
  - Двата блока се во ред, но се различни, се запишува блокот од дискот 1 на дискот 2

Стабилни складишта



Анализа на влијанието на падовите врз стабилните запишувања

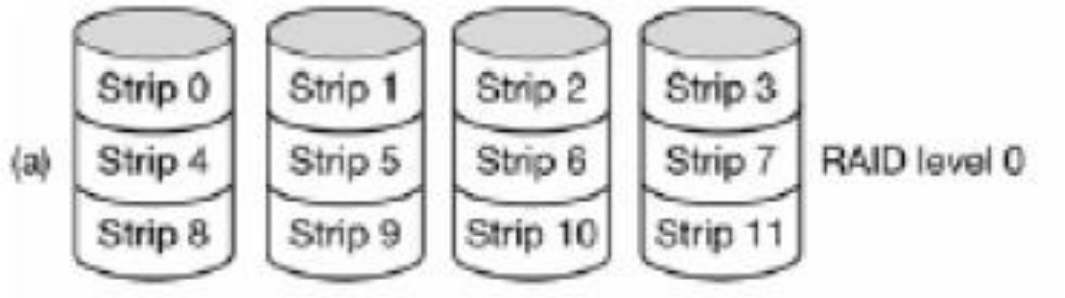
RAID

- За да се задоволат барањата на пазарот за се поголеми брзини и паралелна обработка на податоците, воведени се неколку различни архитектури за организација на дискови – RAID (Redundant Array of Inexpensive/Independent Disks).

- Основната цел на оваа архитектура е повеќе дискови да бидат меѓусебно поврзани на некој начин овозможувајќи редундантност и надежност при што компјутерот ги гледа како една целина не знаејќи како тоа се постигнува.

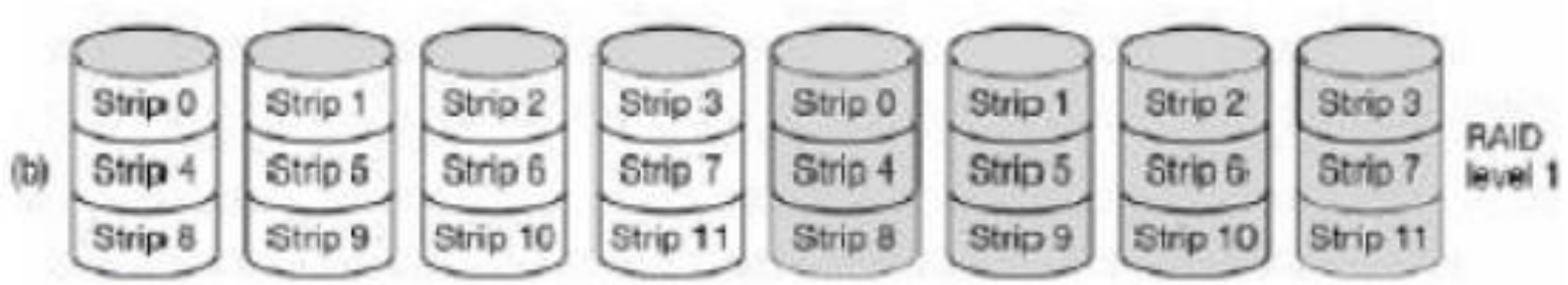
RAID 0 – Stripe Set

- Секој stripe содржи k сектори
- Се користи за зголемување на перформансите – паралелен I/O.
- Капацитет зависи од најмалиот диск (HD1 = 100GB, HD2 = 120 GB, ВКУПНО = 200GB
- Негативен параметар, MTTF = MTTF/#HD.



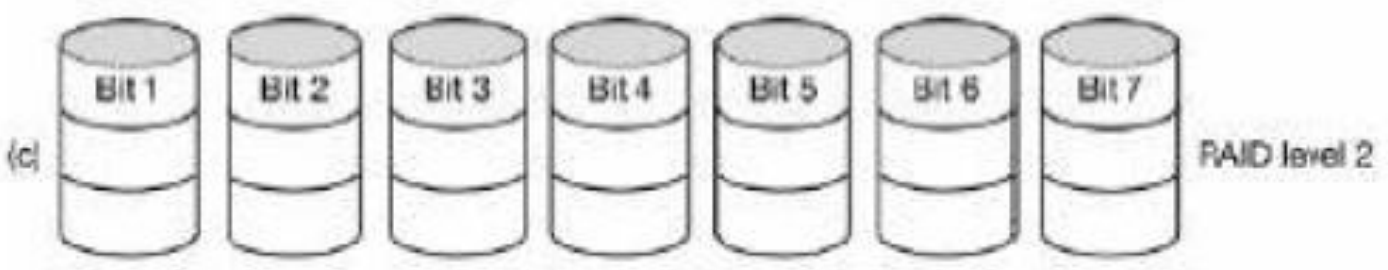
RAID 1

- Постојат дупликати на сите дискови (4 примарни + 4 backup дискови).
- При запишување, секој strip се запишува двапати (слаби перформанси).
- При читање може да се користи било која копија (добри перформанси).
- Одлична отпорност на грешки.



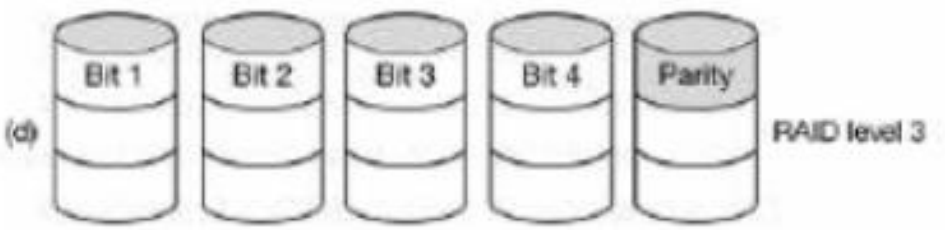
RAID 2

- Податоците се делат на ниво на бит и се користи Хамнингов код корекција на грешка.
- Возможни се големи рати на пренос, но не се користи често.
- Би биле потребни 39 диска (32 диска за збор, 7 за корекција на грешка).
- Бара синхронизација на ротација на дисковите



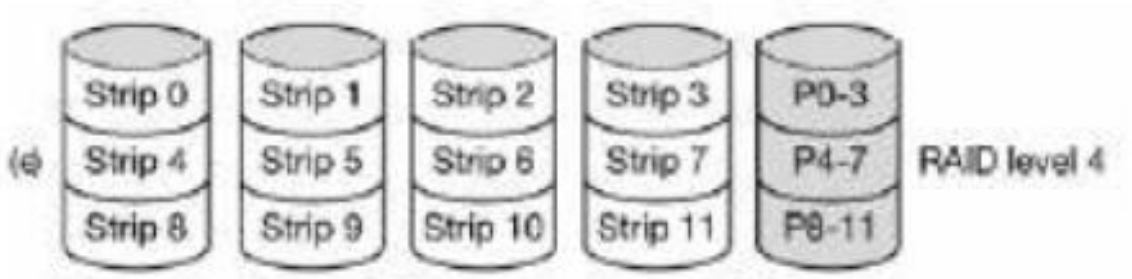
RAID 3

- Делење на ниво на бајт со посветен диск за парност.
- Многу ретко се користи во пракса
- Може да се направи и корекција на грешка
- Бројот на различни I/O барање во секунда е многу мал.



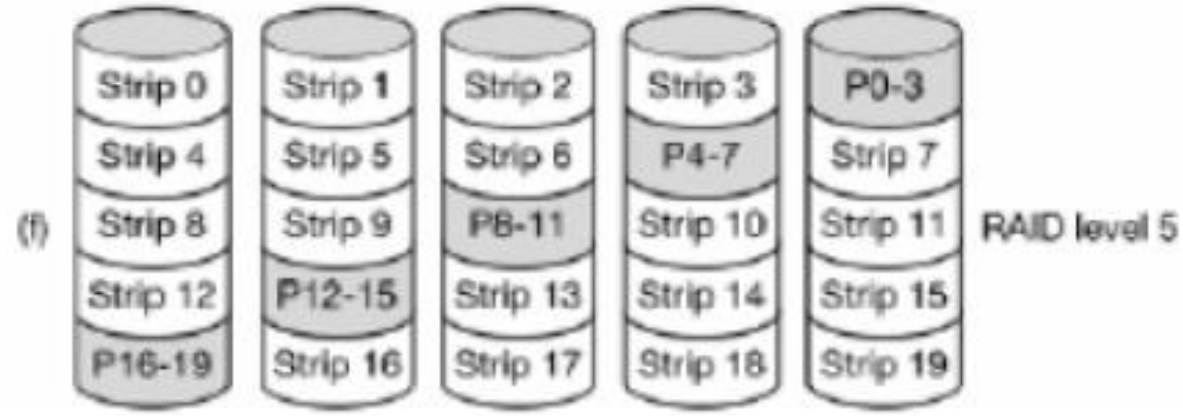
RAID 4

- Делење на ниво на блокови со посветен диск за парност
- Може да сервисира повеќе барања истовремено.
- Блокот за парност се добива со XOR од останатите 4 блока.
- Ако некој диск откаже, загубените бајти се добиваат од соодветниот блок за парност.
- Негативности: При промена на еден сектор мора да се исчитаат сите дискови и да се ажурира соодветниот блок за парност



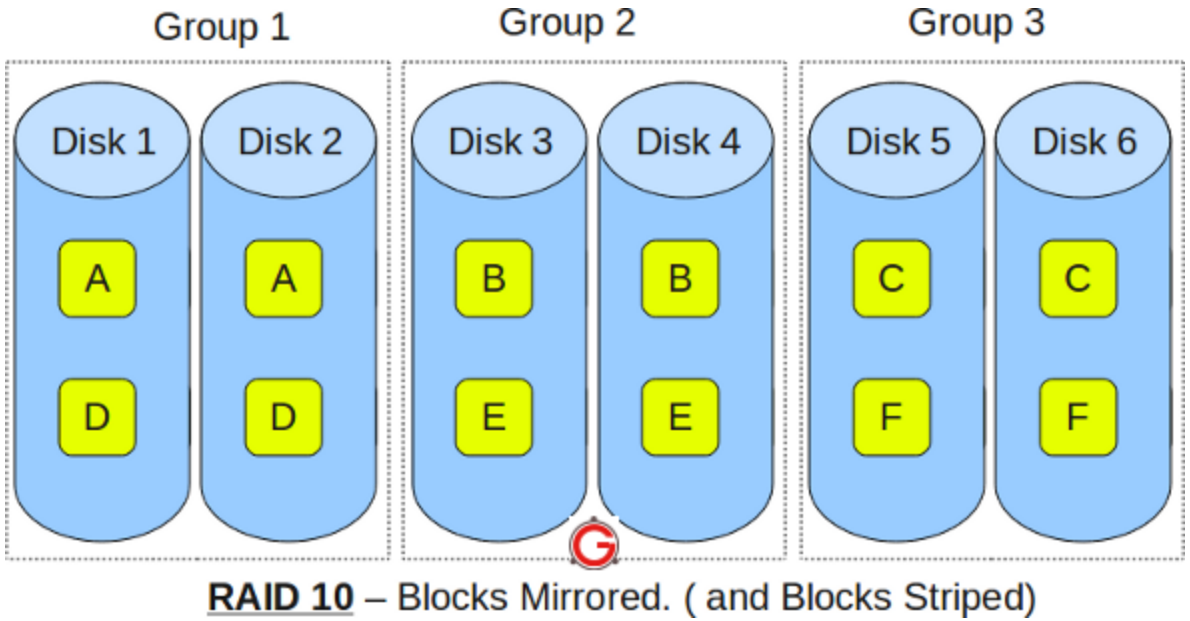
RAID 5

- Битовите за парност се дистрибуираат на Round Robin начин
- Дискот со парност не е повеќе тесно грло
- Покомплексен е процесот на реконструирање на содржината при откажување на еден од дисковите



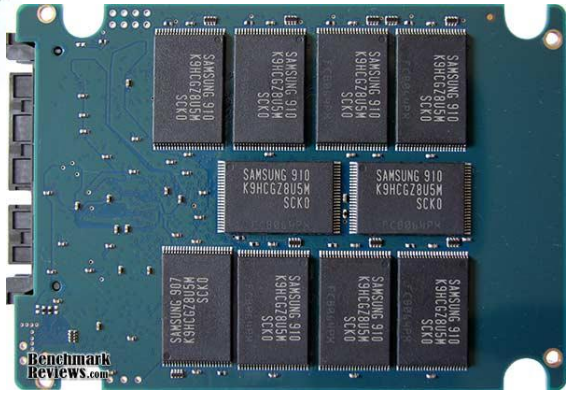
RAID 10 се вика и RAID 1+0

RAID 10



- Му требаат најмалку 4 диска
- Се групираат дисковите во парови како огледални дискови
- За 6 диска во RAID 10, ќе има три групи –Група 1, Група 2, Група 3
- Во една група, податоците се огледални – Диск 1 и Диск 2 се Група 1
- Во групата, податоците се распружени, т.е. Block A се запишува во Група 1, Block B во Група 2, Block C во Група 3.
- Се нарекува “stripe of mirrors”. Т.е дисковите во групата се огледални, а групите се распружени.

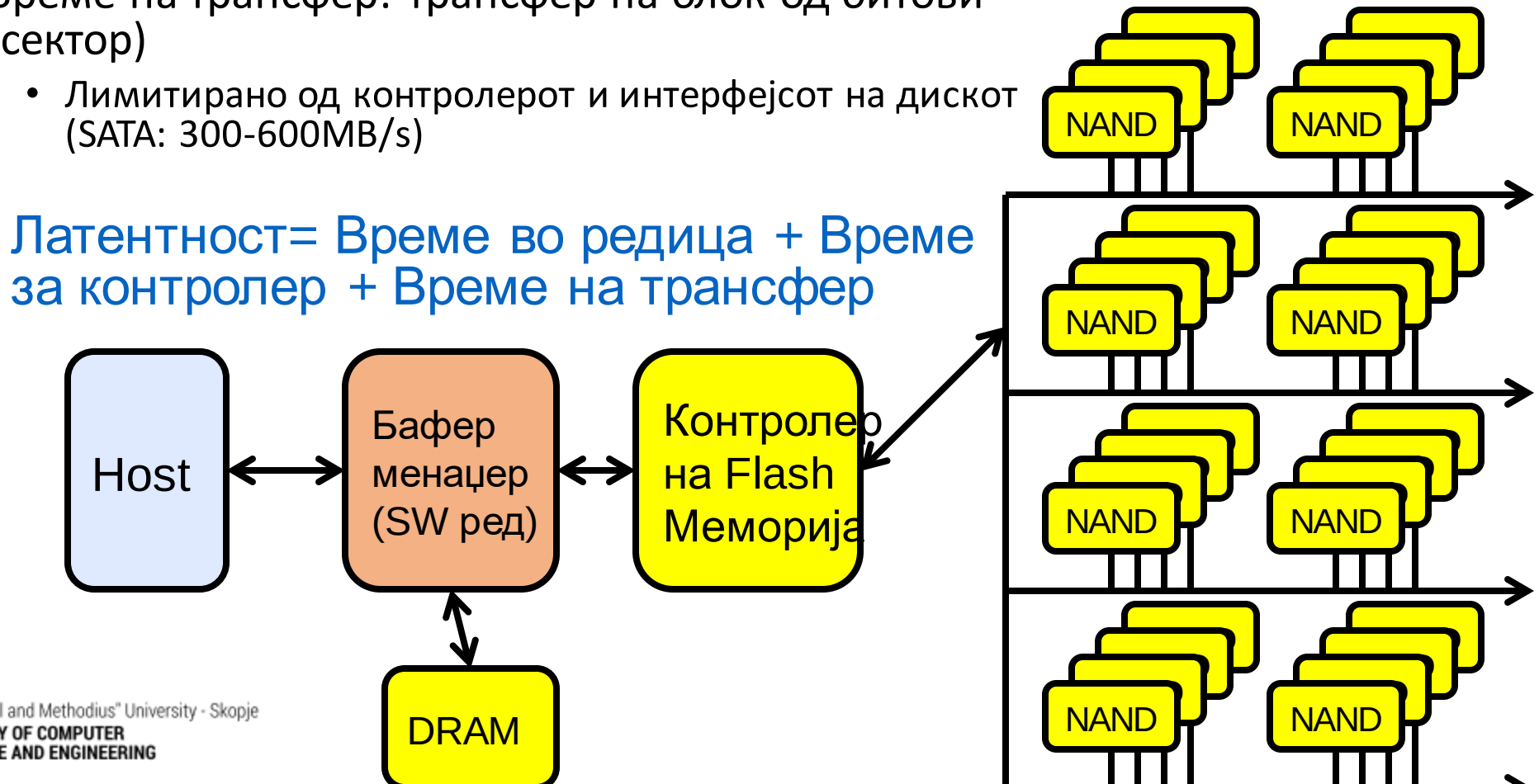
Solid State Disks (SSDs)



- 1995 – Се заменуваат ротирачкиот магнетен медиум со non-volatile меморија (DRAM подржан со батерија)
  - Од 2009, се користи NAND Flash: Single Level Cell (1-bit/cell), Multi-Level Cell (2-bit/cell)
- Сектор адресибилни, но се чуваат 4-64 “сектори” во една мемориска страна
- Нема подвижни делови (нема мотор за ротација и пребарување)
  - Се елиминираат соодветните доцнења (0.1-0.2ms време на пристап)
  - Мала потрошувачка и мали димензии

SSD Архитектура– Читања

- Читањето на податоци е слично со читање од меморија (25µs)
- нема доцнење поради пребарување или ротација
- Време на трансфер: трансфер на блок од битови (сектор)
  - Лимитирано од контролерот и интерфејсот на дискот (SATA: 300-600MB/s)



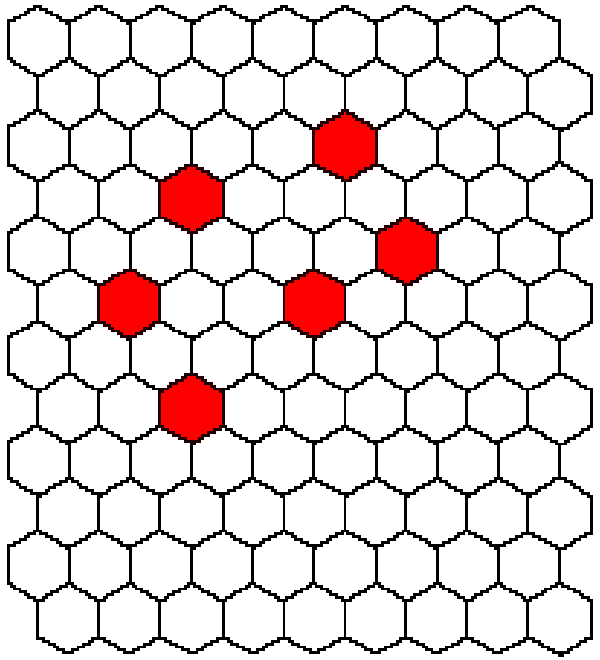
SSD Архитектура – Запишувања

- Запишување е комплексно! (~200µs – 1.7ms)
  - Може да се запишува на празни страни (бришењето ~1.5ms)
  - Контролерот има базен на празни страни, ги спојува користените сектори, исто така резервира одреден % од капацитетот
- Запишувањето и бришењето бара висок напон
  - Се оштетуваат мемориските ќелии, се ограничува животниот век на SSD
  - Контролерот користи ECC, и користи wear leveling
  - Латентност = Време во редица + Време за контролерот (Да пронајде слободни блокови) + Време на трансфер

Правило: запишувањето е 10 пати поскапо од читањето, бришењето е 10 пати поскапо од запишувањето

Како се запишува на SSD

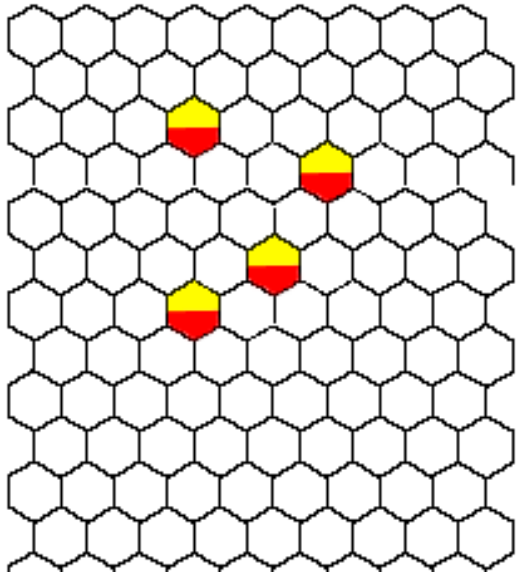
- За SLC (ќелии со едно ниво), ќелијата е или On или Off.



- Запишувањето кај ваков тип на SSD е полесно. Има помалку работа за контролерот

Како се запишува на SSD

- За MLC (ќелии со две нивоа), ќелијата е или On On, On Off, Off On или Off Off. Двојно повеќе капацитет

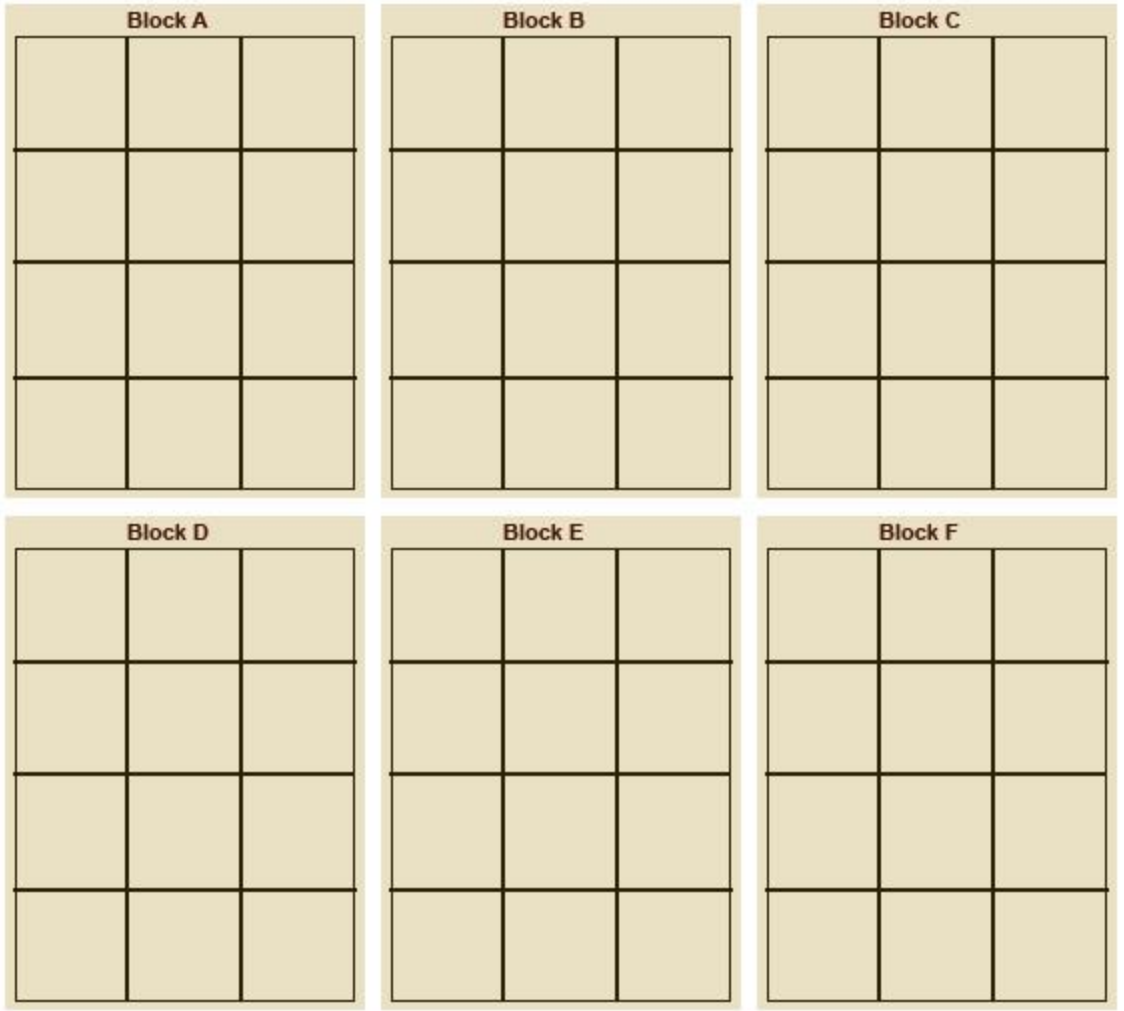


Како се запишува на SSD

- SSD е поделено на блокови, во кои се чуваат страници.
- Страниците се составени од соседни ќелии од NAND Flash меморијата
- Блоковите чуваат страници и бројот на блокови е одреден од големината на SSD.
- Страниците се со големина од 4KB и блоковите содржат 64 страници.
- Податоците се бришат во 256KB (64*4) блокови.

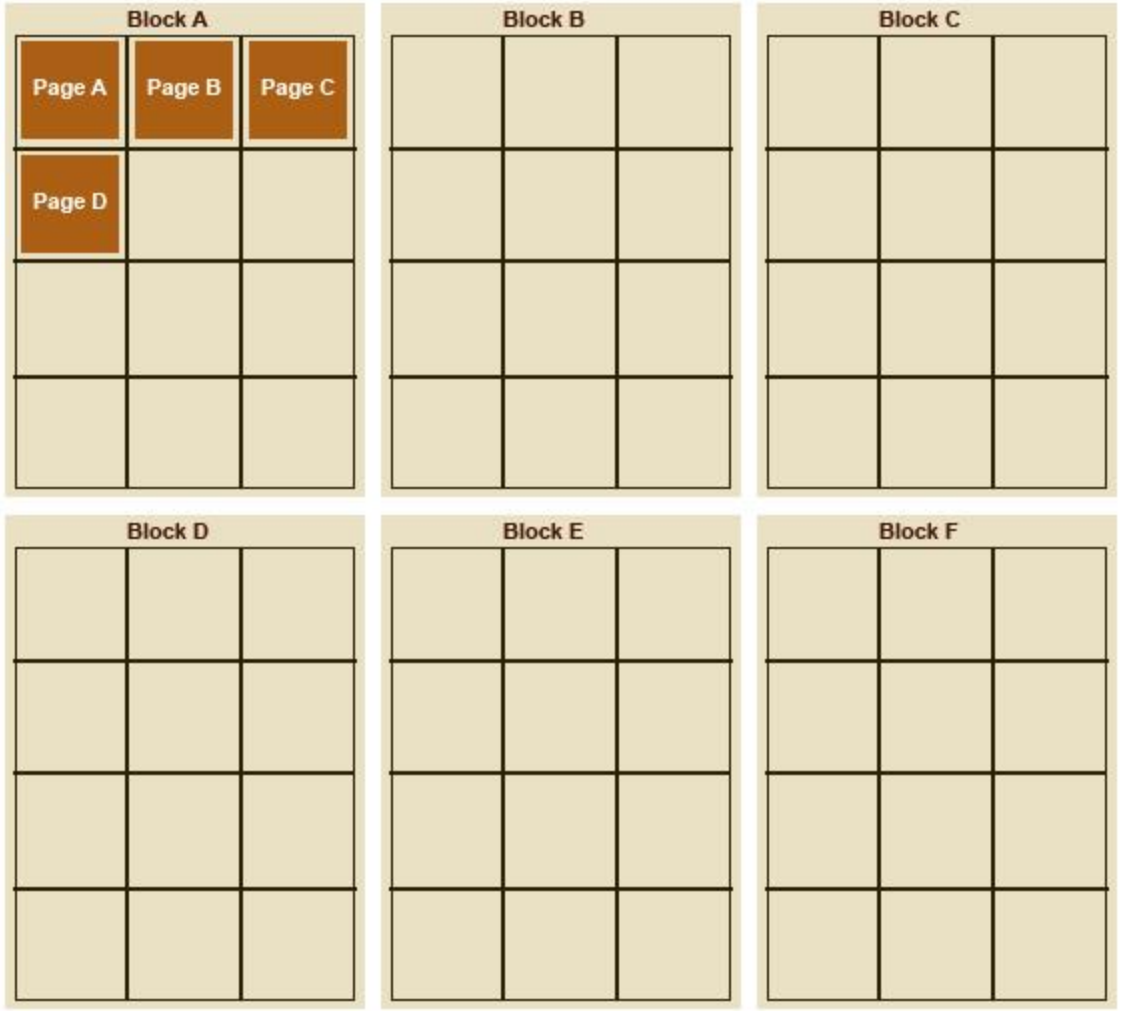
Како се запишува на SSD

- За илустрација, нека блоковите се со големина од 12 страни и секоја страна е со големина од 1 бајт



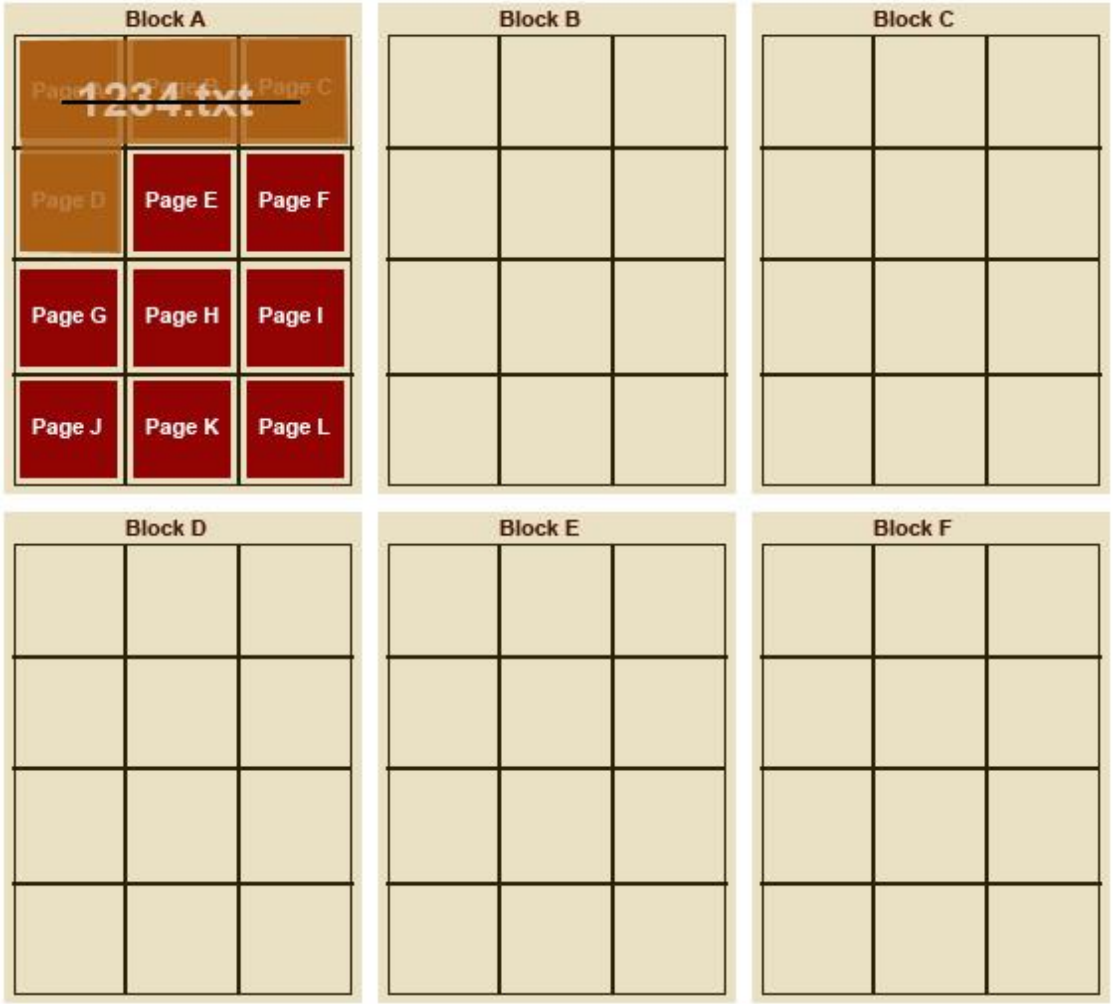
Како се запишува на SSD

- Да претпоставиме дека сме креирале датотека со големина од 4B, тоа значи со големина од 4 страни



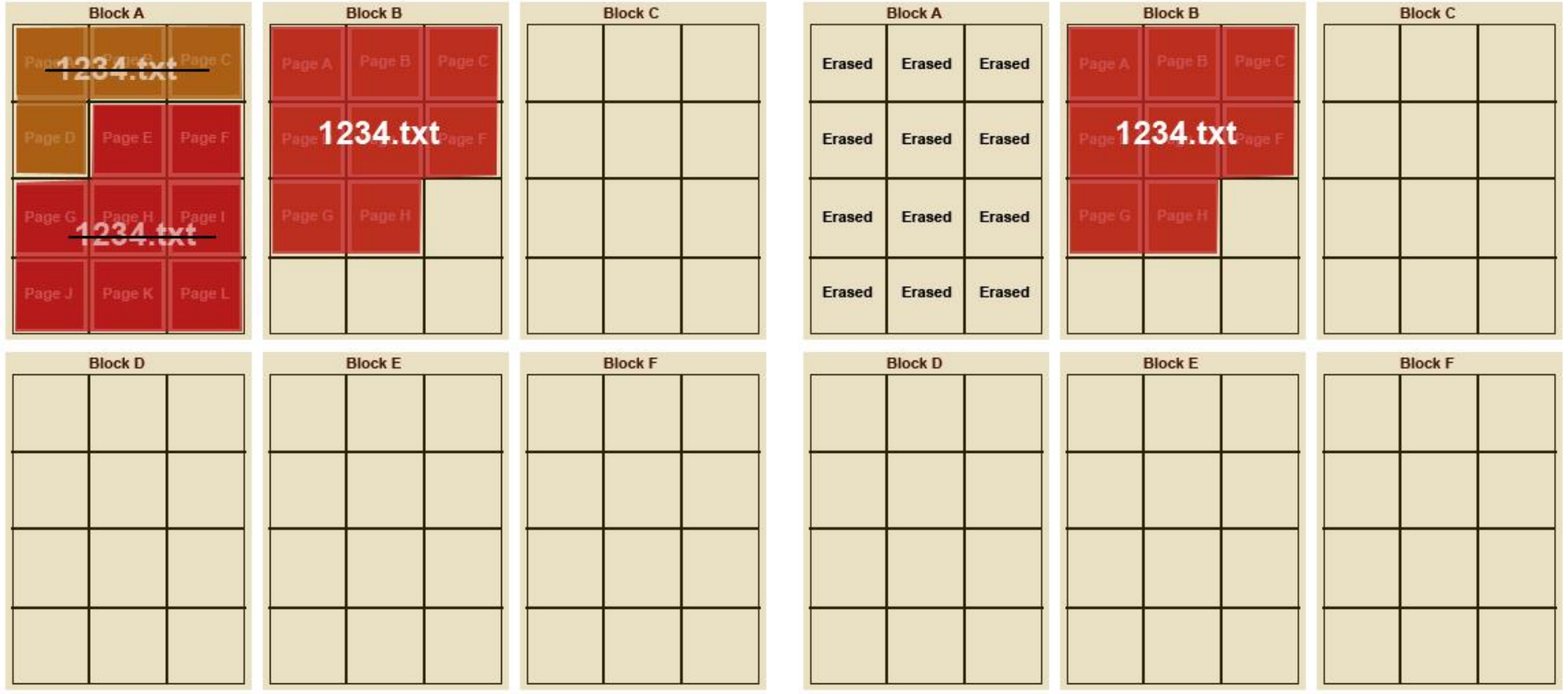
Како се запишува на SSD

- Доколку ја измениме датотеката и додадеме текст со големина од 8B.
- Старите страници се означуваат за бришење, додека, пак, новите страници се запишуваат на блокот



Како се запишува на SSD

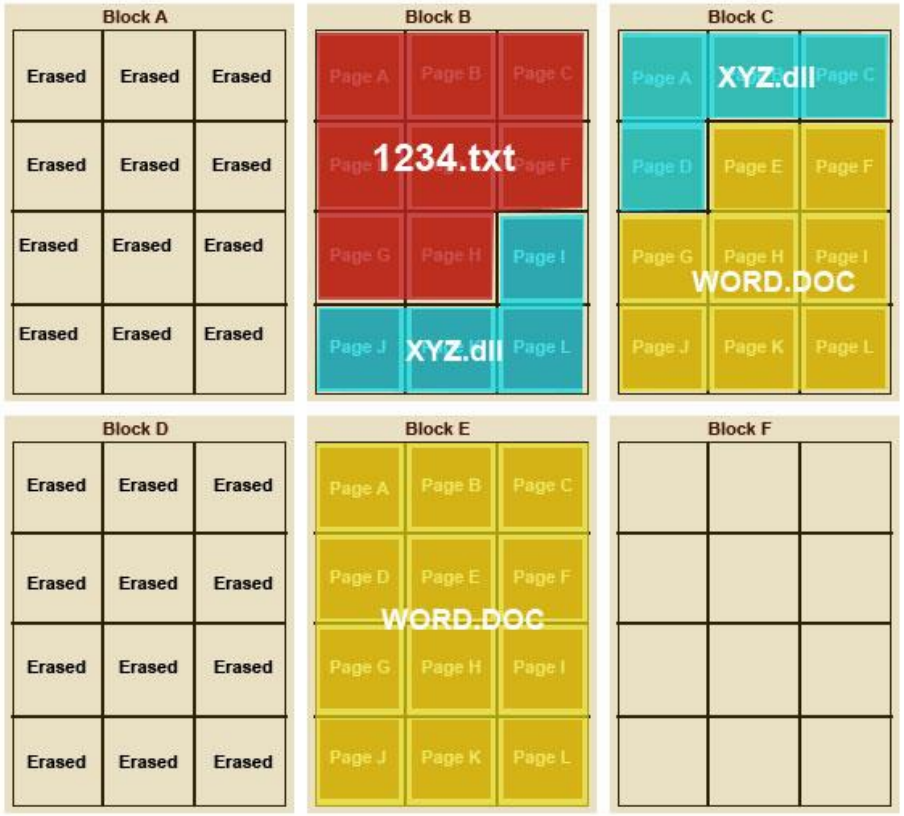
- SSD бришат само цели блокови, но не и страни!
- Затоа за да се избришат старите податоци, мора новите податоци да се запишат на блокот B, па потоа да се избрише блокот A



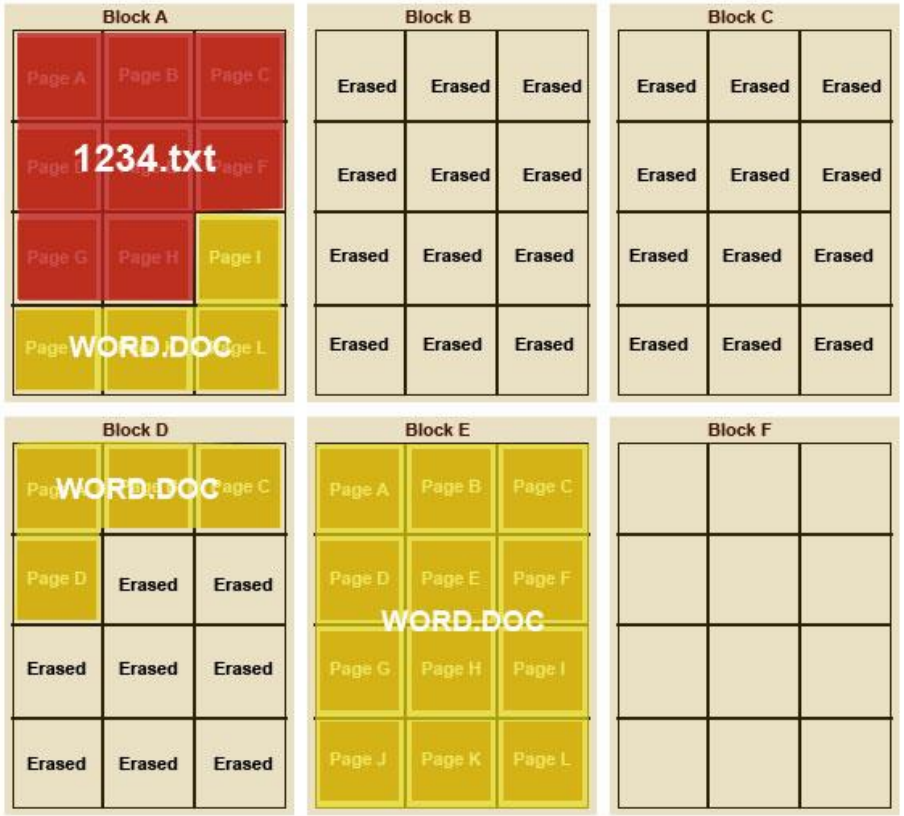
Како се запишува на SSD

- Што се случува доколку ја избришеме xyz.dll датотеката?

Почетна состојба



Резултат



Како се запишува на SSD

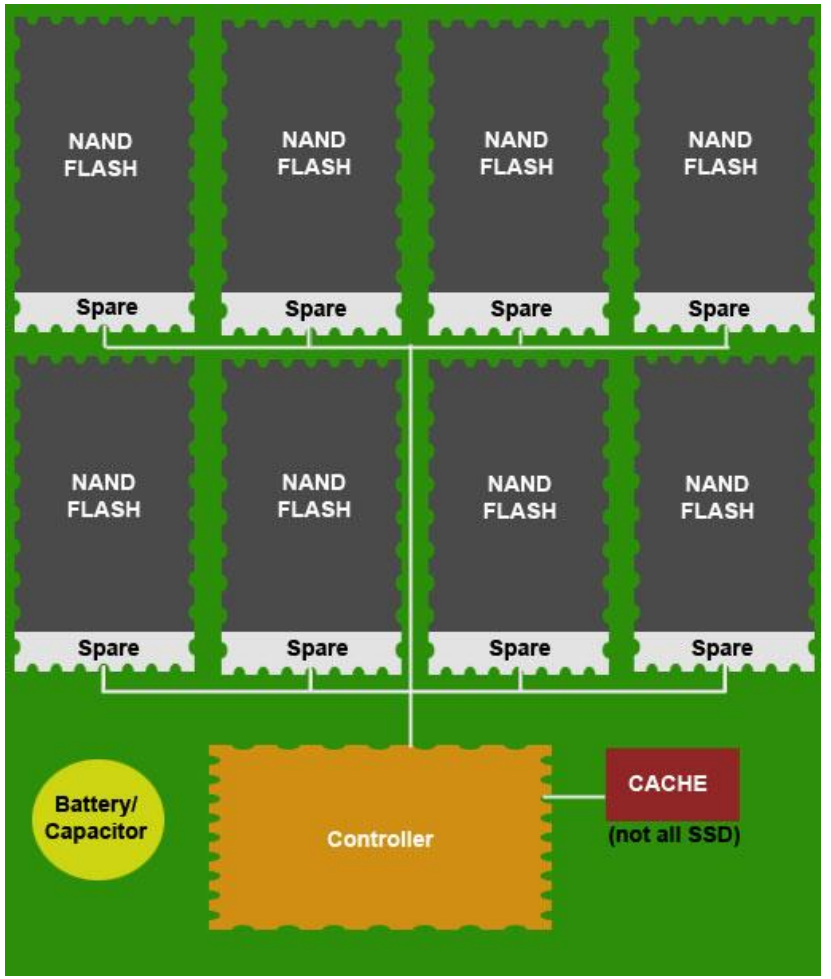
- Чистењето на податоци се нарекува Garbage Collection и ги одржува перформансите на SSD
- SSD никогаш не се дефрагментира!
- Чистењето на SSD, пред да се запишат податоци на него е слаба страна и воведува дополнителни write активности кои ја забавуваат неговата работа.
- На SSD може ограничен број пати да се запишува

Како се запишува на SSD

- Голем дел од SSD меморијата не ја користи корисникот, туку ја користи контролерот за преместување (shuffle) на податоците
  - spare меморија
- Расипаните блокови го намалуваат капацитетот

SSD анатомија

- 128 GB SSD диск со 8 NAND Flash чипови
- Секој чип е со големина од 16GB или вкупно 120 GB, бидејќи 1GB е помошен (spare) чип
- Кешот не задолжителен, и во него има директориум за зафатените блокови и wear leveling (колку пати било запишувано на блокот)
- Најчесто имаат од 4 до 10 канали до NAND чиповите, со што се зголемува неговата пропусна моќ.
- Батеријата/кондензаторот се користи кога има проблем со напојување да може да се доврши запишувањето на SSD-то



Перформанси & цена

	Податочна рата (секвенциален R/W)	Цена/GB	Големина
HDD	50-100 MB/s	$0.025-0.05/GB	2-16 TB
SSD[1]	200-600 MB/s (SATA) 6 GB/s (PCI)	$ 0.2-0.1/GB	200GB-4TB
DRAM	10-16 GB/s	$5-10/GB	64GB-256GB

BW: SSD up to x10 than HDD, DRAM > x10 than SSD

Price: HDD x20 less than SSD, SSD x5 less than DRAM

SSD Заклучок

- Pros (vs. хард диск):
  - Мала латентност, голема податочна рата (се елиминира доцнењето за пребарување/ротација)
  - Нема подвижни делови:
    - Лесен, мала потрошувачка, тивок, отпорен на шок
  - Чита со брзина на меморија (ограничен од котролероти и I/O магистралата)

SSD Заклучок (2)

- Cons
  - Мал капацитет (0.1-0.5x во однос на диск), скап (20x диск)
    - Хибридна алтернатива: комбинација од мал SSD со голем хард диск
  - Асиметрични перформанси при запишување: читај страна /избриши/запиши страна
    - Алгоритмите за garbage collection (GC) во контролерот имаат голем влијание врз перформансите
  - Животен век
    - 50-100K запишувања/страна SLC, 1-10K запишувања/страна MLC