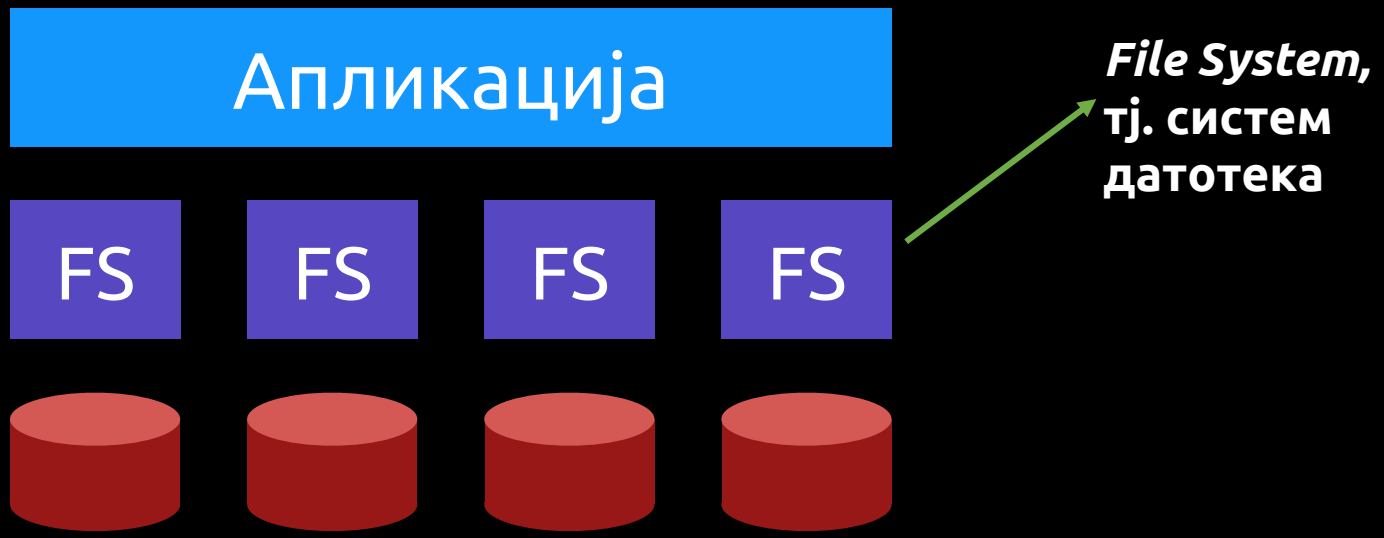
Оперативни системи 2

Факултет техничких наука Косовска Митровица

Драгиша Миљковић

RAID

Први покушај – JBOD



Апликације су „паметне“, оне чувају различите фајлове на различитим системима датотека мада ти системи нису логички повезани.

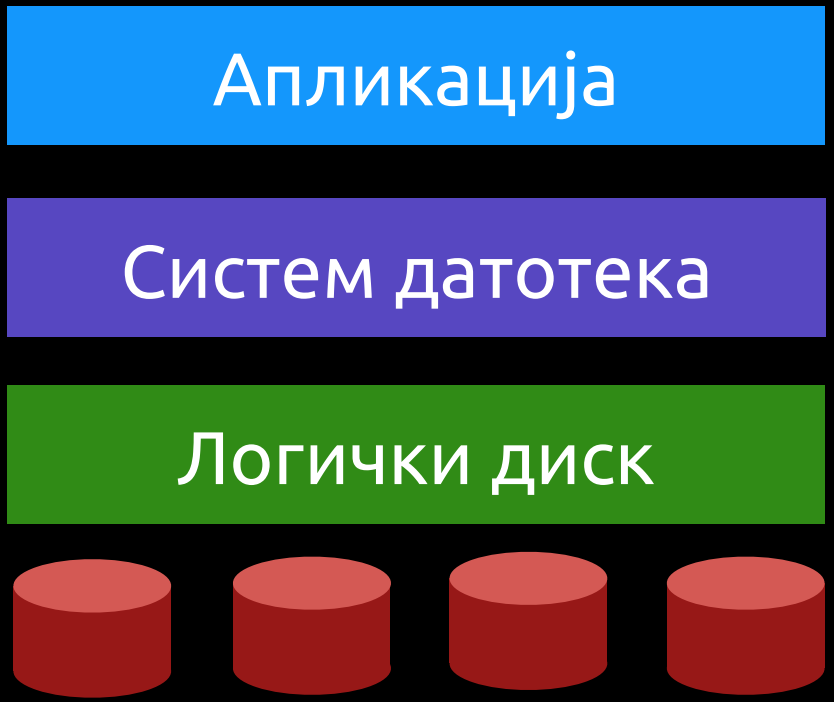
JBOD: Just a Bunch Of Disks („напросто гомила дискова“)

Решење 2 – RAID

RAID је:

-транспарентан,

-лак за прикључивање и коришћење.



Логички диск повећава:

-капацитет,

-перформансе,

-поузданост.

Циљ је изградити један логички диск који се заправо састоји из много физичких дискова.

RAID: Redundant Array of Inexpensive Disks

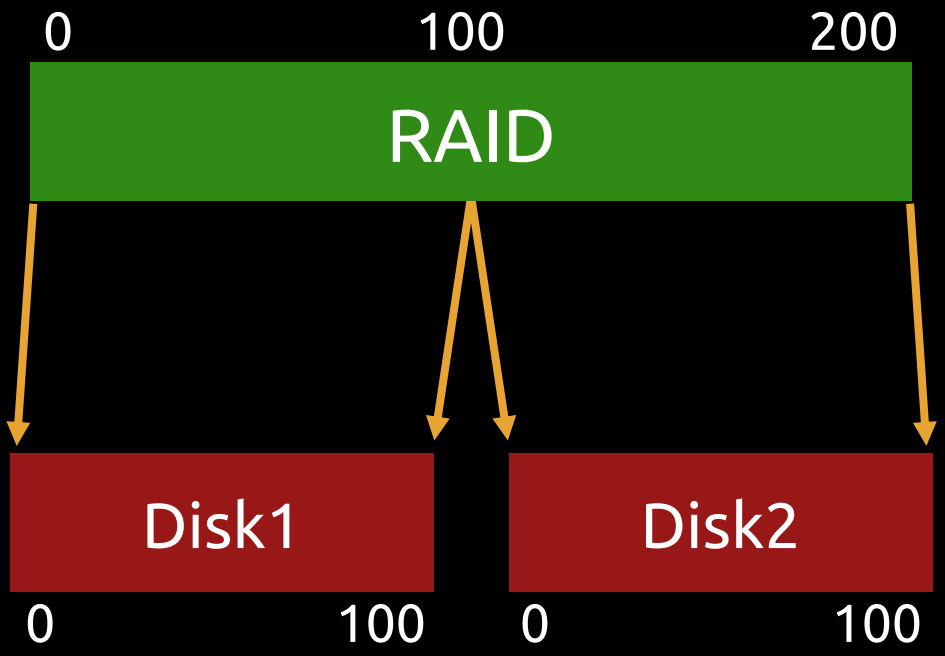
Чему ово?

Економски је исплативо, ови дискови су најјефтинији вид смештања података.

Дакле, циљ је помоћу више јефтиних компоненти софтверски изградити један логички уређај који пружа добре перформансе.

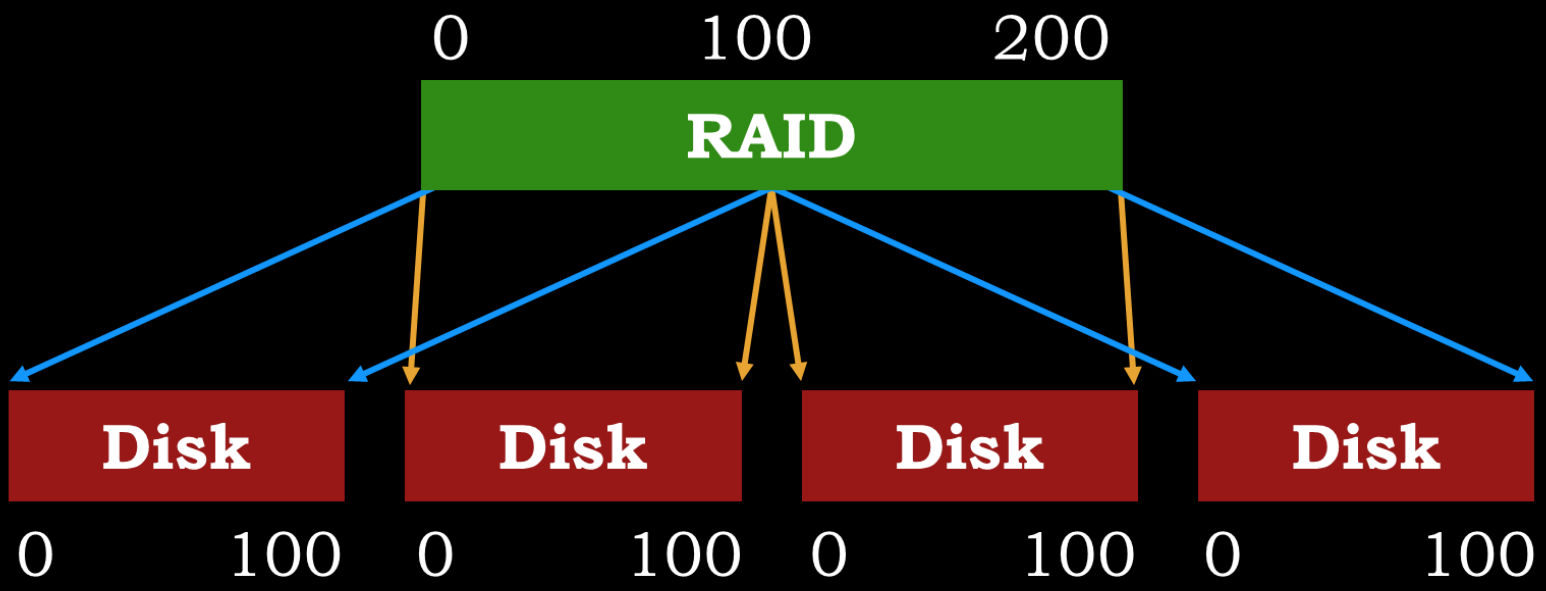
Koja je алтернатива овом приступу? Куповина неког скупог диска са високим перформансама.

Основна стратегија – мапирање



- Изграђује се један велики и брз (логички) диск из више мањих.

Основна стратегија – редундантност



- Додавање још дискова повећава поузданост (коришћењем неке од техника за спречавање губитка података у случају отказа).

Мапирање

Како треба мапирати адресе логичких блокова у адресе физичких блокова?

- На сличан начин оном који сте видели код виртуалне меморије.
  1. **Динамичко мапирање**: Табеле страница.
  2. **Статичко мапирање**: коришћењем једноставне математике. Ово се користи код RAID-а.

Редундантност

Зарад редундантности се морају жртвовати неке ствари, па се нпр. мора бирати између следеће две ствари:

- Повећавањем броја резервних копија се повећава поузданост (а у одређеним случајевима и побољшавају перформансе).
- Смањивањем броја копија повећава се искоришћеност простора.

Још пар ствари

RAID је систем за мапирање логичких у физичке блокове.

Видећемо како се понаша RAID код читања и уписивања које апликације захтевају (и секвенцијално и насумично).

Шта ћемо мерити?

- **Капацитет, поузданост, перформансе**

О чему треба водити рачуна?

- Који логички блокови се мапирају у које физичке блокове,
- Како можемо да додамо још физичких дискова,
- Различити нивои RAID-а имају различите предности и недостатке.

Метрика

Капацитет – Колико простора је доступно апликацијама?

Поузданост – Колико дискова се може користити на безбедан начин? (Претпоставимо да је могуће да се деси отказ хард-диска.)

Перформансе – Колико времена је потребно да се бави задати I/O?

Карактеристике једног диска ћемо обележавати на следећи начин:

- **N** := број дискова
- **C** := капацитет једног диска
- **S** := секвенцијална пропусност једног диска
- **R** := насумична пропусност једног диска
- **D** := време потребно за извршење једне мале I/O операције

Реализација стабилних система – RAID (*Redundant Array of Inexpensive Disks*)

RAID омогућује коришћење вишеструких дискова који су бржи, већи и поузданији.

RAID је организован у више различитих нивоа.

- **RAID ниво 0 –** слагање вишеструких дискова.
- **RAID ниво 1 –** коришћење технике огледала (***mirroring***).
- **RAID нивои 4/5 –** редундантност заснована на парности.
- **RAID 0+1**, **RAID 1+0**.
- Итд.

RAID ниво 0

RAID ниво 0 је најједноставнији облик, овде се подаци између дискова деле на нивоу блокова.

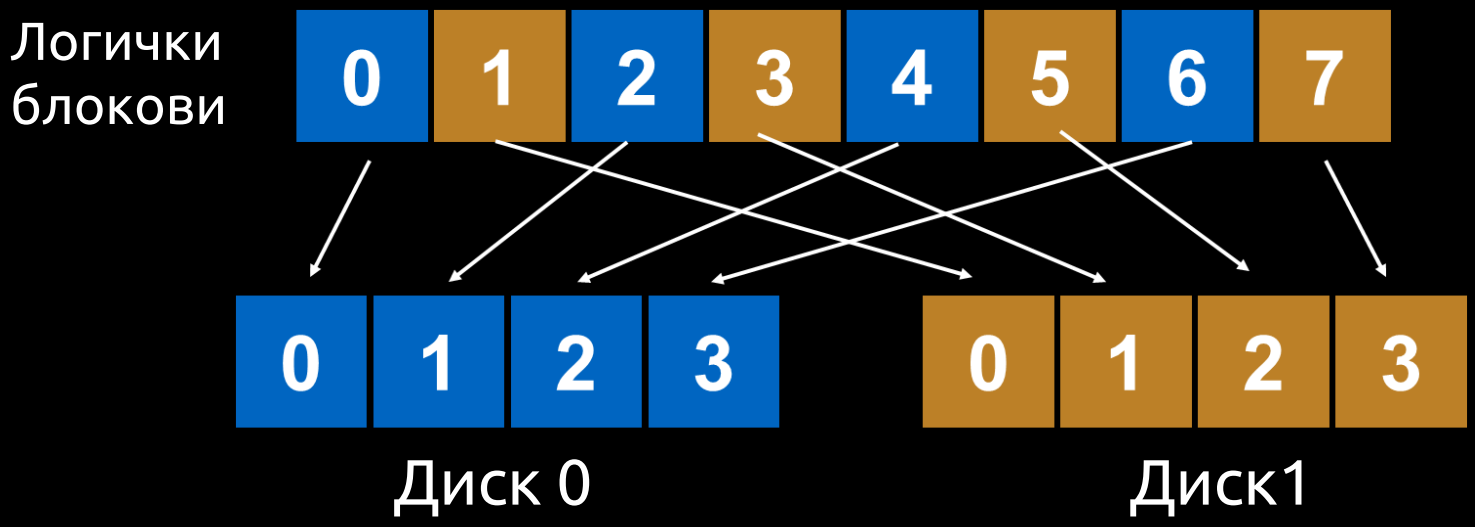
- Блокови су размештени по диску по систему *round robin*.
- **Овај ниво је оптимизован за капацитет. Нема никакве редундантности.**

Disk 0	Disk 1	Disk 2	Disk 3
0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

stripe →

RAID-0: *Striping*

RAID ниво 0



Диск 0	Диск 1
0	1
2	3
4	5
6	7

Ако је дата логичка адреса A, физичка адреса се рачуна на следећи начин:

```
Диск = A % број_дискова
Offset = A / број_дискова
```

RAID ниво 0 (наставак)

Пример RAID-а 0 са већим „тракама“ података

- Величина *траке* – 2 блока (8 килобајта)

Disk 0	Disk 1	Disk 2	Disk 3
0	2	4	6
1	3	5	7
5	10	12	14
9	11	13	15

Величина траке:
2 блока

stripe

Striping са већом величином блока података

Величина траке утиче на перформансе низа.
Ипак, одредити „најбољу“ величину траке није лако.

RAID ниво 0 – анализа

Оценити капацитет, поузданост и перформансе.

- Први начин – кашњење једног захтева
  - Колико паралелизма може да постоји у току једне I/O операције.
- Други начин – *steady-state* пропусност RAID-0
  - Укупна пропусност код великог броја конкурентних захтева.

RAID ниво 0 – анализа

Колики је капацитет?	**N*C**
Колико дискова сме да откаже?	**0**
Кашњење	**D**
Пропусност (секвенцијално, насумично)?	**N*S, N*R**

Уколико се дода још дискова, повећаће се пропусност, али не и кашњење.

- ***N*** *је број дискова*
- ***C*** *је капацитет једног диска*
- ***S*** *је секвенцијална пропусност једног диска*
- ***R*** *је насумична пропусност једног диска*
- ***D*** *је време извршења једне мале I/O операције*

RAID ниво 0 – анализа (пример)

Оптерећење: секвенцијално (**S**) и насумично (**R**)

- Секвенцијално: пребацивање 10 MB континуалних података.
- Насумично: пребацивање 10 KB података.
- Рецимо да је просечно време позиционирања 7ms
- И да је просечно време ротационог кашњења 3ms
- Нека је максимална брзина преноса података датог диска 50 MB/s

Пример: нека је **R** мање од 1 MB/s, а **S** нека буде скоро 50 MB/s.

$$S = \frac{\text{Количина података}}{\text{Време приступа}} = \frac{10\ \text{MB}}{210\ \text{ms}} = 47{,}62\ \frac{\text{MB}}{\text{s}}$$

$$R = \frac{\text{Количина података}}{\text{Време приступа}} = \frac{10\ \text{KB}}{10{,}195\ \text{ms}} = 0{,}981\ \frac{\text{MB}}{\text{s}}$$

RAID ниво 1

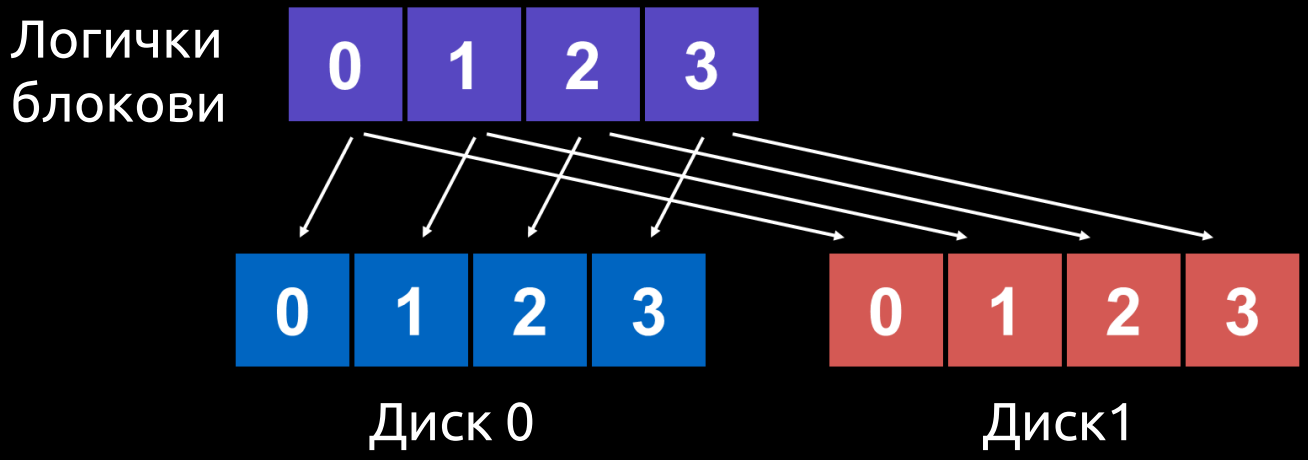
RAID 1 користи **технику огледала** – *mirroring*.

- Прављење копије сваког диска у систему.
  - Дакле, сваки податак је доступан у две копије на два различита диска.
- Копије блокова се смештају на посебне дискове како би се достигла отпорност на отказе дискова.
- Пример за два диска:

Disk 0	Disk 1
0	0
1	1
2	2
3	3

Једноставан RAID-1 – техника огледала

RAID ниво 1- пример за два диска



Disk 0	Disk 1
0	0
1	1
2	2
3	3

RAID ниво 1

Ако размислите – постоји више начина за смештање података на дискове. Нпр. једна могућност је комбинација RAID-1 и RAID-0 приступа, таква комбинација се зове **RAID-10** (илити **RAID 1+0**, „**stripe of mirrors**“) и изгледа као на примеру испод.

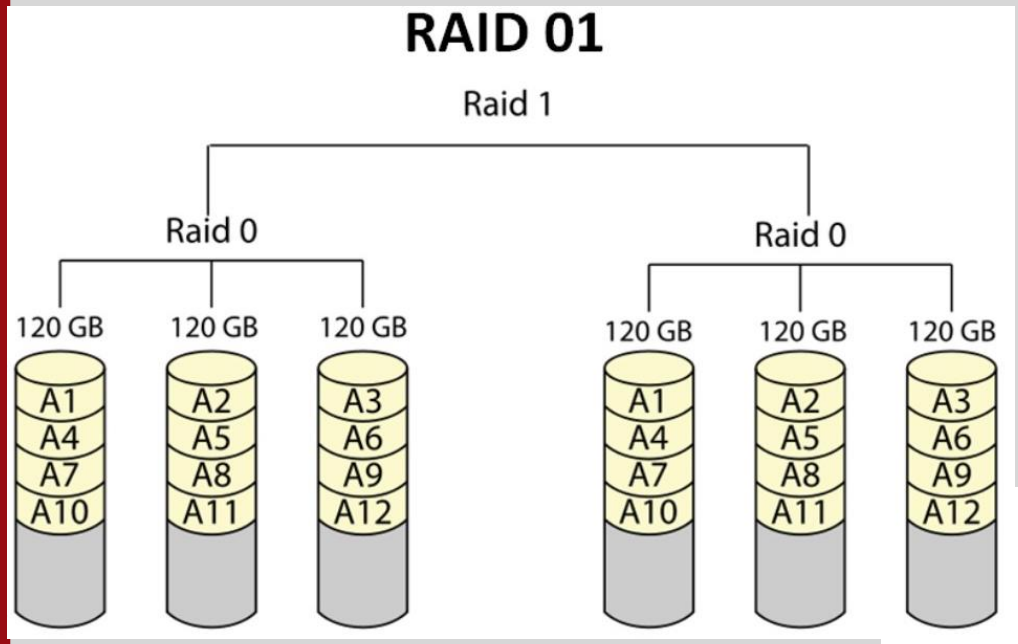
Пример за четири диска:

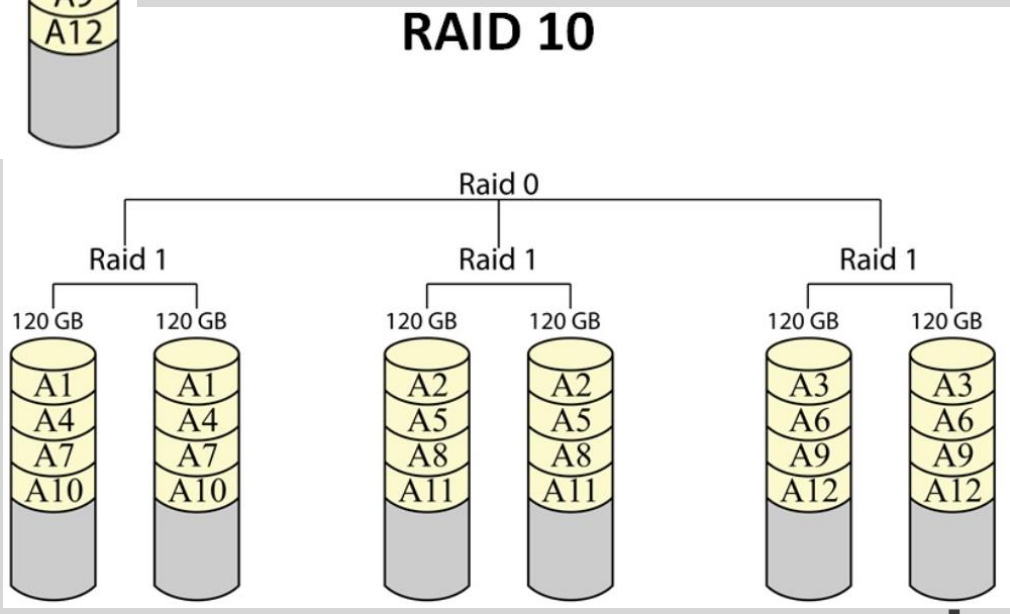
Disk 0	Disk 1	Disk 2	Disk 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

Једноставан RAID-1+0

Могуће је направити и обрнуту комбинацију, **RAID-01** (илити **RAID 0+1**, „**mirror of stripes**“).

Пример RAID-01 и RAID-10 распореда





RAID ниво 1- пример за четири диска

Диск 0	Диск 1	Диск 2	Диск 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

Колико дискова сме да откаже?
Овде ћемо да претпоставимо да диск може или да ради 100% исправно, или да потпуно откаже, као и да систем одмах зна када диск откаже. Ово није сасвим реална претпоставка (често откаже само неки део диска).
А постоје мало сложеније грешке, грешке латетнтних сектора, корупција података… Њих ћемо да занемаримо.

RAID ниво 1 – особине

Колики је капацитет?	**(N/2)*C**
Колико дискова може да откаже?[1]	**од 1 до N/2**
Кашњење (читање, уписивање)	**D**

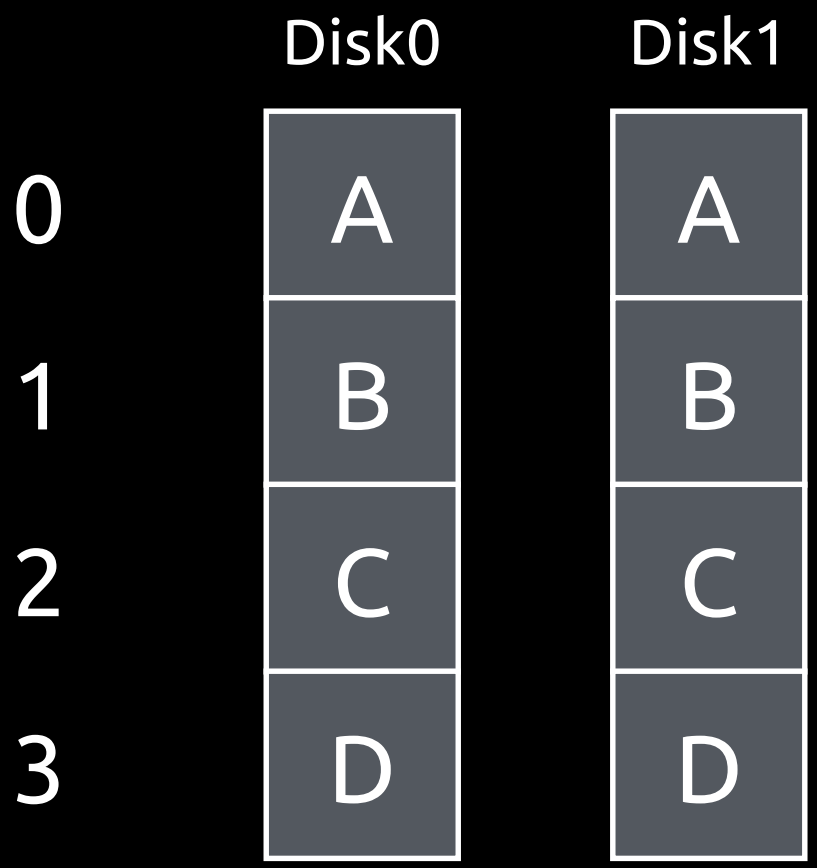
[1] Овде зависи који ће дискови да откажу јер сваки диск има само једну копију. Нема опоравка ако откажу два диска који чувају исте податке. Али, уколико нам откажу различити дискови, онда максимално може чак пола да их откаже – подаци су сачувани на њиховим копијама.

RAID ниво 1

Колика је пропусност:

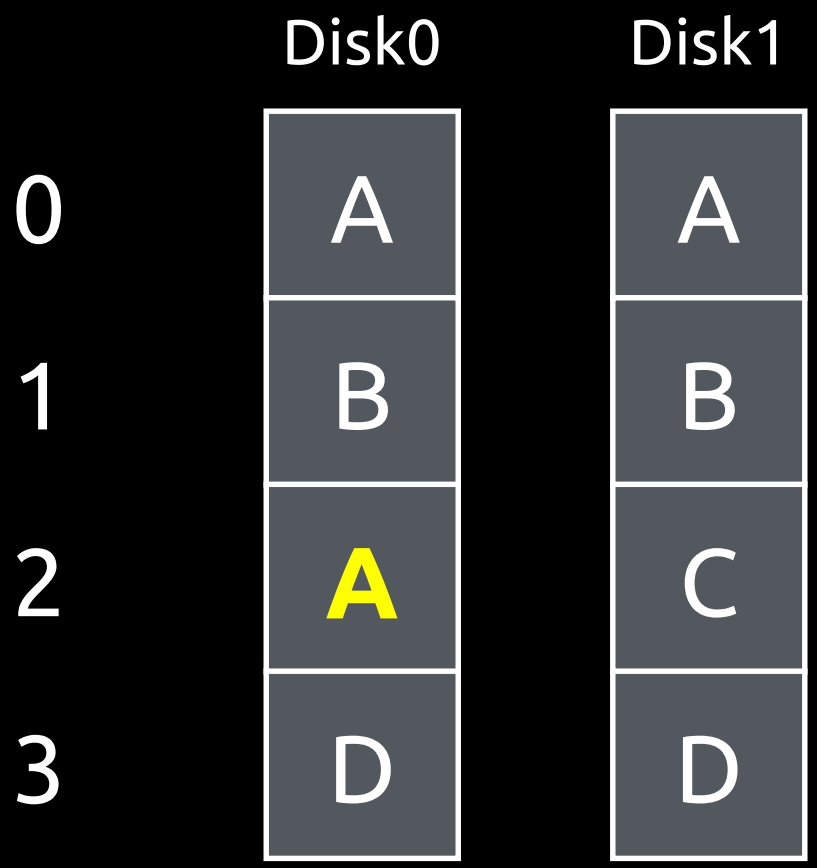
- Насумична читања? **N*R**
- Насумична уписивања? **(N/2)*R**
- Секвенцијална уписивања? **(N/2)*S**
- Секвенцијална читања? **(N/2)*S**

Пример грешке

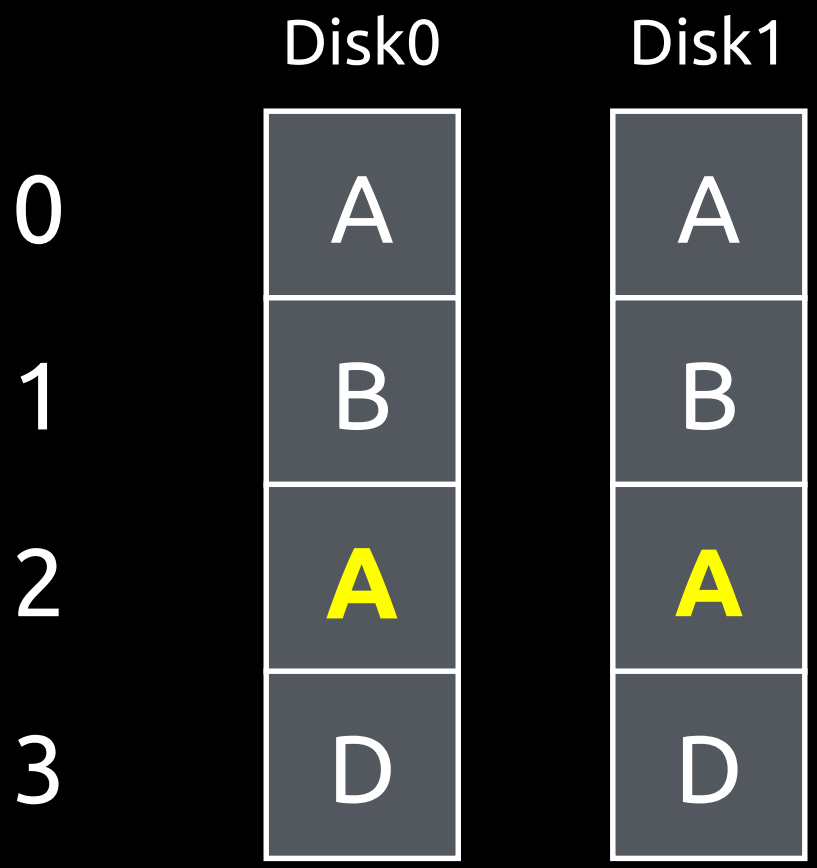


write(A) у блок 2

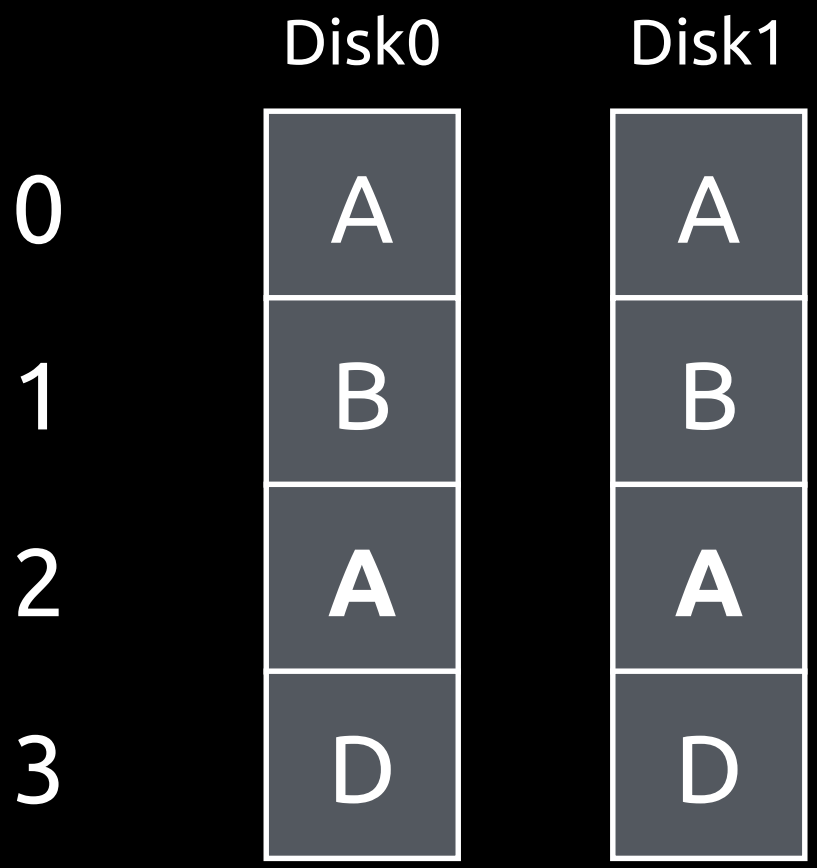
Пример грешке



Пример грешке

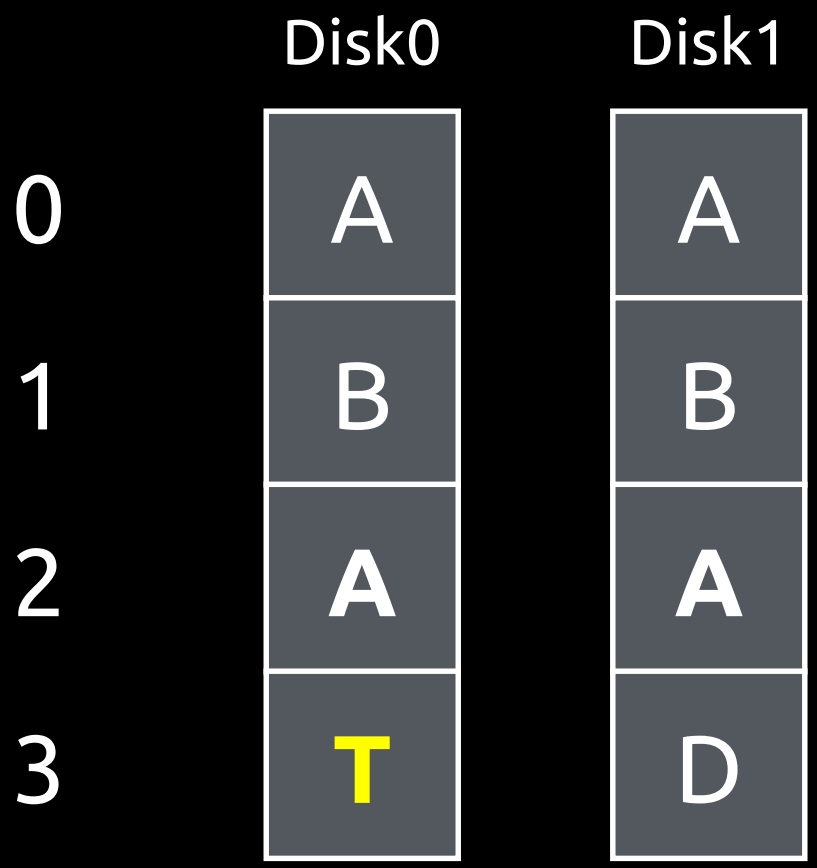


Пример грешке



write(T) у блок 3

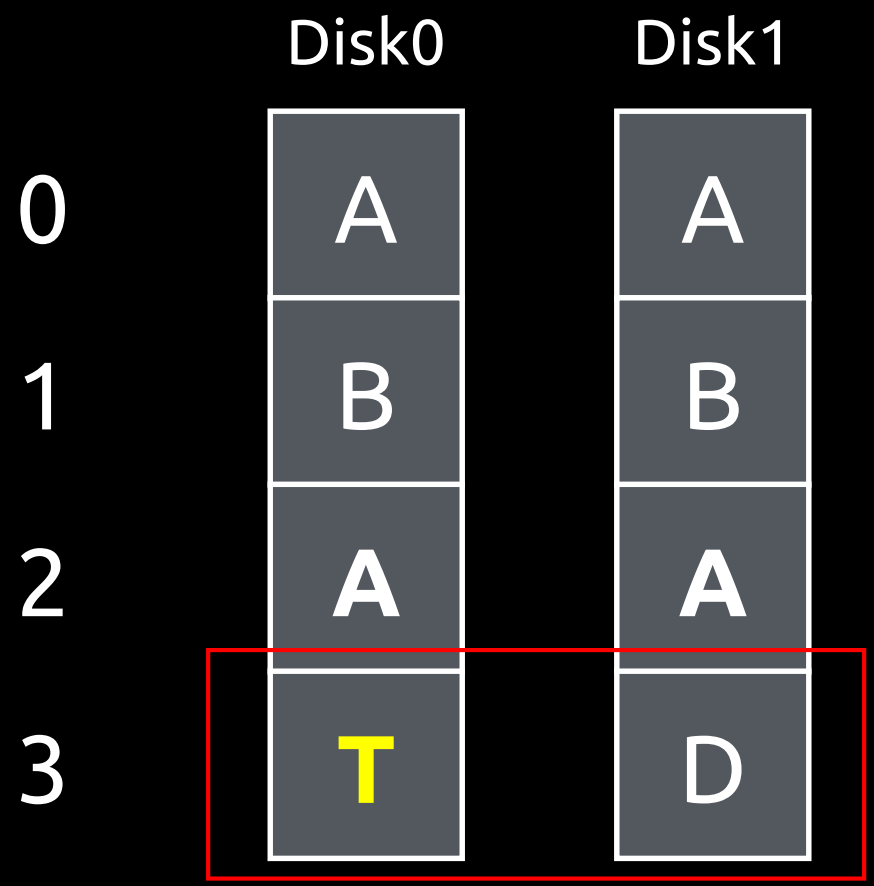
Пример грешке



У овом тренутку се деси

ГРЕШКА!!!

Пример грешке



Након поновног покретања система, како да знамо који је податак исправан?

Напомена: подједнако вероватно је било и да податак буде прво уписан на Диск1, а онда на Диск0.

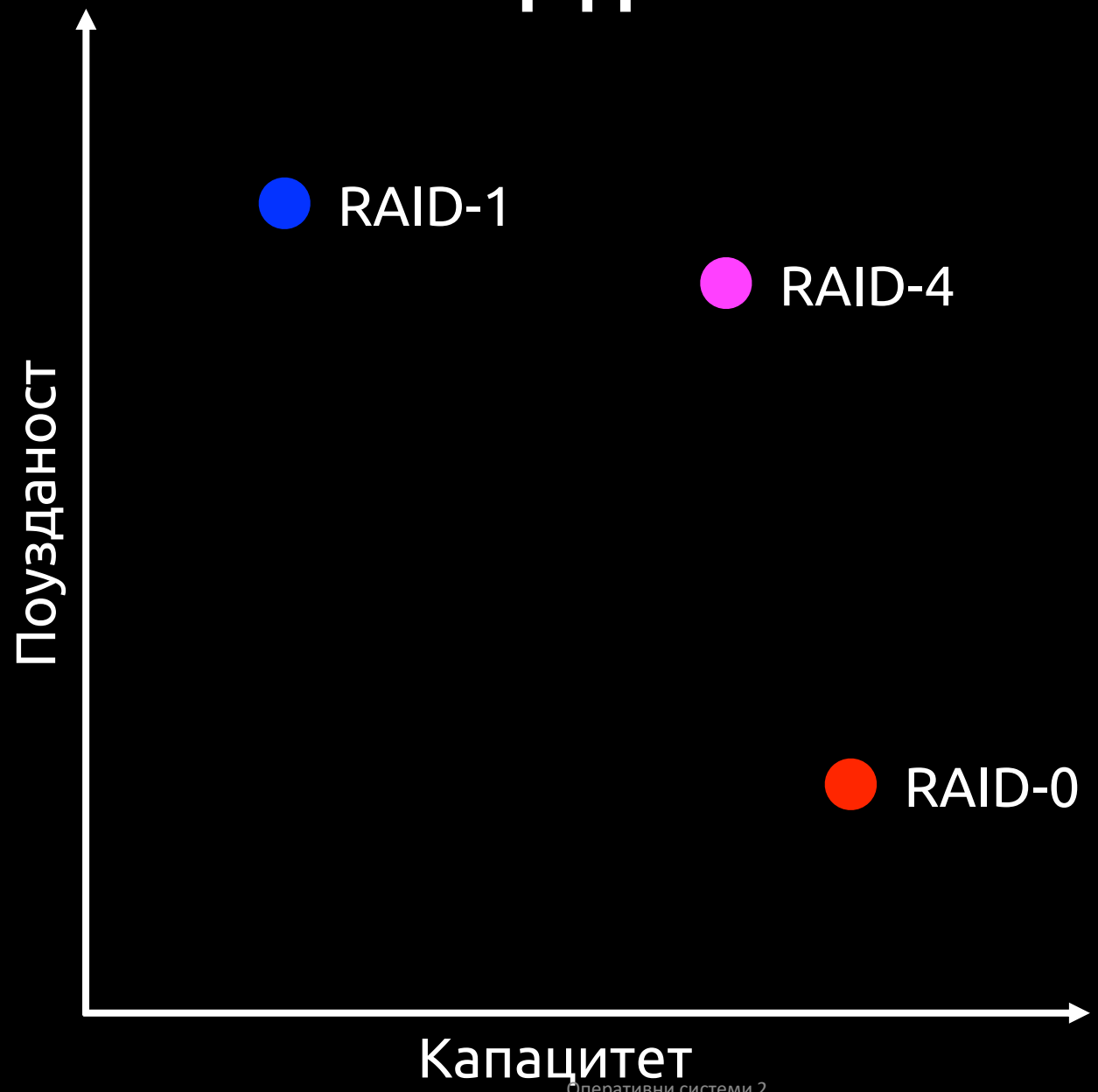
Хардверско решење

Проблем: конзистентна ажурирања (*Consistent-Update*).

Коришћење постојане (*non-volatile*) RAM меморије у RAID контролеру.

Софтверски RAID контролери (нпр. *Linux md*) немају ову опцију.

Перфромансе



RAID ниво 4

RAID ниво 4 се користи да се низовима дискова дода редундантност помоћу парности (која се чува на посебном диску).

- Ово је као у алгебри, ако имате једначину са N непознатих, а N-1 непознатих вам је познато, лако ћете израчунати последњу.
- *Stripe* – ово је као једна једначина. Подаци на диску који откаже су наша непозната у једначини.

* P: Parity

Disk 0	Disk 1	Disk 2	Disk 3	Disk 4
0	1	2	3	P0
4	5	6	7	P1
8	9	10	11	P2
12	13	14	15	P3

Једноставан RAID-4 са парношћу

RAID ниво 4 (наставак)

Оптимизација једноставног RAID-4 је позната под именом ***Full-stripe* уписивање**.

- Треба израчунати нову вредност P0 (*Parity 0*).
- Уписати све блокове у пет дискова у паралели.
- *Full-stripe* уписивања су најефикаснији начин.

Disk 0	Disk 1	Disk 2	Disk 3	Disk 4
0	1	2	3	P0
4	5	6	7	P1
8	9	10	11	P2
12	13	14	15	P3

Full-stripe уписивање код RAID-4

RAID ниво 4 (наставак)

Пример уписивања у RAID 4. нивоа коришћењем методе одузимања.

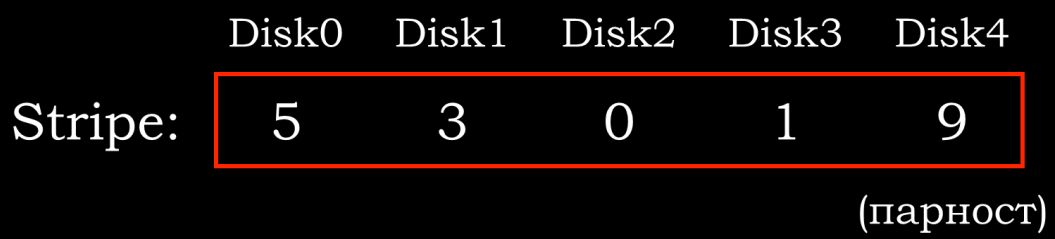
- За свако уписивање, RAID извршава четири физичка I/O (два читања и два уписивања).

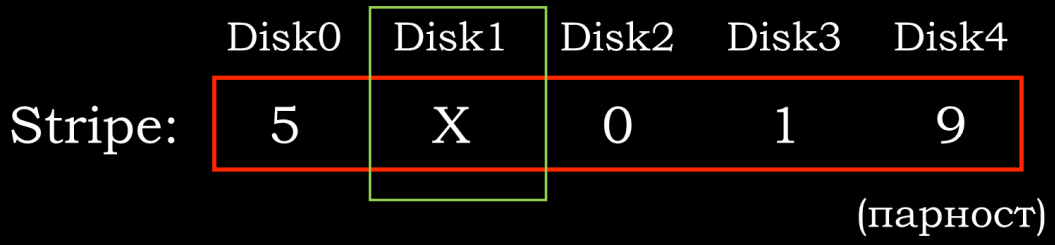
Disk 0	Disk 1	Disk 2	Disk 3	Disk 4
0	1	2	3	P0
*4	5	6	7	+P1
8	9	10	11	P2
12	*13	14	15	+P3

Уписивање у блокове 4, 13, као и у одговарајуће блокове парности

Јавља се проблем код малих уписивања.

Пример отказа





- Лако можемо да израчунамо да је X = 3

Пример израчунавања парности

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	3	0	1	2	X

(парност)

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	3	0	1	2	6

(парност)

- Које функције се користе да се израчуна парност?

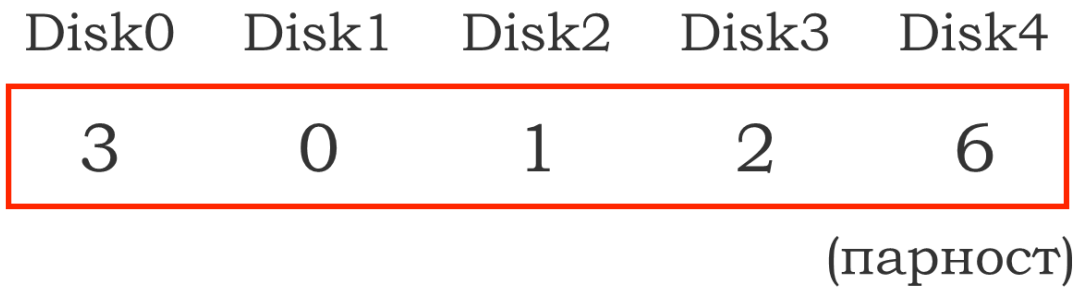
RAID ниво 4 – анализа

Колики је капацитет? **(N-1)*C**

Колико дискова може да откаже? **1**

Кашњење (читање, уписивање)? **D, 2*D**

(због уписивања на диск парности)

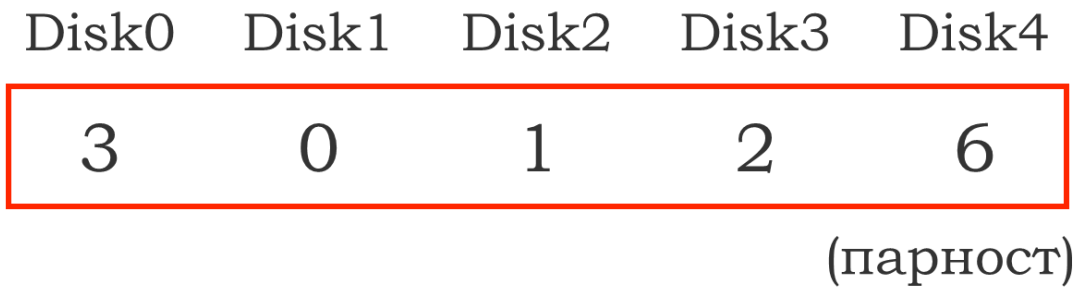


RAID ниво 4 – пропусност

Колика је пропусност:

Секвенцијална читања?	**(N-1)*S**
Секвенцијална уписивања?	**(N-1)*S**
Насумична читања?	**(N-1)*R**
Насумична уписивања?	**R/2**

(због уписивања на диск парности)



Како избећи уско грло које изазива „парност“?

RAID ниво 5

RAID ниво 5 је решење за проблем који настаје код малих уписивања.

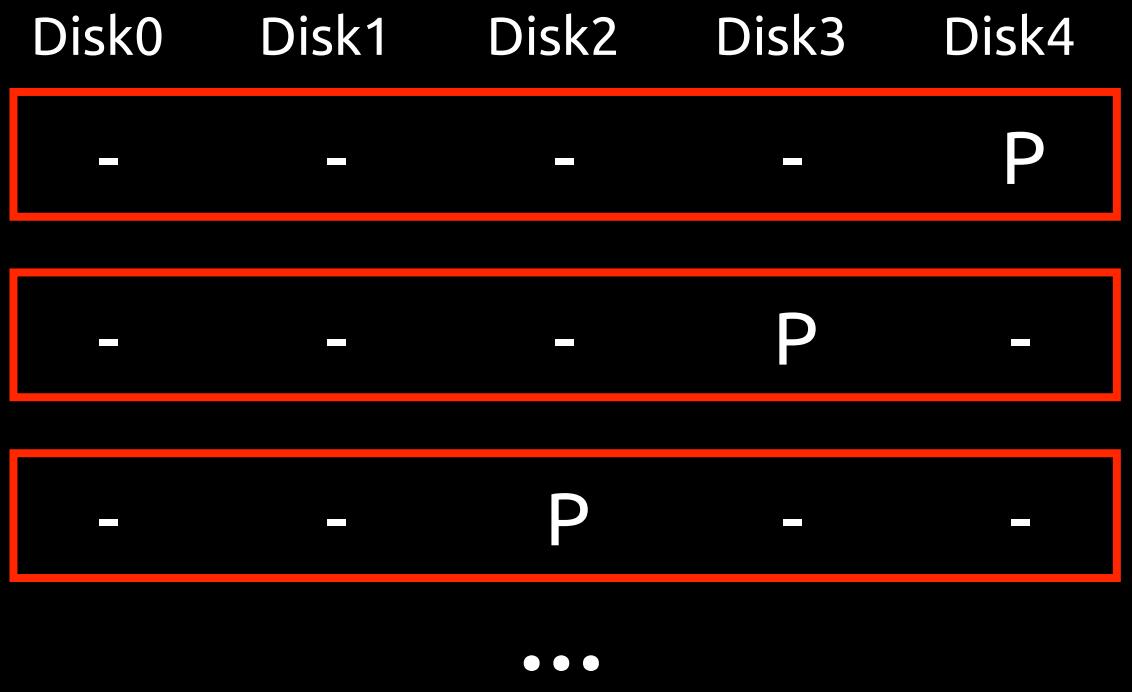
- Проблем малих уписивања изазива лоше перформансе код RAID-а нивоа 4.
- Ниво 5 ради скоро идентично као RAID-4, осим што ротира блокове парности кроз уређаје.

Свака стаза RAID нивоа 5 је ротирана на уређајима (као на слици).

Disk 0	Disk 1	Disk 2	Disk 3	Disk 4
0	1	2	3	P0
5	6	7	P1	4
10	11	P2	8	9
15	P3	12	13	14
P4	16	17	18	19

RAID-5 са ротираном парношћу

RAID-5



- Блокови парности се распоређују укруг по свим дисковима.

RAID ниво 5 – анализа

- Колики је капацитет? **(N-1)*C**
- Колико дискова може да откаже? **1**
- Кашњење (читање, уписивање)? **D, 2*D**

Исто је као код RAID-4.

RAID ниво 5 – пропусност

Колика је пропусност:

- Секвенцијална читања? **(N-1)*S**
- Секвенцијална уписивања? **(N-1)*S**
- Насумична читања? **N*R**
- Насумична уписивања? **N*R/4**

RAID – анализа нивоа

	RAID-0	RAID-1	RAID-4	RAID-5
КАПАЦИТЕТ	**N*C**	**N*C/2**	**(N-1)*C**	**(N-1)*C**
ПОУЗДАНОСТ	**0**	**1 (сигурно)** **N/2 (уколико имамо среће)**	**1**	**1**
ПРОПУСНОСТ				
Секвенцијално читање	**N*S**	**N*S/2**	**(N-1)*S**	**(N-1)*S**
Секвенцијално уписивање	**N*S**	**N*S/2**	**(N-1)*S**	**(N-1)*S**
Насумично читање	**N*R**	**N*R**	**(N-1)*R**	**N*R**
Насумично уписивање	**N*R**	**N*R/2**	**R/2**	**(N/4) *R**
КАШЊЕЊЕ				
Читање	**D**	**D**	**D**	**D**
Уписивање	**D**	**D**	**2D**	**2D**

RAID анализа нивоа

Закључци:

- RAID-0 је увек најбржи и има највећи капацитет (али то иде на рачун поузданости).
- За насумична уписивања је најбољи RAID-1, али по цену капацитета.
- Уколико су главни циљеви капацитет **и** поузданост, победник је RAID-5 (али по цену перформанси код малих уписивања)
- Коначно, уколико се најчешће врше секвенцијална уписивања, а жели се и што већи капацитет, RAID-5 има највише смисла.

Још неки подаци о RAID-у

Постоји још много других дизајна RAID система, укључујући нивое 2 и 3, као и ниво 6 (који допушта отказ више дискова истовремено).

Поставља се и питање шта се ради код отказа хард-диска?

- Понекада се увек у близини држи резервни диск који треба да надомести онај који је отказао. Али како отказ утиче на перформансе? И како утиче на перформансе процес замене и попуњавања диска?

Ми смо причали поједностављено, а заправо постоје неке суптилније грешке: нпр. грешке латентних сектора и грешке покварених блокова.

Коначно, RAID не мора да буде имплементиран хардверски, већ може и софтверски, али онда постаје подложан проблему конзистентног ажурирања.

Закључак

Код RAID-а има много инжењерских компромиса

- Капацитет
- Поузданост
- Перформансе

} За различите случајеве

Интерфејс заснован на блоковима - лако се изграђује и поприлично је популаран због транспарентности