

Keš memorija

Nikola Milenić

Matematička gimnazija

16. 05. 2024.

- Šta je najveće usko grlo računara?
 - Logičke operacije?
 - Celobrojni račun?
 - Operacije sa pomičnim zarezom?

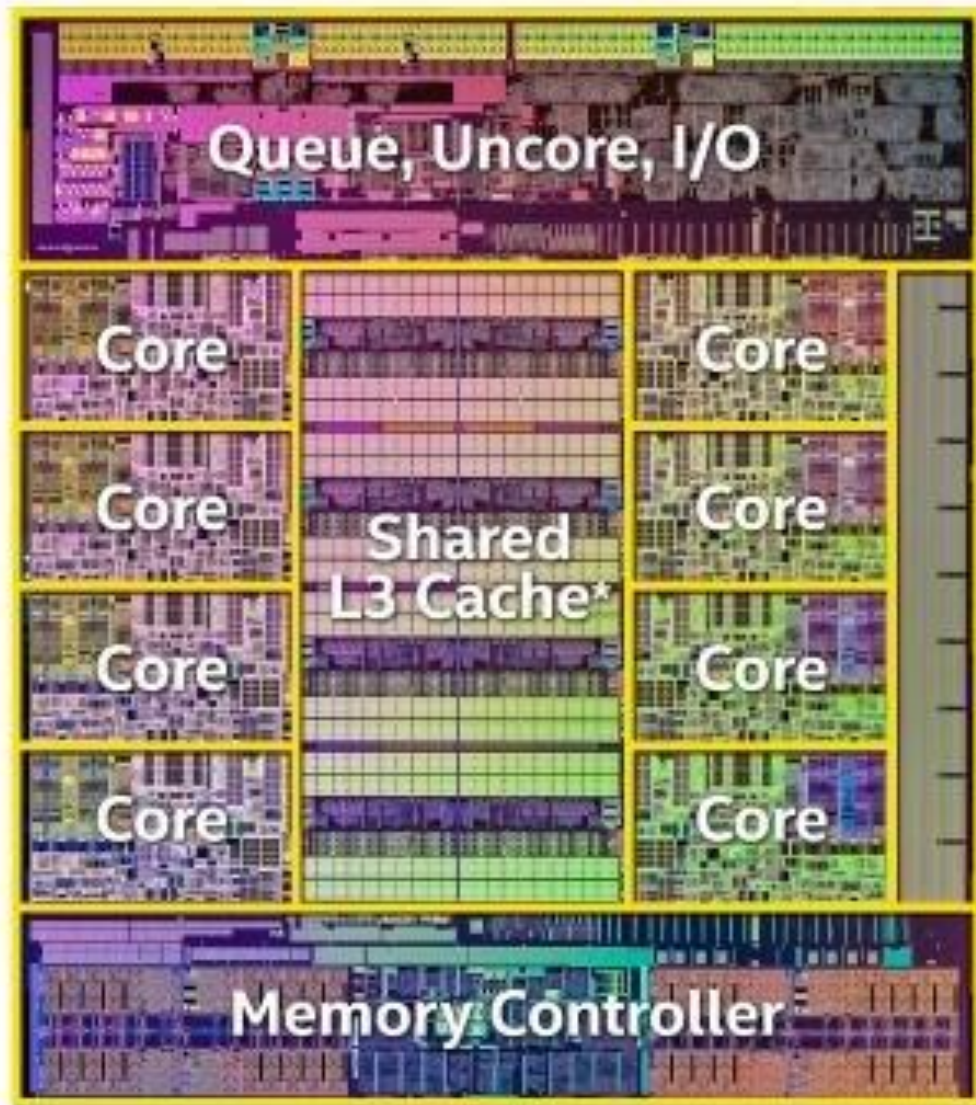
Koji kod je brži?



```
1 #include <stdio.h>
2 int m[10000][10000];
3 int res = 0;
4 void main(){
5     for (int x = 0; x < 10000; x++)
6         for (int y = 0; y < 10000; y++)
7             m[y][x] = x*y;
8     for (int x = 0; x < 10000; x++)
9         for (int y = 0; y < 10000; y++)
10            res += m[y][x];
11     printf("%d\n", res);
12 }
```

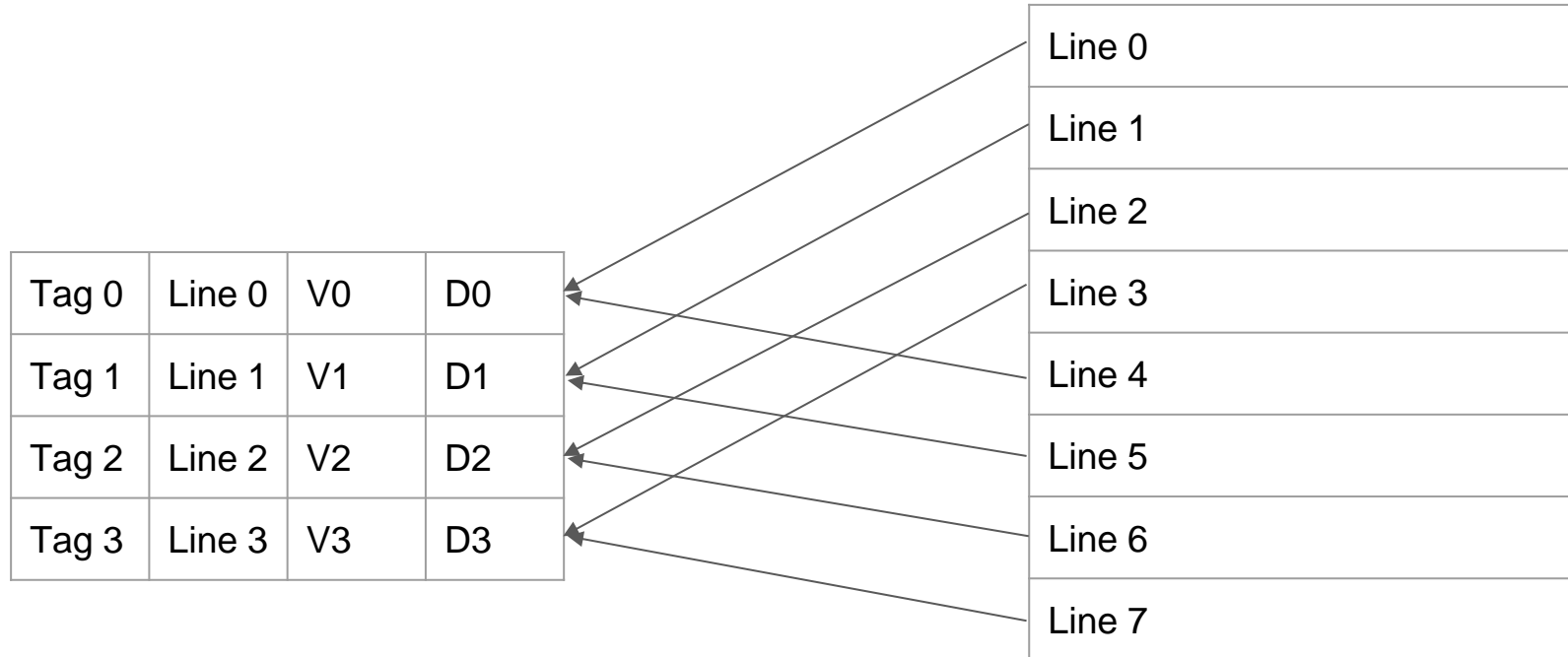
```
1 #include <stdio.h>
2 int m[10000][10000];
3 int res = 0;
4 void main(){
5     for (int y = 0; y < 10000; y++)
6         for (int x = 0; x < 10000; x++)
7             m[y][x] = x*y;
8     for (int y = 0; y < 10000; y++)
9         for (int x = 0; x < 10000; x++)
10            res += m[y][x];
11     printf("%d\n", res);
12 }
```

- Pristup memoriji je glavno usko grlo
 - RAM je daleko od CPU
- Kako smanjiti potrebu za pristupanjem memoriji?
- Pretpostavka: lokalnost pristupa
 - Prostorna
 - Vremenska



- Memorija unutar CPU
- Veličina je ograničena, ali je dosta brži od RAMa
- Prilikom pristupa memoriji, dovlači blok susednih adresa
- Transparentan za programera
- Šta keširati a šta izbaciti?

Direktno mapiran keš



Adresa

Tag	Line	Word
-----	------	------



Asocijativni keš



Tag 0	Line 0	V0	D0
Tag 1	Line 1	V1	D1
Tag 2	Line 2	V2	D2
Tag 3	Line 3	V3	D3



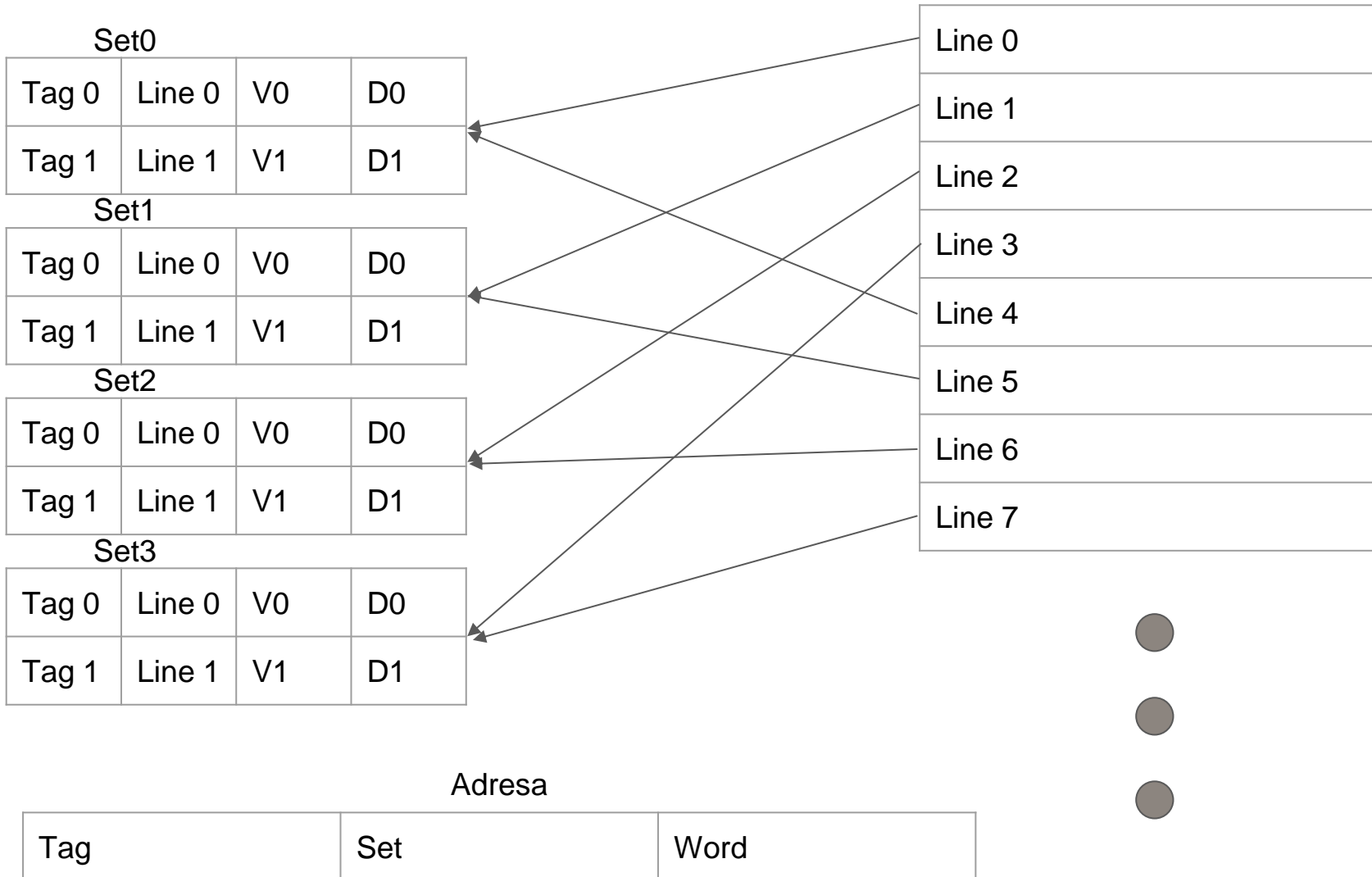
Line 0
Line 1
Line 2
Line 3
Line 4
Line 5
Line 6
Line 7

Adresa

Tag	Word
-----	------

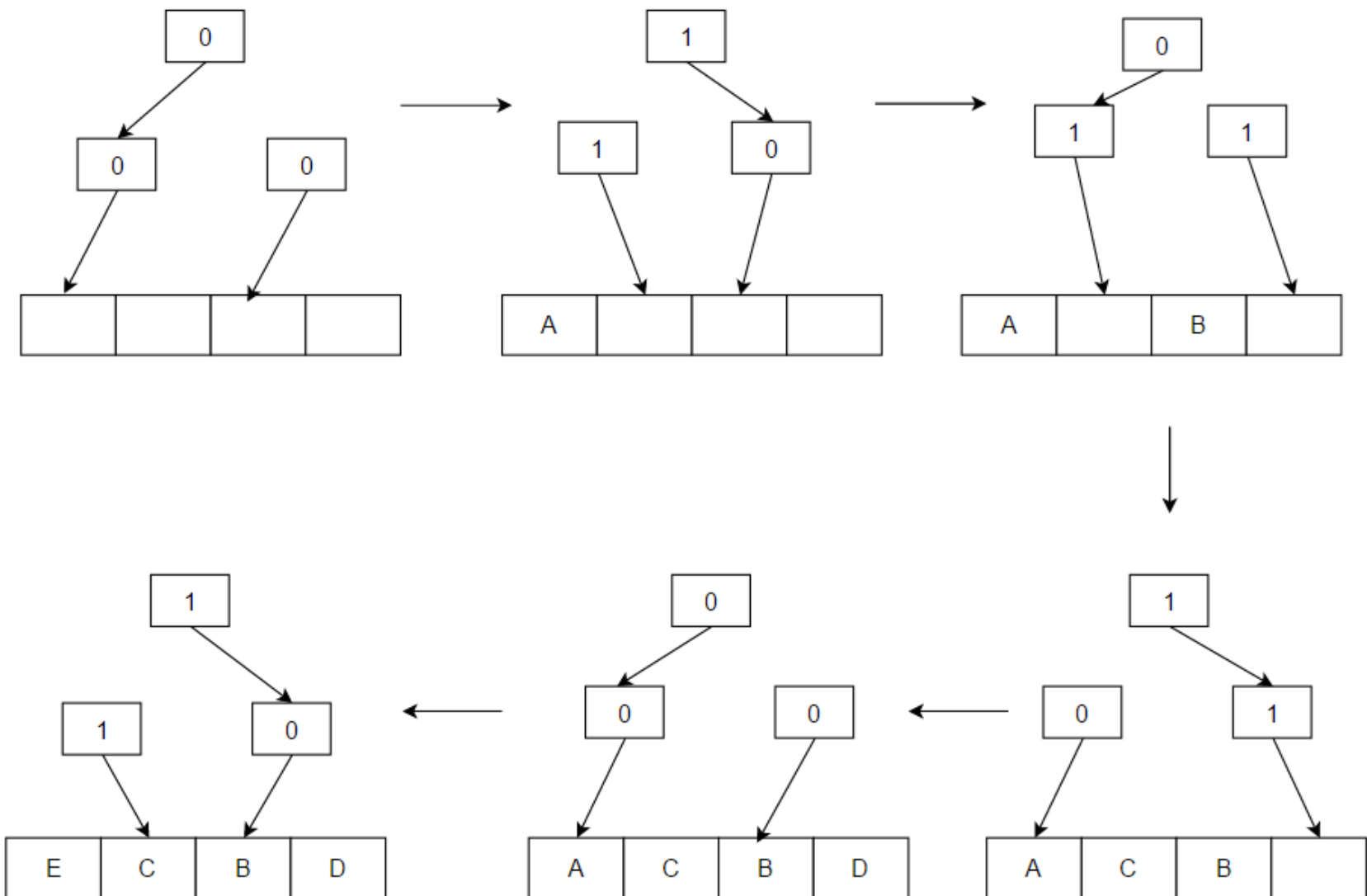


Set-asocijativni keš

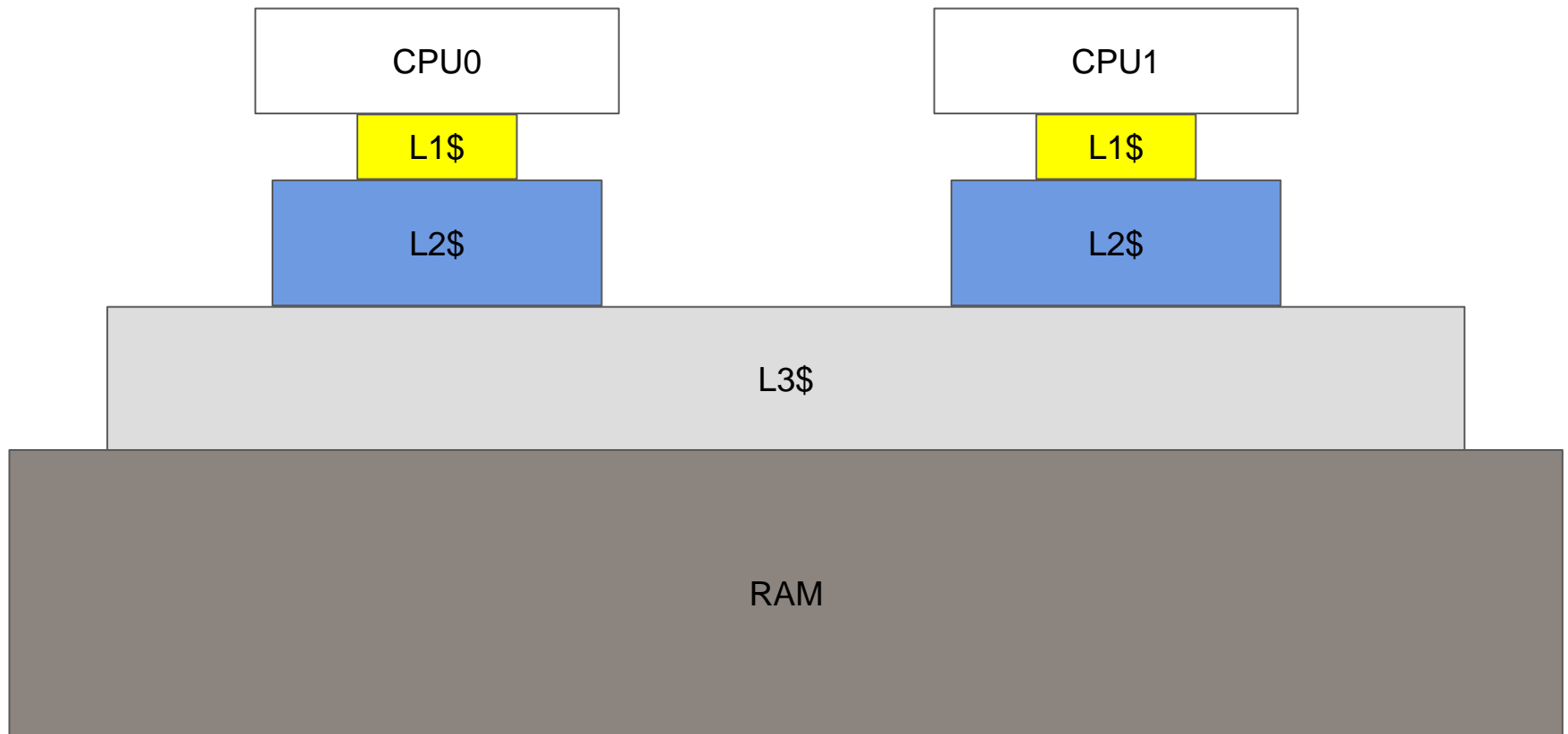


- Kako da odredimo koju linija iz seta da izbacimo?
- LRU - Least Recently Used
- Sa dve linije u setu je trivijalno, ali sa većim setovima postaje sve komplikovanije
- Obično se koriste setovi sa po 4 ili 8 linija
- Česta heuristika: pseudo LRU

Pseudo LRU



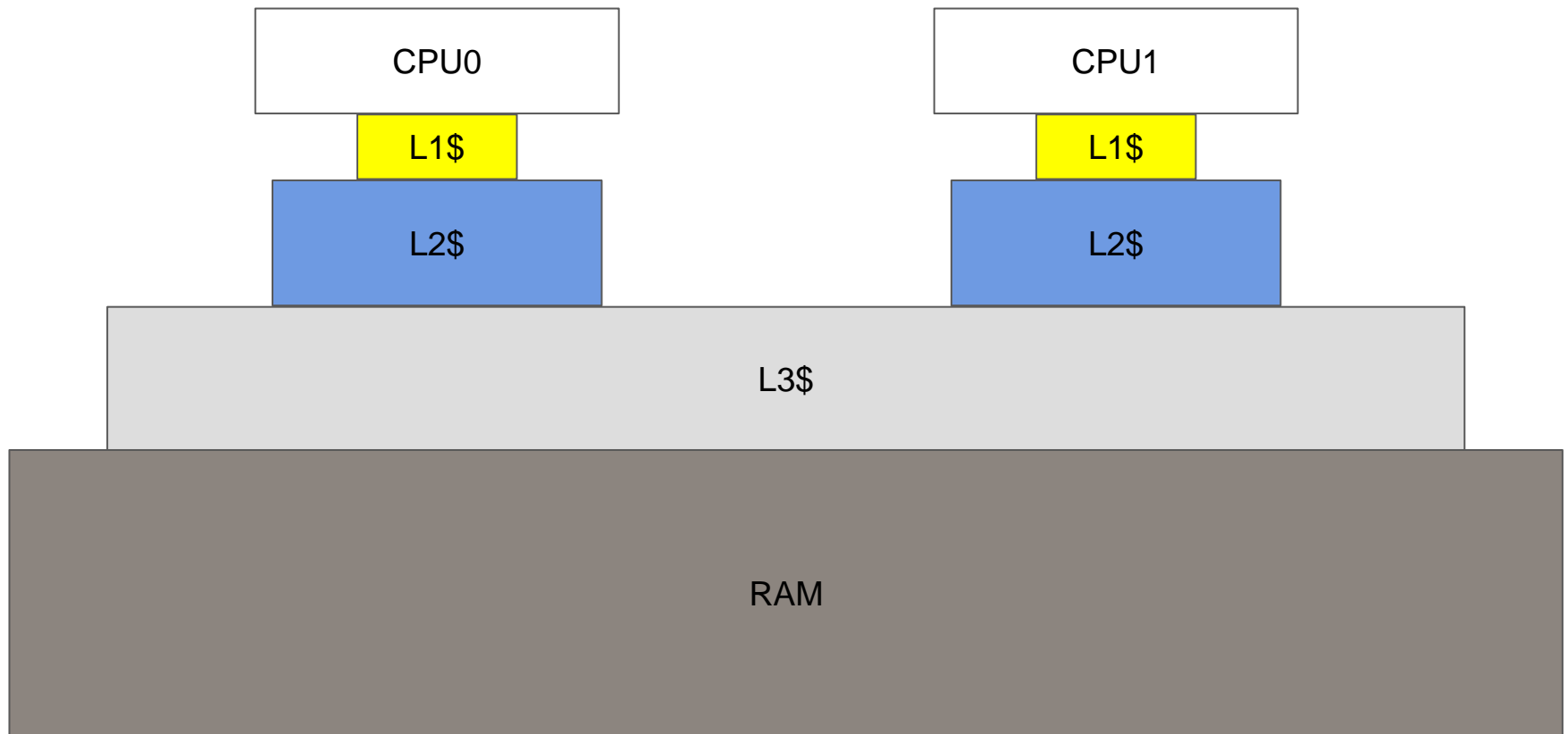
- Što je keš veći, to je sporiji
 - Veća udaljenost od jezgra
 - Sporija propagacija signala
 - Sporije dekodovanje adrese
- Ako je keš premali, imaće nizak hit rate
- Da li želimo mali i brz ili veliki i spor keš?
 - Da





- Više jezgara može da pristupa istoj promenljivoj
- U više nivoa keša može postojati više različitih kopija iste linije
- Kako osigurati da da sva jezgra vide koherentnu sliku memorije?

- Write through na nižim nivoima
- Snooping



Hvala na pažnji!

Pitanja?