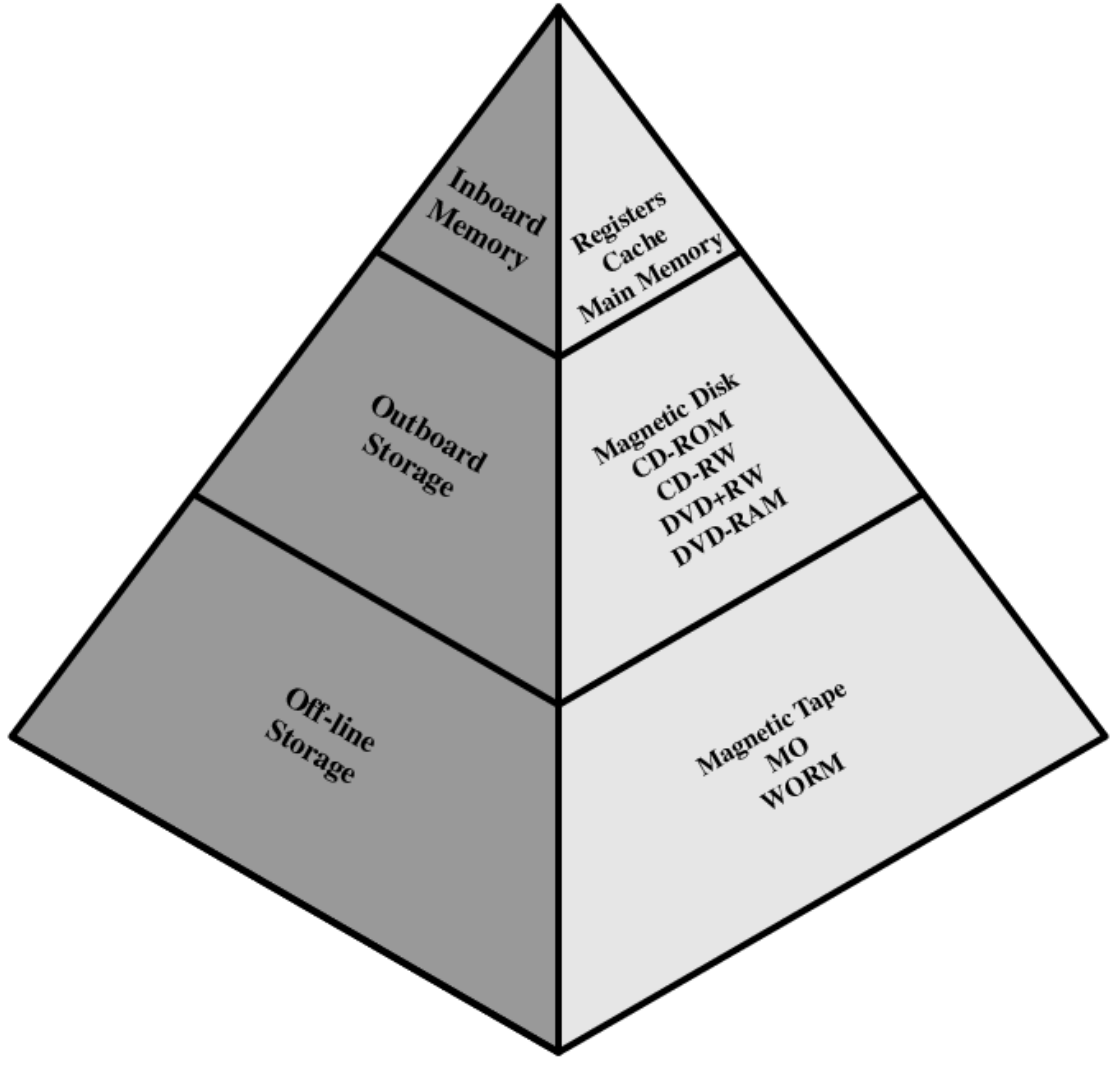
Atminties technologijos

Spartinančiosios atminties technologijos

4 paskaita

Atminties hierarchija



Atminties lygiai ir charakteristikos

Parametrai \ Lygis	Registrai	Spartinan. atmintis L3	Pagrind. atmintis	Išorinė atmintis
Talpa	~8 KB	~24 MB	~1000GB	~ 10 TB
Kreipties laikas (ns)	0.25	1 - 20	50 - 100	5 000 000
Pralaidumas (MB/s)	8000-16000	1600-5000	500-2000	100-600
Kas valdo	Kompilia-torius	Aparatūra	OS	OS / vartot.
Žemesnis lygis	Spart. atmintis	Pagr. Atm.	MD	MJ
1 MB kaina, Lt	???	200-250	0,5 - 1	0,2 - 0,1

Spartinančioji atmintis

Spartinančioji atmintis - tai nedidelės talpos labai sparti atmintis dažniausiai sudaroma iš statinės operatyviosios atminties (SRAM) mikroschemų, kurioje saugomi ypač dažnai naudojami pagrindinės atminties fragmentai.

Kam reikalinga spartinančioji atmintis?

- Instrukcijų / duomenų įkėlimo laikas iš atminties į procesoriaus registrus labai ilgas (SDRAM ~50 ns) , palyginti su instrukcijos vykdymo trukme (1-4 taktai ~ 1 ns).
- Spartinančioji atmintis suteikia galimybę kreiptis į mažus pagrindinės atminties fragmentus **4–50 kartų** greičiau nei į DRAM.

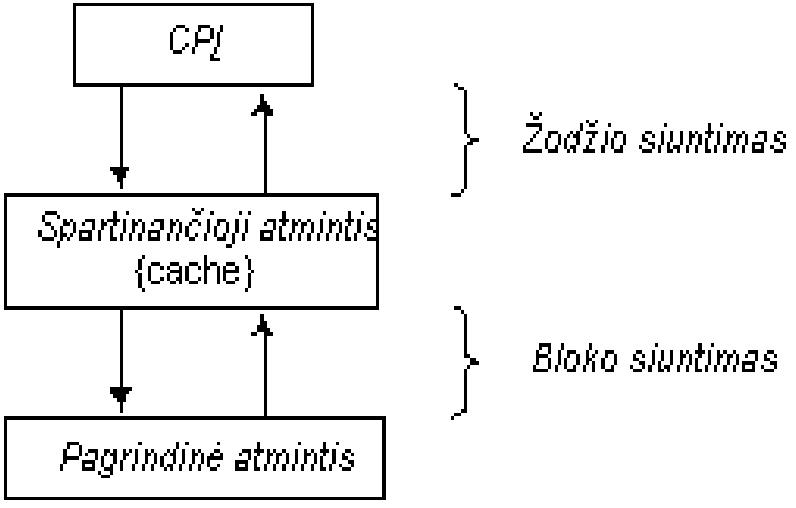
Lokališkumo principas

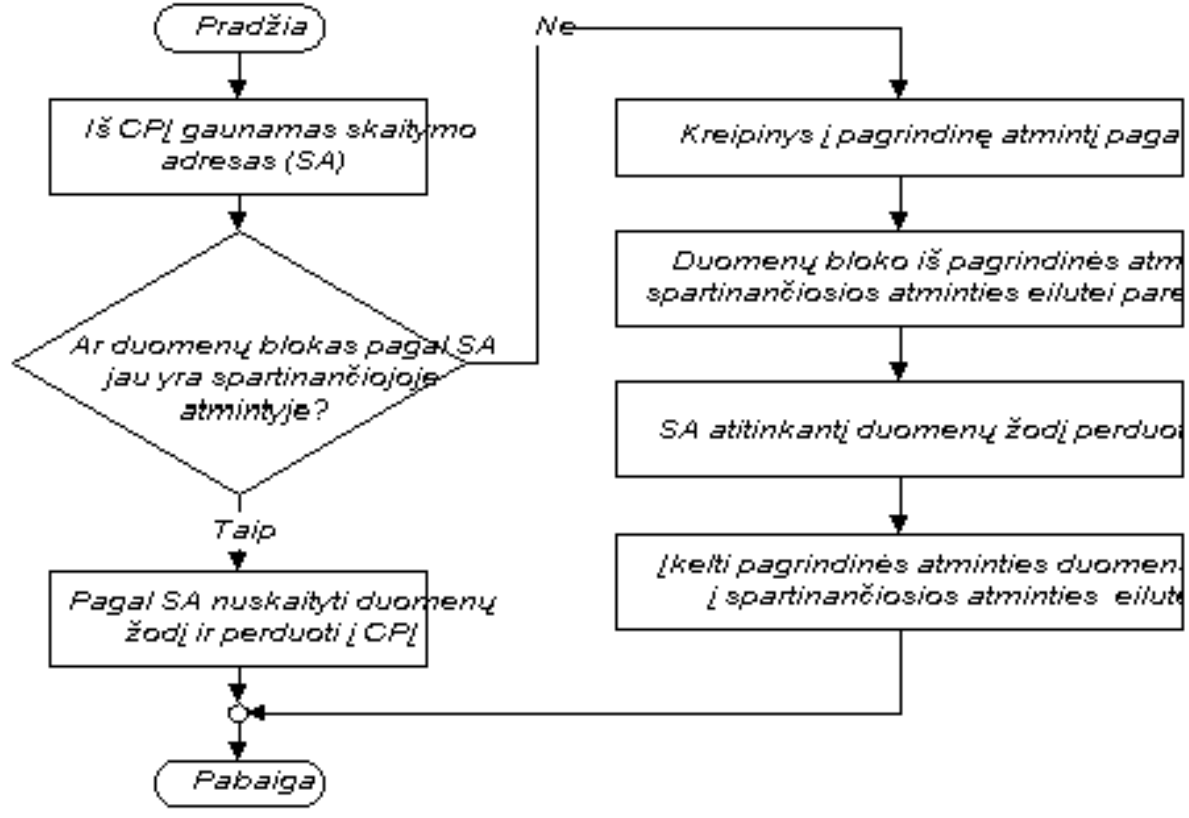
- Tyrimai rodo, kad programos linkę naudoti duomenis ir komandas, kurias jau yra panaudoje. Pavyzdžiui Intel Core2 procesorius vidinėje spartinančiojoje atmintyje saugo per 90% visų procesoriui būtinų adresų. Tai reiškia, kad daugiau kaip 90% kreipčių į atmintį bus atlikta per spartinančiąją atmintį.
- **Locality of Reference** – „kreipčių lokalizavimo“ teorijos pagrindinė koncepcija:
  - ***Bet kuriuo laiko momentu tam tikra pagrindinės atminties dalis (manoma, kad 10–20%) gali būti reikalinga procesoriui (su 80–90% tikimybe)*.**

Lokališkumo principas:

- laiko atžvilgiu: jei dabar reikalingas kuris nors elementas, labai tikėtina, kad netrukus vėl jo prisireiks;
- vietos atžvilgiu: jei dabar reikalingas kuris nors elementas, labai tikėtina, kad netrukus bus reikalingas ir jam gretimas.

Spartinačiosios atminties modelis





Spartinančioji atmintis (CPU cache)

Pagrindinės charakteristikos:

- Dydis
- Skaitymo metodai
- Rašymo metodai (Write policy)
- Pakeitimo metodai (Replacement policy)
- Komunikacijų protokolai tarp spartinančiųjų atminčių (koherencijos protokolai)
- Prastovų mažinimo būdai (HT, out-of-order CPU...)

Fizinė vieta kompiuteryje:

- CPU
- CPU modulis
- Atskirai nuo CPU

Spartinančiosios atminties skaitymo-rašymo architektūra

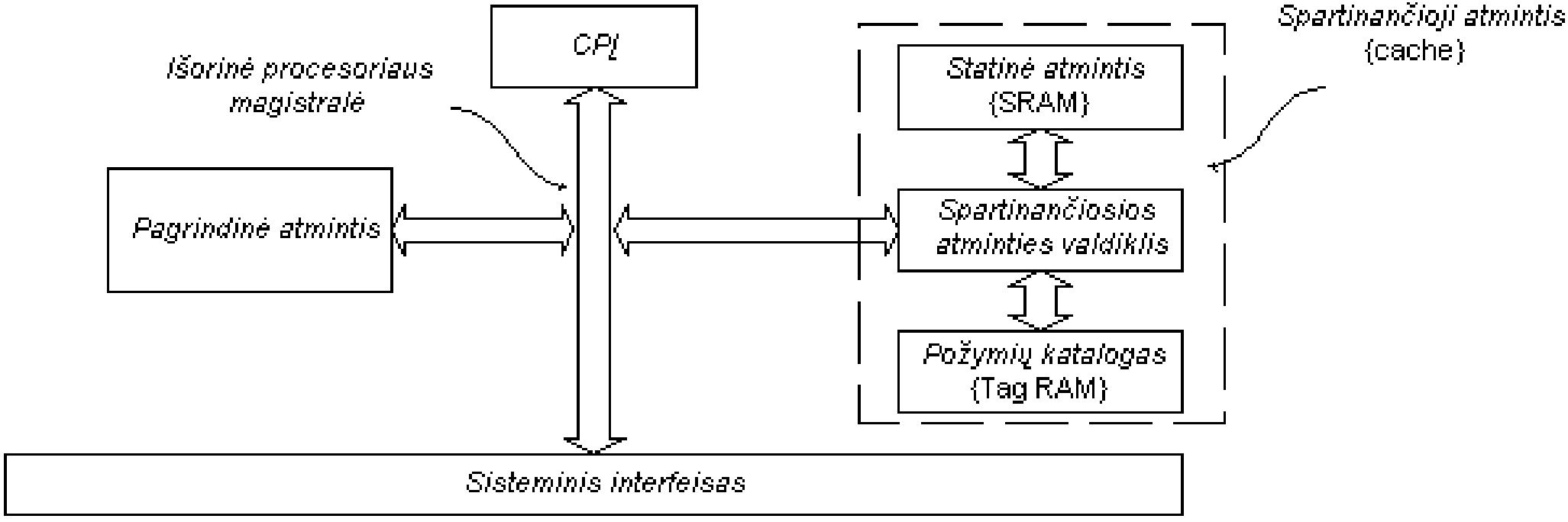
Skaitymo architektūra:

- peržiūros iš šalies (Look Aside),
- ištisinės peržiūros (Look Through)

Rašymo architektūra:

- atgalinis rašymas (Write Back)
- ištisinis rašymas (Write Through).

Peržiūros iš šalies (Look Aside) skaitymo architektūra



Peržiūros iš šalies (Look Aside) skaitymo architektūra

Peržiūros iš šalies architektūros skiriamasis bruožas yra tai, kad spartinančioji atmintis veikia lygiagrečiai su pagrindine atmintimi. Būtina pabrėžti, kad ir pagrindinė atmintis, ir spartinančioji atmintis „pastebi“ magistralės kreipčių į atmintį ciklus tuo pačiu metu. Tai atitinka atminties pavadinimą –„peržiūra iš šalies“.

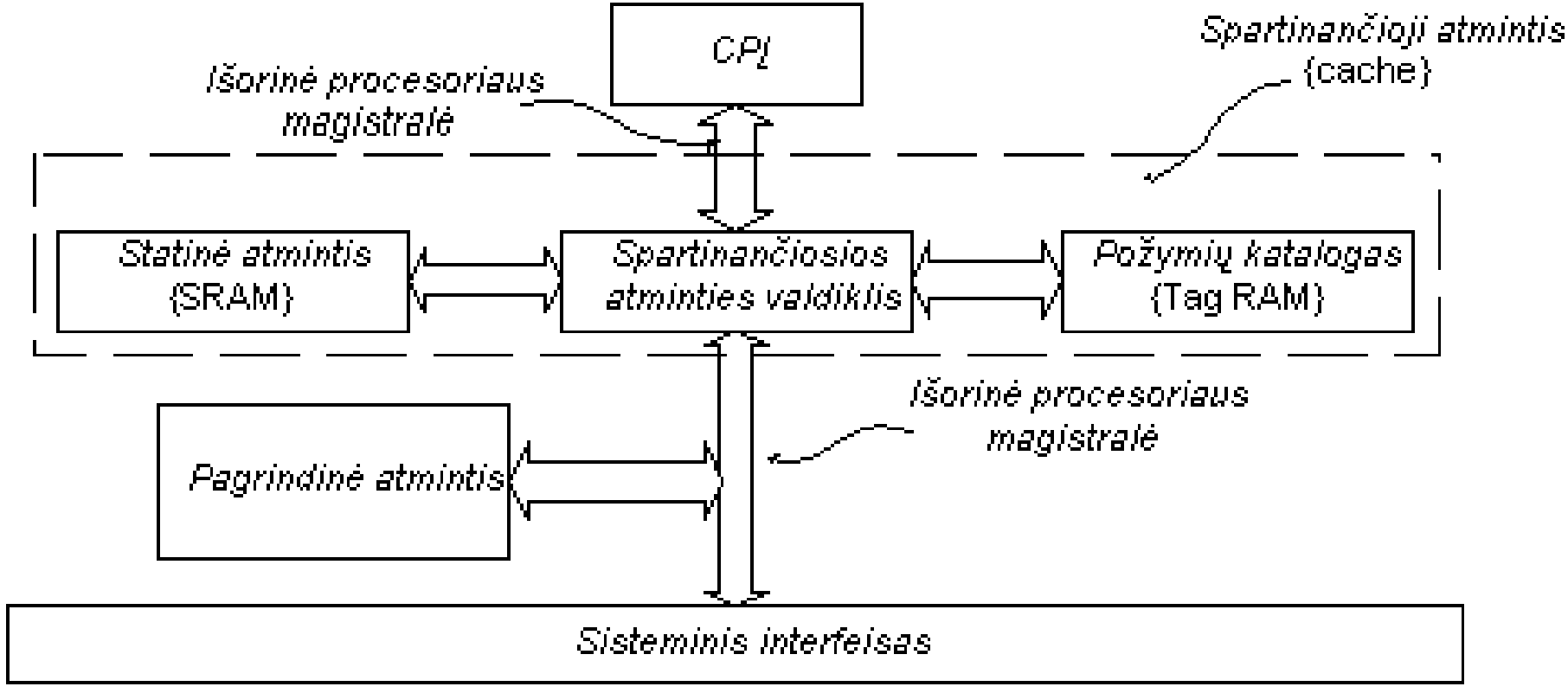
Privalumas

Ši architektūra užtikrina greitesnį sistemos atsaką į neigiamus spartinančiosios atminties rezultatus {*miss*}, kadangi ir dinaminė operatyvioji atmintis, ir spartinančioji atmintis vienu metu pastebi magistralės ciklą

Trūkumai

Procesorius negali kreiptis į spartinančiąją atmintį tuo metu, kai kitas magistralės valdiklis kreipiasi į pagrindinę atmintį.

Ištisinės peržiūros (Look Through) skaitymo architektūra



Ištisinės peržiūros (Look Through) skaitymo architektūra

Ištisinės peržiūros (Look Through) spartinančiosios atminties architektūros schema išdėstyta tarp procesoriaus ir pagrindinės atminties. Spartinančioji atmintis „stebi“ procesoriaus magistralės ciklus prieš jiems patekus į sisteminę magistralę.

Privalumas

Esant šiai architektūrai procesorius gali dirbti su spartinančiąja atmintimi tuo metu, kai kitas kompiuterio magistralės valdiklis kreipiasi į pagrindinę atmintį. Taip procesorius esti „izoliuotas“ nuo kompiuterio sistemos laukimo stadijų.

Trūkumas

Kreipimasis į pagrindinę atmintį, kai spartinančiojoje atmintyje nėra būtinų duomenų, sulėtėja, nes pagrindinė atmintis nepasiekiama, kol nėra patikrinta spartinančioji atmintis (tam reikia papildomų magistralės ciklų).

Įrašymo metodai (Write Policy)

Įrašymo metodai apibrėžia, kaip naudojama spartinančioji atmintis įrašymo į pagrindinę atmintį ciklo metu.

Reikalavimai įrašymo metodams:

- Neturi perrašyti spartinančiosios atminties bloko, jei pagrindinė atmintis nėra pakeičiama
- Daugelio procesorių sistemose, kiekvienas procesorius privalo turėti savo spartinančią atmintį
- I/O sistema gali kreiptis į pagrindinę atmintį tiesiogiai

Įrašymo metodai (Write Policy)

Yra du pagrindiniai rašymo į spart.atmintį metodai:

- **atgalinis įrašymas (Write-Back)**
- **ištisinis įrašymas (Write-Through).**

Pagal **atgalinio įrašymo (**Write-Back**)** metodą spartinančioji atmintis veikia kaip buferis, t. y. procesorius pradeda įrašymo ciklą, spartinančioji atmintis priima duomenis ir nutraukia šį ciklą.

Spartinančioji atmintis įrašo duomenis į pagrindinę atmintį tik tada, kai laisva procesoriaus magistralė. Šis metodas užtikrina kompiuterio sistemos spartą, nes procesorius gali tęsti darbą, o pagrindinė atmintis atnaujinama vėliau.
Dėl tokios įrašymo į pagrindinę atmintį kontrolės spartinančioji atmintis sudėtinga ir jos kaina didėlė.

Įrašymo metodai (Write Policy)

Ištisiniu įrašymu (Write-Through) **metodu** procesorius duomenis į pagrindinę atmintį įrašinėja visada per spartinančiąją atmintį. Spartinančiosios atminties turinys gali atsinaujinti, tačiau įrašymo ciklas nesibaigia tol, kol šie visi duomenys įrašomi į pagrindinę atmintį.

Šis metodas ne tiek sudėtingas ir todėl ne toks brangus. Kompiuterio sistemos, taikančios šį metodą, galimybės mažesnės, nes procesorius turi laukti, kol duomenys pasieks pagrindinę atmintį.

Spartinančiosios atminties komponentės

Spartinančiosios atminties posistemį galima suskirstyti į tris funkcinius blokus:

- statinė atmintis – SRAM
- spartinančiosios atminties požymių (tag) statinė atmintis (požymių katalogas {Tag RAM})
- spartinančiosios atminties valdiklis

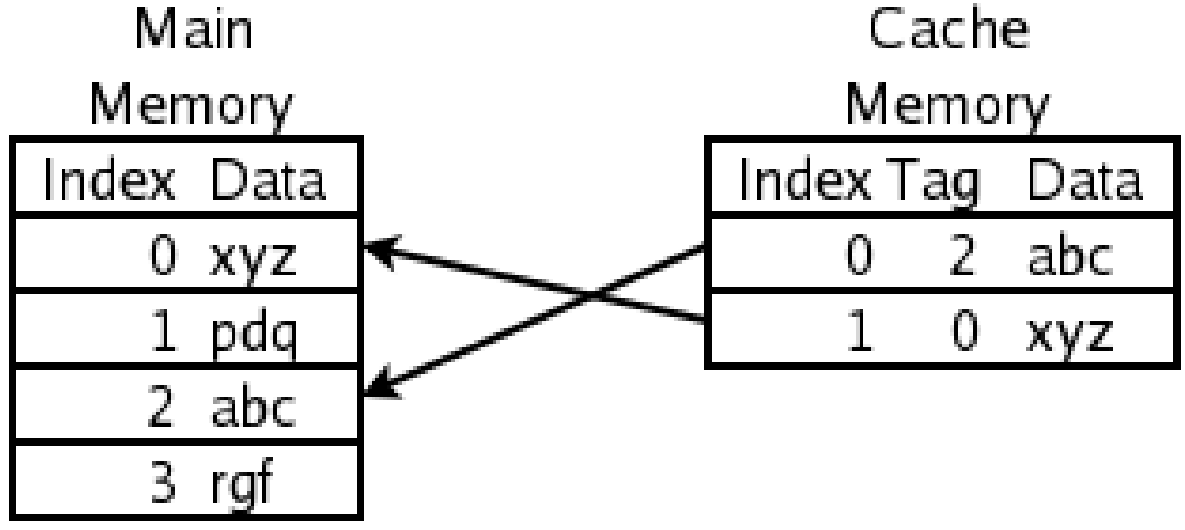
Spartinančiosios atminties komponentės

Statinė atmintis {SRAM}. *Statinė laisvosios kreipties atmintis* {Static Random Access Memory – SRAM} yra atminties, saugančios duomenis, blokas. Nuo SRAM talpos priklauso pačios spartinančiosios atminties talpumas.

Požymių katalogas {Tag RAM}. Tai maža statinės atminties dalis, kur tam tikra savita forma saugomi adresai tų duomenų, kurie šiuo metu yra spartinančiosios atminties SRAM'e.

Spartinančiosios atminties valdiklis. Tai spartinančiosios atminties „smegenys". Valdiklis atnaujina *SRAM* ir *Tag RAM* bei kontroliuoja nustatytą įrašymo metodą. Spartinančiosios atminties valdiklis taip pat nustato, ar iškviestosios pagrindinės atminties turinys talpinamas į spartinančiąją atmintį (ar ji yra „kešuojama").

Spartinančiosios atminties struktūra



Spartinančiosios atminties eilutės (**cache line**) dalys:

- Saugomo pagrindinės atminties bloko adresas
- Duomenys
- Būsenos požymis (galiojanti, tuščioji).

Spartinančiosios atminties struktūra

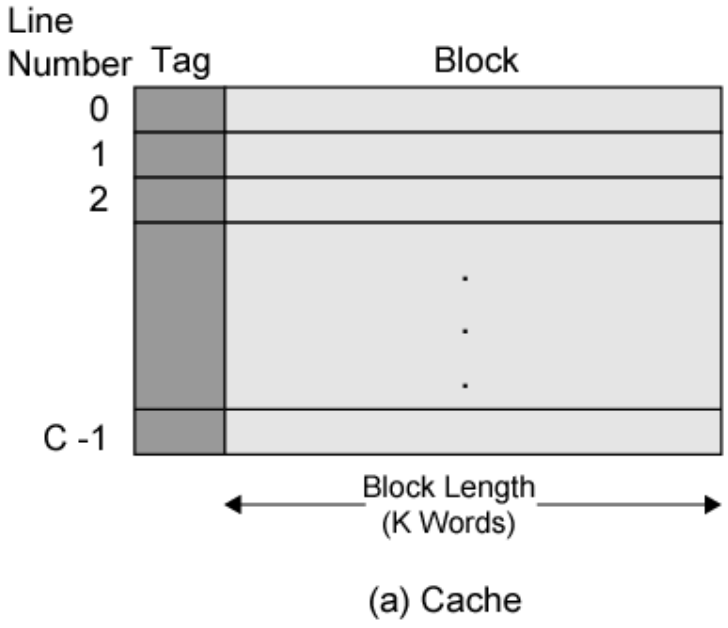
Spartinančiosios atminties struktūros terminai:

- spartinančiosios atminties puslapis (**cache page**)
- spartinančiosios atminties eilutė (**cache line**).

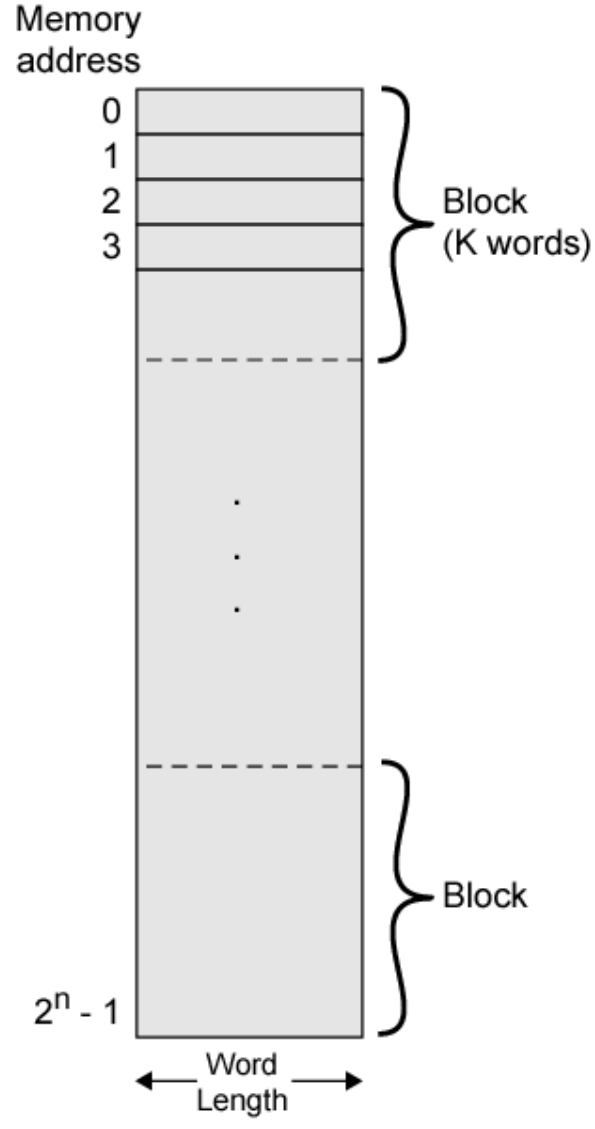
Pagrindinė atmintis tariamai suskaidoma į fragmentus, vadinamus *spartinančiosios atminties puslapiais (sets)*. Šio puslapio dydis priklauso nuo pačios spartinančiosios atminties talpos ir nuo to, kaip spartinančioji atmintis sudaryta.

Spartinančiosios atminties puslapis padalytas į mažesnius fragmentus, vadinamus *spartinančiosios atminties eilutėmis (blokais)*. Spartinančiosios atminties eilutės dydis priklauso ir nuo spartinančiosios atminties, ir nuo procesoriaus architektūros. Dažnai tai būna 4-256 baitų.

Spartinančiosios atminties struktūra



(a) Cache



(b) Main memory

Spartinančiosios atminties struktūra

Spartinančiosios atminties eilutė gali būti ***galiojanti*** *(valid*) – tai reiškia, kad esamu laiko momentu ji tiksliai atvaizduoja atitinkamą pagrindinės atminties bloką.
Priešingu atveju - eilutė yra ***negaliojanti*** (*tuščioji*).

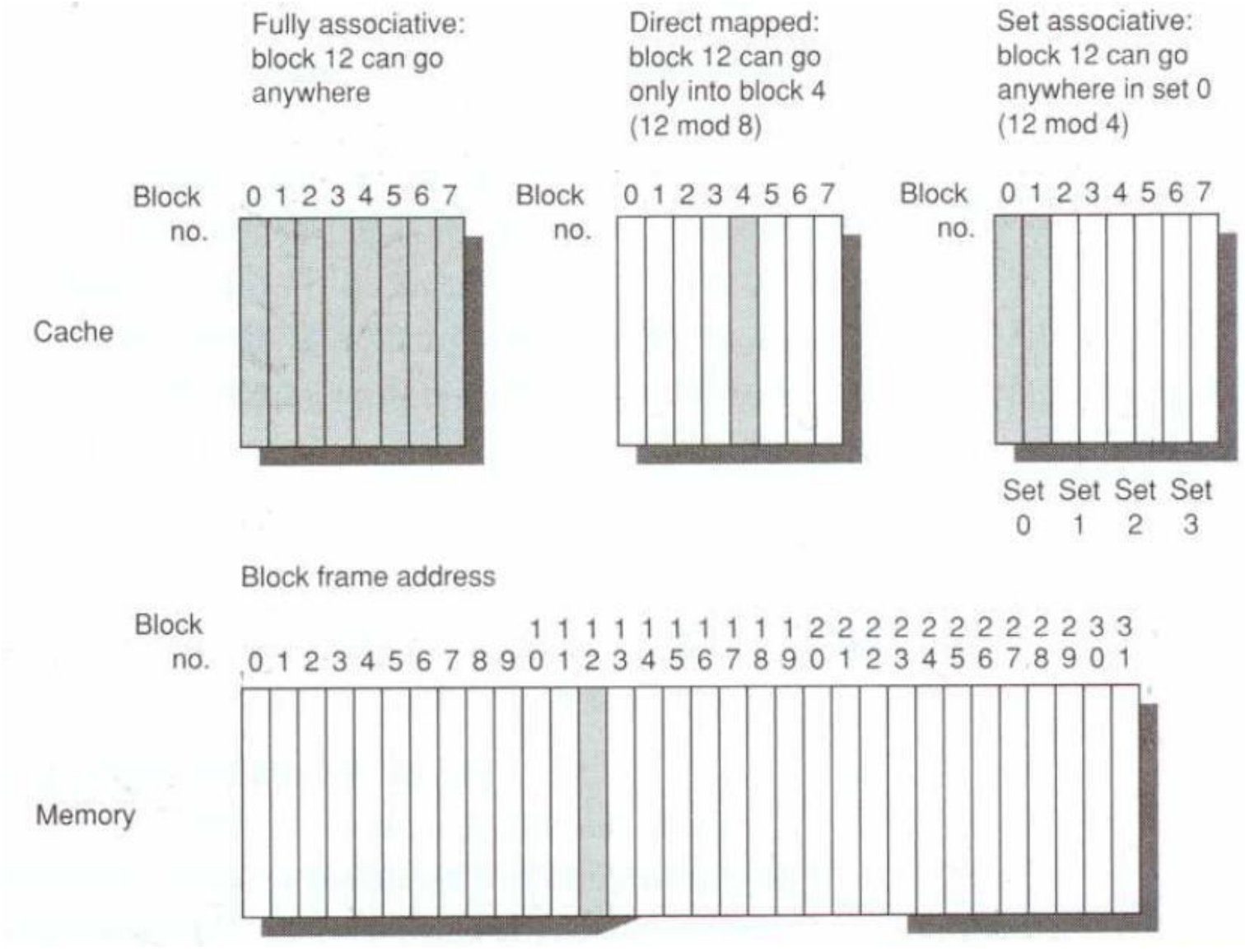
Kompleksinė informacija (**tag**) apie tai, koks būtent pagrindinės atminties blokas užima tam tikrą spartinančiosios atminties eilutę (t. y. adreso aukštesnioji dalis arba puslapio numeris), ir apie jos būseną, saugoma specialioje su konkrečia eilute susijusioje požymių atminties (Tag RAM) ląstelėje (požymių atmintis kartais vadinama katalogu).

Vykstant keitimuisi duomenimis su pagrindine atmintimi spartinančiosios atminties eilutė paprastai naudojama visa iš karto (kai spartinančioji atmintis nesuskaidyta į sektorius).

Spartinančiosios atminties tipai (Associativity)

- **Tiesioginio atitikimo sp. atm. *(Direct Mapped)* -** kiekvienas iš pagrindinės atminties paimtas eilutės dydžio blokas turi vienintelę apibrėžtą vietą spartinančiojoje atmintyje.
- **Pilnai asociatyvus sp. atm. *(Fully-Associative)* -** kiekvienas iš pagrindinės atminties paimtas eilutės dydžio blokas gali būti bet kurioje vietoje spartinančiojoje atmintyje.
- **Dalinai asociatyvus sp. atm. *(Set Associative)* -** kiekvienas iš pagrindinės atminties paimtas eilutės dydžio blokas gali būti bet kurioje iš **k** vietų spartinančiojoje atmintyje.
  Skaičius **k** vadinamas asociatyvumo laipsniu arba krypčių skaičiumi. (2-way, 4-way...)

Spartinančiosios atminties asociatyvumo lygiai

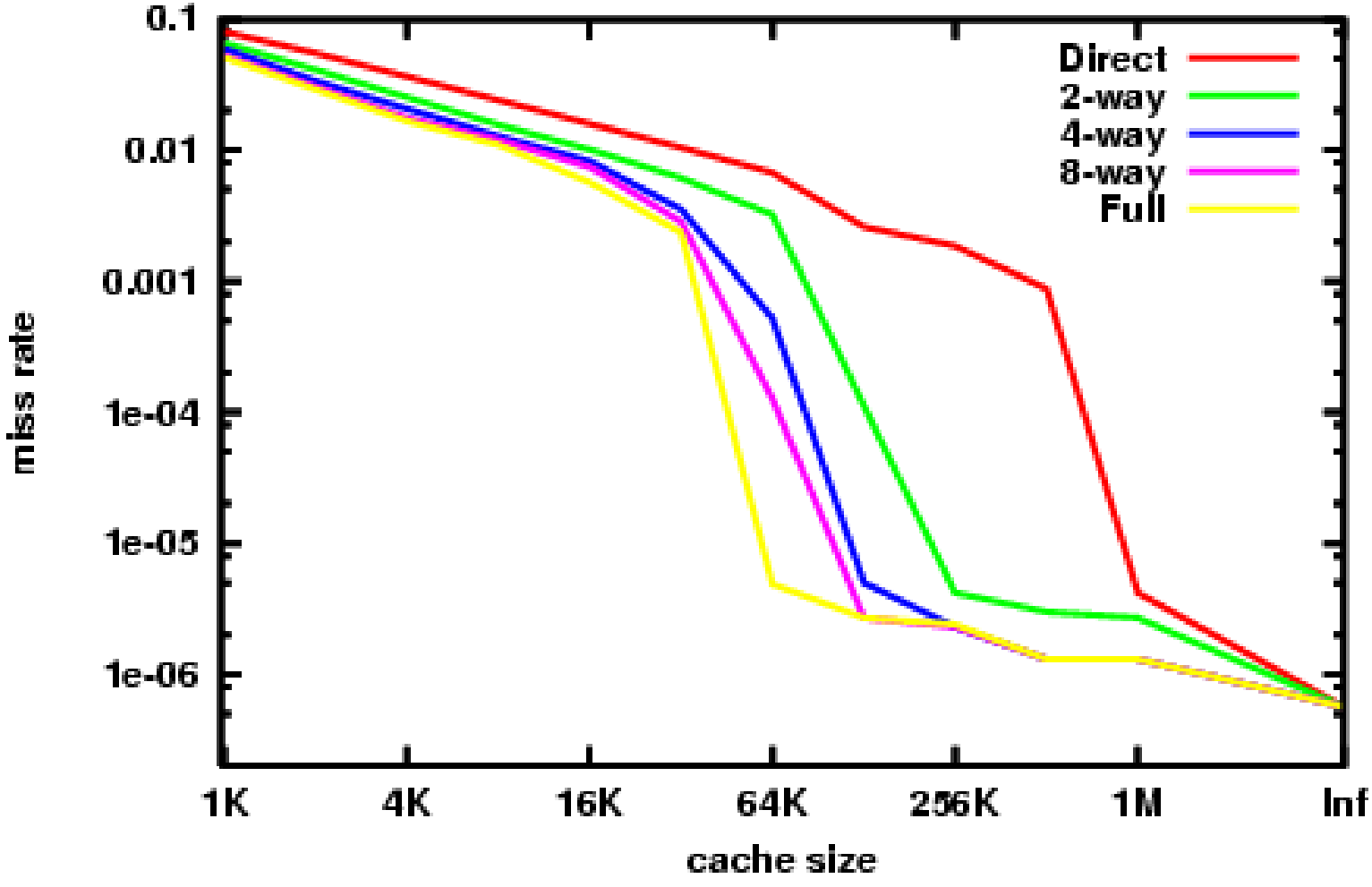


Spart.atminties organizavimas

Spartinančiosios atminties parametrai

- nuskaitymo pataikymas (*cache hit*)
- nuskaitymo nepataikymas (*cache miss*)
- pataikymo dažnis – **fh** (*hit ratio)*
- nepataikymo dažnis – (miss ratio) **fm***(fm = 1 - fh*)
- išrinkimo laikas pataikius – (*hit time)*
- uždelsimas nepataikius - (*miss penalty*)

Spart.atminties efektyvumas



Spartinančios atminties struktūros procesoriuose

Modelis	Dydis	Eilutės ilgis	Organizacija
VAX-11/780	8 KB	8	2 krypčių dalinai asociatyvus
Intel 80486	8 KB	16	4 krypčių dalinai asociatyvus
Pentium	2 × 8 KB	32	2 krypčių dalinai asociatyvus
Pentium II	2 × 16 KB	32	4 krypčių dalinai asociatyvus
Pentium III	2 × 32 KB	32	4 krypčių dalinai asociatyvus
PowerPC 601	32 KB	32	8 krypčių dalinai asociatyvus
PowerPC 604	2 × 32 KB	32	4 krypčių dalinai asociatyvus
Alpha 21164	2 × 32 KB	32	tiesioginio atitikimo
Pentium 4	8 KB + 12 KB	64	4 krypčių dalinai asociatyvus
AMD Athlon	2 × 64 KB	64	2 krypčių dalinai asociatyvus

Spart. atminties eilutės pakeitimo algoritmai (replacement policy)

- Atsitiktinis pakeitimas
- LRU (*Least-recently used*) - seniausiai panaudota
- LFU (Least frequently used) - mažiausiai kartų naudota
- FIFO (First in First out)

Eilutės pakeitimo algoritmų lyginimas (*Cache misses 1000 instrukcijų*)

	Associativity								
	Two-way			Four-way			Eight-way		
Size	LRU	Random	FIFO	LRU	Random	FIFO	LRU	Random	FIFO
16 KB	114.1	117.3	115.5	111.7	115.1	113.3	109.0	111.8	110.4
64 KB	103.4	104.3	103.9	102.4	102.3	103.1	99.7	100.5	100.3
256 KB	92.2	92.1	92.5	92.1	92.1	92.5	92.1	92.1	92.5

Spart. atminties našumas

CPU darbo laikas =

(CPU ciklai + Atminties prastovos ciklai) x periodas

Atminties prastovos ciklai =

Nepataikymų skaičius x Uždelsimas nepataikius =

IC x (Nepataikymai/Instrukcija) x Uždelsimas nepataikius **=**

IC x (Atminties kreiptis/Instrukcija) x Nepataikymų dažnis x Uždelsimas nepataikius

Vidutinis atminties pasiekimo laikas =

Nuskaitymas pataikius + Nepataikymų dažnis x uždelsimas nepataikius

Pavyzdys

Užduotis

Turime kompiuterį, kurio CPI lygus 1, kai visi kreipiniai į spartinančią atmintį yra sėkmingi. Duomenys yra skaitomi+rašomi (load, store) ir tai sudaro 50% instrukcijų. Nepataikymo prastova 25 ciklai, nepataikymo dažnis 2%. Kaip padidėtų kompiuterio našumas, jei visos instrukcijos būtų spart. atminyje.

Sprendimas

Jei visi duomenys būtų spartinančioje atmintyje, tai CPU darbo laikas:

CPU darbo laikas = (CPU ciklai + Atminties prastovos ciklai) x periodas =
= (IC x CPI + 0) x periodas = IC x 1.0 x periodas

Pavyzdžio tęsinys

Jei nepataikymo dažnis lygus 2%, tuomet:

Atminties prastovos ciklai = IC x (Atminties kreipiniai / Instrukcija) x nepataikymų dažnis x prastova nepataikius = IC x 0.5 x 0.02 x 25 = IC x 0.25

CPU darbo laikas$_{cache}$ = (CPU ciklai + Atminties prastovos ciklai) x periodas = IC x (1 + 0.25) x periodas

Našumų santykis = (IC x 1.25 x periodas) / (IC x 1.0 x periodas) = 1.25

Spart. atminties optimizavimas

Yra nubrėžtos 6 spartinančios atminties optimizavimo gairės. Jos sugrupuotos į tokias 3 grupes:

- **Nepataikymo dažnio mažinimas**
  - didinamas bloko dydis
  - didinamas spartinančiosios atminties dydis
  - didinamas asociatyvumo lygis
- **Nepataikymo prastovos mažinamas**
  - kelių lygių spartinančioji atmintis
  - skaitymo prioritetas prieš rašymą
- **Pataikymo laiko mažinimas**
  - vengiamas adreso transliavimas kai indeksuojami puslapiai

Bloko (eilutės) didinimas

	Cache size			
Block size	4K	16K	64K	256K
16	8.57%	3.94%	2.04%	1.09%
32	7.24%	2.87%	1.35%	0.70%
64	7.00%	2.64%	1.06%	0.51%
128	7.78%	2.77%	1.02%	0.49%
256	9.51%	3.29%	1.15%	0.49%

Didinant bloko dydį, nepataikymo dažnis daugeliu atvejų mažėja.

Procentais nurodytas nepataikymų dažnis

Bloko (eilutės) didinimas

Block size	Miss penalty	Cache size			
		4K	16K	64K	256K
16	82	8.027	4.231	2.673	1.894
32	84	**7.082**	3.411	2.134	1.588
64	88	7.160	**3.323**	**1.933**	**1.449**
128	96	8.469	3.659	1.979	1.470
256	112	11.651	4.685	2.288	1.549

Didinant bloko dydį, nepataikymų pratova daugeliu atvejų didėja.

Pateiktas atminties pasiekimo laikas ciklais ir ns

Spart.atminties lygmenys

Siekiant rasti optimalų sprendimą tarp nepataikymo dažnio (*hit ratio)* ir uždelsimo (*latency)*, spartinančioji atmintis dalinama į lygmenis.

Procesoriaus spart. atmintis dalinama į lygmenis:

1) L1 spart.atmintis
   - L1 duomenų lygmuo (8 KB-16KB)
   - L1 instrukcijų lygmuo (16-32 KB)
2) L2 spartinančioji atmintis (256-2MB)

3) L3 – dažniausiai atskiroje shemoje esanti spartinančioji atmintis (iki 9MB Itanium2, iki 256MB IBM POWER5)

Vidutinis atminties pasiekimo laikas =
Nuskaitymas pataikius(L1) + Nepataikymų dažnis(L1) x [Nuskaitymas pataikius(L2) + Nepataikymų dažnis(L2) x uždelsimas nepataikius (L2)]

Pavyzdys

Užduotis

Iš 1000 kreipinių į atmintį (RAM), 40 nepataikymų yra L1 spart.atmintyje, 20 nepataikymų L2 atmintyje. Koks yra nepataikymo dažnis?

Duota: nepataikymo prastova kreipiantis iš L2 į RAM 200 ciklų, pataikymo laikas į L2 spart, atmintį 10 ciklų, pataikymo laikas į L1 yra 1 ciklas, santykis Atminties kreipiniai / Instrukcija = 1.5. Koks vidutinis atminties pasiekimo laikas?

Sprendimas

Nepataikymo dažnis (L1) = 40 /1000 = 4%

Nepataikymo dažnis (L2) = 20 / 40 = 50% (lokalus)

Nepataikymo dažnis (L2) = 20 / 1000 = 2% (globalus)

Vidutinis atminties pasiekimo laikas =

Nuskaitymas pataikius(L1) + Nepataikymų dažnis(L1) x [Nuskaitymas pataikius(L2) + Nepataikymų dažnis(L2) x uždelsimas nepataikius (L2)] =

= 1 + 40/1000 x (10 + 20 / 40 x 200) = 5.4 ciklai

Spart. atminties optimizavimas

Technique	Hit time	Miss penalty	Miss rate	Hardware complexity	Comment
Larger block size		–	+	0	Trivial; Pentium 4 L2 uses 128 bytes
Larger cache size	–		+	1	Widely used, especially for L2 caches
Higher associativity	–		+	1	Widely used
Multilevel caches		+		2	Costly hardware; harder if L1 block size ≠ L2 block size; widely used
Read priority over writes		+		1	Widely used
Avoiding address translation during cache indexing	+			1	Widely used

Optimizavimo technikų įtaka spartinančiosios atminties charakteristikoms:

- \- kenkia
- \+ padeda
- 0 .. 3 sudėtingumo lygis