

Технически Университет - Габрово	Организация на компютъра
Тема: Кеш памет	Лабораторно № 5
Цел: Запознаване с предназначението, архитектурните особености и използването на кеш паметта при персоналните компютри	

I. Теоретична част

Когато при комуникация между две устройства, от които едното е в пъти по-бързо от другото, обменът на информация от гледна точка на бързодействие не е ефективен. Причината за това е, че „бързото“ устройство трябва да изчака „бавното“ устройство всеки път, когато очаква да получи данни от него. В този случай бързодействието на системата е съизмеримо с бързодействието на „бавното“ устройство. Едно възможно решение е използването на буферна памет, която физически се намира между „бързото“ и „бавното“ устройства. Тази памет се нарича кеш (cache) и е от тип статична RAM (SRAM). Целта е „бързото“ устройство да комуникира само с достатъчно бързата кеш памет. В идеалният случай, бързодействието на кеш паметта и на „бързото“ устройство трябва да съвпадат. Така не се налага изчакване и обменът е с максимална скорост.

Основните проблеми при системите, които използват кеш памет, са няколко:

1. Колкото времето за достъп до кеш паметта е по-малко, толкова кеш паметта е по-скъпа. Това налага да се работи с кеш памет с малък капацитет.
2. Когато разликата между бързодействието между двете устройства е много голяма може да се наложи да се използват няколко нива на кеширане на информацията. Колкото тази разлика е по-голяма, толкова повече нива са необходими. В този случай, кеш паметта с която „бързото“ устройство комуникира директно се нарича кеш от ниво 1 (L1 cache). Тя е най-бърза и най-скъпа и затова е с малък обем. Последното ниво на кеш паметта (Ln cache) има директна връзка с „бавното“ устройство. Тази памет е най-бавна и най-евтина и затова е с най-голям обем при сравнение с останалите нива на кеш. При компютърните системи се използват 2, 3 и повече нива на кеширане.
3. Необходима е допълнителна апаратна част, която да синхронизира обмена между кеш паметта и „бавното“ устройство. Това се налага, тъй като „бързото“ устройство няма физическа връзка с бавното устройство, а само с кеш паметта. Колкото повече нива на кеширане се използват, толкова по-сложна и по-скъпа е необходимата апаратна част.

1.1. Кеш памет при персоналните компютри

При персоналните компютри кеш паметта се използва основно с цел буфериране на обмена между микропроцесора (CPU) и оперативната памет (DRAM). Кеш памет се използва и при всички периферни устройства при които има съществена разлика в бързодействието на модулите в тях, които си комуникират. Например, при твърдите дискове се използва кеш памет, тъй като има съществена разлика в бързодействието на интерфейса, чрез който диска се свързва към дънната платка (хост адаптер), и механиката на диска (четящо-записващи глави). В това упражнение се разглежда комуникацията между CPU и DRAM, посредством помощта на чипсета.

1.2. Архитектурни особености на кеш паметите

При съвременните компютърни конфигурации микропроцесорът има връзка само с L1 кеш. Презареждането на L1 кеш с данни и инструкции, които са необходими на CPU се реализира от чипсета (северен мост или MCH). При някои системи, L1 кеш е общ за инструкции и данни. При други системи, с цел повишаване на производителността, L1 кеш е физически разделен на две: L1 кеш за данни и L1 кеш за инструкции. В този случай се използва Харвардска архитектура на кеша, което подобрява производителността на системата. Възможно е кеш паметта за данни да е двупортова, за да е възможно едновременно извличане на два операнда.

1.3. Асоциативна и неасоциативна кеш

Тъй като кеш паметта е с много по-малък капацитет при сравнение с капацитета на DRAM, е много важно от колко базови физически адреси от DRAM може да се кешира (записва в кеш паметта). Микропроцесорът работи с кеш паметта в блоков режим. Най-малкото количество информация е един кеш ред (линия). Нейният размер в байтове е кратен на степените на 2, например 16 байта, 32 байта, 64 байта. Следователно, CPU чете или записва в кеш паметта един кеш ред. Чипсетът работи аналогично с останалите нива на кеш и следователно чете/записва от/в DRAM минимум един кеш ред.

Когато може да се кешира информация, която започват от един базов физически адрес от DRAM се казва, че кеш паметта е с директна адресация или неасоциативна. Тази организация на кеш не е особено ефективна, тъй като ако трябва да се кешира информация, която започва от друг физически адрес се налага изтриване на текущата информация в кеш, за да се запише новата информация. Най-добрият вариант, от гледна точка на ефективност, е кеш паметта да позволява кеширане от произволен брой физически адреси, до запълване на кеша. Кеш паметта, която позволява това, се нарича напълно асоциативна. Тя обаче е и най-скъпа, тъй като за реализацията ѝ са необходими най-много интегрални транзистори. На практика, най-често използваната кеш памет, е асоциативната в няколко направления. При нея е възможно кеширане от определен брой физически адреси, най-често 2, 4, 6 или 8. Тази памет е компромисна от гледна точка на цена и производителност.

1.4. Съгласуваност на паметта

Съвременните операционни системи са многопотребителски и многозадачни, които се стартират на многопроцесорни и/или многоядрени платформи. В този случай основният проблем при системите, които използват кеш памет, е гарантиране на съгласуваност на паметта. Паметта в една компютърна конфигурация е съгласувана, ако съдържанието на всички копия на части от DRAM в кеш паметта и DRAM съвпадат. Постигането на съгласуваност е задължително, тъй като е възможно някои споделени ресурси (променливи, обекти) от един процес да са кеширани в различни кеш модули (на различни физически процесори или различни физически ядра на един многоядрен процесор). В този случай промяната на един ресурс води до запис в конкретен кеш. В този момент паметта вече не е съгласувана. Ако се достъпи (чете) същият ресурс от друг кеш, той няма да е със същата стойност, а това е недопустимо. Най-лесният начин за гарантиране на съгласуваност на паметта е всеки запис в кеш да води до запис в DRAM и обновяване на всички останали кеш. Тази стратегия е много бавна, особено, ако има

множество последователни записи. Съгласуваността може да се гарантира със забавяне във времето. В този случай се преминава към гарантиране на съгласуваност в момента когато тази съгласуваност е наложителна. Колкото повече нива на кеширане се използват и колкото повече ядра има един процесор, толкова е по-сложно и бавно е да се гарантира съгласуваност на паметта. За целта се използват различни стратегии за запис в кеш наречени методи за кеширане. Например, при процесорите на Intel най-често използваните методи за кеширане са WT, WB, WP, WC и UC.

II. Задачи за изпълнение

Задача 1.

Направете сравнителен анализ на различните видове кеш памет като попълните следната таблица:

Памет	Вероятност за попадение в кеш	Брой на адресните компаратори
Неасоциативна		
Напълно асоциативна		
Асоциативна в N направления		

Задача 2.

Една компютърна конфигурация има следната кеш памет:

L1 Data cache	32 KBytes, 8-way set associative, 64-byte line size
L1 Instruction cache	32 KBytes, 8-way set associative, 64-byte line size
L2 cache	2 x 256 KBytes, 8-way set associative, 64-byte line size
L3 cache	3 MBytes, 12-way set associative, 64-byte line size

Анализирайте какво представлява всяка една от тези памети. С каква архитектура е L1 кеш?

Задача 3.

Програмното разпознаване на количеството и типа на инсталираната кеш памет се реализира чрез инструкция CPUID. Потърсете информация коя функция от CPUID се използва за целта и какъв е форматът на връщаната от нея информация. Тествайте количеството и типа на инсталираната кеш памет в компютърна конфигурация с която работите в момента. Тествайте скоростта на обмен в GB/s при четене, запис и копиране от и в кеш паметта. Какво е времето на достъп в ns до L1 и L2 кеш.

Задача 4.

Колко нива на кеширане се използват при процесори Itanium. Какво е количеството на кеш паметта и какъв е нейния тип при всяко ниво на кеширане.

Задача 5.

Направете сравнителен анализ на следните методи за кеширане:

Метод	Кога се обновява оперативната памет	Възможно ли е използване при многопроцесорни конфигурации	Поддържа ли съгласуваност на паметта
WT			
WB			
WP			
WC			

Задача 6.

Опишете характеристиките на протокол MESI:

Състояние на кеш ред	Има ли съгласуваност на паметта	Редът има ли копие в кеш на други процесори	Какво се случва при запис в такъв ред
Modified			
Exclusive			
Shared			
Invalid			

III. Допълнителни въпроси

1. За каква област от паметта би трябвало кеширането да се забрани (да се зададе UC)?
2. При работа с кеш паметта се използват следните инструкции: INVD, WBINVD и INVLPG. Опишете действието на всяка от тях.
3. Коя кеш памет, L1 или L2, е с по-голямо значение за производителността на една компютърна система? Обосновайте своя отговор.