ЛЕКЦІЯ 4

Архітектура системи команд та пам’ять

Системою команд обчислювальної машини називають повний набір команд, які може виконувати дана ОМ. В свою чергу, під *архітектурою системи команд* (АСК) прийнято визначати ті засоби обчислювальної машини, які видні та доступні програмісту. АСК можна розглядати як лінію узгодження потреб розробників програмного забезпечення з можливостями творців апаратури обчислювальної машини (рис. 4.1).

Мета одних та інших – реалізація обчислень найбільш ефективним способом, тобто за мінімальний час, і тут важливу роль грає правильний вибір архітектури систем команд.

В спрощеній трактовці час виконання програми (T) можна визначити через число команд в програмі (N), середню кількість тактів процесора, припадаючих на одну команду (CPI), та довжину тактового періоду τ;

$$T = N \times CPI \times \tau.$$

Рисунок 4.1 - Архітектура системи команд як інтерфейс між програмним і апаратним забезпеченням

Класифікація архітектури системи команд

В історії розвитку обчислювальної техніки як в дзеркалі відображуються зміни, що відбувалися в поглядах розробників на перспективність однієї чи іншої архітектури системи команд. Ситуацію, що склалася в даний час в області АСК, ілюструє рис. 4.2.

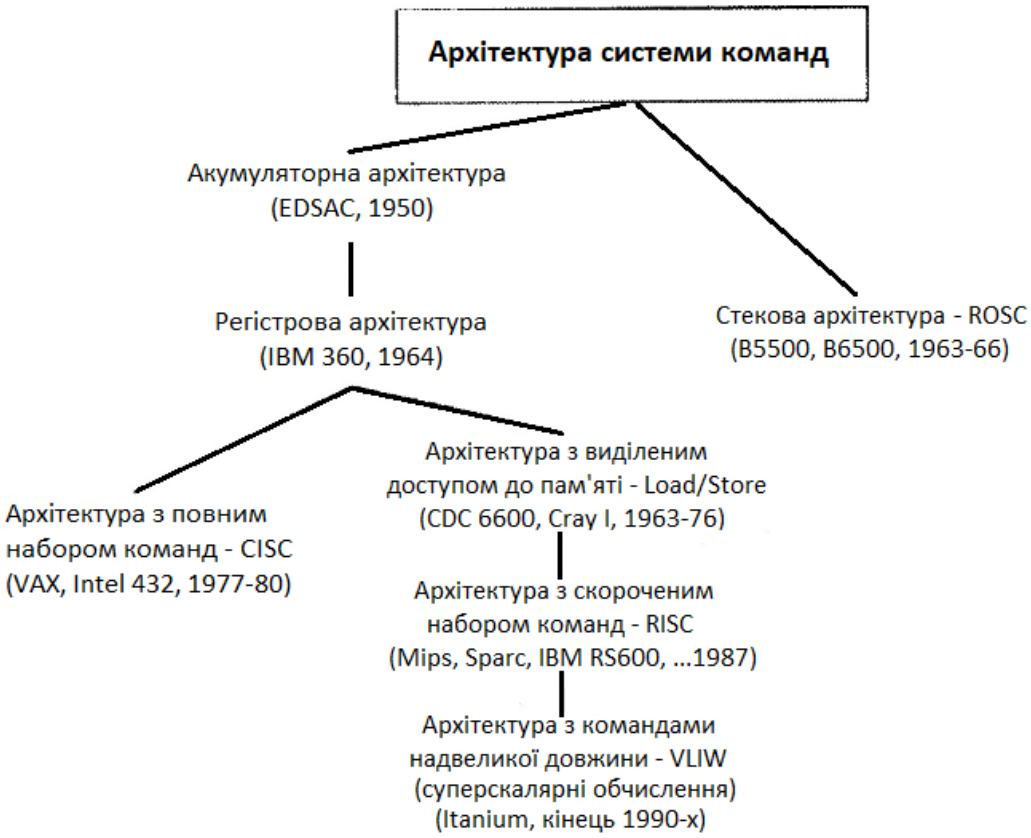

Рисунок 4.2. - Хронологія розвитку архітектур системи команд

Серед мотивів, що частіше за все зумовлюють перехід до нового типу АСК, зупинимося на двох найбільш суттєвих. Перший – це склад операцій, що виконуються обчислювальною машиною, та їх складність. Другий – місце зберігання операндів, що впливають на кількість та довжину адрес, вказаних в адресній частині команд обробки даних. Саме ці характеристики взяті як показники класифікації архітектур систем команд.

Класифікація по складу та складності команд

Сучасна технологія програмування орієнтована на мови високого рівня (МВР), головна мета яких – полегшити процес програмування. Але перехід до МВР породив серйозну проблему: складні оператори, характерні для МВР, суттєво відмінні від простих машинних операцій, які реалізовані в більшості обчислювальних машинах. Наслідком такої невідповідності є недостатньо ефективне виконання програм на ОМ. Проблема отримала назву семантичний

розрив, а для її розв'язку розробники обчислювальних машин в даний час вибирають один з трьох типів АСК:

- архітектура з повним набором команд: CISC (Complex Instruction Set Computer);

- архітектура з скороченим набором команд: RISC (Reduced Instruction Set Computer);

- архітектура з командними словами надвеликої довжини: VLIW (Very Long Instruction Word).

В CISC-архітектурі семантичний розрив долається за рахунок розширення системи команд, доповнення її складними командами, семантично аналогічними операторам МВР. Основоположником CISC-архітектури вважається компанія IBM, яка почала застосовувати даний підхід з родиною машин IBM 360 і продовжує його в своїх потужних сучасних універсальних ОМ (мейнфреймах). Аналогічний підхід характерний і для компанії Intel в її мікропроцесорах серії x86. Для CISC-архітектури характерні:

- наявність в процесорі порівняно невеликої кількості регістрів загального призначення;

- велика кількість машинних команд, частина яких апаратно реалізують складні оператори МВР;

- різноманітність способів адресації операндів;

- багато форматів команд різної розрядності;

- наявність команд, де обробка здійснюється зі зверненням до пам'яті.

До типу CISC можна віднести майже всі ОМ, які випускались до середини 1980-х років, і значну частину сучасних. Розглянутий спосіб рішення проблеми семантичного розриву разом з тим веде до ускладнення апаратури ОМ, головним чином, пристроїв керування, що, в свою чергу, негативно впливає на продуктивність ОМ вцілому. Ця обставина спонукала більш уважно проаналізувати програми, отримані після компіляції з МВР. Було зроблено комплекс випробувань, в результаті яких з'ясувалося, що частина додаткових команд, еквівалентних операторам МВР, в загальному об'ємі програм не перевищує 10 – 20%, а для деяких найбільш складних команд навіть 0,2%. В той же час об'єм апаратних засобів, необхідних для реалізації таких додаткових команд, зростає досить суттєво. Так, ємність мікропрограмної пам'яті, що зберігає мікропрограми виконання всіх команд ОМ, із-за введення складних команд може збільшуватись на 60%.

Детальний аналіз результатів згаданих випробувань призвів до серйозного перегляду традиційних рішень, результатом чого стала поява RISC-архітектури. Термін RISC вперше був застосований Д. Патерсоном та Д. Дитцелем в 1880 році. Суть полягає в обмежені списку команд ОМ найчастіше використовуваними простими командами, що оперують даними, розміщеними лише в регістрах процесорів. Звернення до пам'яті допускається лише за допомогою спеціальних команд читання та запису. Різко зменшується кількість форматів команд і способів вказання адреси операндів. Це дозволило суттєво спростити апаратні засоби ОМ та підвищити їх швидкодію. RISC-архітектуру розробляли таким чином, щоб зменшити Т за рахунок скорочення CPI та τ. Виявилося, що реалізація складних команд за рахунок послідовності із простих, але швидких RISC-команд не менш ефективна, ніж апаратний варіант складних команд в CISC-архітектурі.

Елементи RISC-архітектури вперше з'явилися в обчислювальних машинах CBC 6600 та суперЕОМ компанії Cray Research. Достатньо успішно реалізується RISC-архітектура і в сучасних ОС.

Відмітимо, що в останній час в мікропроцесорах компаній Intel та AMD широко використовуються ідеї, притаманні RISC-архітектурі, так що багато відмінностей між CISC та RISC поступово стираються.

Окрім CISC- та RISC-архітектур, в загальній класифікації був згаданий ще один тип АСК – архітектура з командними словами надвеликої довжини (VLIW). Концепція VLIW базується на RISC-архітектурі, але декілька простих RISC-команд об'єднуються в одну наддовгу команду та виконується паралельно. В плані АСК архітектура VLIW порівняно мало відмінна від RISC. З'явилося лише додатковий рівень паралельності обчислень, в результаті чого архітектуру VLIW логічніше адресувати не до обчислювальних машин, а до обчислювальних систем.

Порівняльна оцінка CISC-, RISC- та VLIW- архітектур

Характеристика	**CISC**	**RISC**	**VLIW**
Довжина команди	Варіюється	Єдина	Єдина
Положення полів в команді	Варіюється	Незмінне	Незмінне
Кількість регістрів	Декілька (часто спеціалізованих)	Багато регістрів загального призначення	Багато регістрів загального призначення
Доступ до пам'яті	Може виконуватися як частина команд різних типів	Виконується лише спеціальними командами	Виконується лише спеціальними командами

У будь-якій ЕОМ, незалежно від її архітектури, дані зберігаються в запам'ятовуючих пристроях (ЗП), які, з огляду на їх характеристики, місце

розташування і спосіб взаємодії з процесором, відносять до внутрішньої або зовнішньої пам'яті. Внутрішня пам'ять розташовується частково на загальному кристалі з процесором (регістри і кеш-пам'ять), а частково – на системній платі (основна пам'ять і, можливо, кеш-пам'ять 3-го і більш високих рівнів). Повільні ЗП великої ємності (твердотільні, магнітні та оптичні диски, магнітні стрічки) називають зовнішньою пам'яттю, оскільки до ЕОМ вони підключаються аналогічно пристроям вводу-виводу. Основними функціями ЗП є прийом, зберігання і видача даних в процесі роботи ЕОМ. Процес прийому даних, в ході якого здійснюється їх занесення в ЗП, називається записом, процес видачі дани – читанням або зчитуванням, а спільно їх визначають як процеси звернення до ЗП.

Характеристики запам'ятовуючих пристроїв внутрішньої пам'яті

Перелік основних характеристик, які необхідно враховувати, розглядаючиконкретний вид ЗП, включає в себе:

- ємність;

- одиницю пересилання;

- метод доступу;

- швидкодію;

- фізичний тип;

- фізичні особливості;

- вартість.

Ємність ЗП характеризують числом бітів, або байтів, які одночасно можуть зберігатися в запам'ятовуючому пристрої. На практиці застосовуються більші одиниці, і для їх позначення до слів «біт» або «байт» додають приставки: кіло, мега, гіга, тера, пета, екса, зетта, йотта (kilo, mega, giga, tera, peta, exa , zetta, yotta). Стандартно ці приставки означають множення основної одиниці вимірювань на 10^3, 10^6, 10^9, 10^{12}, 10^{15}, 10^{18}, 10^{21} і 10^{24} відповідно. У обчислювальної техніці, орієнтованої на двійкову систему числення, вони відповідають значенням, досить близьким до стандартних, але представляє собою цілу ступінь числа 2, тобто 2^{10}, 2^{20}, 2^{30}, 2^{40}, 2^{50},2^{60}, 2^{70}, 2^{80}. Щоб уникнути різночитань 2000 року МЕК – Міжнародна електротехнічна комісія (IEC – International Electrotechnical Comission) затвердила стандарт IEC 60027-2, що передбачає нові позначення, в яких до основної приставці додається склад bi (від англійського binary – двійковий). Так, якщо одиниця вимірювання ємності кратна байту, пропонуються наступні назви та позначення: kibibyte (KiB), mebibyte (MiB), gibibyte (GiB), tebibyte (TiB), pebibyte (PiB), exbibyte (EiB), zettabyte (ZiB), yottabvte (YiB). В

українській транскрипції – кібібайт (КіБ), мебібайт (МіБ), гібібайт (ГіБ), тебібайт (Тіб) і т. д.

Важливою характеристикою ЗП є одиниця пересилання. Для основної пам'яті одиниця пересилання визначається шириною шини даних, тобто кількістю бітів, що передаються по лініях шини паралельно. Зазвичай одиниця пересилання дорівнює довжині слова, але не обов'язково. Так, при пересиланні інформації між основною пам'яттю і кеш-пам’яттю дані передаються одиницями, що перевищують розмір слова, і такі одиниці називаються блоками.

При оцінці швидкодії необхідно враховувати застосовуваний в даному типі ЗП метод доступу до даних.

Розрізняють чотири основні методи доступу:

- послідовний;

- прямий;

- довільний;

- асоціативний,

з яких для внутрішньої пам'яті характерні два останніх. У ЗП з довільним доступом комірка має унікальну фізичну адресу. Звернення до будь-якого осередку займає один і той самий час і може проводитися в будь-якій послідовності (довільної черговості). Прикладом можуть слугувати запам'ятовуючі пристрої основної пам'яті. Асоціативний доступ дозволяє звертатися до осередків ЗП відповідно до ознак збережених в них даних, він забезпечує пошук осередків, що містять таку інформацію, в якій значення окремих бітів збігається зі станом однойменних бітів в заданому зразку. Порівняння здійснюється паралельно для всіх осередків пам'яті. За асоціативним принципом побудовані деякі блоки кеш-пам'яті.

Швидкодія ЗУ є одним з найважливіших його показників. Для кількісної оцінки швидкодії зазвичай використовують чотири параметри:

- Час вибірки даних. Він відповідає інтервалу часу між початком операції зчитування і видачею зчитаних даних з ЗП.

- Час зберігання даних – інтервал часу, протягом якого пристрій в заданому режимі зберігає дані без регенерації.

- Цикл звернення до ЗП або період звернення. Їм називають мінімальний інтервал часу між двома послідовними доступами до ЗП. Період звернення включає в себе власне час доступу плюс деякий додатковий час. Додатковий час може вимагатися для загасання сигналів на лініях, а в деяких типах ЗП, де

зчитування інформації призводить до її руйнування, – для відновлення прочитаної інформації.

- Швидкість передачі даних – кількість даних, що зчитуються (записуються) запам'ятовуючим пристроєм за одиницю часу.

Говорячи про фізичний тип ЗП, необхідно відзначити, що ЗП внутрішньої пам'яті сучасних обчислювальних машин базуються на напівпровідниковій технології.

Залежно від застосованої технології слід враховувати і ряд фізичних особливостей ЗП. Так, для напівпровідникової технології доводиться враховувати фактор енергозалежності. В енергозалежній пам'яті інформація може бути викривлена або втрачена при відключенні джерела живлення, в той час як в енергонезалежних ЗП записана інформація зберігається і при відключенні напруги живлення. Напівпровідникова пам'ять може бути як енергозалежною, так і навпаки, в залежності від її типу. Крім енергозалежності потрібно враховувати, чи приводить зчитування інформації до її руйнування.

Вартість ЗП прийнято оцінювати відношенням загальної вартості ЗП до його ємності в бітах, тобто вартістю зберігання одного біта інформації.

Ієрархія запам'ятовуючих пристроїв

Пам'ять часто називають «вузьким місцем» фон-неймановської ОМ через її значне відставання по швидкодії від процесорів. Так, якщо продуктивність процесорів зростає вдвічі приблизно кожні 1,5 року, то для мікросхем пам'яті приріст швидкодії не перевищує 9% в рік (подвоєння за 10

років), що виражається в збільшенні розриву в швидкодії між процесором і пам'яттю приблизно на 50% в рік .

Якщо проаналізувати використовувані в даний час типи ЗП, виявляється наступна закономірність:

- чим менше час вибірки, тим вище вартість зберігання біта;

- чим більше ємність, тим нижче вартість зберігання біта, але більше час вибірки.

При створенні системи пам'яті постійно доводиться вирішувати завдання забезпечення необхідної ємності і високої швидкодії за прийнятну ціну. Найбільш поширеним підходом тут є побудова системи пам'яті ЕОМ за ієрархічним принципом. Ієрархічна пам'ять складається з ЗП різних типів (рис. 4.3), які, в залежності від характеристик, відносять до певного рівня ієрархії. Більш високий рівень менше по місткості, швидше і має велику вартість в перерахунку на біт, ніж нижчий рівень. Рівні ієрархії взаємопов'язані: всі дані

на одному рівні можуть бути також знайдені на більш низькому рівні, і всі дані на цьому нижчому рівні можуть бути знайдені на наступному нижчому рівні і т. д.

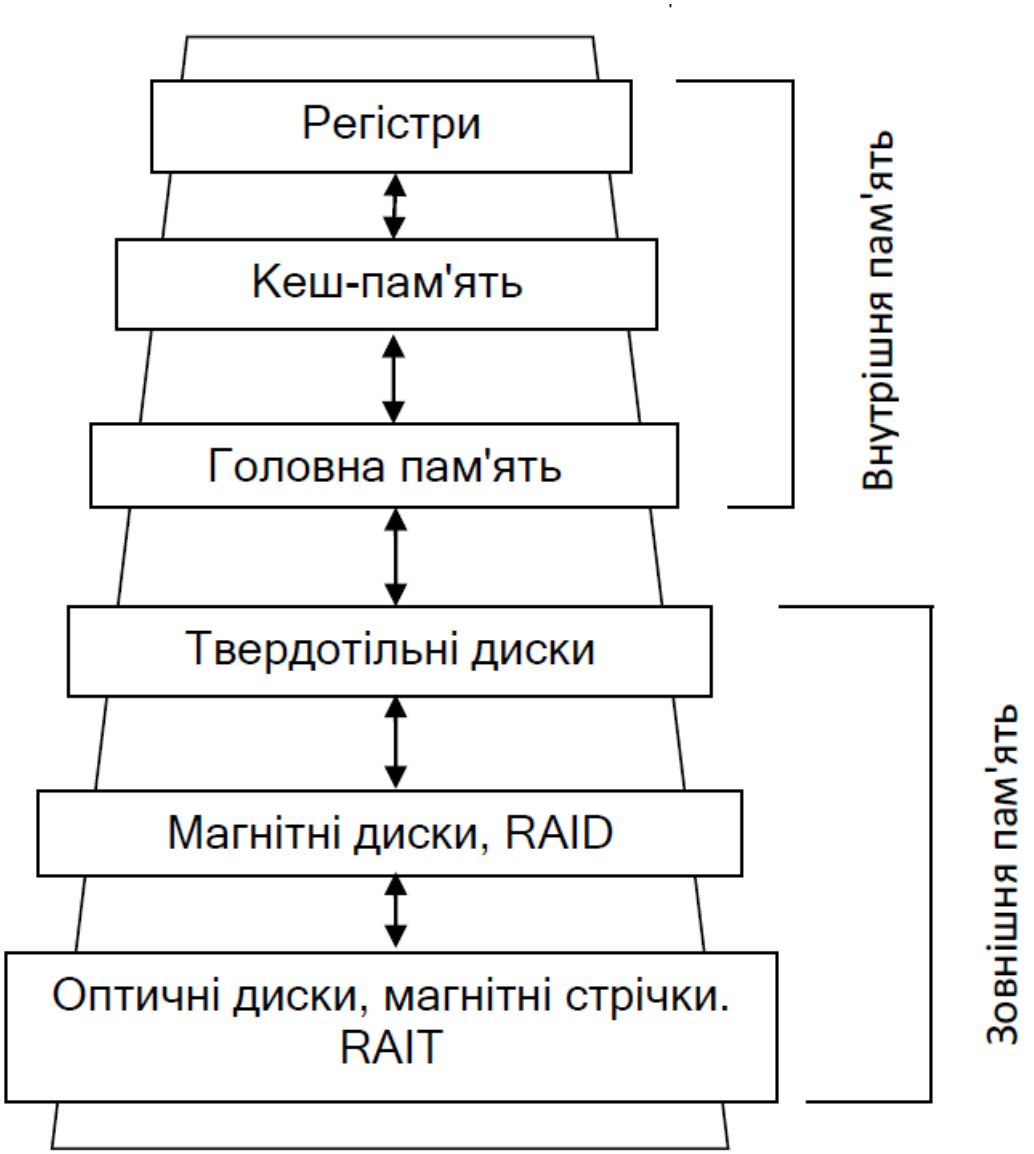

Рисунок 4.3 - Ієрархія запам'ятовуючих пристроїв

Три верхніх рівня ієрархії утворюють внутрішню пам'ять ЕОМ. а всі нижні рівні – це зовнішня або вторинна пам'ять. В міру руху вниз по ієрархічній структурі:

1. зменшується співвідношення «ціна/біт»;
2. зростає ємність;
3. росте час вибірки;
4. зменшується частота звернення до пам'яті з боку центрального процесора.

Якщо пам'ять організована відповідно до пунктів 1-3, а характер розміщення в ній даних задовольняє пункту 4, ієрархічна організація веде до зменшення загальної вартості при заданому рівні продуктивності.

Справедливість цього твердження випливає з принципу локальності за зверненням. Якщо розглянути процес виконання більшості програм, то можна помітити, що з дуже високою ймовірністю адреса чергової команди програми або слідує безпосередньо за адресою, за якою була зчитана поточна команда, або розташована поблизу неї. Таке розташування адрес називається просторовою локальністю програми. Оброблювані дані, як правило, структуровані, і такі структури зазвичай зберігаються в послідовності суміжних осередків пам'яті. Крім того, програми містять безліч невеликих циклів і підпрограм. Це означає, що невеликі набори команд можуть багаторазово повторюватися протягом деякого інтервалу часу, тобто має місце тимчасова локальність. Всі три види локальності об'єднує поняття локальність за зверненням. Принципу локальності часто надають чисельну форму і подають як так зване правило «90/10»: 90% часу роботи програми пов'язано з доступом до 10% адресного простору цієї програми.

З властивості локальності випливає, що програму розумно представити у вигляді послідовно оброблюваних фрагментів – компактних груп команд і оброблюваних ними даних. Помістивши такі фрагменти в більш швидку пам'ять, можна істотно знизити загальні затримки на звернення, оскільки команди і вихідні дані, будучи один раз передані з повільного ЗП в швидкий, потім можуть використовуватися багаторазово, і середній час доступу до них в цьому випадку визначається вже більш швидким ЗП. Це дозволяє зберігати великі програми і масиви даних на повільних, ємних, але дешевих ЗП, а в процесі обробки активно використовувати порівняно невелику швидку пам'ять, збільшення ємності якої пов'язане з високими витратами.

На кожному рівні ієрархії інформація (дані) розбивається на блоки, які виступають як найменша інформаційна одиниця, що пересилається між двома сусідніми рівнями ієрархії. Розмір блоків може бути фіксованим або змінним. При фіксованому розмірі блоку ємність пам'яті зазвичай кратна його розміру. Розмір блоків на кожному рівні ієрархії найчастіше різний і збільшується від верхніх рівнів до нижніх.

При доступі до команд і вихідних даних, наприклад для їх зчитування, спочатку проводиться пошук в пам'яті верхнього рівня. Факт виявлення потрібної інформації називають попаданням (hit), в іншому випадку говорять про промах (miss). При промаху проводиться пошук в ЗП наступного нижчого рівня, де також можливі потрапляння або промах. Після виявлення необхідної інформації виконується пересилання блоку, що містить потрібну інформацію, з нижніх рівнів на верхні.

При оцінці ефективності подібної організації пам'яті зазвичай використовують такі характеристики:

- коефіцієнт попадання (hit rate) – відношення числа звернень до пам'яті, при яких відбулося потрапляння, до загальної кількості звернень до ЗП даного рівня ієрархії;
- коефіцієнт промахів (miss rate) – відношення числа звернень до пам'яті, при яких мав місце промах, до загальної кількості звернень до ЗП даного рівня ієрархії;

- час звернення при потраплянні (hit time) – час, необхідний для пошуку потрібної інформації в пам'яті верхнього рівня (включаючи з'ясування, чи є звернення потраплянням), плюс час на фактичне зчитування даних;
- витрати на промах (miss penalty) – час, необхідний для заміни блоку в пам'яті більш високого рівня на блок з потрібними даними, розташований в ЗП наступного (більш низького) рівня. Витрати на промах включають в себе: час доступу (access time) – час звернення до першого слова блоку при промаху і час пересилки (transfer time) – додатковий час для пересилки залишкових слів блоку. Час доступу обумовлений затримкою пам'яті нижчого рівня, в той час як час пересилки пов'язаний з пропускною здатністю каналу між ЗП двох суміжних рівнів.

Найшвидший, але і мінімальний по ємності тип пам'яті – це внутрішні регістри ЦП, які іноді об'єднують поняттям надоперативний пристрій (НОЗП) або регістровий файл. Як правило, кількість регістрів невелика, хоча в архітектурі зі скороченим набором команд їх число може доходити до декількох сотень. Основна пам'ять, значно більшої місткості, розташовується нижче. Між регістрами ЦП і основною пам'яттю часто розміщують кеш-пам'ять, яка по ємності відчутно програє основній пам’яті, але істотно перевершує останню за швидкодією, поступаючись в той же час НОЗП. Всі види внутрішньої пам'яті реалізуються на основі напівпровідникових технологій і в основному є енергозалежними. Довготривале зберігання великих обсягів даних забезпечується зовнішніми ЗП, серед яких найбільш поширені ЗП на основі магнітних і оптичних дисків. Останнім часом все більшої популярності набувають твердотільні диски на базі флеш пам'яті. Ще один рівень ієрархії може бути доданий між основною пам'яттю і магнітними дисками. Цей рівень має назву дискової кеш-пам'яті і реалізується у вигляді самостійного ЗП, що включається до складу магнітного диска. Дискова кеш-пам'ять суттєво підвищує продуктивність при обміні інформацією між дисками і основною пам'яттю.