**7. Организация памяти современного процессора (L1, L2, L3 Cache, TLB, Virtual memory)**

Процессор — быстрый, память — медленная. Чем быстрее память, тем она дороже.

Кэш-память — быстрая, но маленького объема. В нее надо помещать наиболее часто используемые слова.

Кэш базируется на принципе локальности:

* Временная — данные, которые были использованы скорее всего будут использованы еще раз.
* Пространственная — соседние с использованным элементом данные скорее всего будут использоваться

Если требуемого слова нет в кэше, оно подгружается из основной памяти вместе со соседними, вытесняя при этом часть загруженных в кэш слов. Один из принципов вытеснения — Least recently used — вытсенение самых старых данных из кэша.

Префетчинг — подвыборка данных, программа начинает их читать раньше, чем их используют.

Хорошая статья:

<http://iproc.ru/parallel-programming/lection-7/2/>

N — ассоциативный кэш:

Память состоит из кусочков - «банков». Кэш состоит из отдельных ячеек — слов. N — количество ячеек памяти, сопоставляющихся одному банку. При увеличении N увеличивается число проводов => увеличивается их длина. Нужен компромисс.

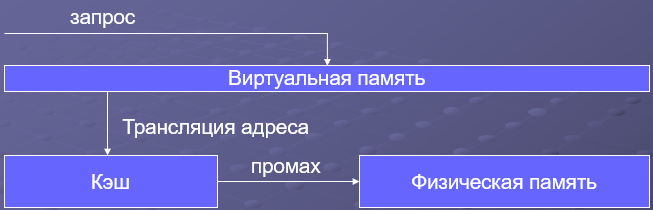
При доступе процессора в память сначала производится проверка, хранит ли кэш запрашиваемые из памяти данные. Для этого производится сравнение адреса запроса со значениями всех тегов кэша, в которых эти данные могут храниться. Случай совпадения с тегом какой-либо кэш-линии называется попаданием в кэш, обратный же случай называется кэш-промахом. Попадание в кэш позволяет процессору немедленно произвести чтение или запись данных в кэш-линии с совпавшим тегом.

Одной из проблем является фундаментальная проблема баланса между задержками кэша и интенсивностью попаданий. Большие кэши имеют более высокий процент попаданий но, вместе с тем, и большую задержку. Чтобы ослабить противоречие между этими двумя параметрами, большинство компьютеров использует несколько уровней кэша, когда после маленьких и быстрых кэшей находятся более медленные большие кэши.

Многоуровневые кэши обычно работают в последовательности от меньших кэшей к большим. Сначала происходит проверка наименьшего и наибыстрейшего кэша первого уровня (L1), в случае попадания процессор продолжает работу на высокой скорости. Если меньший кэш дал промах, проверяется следующий, чуть больший и более медленный кэш второго уровня (L2), и так далее, пока не будет запроса к основному ОЗУ.

Несколько уровней кэша, работающих с разной скоростью (чтение из)

* L1 — 3 такта
* L2 — 9 – 15 тактов
* Память — 30 – 180 тактов



*Виртуальная память* — абстракция, создающая иллюзию владения одним процессом всего адресного пространства

Виртуальная память — метод, отображающий адреса памяти, используемые программой, названные виртуальными адресами в физические адреса в памяти компьютера. Для программы адресное пространство выглядит как доступное и непрерывное адресное пространство, либо как набор непрерывных сегментов

Страничная организация виртуальной памяти:

Оперативная память делится на страницы: области памяти фиксированной длины (например, 4096 байт), которые являются минимальной единицей выделяемой памяти (то есть даже запрос на 1 байт от приложения приведёт к выделению ему страницы памяти). Исполняемый процессором пользовательский поток обращается к памяти с помощью адреса виртуальной памяти, который делится на номер страницы и смещение внутри страницы. Процессор преобразует номер виртуальной страницы в адрес соответствующей ей физической страницы при помощи буфера ассоциативной трансляции (TLB). Если ему не удалось это сделать, то требуется дозаполнение буфера путём обращения к таблице страниц (так называемый Page Walk), что может сделать либо сам процессор, либо операционная система (в зависимости от архитектуры). Если страница была выгружена из оперативной памяти, то операционная система подкачивает страницу с жёсткого диска в ходе обработки события Page fault. При запросе на выделение памяти операционная система может «сбросить» на жёсткий диск страницы, к которым давно не было обращений. Критические данные (например, код запущенных и работающих программ, код и память ядра системы) обычно находятся в оперативной памяти.

*TLB(translation lookaside buffer)* - это специализированный кэш центрального процессора, используемый для ускорения трансляции адреса виртуальной памяти в адрес физической памяти. TLB используется всеми современными процессорами с поддержкой страничной организации памяти. TLB содержит фиксированный набор записей (от 8 до 4096) и является ассоциативной памятью (Особый вид машинной памяти, используемый в приложениях очень быстрого поиска). Каждая запись содержит соответствие адреса страницы виртуальной памяти адресу физической памяти. Если адрес отсутствует в TLB, процессор обходит таблицы страниц и сохраняет полученный адрес в TLB, что занимает в 10—60 раз больше времени, чем получение адреса из записи, уже закэшированной TLB. Вероятность промаха TLB невысока и составляет в среднем от 0,01 % до 1 %.