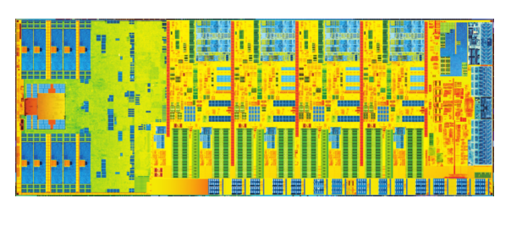
MICROARCHITECTURE CHEAT SHEET

X86 CPUs & Performance


Die photo of a quadcore CPU , Copyright of Intel

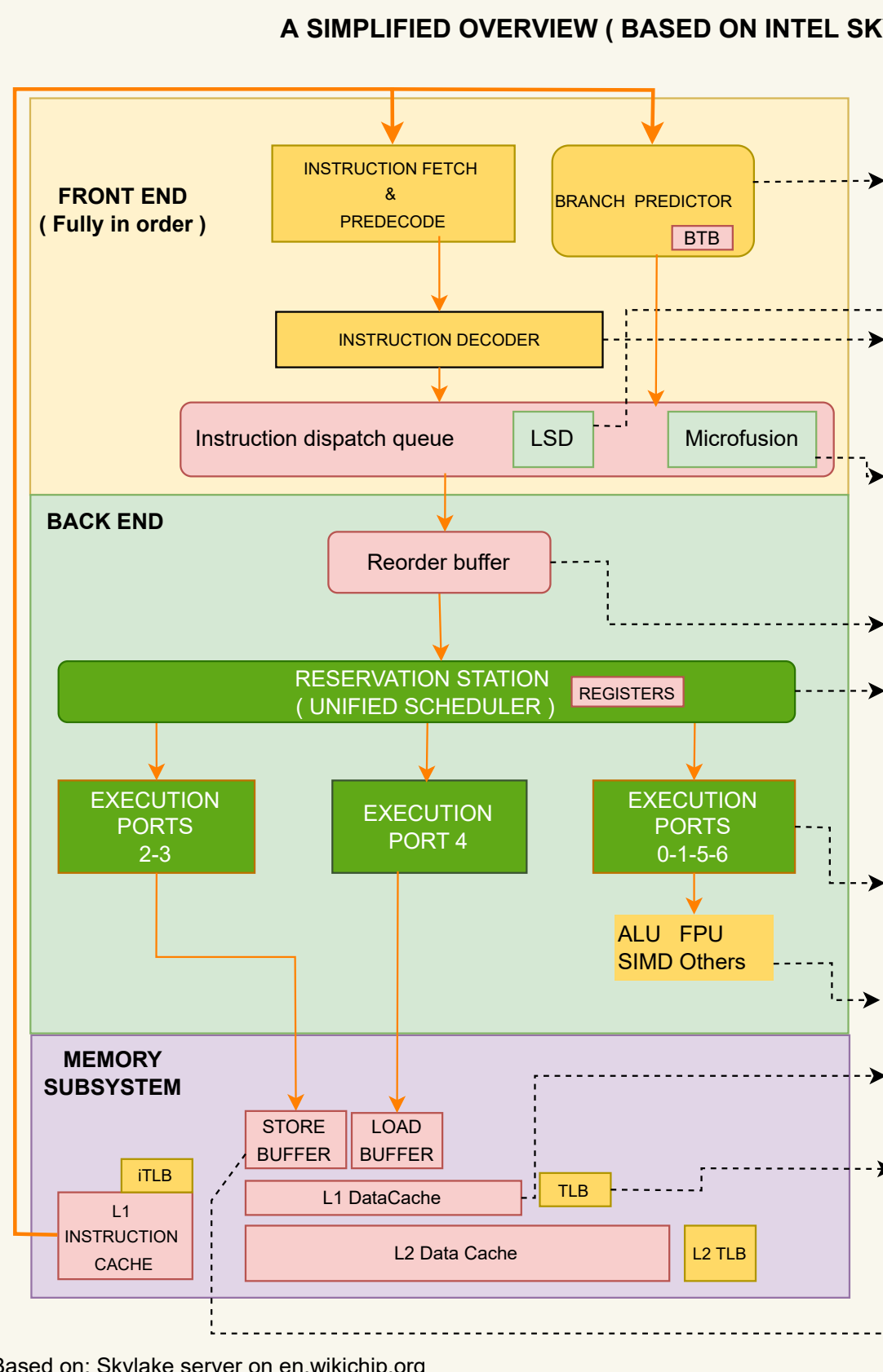
LAST UPDATE DATE : 03 JAN 2023

FOR LATEST VERSION: www.github.com/akhin/microarchitecture-cheatsheet

AUTHOR: AKIN OCAL akin_ocal@hotmail.com

PIPELINE REALM : INSIDE AN INDIVIDUAL CORE

A SIMPLIFIED OVERVIEW (BASED ON INTEL SKYLAKE)



Speculative execution
See the branch prediction realm for branch predictor, BTB and LSD.

UOPS/MicroOps
CISC instructions are split into smaller RISC instructions called as uops for more throughput.

Microfusion
Memory-write and read-modify uops are combined into a single uop for better throughput.

Out of order execution
This is done at the backend to improve the throughput.

However the results have to be output in order. So reordering at the end done by help of the reorder buffer.

Super scalar execution
Multiple uops executed in parallel.

Arithmetic operations
See the arithmetic realm.

Caches
See the cache memory realm .

TLBs
See the virtual memory realm .

Load,Store buffers
See the load-store realm .

Based on: Skylake server on en.wikichip.org

AMD pipelines : The main difference in AMD architectures is that there are parallel pipelines for integers and floating points : AMD Zen2 pipeline diagram on en.wikichip.org

PIPELINE PARALLELISM & PERFORMANCE

Pipeline diagrams : The diagrams below in the following topics are outputs from an online microarchitecture analysis tool UICA and they represent parallel execution through cycles.

Rows are multiple instructions being executed at the same time.

Columns display how instruction state changes through cycles.

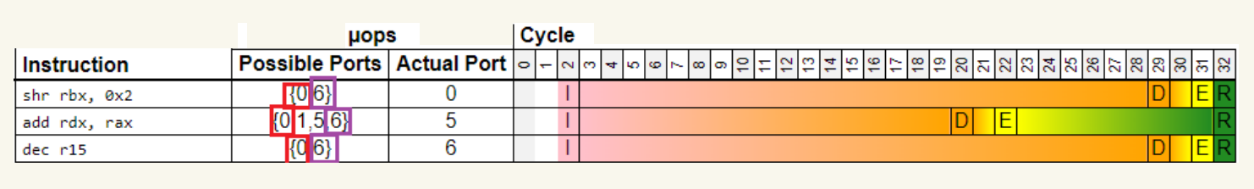
IPC : As for pipeline performance, typically IPC is used. It stands for "instructions per cyle".
A higher IPC value usually means a better throughput.
You can measure IPC with perf : https://perf.wiki.kernel.org/index.php/Tutorial

Rate of retired instructions : Apart from IPC, number of retired instructions should be checked. Retired instructions are not committed/finalised as they were wrongly speculated. On the other hand executed instructions are the ones which were finalised. Therefore a high rate of retired instructions indicates low branch prediction rate.

P	Predecoded
Q	Added to IDQ
I	Issued
r	Ready for dispatch
D	Dispatched
E	Executed
R	Retired

Instruction lifecycle states in UICA diagrams

CONTENTION FOR EXECUTION PORTS IN THE PIPELINE



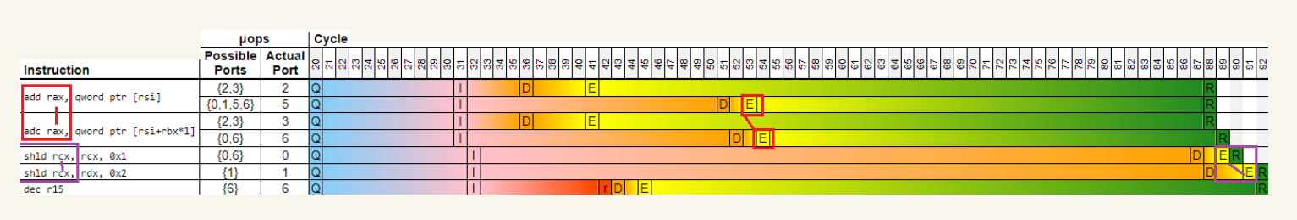
Instruction	μops Possible Ports	μops Actual Port
shr rbx, 0x2	{0,6}	0
add rdx, rax	{0,1,5,6}	5
dec r15	{0,6}	6

In the example above , all instructions are working on different registers, but SHR, ADD, DEC instructions are competing for ports 0 and 6. SHR and DEC are getting executed after ADD instruction.

Also notice that there is longer time between E(executed) and R(retired) states of instruction ADD as retirement has to be done in-order whereas execution is out-of-order.

Reference : Denis Bakhvalov`s article

INSTRUCTION STALLS DUE TO DATA DEPENDENCY



Instruction		μops Possible Ports	μops Actual Port
add rax,	qword ptr [rsi]	{2,3}	2
		{0,1,5,6}	5
adc rax,	qword ptr [rsi+rbx*1]	{2,3}	3
		{0,6}	6
shld rcx,	rcx, 0x1	{0,6}	0
shld rcx,	rdx, 0x2	{1}	1
dec r15		{6}	6

In the example above , there are 2 dependency chains , each marked with a different colour. In the first red coloured one , 2 instructions are competing for RAX register and notice that the second instruction gets executed after the first one. And the same applies to the 2nd purple pair.

Reference : Denis Bakhvalov`s article

RDTSCP INSTRUCTION FOR MEASUREMENTS

TSC (time stamp counter) is a special register that counts CPU cycles. RDTSCP can be used to read the TSC value which then can be used for measurements.. It can also avoid out-of-order execution effects to a degree :

it does wait until all previous instructions have executed and all previous loads are globally visible.
(From Intel Software Developer`s Manual Volume2 4.3 , April 2022)

Intel`s How to benchmark code execution times whitepaper has details of using RDTSCP instruction.

AMD Programmers Manual Vol3 states : *RDTSCP forces all older instructions to retire before reading the time-stamp counter*

ESTIMATING INSTRUCTION LATENCIES

You can use Agner Fog`s Instruction tables to find out instructions` reciprocal throughputs (clock cycle per instruction). As an example, reciprocal throughput of instruction RDTSCP is 32 on Skylake microarchitecture :

-> 1 cycle @4.5GHZ (highest frequency on Skylake) is 0.22 nanoseconds
-> 32*0.22=7.04 nanoseconds

So its resolution estimation is about 7 nanoseconds on a 4.5 GHz Skylake CPU. You have to recalculate it for different microarchitectures and clock speeds.

HYPERTHREADING / SIMULTANEOUS MULTITHREADING

Hyperthreading name is used by Intel and it is called as "Simultaneous multithreading" by AMD. In both resources including caches and execution units are shared.

Reference : Agner Fog`s microarchitecture book has "multithreading" sections for each of Intel and AMD microarchitectures

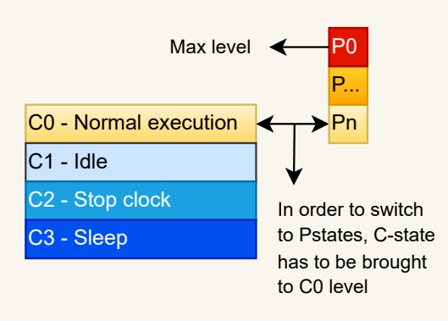
Regarding using it , if your app is data-intensive , halved caches won`t help. Therefore it can be disabled it via BIOS settings.In general, it moves the control of resources from software to hardware and that is usually not desired for performance critical applications.

DYNAMIC FREQUENCIES

Modern CPUs employ dynamic frequency scaling which means there is a min and a max frequency per CPU core.

ACPI : ACPI defines multiple power states and modern CPUs implement those. P-State`s are for performance and C states are for energy efficiency.

Intel has various tunability options and the most well known is TurboBoost. On AMD side there is Turbocore . You can use those to maximise the CPU usage.



Number of active cores & SIMD AVX2/512 on Intel CPUs: Intel`s power management policies are complex. See the arithmetic and the multicore realms as number of active cores and some of AVX2/512 extensions also may affect the frequency while in Turboboost.

Varying max clock speeds on AMD CPUs: Some AMD CPUs` cores have slightly varying max frequencies.
Reference : AMD`s GDC22 presentation page6

LOAD STORE REALM

LOAD & STORE BUFFERS

Load and store buffers allow CPU to do out-of-order execution on loads and stores by decoupling speculative execution and commiting the results to the cache memory.

Reference : https://en.wikipedia.org/wiki/Memory_disambiguation

STORE-TO-LOAD FORWARDING

Using buffers for stores and loads to support out of order execution leads to a data syncronisation issue. That issue is described in en.wikipedia.org/wiki/Memory_disambiguation#Store_to_load_forwarding

As a solution, CPU can forward a memory store operation to a following load, if they are both operating on the same address.

An example store and load sequence :

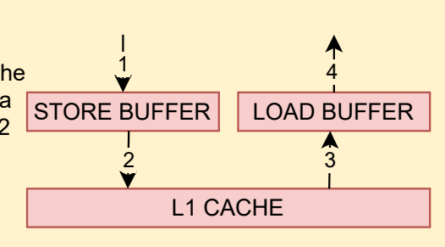
```
mov [eax],ecx  ; STORE , Write the value of ECX register to the memory
               ; address which is stored in EAX register
mov ecx,[eax]  ; LOAD, Read the value from that memory address
               ; ( which was just used) and write it to ECX register
```

STORE-TO-LOAD FORWARDING & LHS & PERFORMANCE

Based on Intel Optimization Manual 3.6.4, store-to-load forwarding may improve combined latency of those 2 operations. The reason is not specified however it is potentially LHS (Load-Hit-Store) problem in which the penalty is a round trip to the cache memory :

https://en.wikipedia.org/wiki/Load-Hit-Store

There are several conditions for the forwarding to happen. In case of a successful forwarding, the steps 2 and 3 (a roundtrip to the cache) will be bypassed.



The conditions for a successful forwarding and latency penalties in case of no-forwarding can be found in Agner Fog`s microarchitecture book.

What would happen without forwarding ? : In the past, game consoles PlayStation3 and Xbox360 had PowerPC based processors which used in-order-execution rather than out-of-order execution. Therefore developers had to separately handle LHS by using restrict keyword and other methods : Elan Ruskin`s article

ARITHMETIC REALM

ARITHMETIC INSTRUCTION LATENCIES

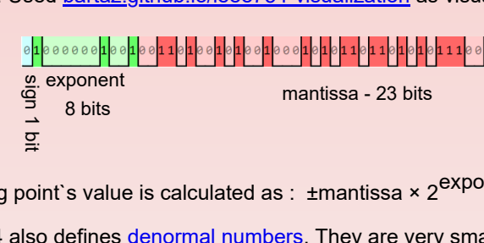
You can see a set of arithmetic opertions from fast to slow below.

The clock cycles are based on Agner Fog`s Instruction tables & Skylake architecture on 64 bit registers.

- Bitwise operations , integer add/sub : 0.25 to 1 clock cycle
- Floating point add : 3 clock cycles
- Floating point multiplication : about 5 clock cycles
- Floating point division: about 14-16 clock cycles
- Integer division : 24-90 clock cycles

FLOATING POINTS

X86 uses IEEE 754 standard for floating points. A 32 bit floating point consists of 3 parts in the memory layout. Below you can see all bits of 1234.5678 FP number. Used bartaz.github.io/ieee754-visualization as visualiser :



A floating point`s value is calculated as : $\pm\text{mantissa} \times 2^{\text{exponent}}$

IEEE754 also defines denormal numbers. They are very small / near zero numbers.

As floating points are approximations, denormal numbers are needed to avoid an undesired case of : a!=b but a-b=0
Without denormals the code to the right would invoke a divide-by-zero exception.
Reference : Bruce Dawson`s article

```
float GetInverseOfDiff(float a, float b)
{
   if (a != b)
   return 1.0f / (a - b);
   return 0.0f;
}
```

Based on Agner Fog`s microarchitecture book, Intel CPUs have a penalty for denormal numbers, for ex: 129 clock cycles on Skylake. They also can be turned off on Intel CPUs.

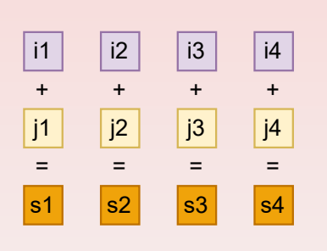
As for AMD side, the recent Zen architecture CPUs seemingly don`t have the same performance degradation.

X86 EXTENSIONS

x86 extensions are specialised instructions. They have various categories from cryptography to neural network operations.

Intel Intrinsics Guide is a good page to explore those extensions.

SSE (Streaming SIMD Extensions) is one of the most important ones. SIMD stands for "single instruction multiple data". SIMD instructions use wider registers to execute more work in a single go :



In the example above, an array 4 integers (i1 to i4) are added to another array of integers (j1 to j4). The result is also an array of sums (s1 to s4). In this example, 4 add operations are executed by a single instruction.

They play key role in compilers` vectorisation optimisations : GCC auto vectorisation

Apart from arithmetic operations, they can be utilised for string operations as well : Daniel Lemire's SIMD based JSON parser : https://github.com/simdjson/simdjson

X86 EXTENSIONS : SIMD DETAILS

The most recent SIMD instruction sets for Intel CPUs are :

AVX : Up to 256 bits
AVX2 : Up to 256 bits
AVX512 : Up to 512 bits

Recent AMD CPUs support AVX & AVX2. Only the latest Zen4 architecture supports AVX512 .

As for programming, there are also wider data types. The data type diagrams below are for 128 bit operations:

__m128 , 4 x 32 bit floating points	Float	Float	Float	Float
__m128d , 2 x 64 bit doubles	Double		Double	
__m128i , 4 x 32 bit ints	int	int	int	int
__m128l , 2 x 64 bit long longs	long long		long long	

Note that as SIMD instructions require more power, therefore usage of some AVX2/512 extensions may introduce downclocking. They should be benchmarked.
For details : Daniel Lemire`s article

BRANCH PREDICTION REALM

BRANCH PREDICTION BASICS

Why : CPUs proactively fetch instructions of potentially upcoming branches to utilise the pipeline as much as possible.

Gain if predicted correctly : If the right branch was predicted that will increase the throughput as it completed fetching a set of instructions in advance.

Penalty in case of misprediction : If the prediction was wrong , that prefetch will be a waste and the cost will be flushing the pipeline.

What are branch instructions ? : Unconditional ones (jmp), conditional ones (eg: jne) , call/ret

How : There are auxilliary hardware buffers.

Branch target buffer stores target addresses (instruction pointers) of branches. AMD uses multiple level of BTBs : L1 BTB, L2 BTB etc.

Pattern history tables track the history of results (whether it was taken or not) per branch.

	1	2	3	4	...
branch 1	T	T	NT	T	...
branch ...	NT	NT	NT	T	...
branch n	T	NT	NT	NT	...

A hypothetical pattern history table
T: taken, NT : not taken

CONDITIONAL MOVE INSTRUCTION

CMOV (Conditional move) instruction also computes the conditions for some additional time. Therefore they don`t introduce extra load to the branch prediction mechanism. They can be used to eliminate branches.

Reference : Intel Optimisation Manual 3.4.1.1

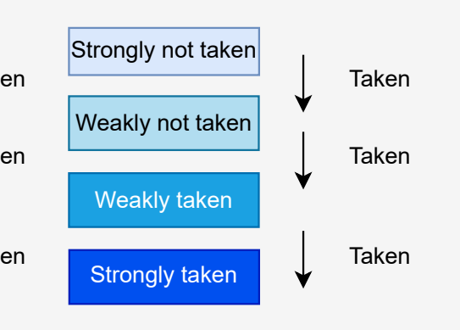
BP METHODS : 2-LEVEL ADAPTIVE BRANCH PREDICTION

Saturating counter as a building block

A 2-bit saturating counter can store 4 strength states.

Whenever a branch is taken it goes stronger.

And whenever a branch is not taken it goes weaker.



2 level adaptive predictor

In this method, the pattern history table keeps 2^n rows and each row will have a saturating counter.

A branch history register which has the history of last n occurences, will be used to choose which row will be used from the pattern history table.

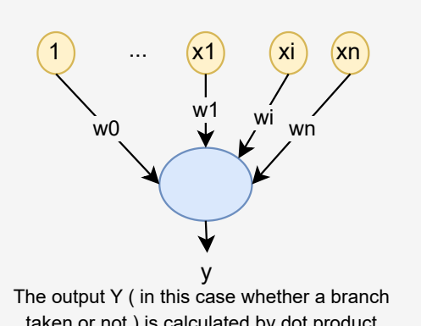
Reference : Agner Fog`s microarchitecture book 3.1.

BP METHODS : AMD PERCEPTRONS

A perceptron is basically the simplest form of machine learning. It can be considered as a linear array of weights.

Agner Fog mentions that they are good at predicting very long branches compared to 2-level adaptive branch prediction in his microarchitecture book 3.12.

For details of perceptron based branch prediction: Dynamic Branch Prediction with Perceptrons by Daniel Jimenez and Calvin Lin



The output Y (in this case whether a branch taken or not) is calculated by dot product of the weight vector and the input vector.

INTEL LSD (LOOP STREAM DETECTOR)

Intel LSD will detect a loop and stop fetching instructions to improve the frontend bandwidth. Several conditions mentioned in Intel Optimisation Manual 3.4.2.4 :

- *Loop body size up to 60 μops, with up to 15 taken branches, and up to 15 64-byte fetch lines.*
- *No CALL or RET.*
- *No mismatched stack operations (e.g., more PUSH than POP).*
- *More than ~20 iterations.*

Note that LSD is disabled on Skylake Server CPUs. Reference : https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server)#Front-end

DISABLING SPECULATIVE EXECUTION PATCHES

You can consider disabling system patches for speculative execution related vulnerabilities such as Meltdown and Spectre for performance, if that is doable in your system. Those patches are not only microcode updates but they also need OS support.

Kernel.org documentation : https://www.kernel.org/doc/html/latest/admin-guide/kernel-parameters.html

Red Hat Enterprise documentation : https://access.redhat.com/articles/3311301

Meltdown paper : https://meltdownattack.com/meltdown.pdf

Spectre paper : https://spectreattack.com/spectre.pdf

ESTIMATED LIMITS : HOW MANY IFs ARE TOO MANY ?

As for max number of entries in BTBs, there are estimations made by stress testing the BTB with sequences of branch instructions :

Intel Xeon Gold 6262 -> roughly 4K
AMD EPYC 7713 -> roughly 3K

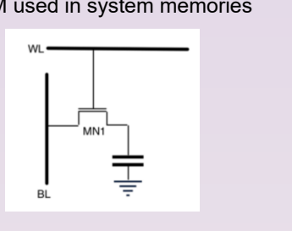
Reference : Marek Majkovski`s article on Cloudflare blog

CACHE MEMORY REALM

CACHE MEMORY VS SYSTEM MEMORY

System memory is made of DRAM cells. Cache memory on the other hand are made of SRAM cells which is much faster than DRAMs. But also they are more expensive:

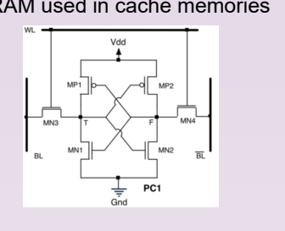
DRAM used in system memories



Access time : 50-150 nanoseconds due to capacitor charge/discharge times and other steps

Cost : Cheaper in the price as it has less components

SRAM used in cache memories



Access time : Under 1 nanosecond

Cost : Expensive in the price due to 6 transistors

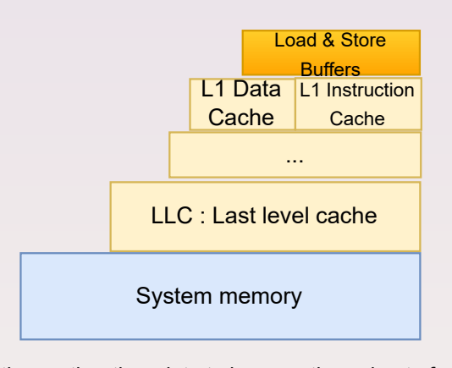
Reference : Ulrich Drepper`s What every programmer should know about memory

CACHE ORGANISATION

Caches are organised in multiple levels. As you go upper in that hierarchy , the capacity increases. Therefore **LLC** term used to indicate the last level of cache.

3 level data caches are currently the most common ones. Intel Broadwell architecture had 4 level caches in the past. Also upcoming AMD CPUs may come with 4 level of caches.

Cache line size is the unit of data transfer between the cache and the system memory. It is typically 64 bytes. And the caches are organised according to the cache line size.



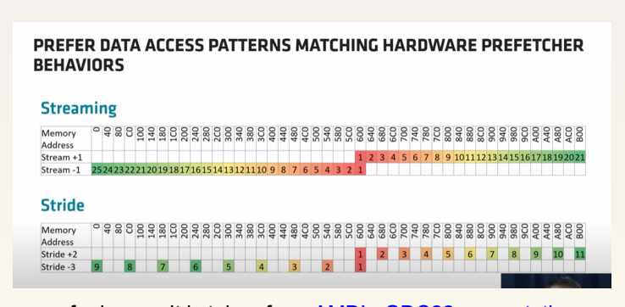
There is also **instruction cache** (iCache) which stores program instructions rather than data to improve throughput of CPU frontend.

In case of a cache hit, the latency is typically single digit nanoseconds. And in case of a cache miss, we need a round trip to the system memory and total latency becomes 3 digit nanoseconds.

HARDWARE AND SOFTWARE PREFETCHING

Hardware prefetchers detect patterns like **streams** (Ex: accessing to contiguous array members) and **strides** (Ex: accessing specific members in arrays of structs) and prefetch data and instruction to cache lines automatically.

Developers can also use instruction **_mm_prefetch** to prefetch data explicitly for cases when hardware can`t predict. That is called as software prefetching.



Reference for image: It is taken from AMD`s GDC22 presentation page44

N-WAY SET ASSOCIATIVITY

Why : Cache capacities are much smaller than the system memory. Moreover , software can use various regions of their address space. So if there was one to one mapping of a fully sequential memory that would lead to cache misses most of the time. Therefore there is a need for efficient mapping between the cache memory and the system memory.

How : In N-Way set associativity, caches are divided to groups of sets. And each set will have N cache blocks. The mapping information is stored in bits of addresses which has 3 parts :

TAG	SET	OFFSET
used as a unique identifier per cache block	used to determine the set in a cache	used to determine the actual bytes in the target cache block

The pseudocode below shows steps for searching a single byte in the cache memory :

```
Get tag, set and offset from the address

For each block in the current set ( which we have just found out )
    if tag of the current block equals to tag ( which we just have found out )
            read and return data using offset , it is a cache hit
If there was no matching tag, it is a cache miss
```

The level of associativity (the number of ways) is a trade off between the search time and the amount of system memory we can map.

BYPASSING THE CACHE : NON-TEMPORAL STORES & WRITE-COMBINING

Temporal data is data that will be accessed in a short period of time. The term **non-temporal** data indicates that data will not be accessed any soon. (cold data). If the amount of non-temporal data is excessive in the cache, that is called as **cache pollution**. Non-temporal store instructions are introduced for this problem and they store data directly to the system memory by bypassing the cache.

Write combining buffers are used with non-temporal stores. CPU will try to fill a whole cache line (typically 64byte) before commit to the system memory and only sends to the system memory when that buffer is filled. That is for reducing the load on the bus between CPU and system memory.

DIRECT CACHE ACCESS

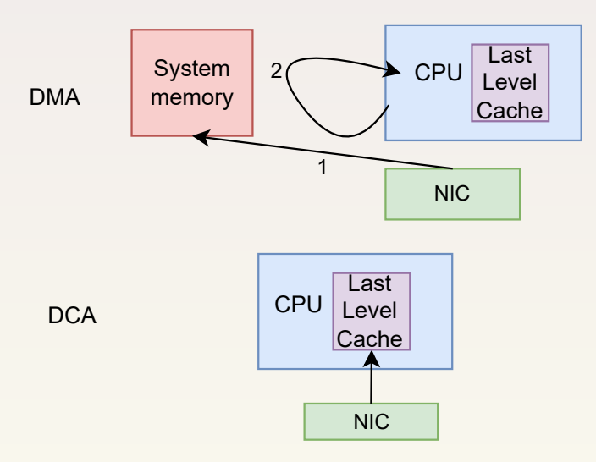
Modern NICs come with a DMA (Direct Memory Access) engine and can transfer data directly to drivers' ring buffers which reside on the system memory.

DMA mechanism doesn`t require CPU involvement. Though mechanism initiated by CPU , therefore CPU support needed.

DCA bypasses the system memory and can transfer to directly LLC of CPUs that support this feature.

Intel refers to their technology as DDIO (Direct I/O).

Reference : Intel documentation



VIRTUAL MEMORY REALM

VIRTUAL MEMORY ORGANISATION

Why virtual memory ?
Because cumulative memory requirement of multiple processes in an OS can be more than the system capacity.

It is basically for sharing memory resources between multiple processes.

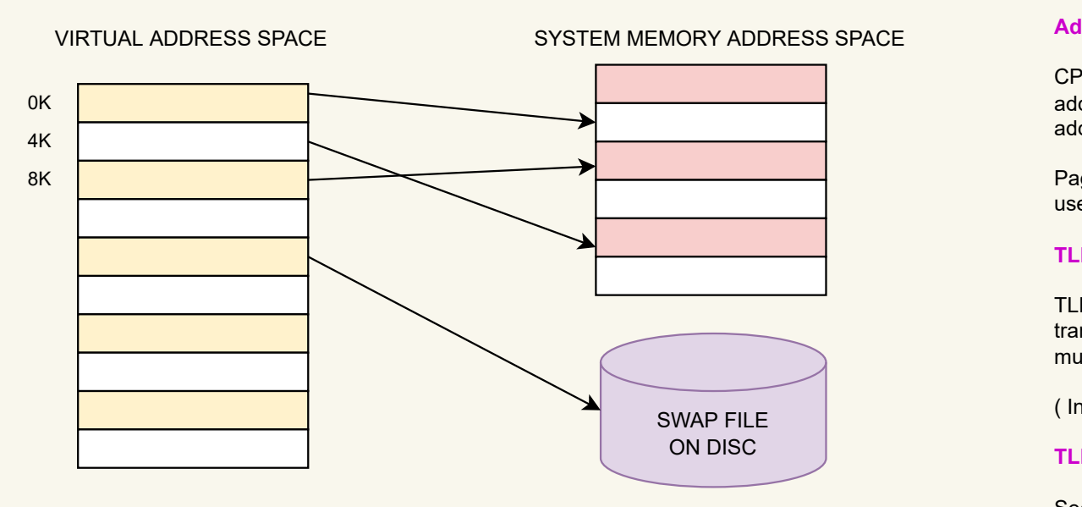
It also provides security by isolating processes.

Pages
- Minimum addressable virtual space that can be requsted from OS.

- Typically 4KB

Swapping
In case of out of memory, process memory will be evicted to the disc.

Page faults
Happens when the page is not on physical memory but on the swap file which is on the harddrive.



VIRTUAL MEMORY ADDRESS TRANSLATION

Address translation & Page table

CPUs work with virtual addresses and those addresses need to be converted to physical addresses.

Page table structures on system memory are used for this purpose.

TLB (Translation lookaside buffer)

TLBs are caches in CPUs to make the translation process faster. Modern CPUs have multiple levels of TLBs.

(Intel refers to L2 TLB as sTLB)

TLB Shootdowns

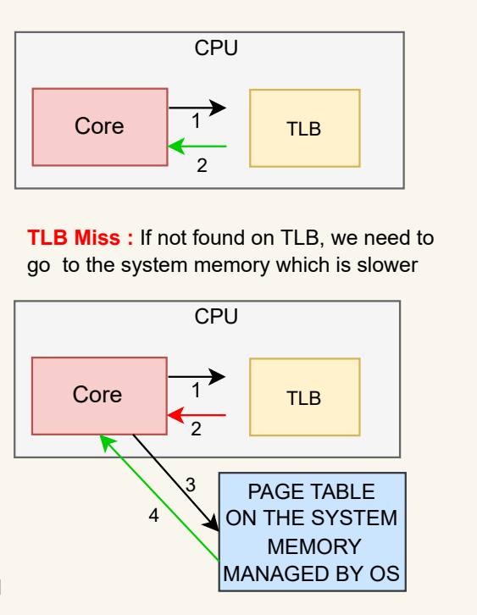
See the multicore realm below.

iTLB

Apart from data TLB , there is also iTLB for caching addresses of instructions on both Intel and AMD architectures.

TLB Hit No need to access page table

TLB Miss : If not found on TLB, we need to go to the system memory which is slower



TLB PRESSURE & HUGE PAGES

TLB pressure

If each page is 4K , that increases the load on the TLB buffer.

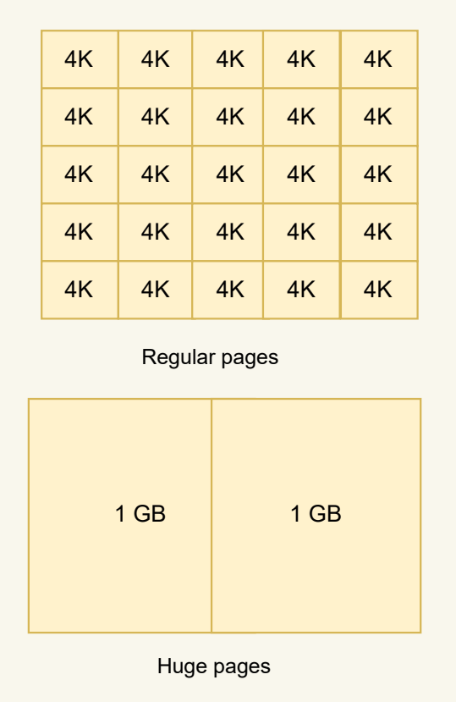
CPU support for larger pages

x86-x64 CPUs support huge pages from 2MB to 1GB to reduce the pressure on TLB.

OS support

Linux implementation refers to them as huge pages and Windows calls them as large pages.

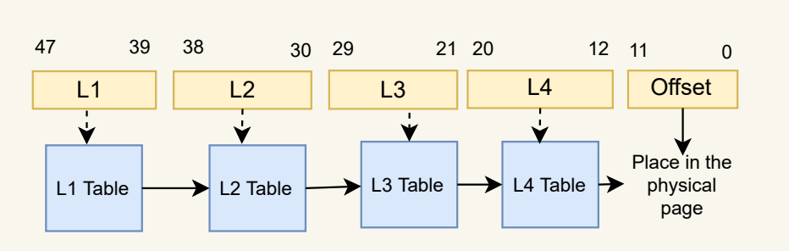
You shall check your OS and CPU in combination to find out the supported sizes.



PAGE TABLE WALKING

Even with pages which group addresses, having all pages in a page table would still need too much storage on 64 bit systems. Therefore page tables are implemented hierarchically.

Memory is divided into address spaces. And there is a tree data structure for each address space in the page table. Processes have to "walk the page table" level by level in the hierarchy to find out the actual address :



4 level page table is the most common one. In the diagram above, the first 48 bits of a 64 bit address are used for page table walking. All of 48 bits have to be used in order to find out the final actual address. (For all details : Intel Software Developer`s Manual Volume3 4.5)

Intel CPUs started to support 5 level tables since Ice Lake.
The advantage of another level is that you can address even more space.

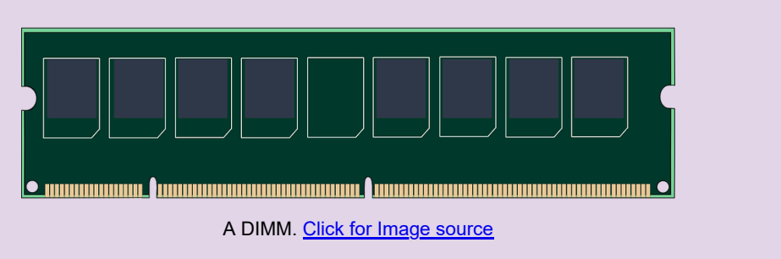
The disadvantage is that the time needed to walk the page table increases due to a new level of indirection.

SYSTEM MEMORY REALM

DDR RAMs

DDR RAMs are the most common commodity hardware as system memory.

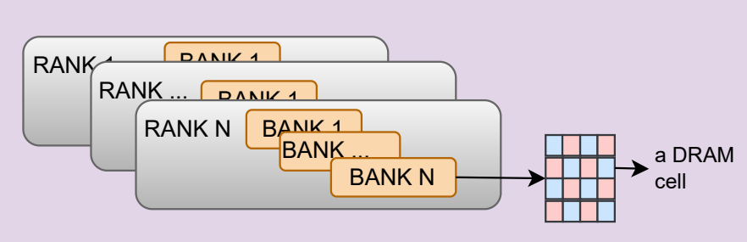
They are found in forms of DIMMs (Dual inline memory module) / RAMSticks.


A DIMM. Click for Image source

Organisation

System memory / RAM is organised as collection of ranks.

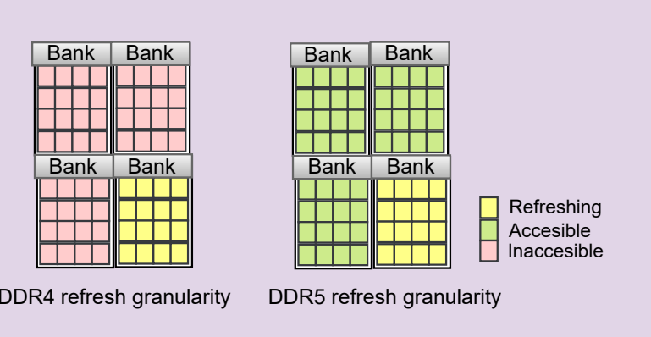
Each rank have banks which are collection of DRAM cells per bit.



DRAM refreshes

DRAM circuits use capacitors which lose their charge over time. (See the cache memory realm). So RAMs have to refresh their DRAM cells periodically.

As for DDR4, refreshing is rank-level which means the other banks in the same rank become inaccessible. DDR5 comes with same-bank-refresh feature which allows a more fine-grained bank-level refresh. Therefore it can offer a higher throughput.



MULTICORE REALM

TOPOLOGIES

TOPOLOGICAL OVERVIEW - INTEL CPUS

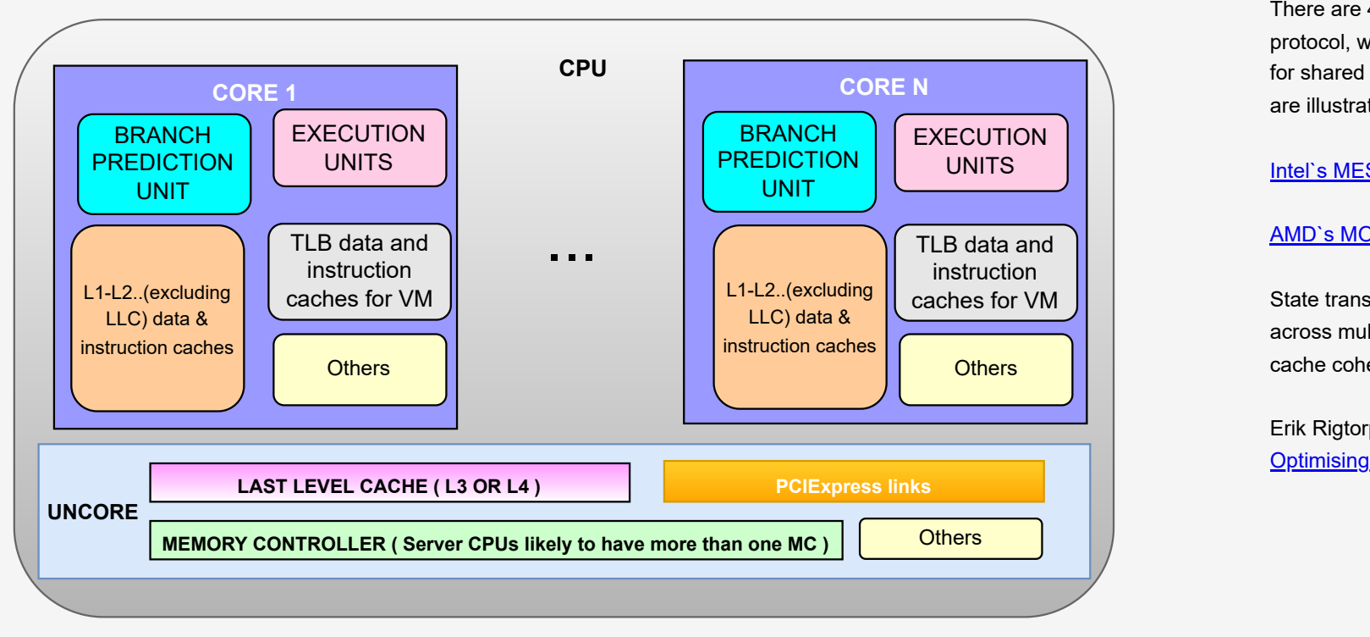

CPU
CORE 1
BRANCH PREDICTION UNIT
EXECUTION UNITS
L1-L2..(excluding LLC) data & instruction caches
TLB data and instruction caches for VM
Others
CORE N
UNCORE
LAST LEVEL CACHE (L3 OR L4)
PCIExpress links
MEMORY CONTROLLER (Server CPUs likely to have more than one MC)
Others

Diagram above aims to show resource per core and shared resources. Note that uncore in an Intel-only term to refer to CPU functionality which are not per core.

Exception of E-cores : An exception to the above diagram is Intel`s recent E-cores. E-cores are meant for power efficiency and paired with less resources. For ex: Alder Lake CPUs` E-cores also share L2 cache.

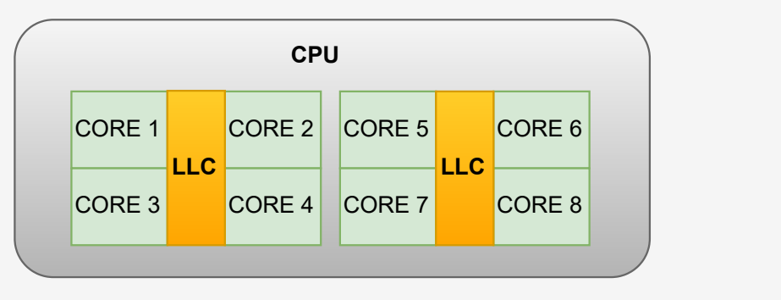
Reference : https://www.anandtech.com/show/16959/intel-innovation-alder-lake-november-4th

TOPOLOGICAL OVERVIEW - AMD CPUS

Most of AMD topology is similar to Intel. However starting from Zen microarchitecture, one key difference is CCXs.

AMD CPUs are designed as group of 4 cores which is called as CCX (Core complex) , and there is one LLC per each CCX/quad core.

Practically the maximum number of cores competing for the LLC (without simultanenous multithreading) is 4 in recent AMD CPUs. An example 8 core CPU with 2 CCXs :



Reference : https://en.wikichip.org/wiki/amd/microarchitectures/zen#CPU_Complex_.28CCX.29

COHERENCY

CACHE COHERENCY : PROTOCOLS

Cache coherency protocols are needed to avoid data hazards. Intel CPUs use MESIFand AMD CPUs use MOESI, however both heavily depend on MESI protocol.

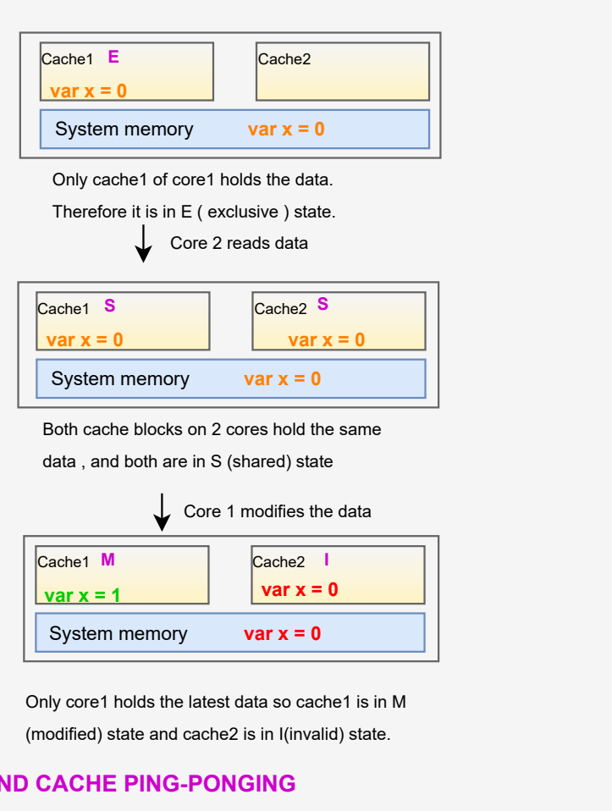
There are 4 states for a CPU cache line in MESI protocol, which are M for modified , E for exclusive , S for shared and I for invalid. The 3 diagrams to the right are illustrating the simplest cases for all 4 states.

Intel`s MESIF on Wikipedia

AMD`s MOESI on Wikipedia

State transition can trigger cache coherency protocol across multiple cores. Variables can be cached to avoid cache coherency traffic whereever applicable :

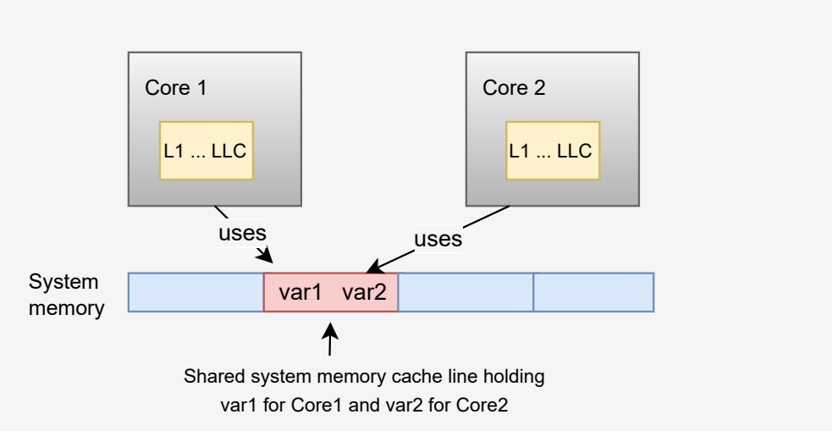
Erik Rigtorp`s article :
Optimising a ring buffer for throughput


Cache1 E
var x = 0
Cache2
System memory var x = 0
Only cache1 of core1 holds the data. Therefore it is in E (exclusive) state.
Core 2 reads data
Cache1 S
var x = 0
Cache2 S
var x = 0
System memory var x = 0
Both cache blocks on 2 cores hold the same data , and both are in S (shared) state
Core 1 modifies the data
Cache1 M
var x = 1
Cache2 I
var x = 0
System memory var x = 0
Only core1 holds the latest data so cache1 is in M (modified) state and cache2 is in I(invalid) state.

CACHE COHERENCY : FALSE SHARING AND CACHE PING-PONGING

In the diagram to the right , if Core1 changes its var1, that change will need to be propagated to all other cores by the cache coherency protocol. That will lead to invalidation of cache areas associated with the shared cache line across all cores, even though it is used by only one core.

That situation is called false sharing.

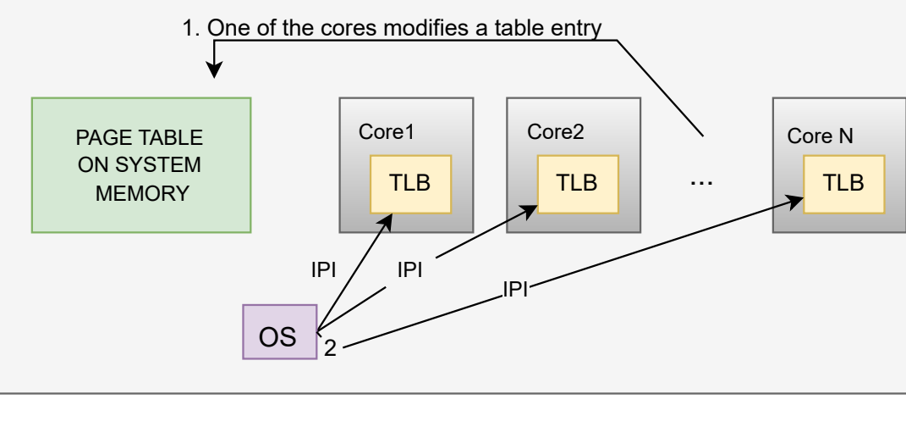

Core 1
L1 ... LLC
Core 2
L1 ... LLC
uses
uses
System memory
var1 var2
Shared system memory cache line holding var1 for Core1 and var2 for Core2

If those happen in higher rates and if cache lines from system memory transferred between cores rapidly , that situation is called as cache ping-pong.

VIRTUAL MEMORY PAGE TABLE COHERENCY : TLB SHOOTDOWNS

Whenever a page table entry is modified by any of the cores, that particular TLB entry is invalidated in all cores via IPIs. This one is not done by hardware but initiated by operating system.

IPI : Interprocessor interrupt , you can take "processor" as core in this context.


1. One of the cores modifies a table entry
PAGE TABLE ON SYSTEM MEMORY
Core1
TLB
Core2
TLB
...
Core N
TLB
IPI
IPI
IPI
OS
2

MEMORY REORDERINGS & SYNCRONISATION

MEMORY REORDERINGS

The term memory ordering refers to the order in which the processor issues reads (loads) and writes (stores) . Based on Intel Software Developer`s Manual Volume3 8.2.3.4 , there is only one kind of memory reordering that can happen. Loads can be reordered with earlier stores if they use different memory locations. That reordering will not happen if they use the same address :

```
CORE1                                  CORE2

; x and y initially 0                  ; x and y initially 0
mov [x], 1 ; STORE TO X                mov [y], 1 ; STORE TO Y
mov [result1], y ; LOAD FROM Y         mov [result2], x ; LOAD FROM X
```

In case of reordering, result1 and result2 above can both end up as zero in both cores. Note that, apart from CPUs , also compilers can do memory reordering : Jeff Preshing`s article : Memory Ordering at Compile Time

INSTRUCTIONS TO AVOID REORDERINGS

Reorderings can be avoided by using serialising instructions such as SFENCE, LFENCE, and MFENCE : Intel Software Developer`s Manual Volume3 8.3 defines them as :

These instructions force the processor to complete all modifications to flags, registers, and memory by previous instructions and to drain all buffered writes to memory before the next instruction is fetched and executed.

There is also bus locking "LOCK" prefix (Intel Software Developer`s Manual Volume3 8.1.2) which can be used as well to avoid reorderings.

ATOMIC OPERATIONS

An atomic operation means that there will be no other operations going on during the execution. From point of execution , an atomic operation is indivisible and nothing can affect its execution.

The most common type of atomic operations are RMW (read-modify-write) operations.

ATOMIC OPERATIONS & MEMORY ALIGNMENT

If an atomic instruction is used for a memory range which is split to multiple cache lines, that will lead to locking the whole memory bus , instead of just the cache line.

Reference : Detecting and handling spit locks (in Linux kernel) on lwn.net

ATOMIC RMW OPERATIONS : COMPARE-AND-SWAP

CAS instruction (CMPXCHG) reads values of 2 operands. It then compares them and if they are equal , it swaps values. All the operations are atomic / uninterruptible. It can be used to implement lock free data structures.

ATOMIC RMW OPERATIONS : TEST-AND-SET

Test-and-set is an atomic operation which writes to a target memory and returns its old value. It is typically used to implement spin locks.

PAUSE INSTRUCTIONS

Busy spinning applications (ex: user space spin locks) can degrade hyperthreading efficiency. There are several pause instructions (PAUSE/ TPAUSE/UMWAIT/UMMONITOR) to help that.

Note that the latency for PAUSE instruction on Skylake clients is an order of magnitude slower than other architectures : Intel Optimisation Manual 2.5.4

TRANSACTIONAL MEMORY

Transactional memory areas are programmer specified critical sections. Reads and writes in those areas are done atomically. (Intel Optimization Manual section 16)

However due to another hardware security issue, Intel disabled them from Skylake to Coffee Lake CPUs : https://www.theregister.com/2021/06/29/intel_tsx_disabled/

AMD equivalent is called as "Advanced Syncronisation Facility". According to Wikipedia article, there are no AMD processors using it yet.

LIMITING CONTENTION BETWEEN CORES

DISABLING UNUSED CORES TO MAXIMISE FREQUENCY (INTEL)

Number of active cores may introduce downclocking : Wikichip article
Therefore disabling unused cores may improve frequency for perf-critical cores, depending on your CPU. You shall refer to your CPU`s frequency table :

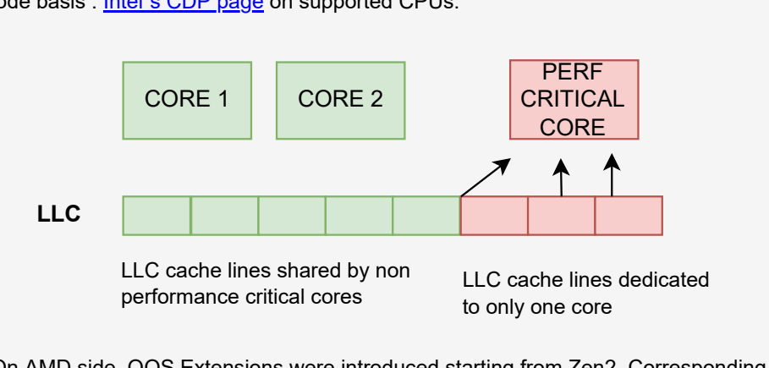
An example frequency table : Wikichip XeonGold5120 article

ALLOCATING A PARTITION OF LLC (SERVER CLASS CPUS)

You can allocate a partition of the shared CPU last level cache for your performance sensitive application to avoid evictions on Intel CPUs that support ***CAT*** feature.

CAT : Cache allocation tech , reference : Intel CAT page

CDP (Code and data prioritisation) allows developers to allocate LLC on code basis : Intel`s CDP page on supported CPUs.


CORE 1
CORE 2
PERF CRITICAL CORE
LLC
LLC cache lines shared by non performance critical cores
LLC cache lines dedicated to only one core

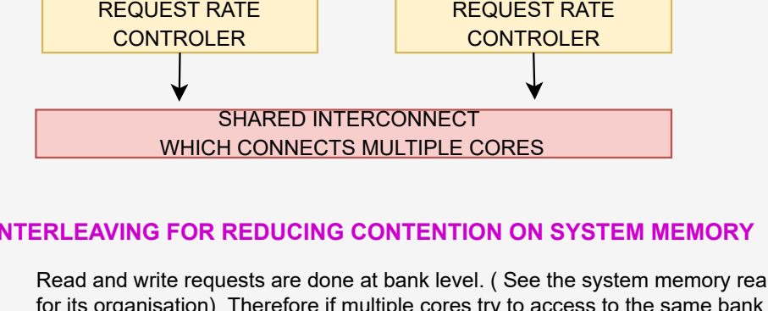
On AMD side, QOS Extensions were introduced starting from Zen2. Corresponding technologies are called as "Cache allocation enforcement" and "Code and data prioritisation": https://developer.amd.com/wp-content/resources/56375_1.00.pdf

MEMORY BANDWIDTH THROTTLING (SERVER CLASS CPUS)

You can throttle memory bandwidth per CPU core on Intel CPUs that support MBA. Each core can be throttled with their request rate controller units.

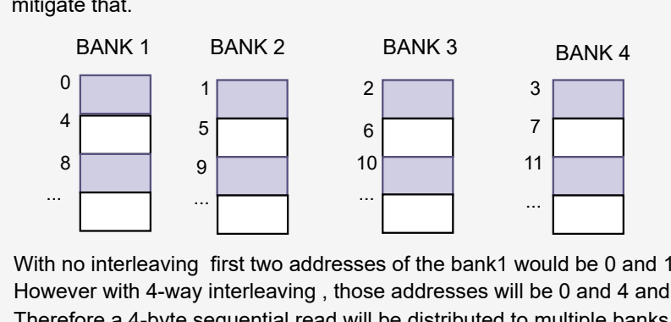
MBA : Memory bandwidth allocation , reference : Intel MBA page

For AMD equivalent, QOS Extensions were introduced starting from Zen2: https://developer.amd.com/wp-content/resources/56375_1.00.pdf


CORE 1
PROGRAMMABLE REQUEST RATE CONTROLER
...
CORE N
PROGRAMMABLE REQUEST RATE CONTROLER
SHARED INTERCONNECT WHICH CONNECTS MULTIPLE CORES

INTERLEAVING FOR REDUCING CONTENTION ON SYSTEM MEMORY

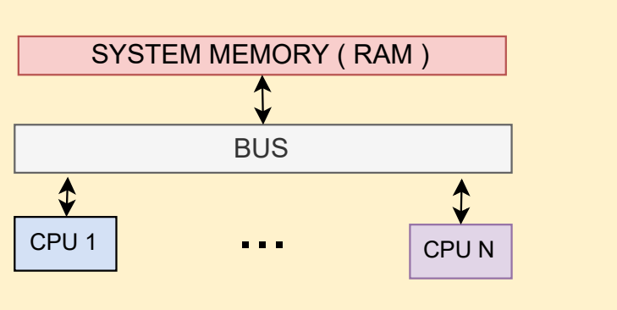
Read and write requests are done at bank level. (See the system memory realm for its organisation) Therefore if multiple cores try to access to the same bank , there will be a contention. Interlaving bank address spaces is one method to mitigate that.



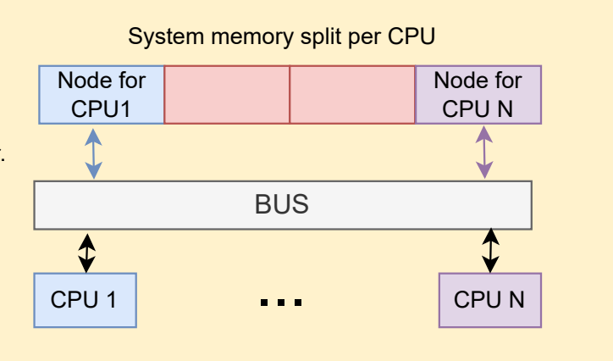
With no interleaving first two addresses of the bank1 would be 0 and 1.
However with 4-way interleaving , those addresses will be 0 and 4 and so on.
Therefore a 4-byte sequential read will be distributed to multiple banks.

MULTICPU REALM (SERVER CLASS CPUS)

SMP (Symmetric multiprocessing)
All CPUs use a single bus to access the same system memory. CPUs may slow down each other as there may be a contention for access to the banks.

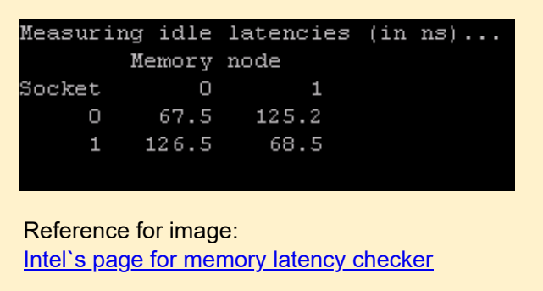


NUMA (Non uniform memory access)
The system memory is organised as nodes and each node is meant to be accessed by only one CPU, therefore memory access is split. So CPUs don`t compete with each other.



Cross NUMA accesses
A CPU can access to its local memory faster. However that also means that a cross-access will be slower. It is typically displayed with a table of distances between nodes.

The side screenshot is from Intel`s memory latency checker. Notice that cross acceses are two times slower than local accesses.



```
Measuring idle latencies (in ns)...
        Memory node
Socket      0      1
     0   67.5  125.2
     1  126.5   68.5
```

Reference for image:
Intel`s page for memory latency checker

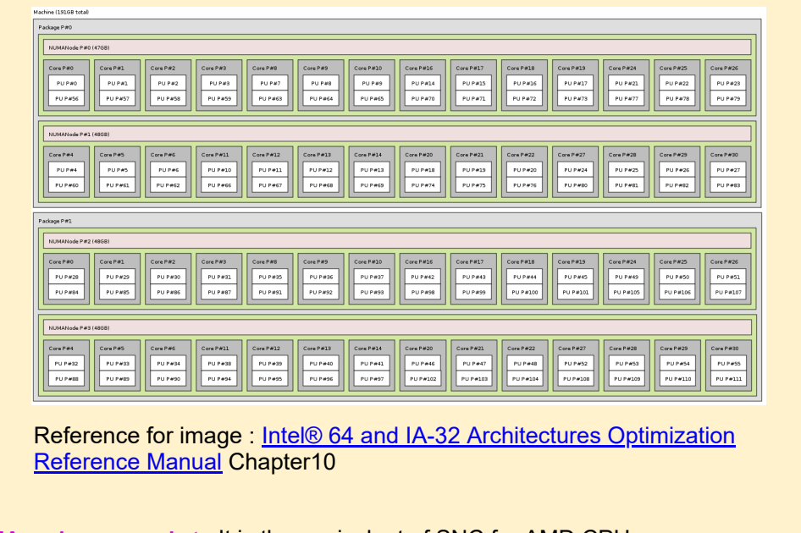
Intel Sub NUMA Cluster mode (starting from Skylake , formerly COD)

Intel`s SNC splits LLC into disjoint clusters based on address range. You can bind each cluster to a set of memory controllers.

Scalable NUMA-aware apps can benefit from this extra LLC & MC based resource-partitioning.

The lstopo image below is from a 2 Intel Xeon CPU system, each with 28 cores. (Since hyperthreading is enabled it displays 112 cores instead of 56)

That configuration gives 2 more NUMA nodes which means further scalability to benefit from :



Reference for image : Intel® 64 and IA-32 Architectures Optimization Reference Manual Chapter10

AMD NUMA nodes per socket : It is the equivalent of SNC for AMD CPUs.
Reference : Dell article about AMD NUMA nodes per socket feature

ACROSS REALMS

INTEL`S TOPDOWN MICROARCHITECTURE ANALYSIS METHOD

Intel`s Top Down analysis is hierarchical organisation of microarchitecture events. It is documented in Intel Optimisation Manual Appendix B1.

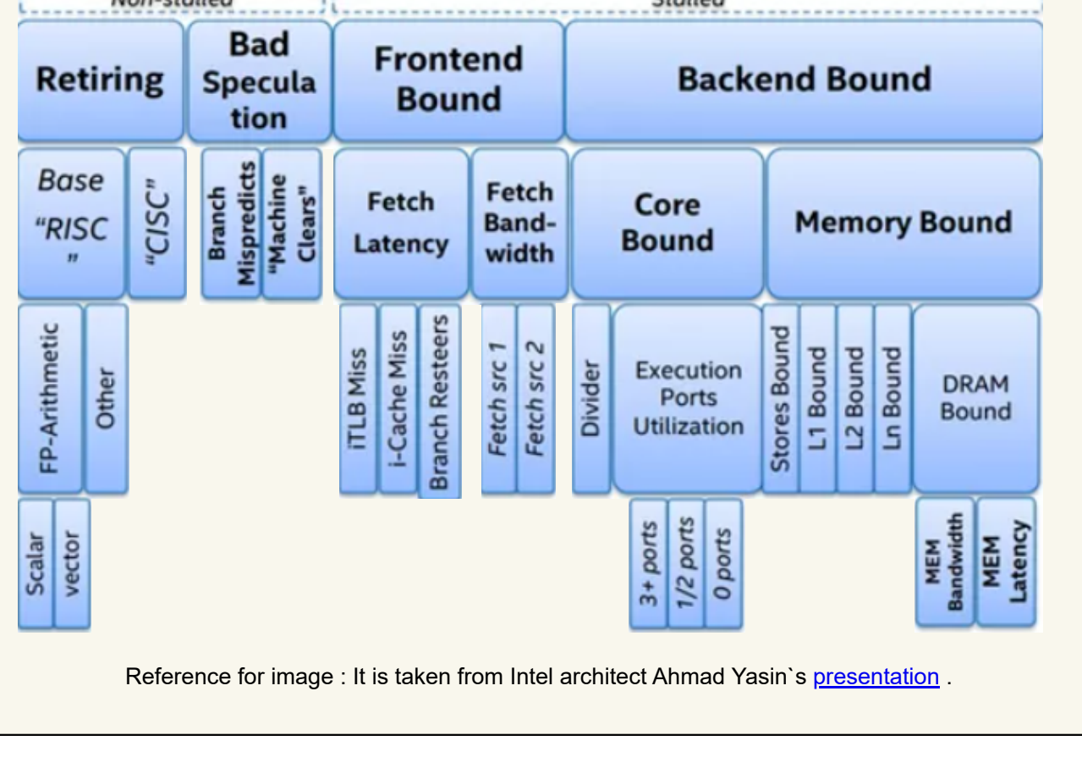
There are 2 main categories for stalls :

1. Frontend : A typical example is large code sizes leading to instruction cache misses.

2. Backend : Usually either memory bound or compute bound

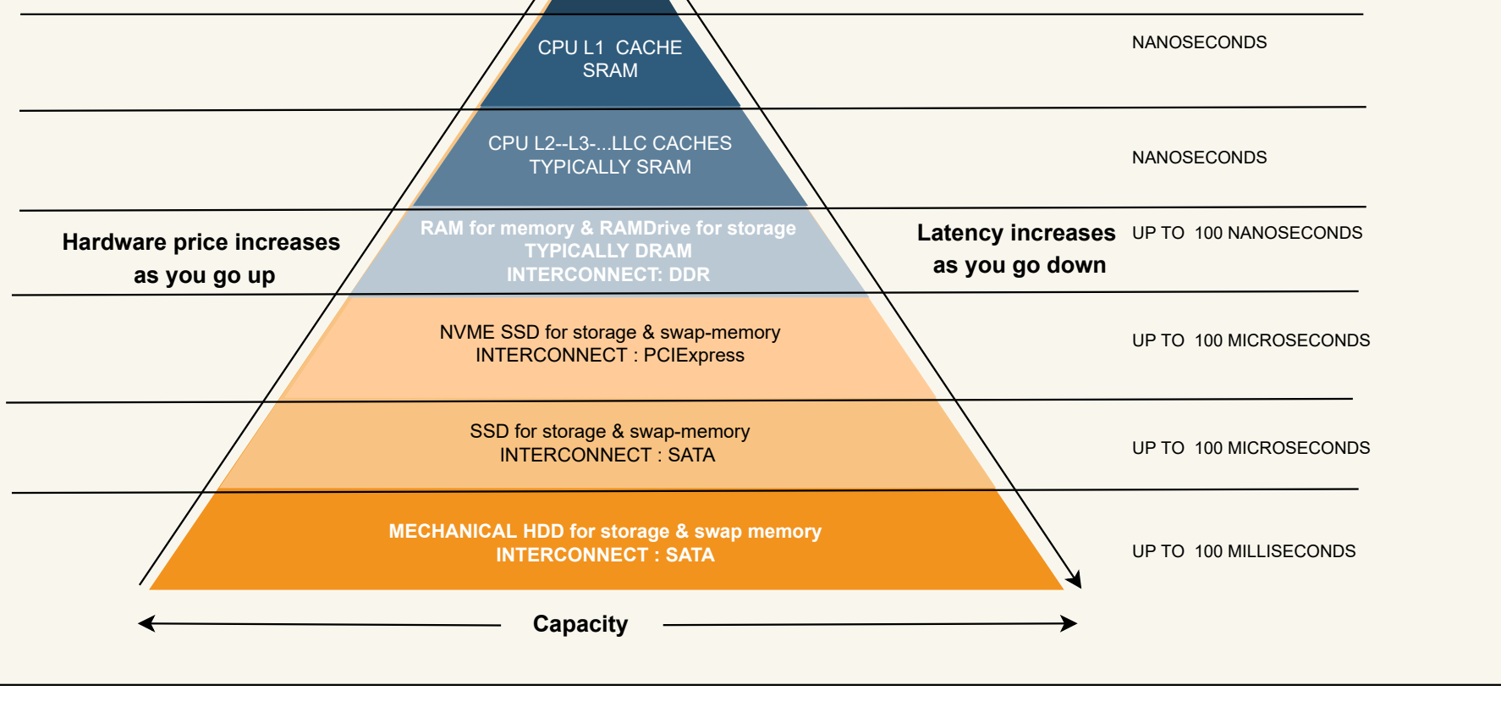
Branch mispredictions are categorised under "bad speculation".

You can use either Intel`s Vtune or Andi Kleen`s Toplev tool to make a top down analysis. Both utilise Intel CPUs` performance monitoring counters.



Reference for image : It is taken from Intel architect Ahmad Yasin`s presentation .

OVERALL MEMORY & STORAGE HIERARCHY


LOAD STORE
CPU L1 CACHE SRAM
CPU L2--L3-...LLC CACHES TYPICALLY SRAM
RAM for memory & RAMDrive for storage TYPICALLY DRAM INTERCONNECT: DDR
NVME SSD for storage & swap-memory INTERCONNECT : PCIExpress
SSD for storage & swap-memory INTERCONNECT : SATA
MECHANICAL HDD for storage & swap memory INTERCONNECT : SATA
Hardware price increases as you go up
Latency increases as you go down
NANOSECONDS
NANOSECONDS
UP TO 100 NANOSECONDS
UP TO 100 MICROSECONDS
UP TO 100 MICROSECONDS
UP TO 100 MILLISECONDS
Capacity

ESTIMATED LATENCY NUMBERS IN CLOCK CYCLES

Simple register operation (ADD OR etc)	less than 1 clock cycle	Floating point division	10-40 clock cycles
Branch prediction	1-2 clock cycles	Compare and swap atomic op.	15-30 clock cycles
Floating point addition	1-3 clock cycles	Integer division	15-40 clock cycles
Multiplication (int, floating point)	1-7 clock cycles	L3 data cache read	30-70 clock cycles
L1 data cache read	3-4 clock cycles	System memory read	100-150 clock cycles
TLB miss	7-21 clock cycles	Cross-NUMA L3 read	100-300 clock cycles
L2 data cache read	10-12 clock cycles	Cross-NUMA system memory read	300-500 clock cycles
Branch misprediction	10-20 clock cycles		

Reference for numbers : http://ithare.com/infographics-operation-costs-in-cpu-clock-cycles/

Note : The reference is from 2016, however it still is good to show proportions between different events.