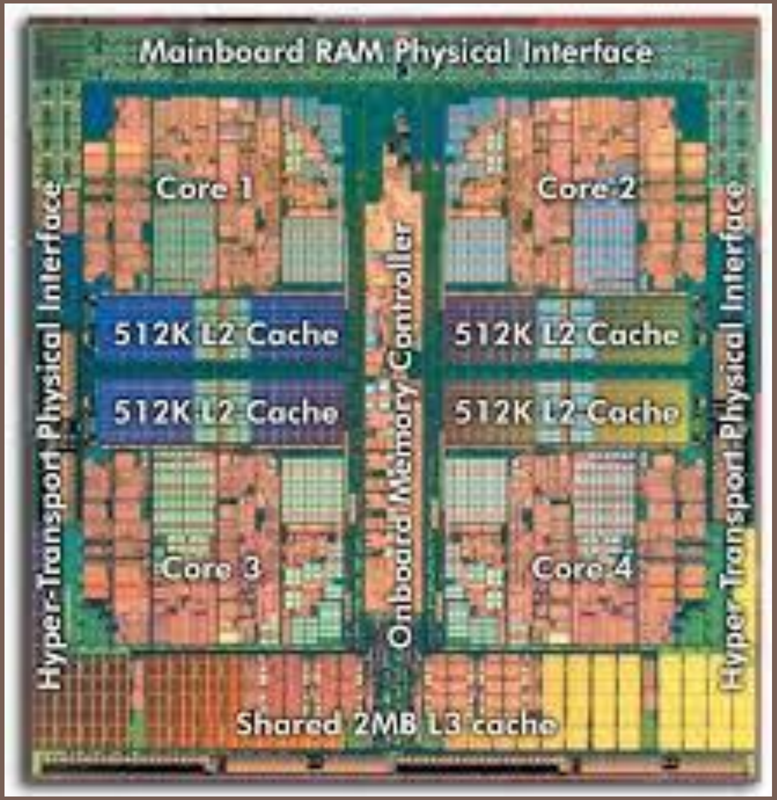
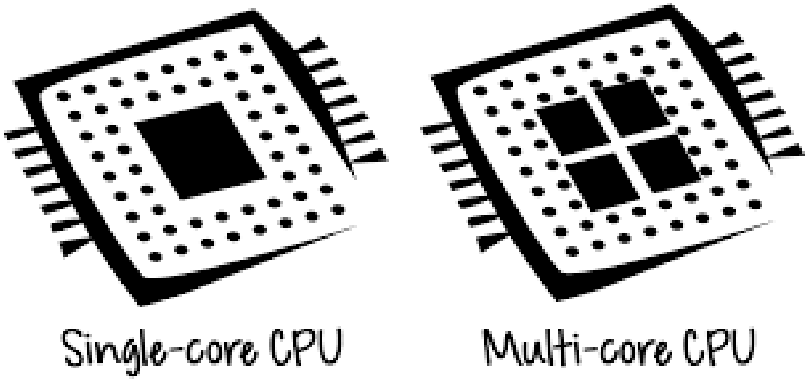

Mainboard RAM Physical Interface
Core 1
Core 2
Hyper-Transport Physical Interface
Onboard Memory Controller
Hyper-Transport Physical Interface
512K L2 Cache
512K L2 Cache
512K L2 Cache
512K L2 Cache
Core 3
Core 4
Shared 2MB L3 cache

INTRODUCTION TO MULTICORE ARCHITECTURE

Multicore Computer

- A multicore computer, also known as chip multi-processor, combines two or more processors (called cores) on a single piece of silcon (called a die)
- Typically, each core consists of all the components of an independent processor, such as registers, ALU, Pipeline hardware and control unit, Plus L1 instruction and data caches
- In addition to the multiple cores, contemporary multicore chips also include L2 Cache and increasingly L3 Cache.

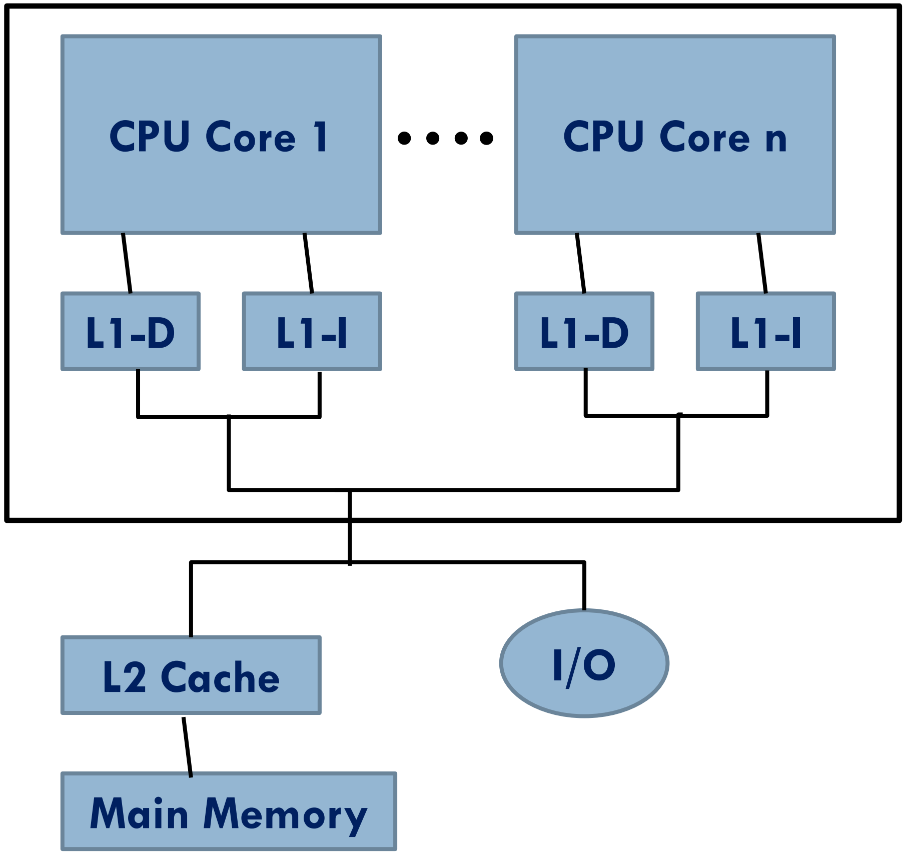


Multicore Organization

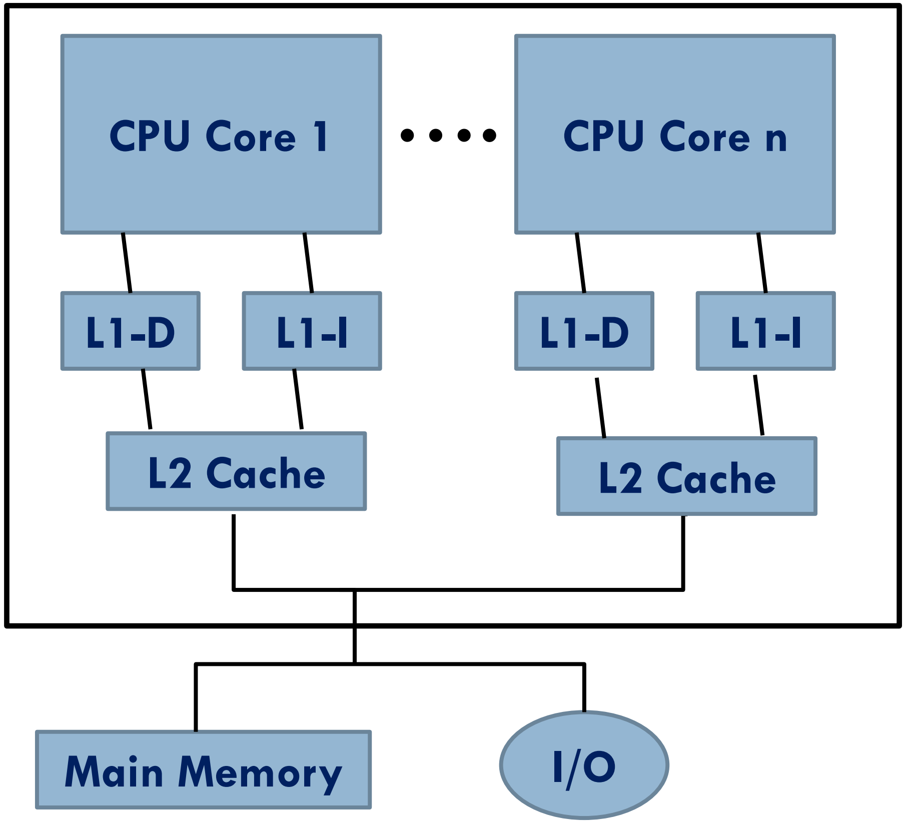
- Use of Multiple cores and multi level cache
- Use of Multi-threading by cores

- The main variables in multicore organization are as follows
  - The number of core processors on the chip
  - The number of levels of cache memory
  - The amount of cache memory that is shared

Multicore Organization Alternatives

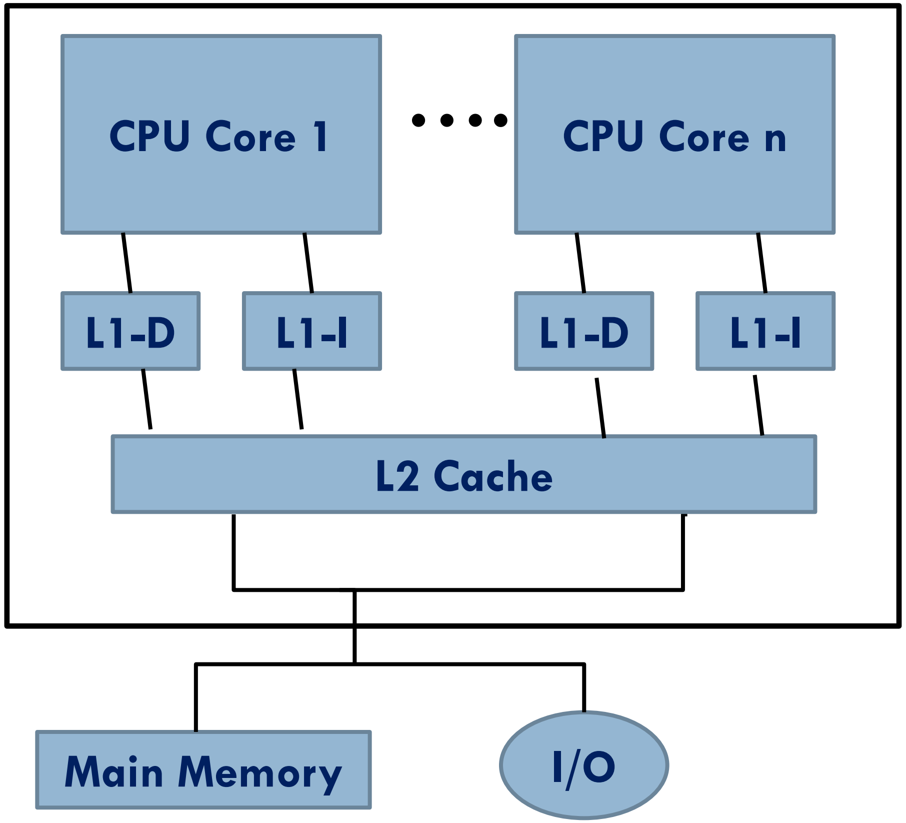


Dedicated L1 Cache

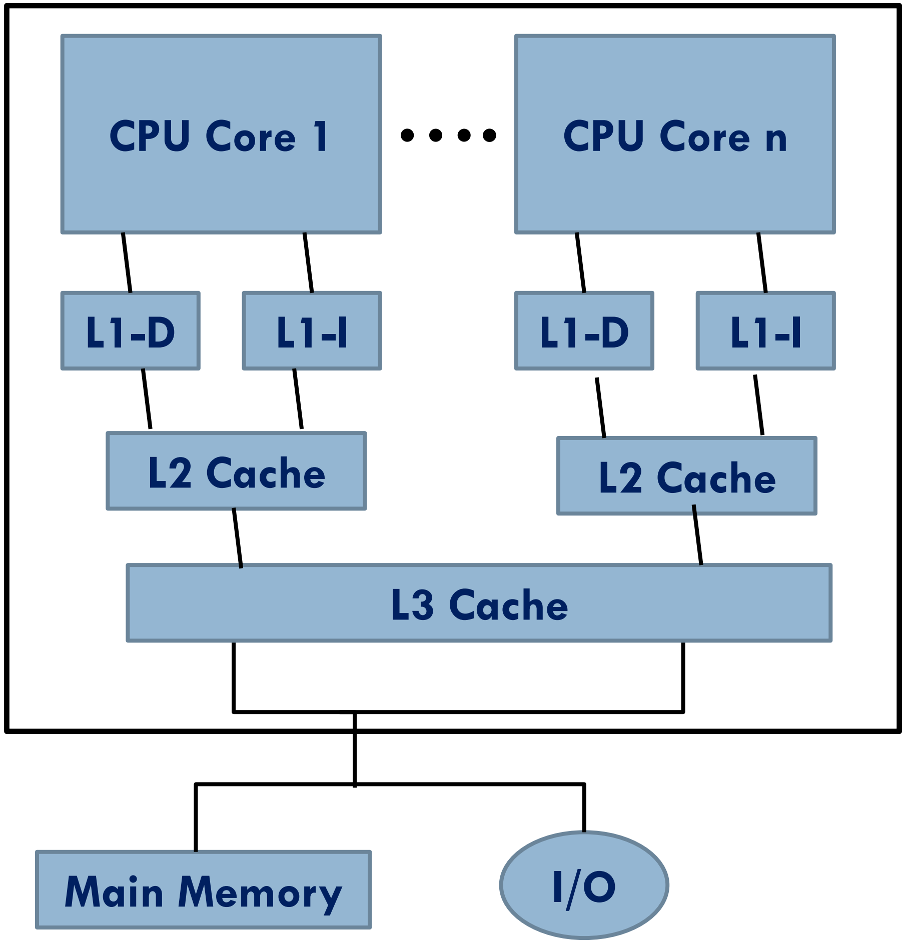


Dedicated L2 Cache

Multicore Organization Alternatives



Shared L2 Cache



Shared L3 Cache

Evolution of Multicore

- Driven by a performance hungry market, microprocessors have always been designed keeping **performance and cost** in mind.
- Gordon Moore, founder of Intel Corporation predicted that the number of transistors on a chip will double once in every 18 months to meet this ever growing demand which is popularly known as Moore's Law in the semiconductor industry . Advanced chip fabrication technology alongside with integrated circuit processing technology offers increasing integration density which has made it possible to integrate one billion transistors on a chip to improve performance.
- However, the performance increase by micro-architecture governed by Pollack's rule is roughly proportional to square root of increase in complexity. This would mean that doubling the logic on a processor core would only improve the performance by 40%.
- With advanced chip fabrication techniques comes along another major bottleneck, **power dissipation** issue. Studies have shown that transistor leakage current increases as the chip size shrinks further and further which increases static power dissipation to large values.

Evolution of Multicore

- One alternate means of improving performance is to increase the frequency of operation which enables faster execution of programs . However the frequency is again limited as any increase beyond a certain frequency increases power dissipation again .
- Power consumption has increased to such high levels that traditional air-cooled microprocessor server boxes may require budgets for liquid-cooling or refrigeration hardware. Designers eventually hit what is referred to as the power wall, the limit on the amount of power a microprocessor could dissipate
- Semiconductor industry once driven by performance being the major design objective, is today being driven by other important considerations such chip fabrication costs, fault tolerance, power efficiency and heat dissipation. This led to the development of multi-core processors which have been effective in addressing these challenges.

https://nptel.ac.in/courses/106/103/106103183/

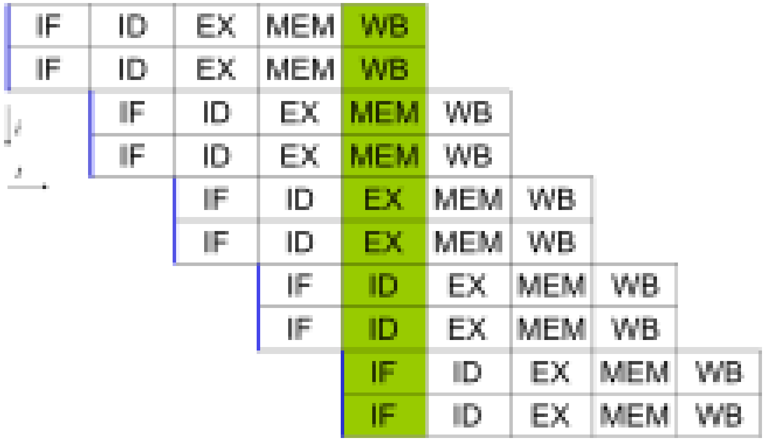
Factors that led to the development of Multicore

- The Organizational change in processor design have primarily been focused on increasing instruction level parallelism, so that more work could be done in each clock cycle.
  - These changes include
    - Pipelining
    - Superscaler
    - Simultaneous Multi threading
    - Multi core

Pipelining (ILP Wall)
Clock speed (Freq Wall)
Power dissipation (Power Wall)
complexity

Superscaler

- The term superscaler was first coined in 1987. It is a machine that is designed to improve the performance of execution of scalar instructions.
- The essence of superscalar approach is the ability to execute instructions independently and concurrently in different pipelines



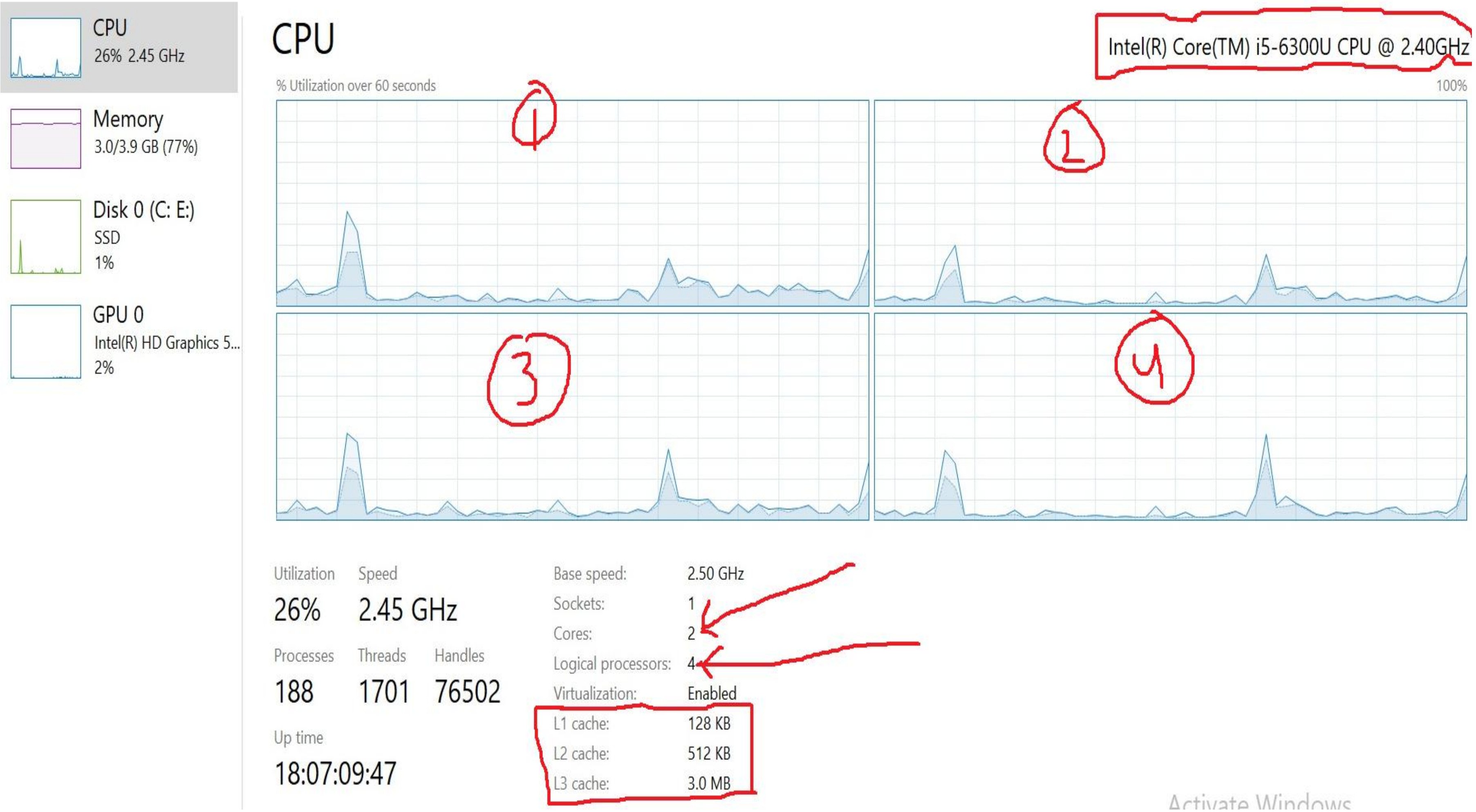
IF	ID	EX	MEM	WB				
IF	ID	EX	MEM	WB				
	IF	ID	EX	MEM	WB			
	IF	ID	EX	MEM	WB			
		IF	ID	EX	MEM	WB		
		IF	ID	EX	MEM	WB		
			IF	ID	EX	MEM	WB	
			IF	ID	EX	MEM	WB	
				IF	ID	EX	MEM	WB
				IF	ID	EX	MEM	WB

- In traditional scalar organization, there is a single pipelined functional unit for integer operations and one for floating point operations. Parallelism is achieved by enabling multiple instructions to be at different stages of the pipeline at one time.

- In Superscalar Organization, there are multiple functional units, each of which is implemented as a pipeline. Each individual functional unit provides a degree of parallelism by virtue of its pipeline structure. The use of multiple functional units enables the processor to execute streams of instructions in parallel, one stream for each pipeline.

- It is the responsibility of hardware in conjunction with compiler to assure that parallel execution does not violate the intent of the program

Windows Task Manager



CPU
26% 2.45 GHz

Memory
3.0/3.9 GB (77%)

Disk 0 (C: E:)
SSD
1%

GPU 0
Intel(R) HD Graphics 5...
2%

CPU

Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz

% Utilization over 60 seconds

100%

Utilization	Speed	
26%	2.45 GHz	
Processes	Threads	Handles
188	1701	76502
Up time		
18:07:09:47		

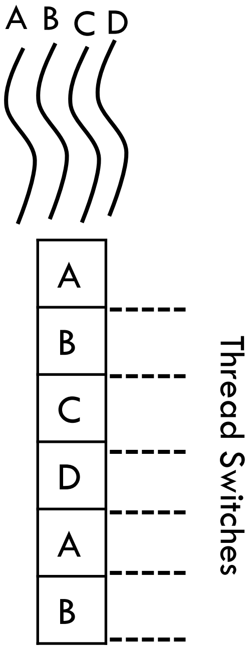
Base speed:	2.50 GHz
Sockets:	1
Cores:	2
Logical processors:	4
Virtualization:	Enabled
L1 cache:	128 KB
L2 cache:	512 KB
L3 cache:	3.0 MB

Multithreading

- Multithreading is an approach in which, the instruction stream is divided into several smaller streams, known as threads, such that the threads can be executed in parallel.
- Multi threads within a process share the same resources
- Thread switch is much less time consuming then process switch
- Approaches of Multi threading
  - Interleaved multithreading (Fine grained multithreading)
  - Blocked Multithreading (Coarse grained multi threading)
  - Simultaneous multithreading (SMT)

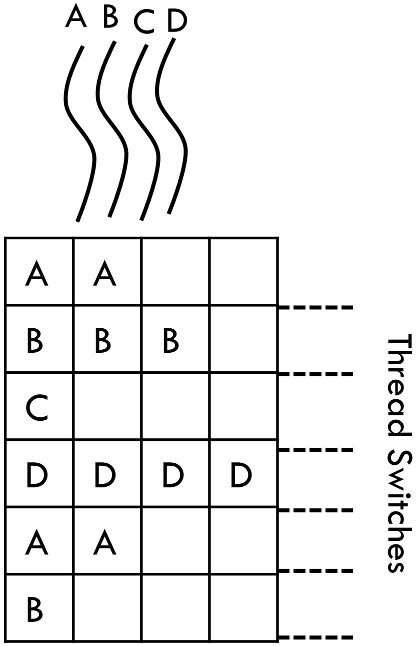
Fine Grained Multithreading

- The processor deals with two or more thread contexts at a time, switching from one thread to another at each clock cycle. If a thread is blocked because of data dependency or memory latencies, that thread is skipped and ready thread is executed.



A
B
C
D
A
B

Interleaved multithreading scalar

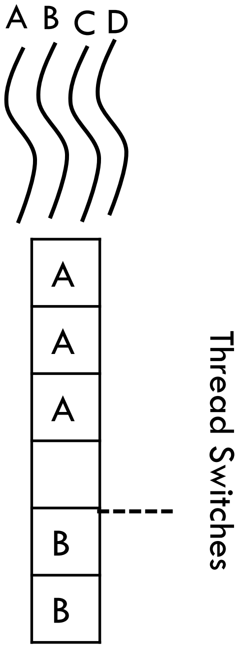


A	A		
B	B	B	
C			
D	D	D	D
A	A		
B			

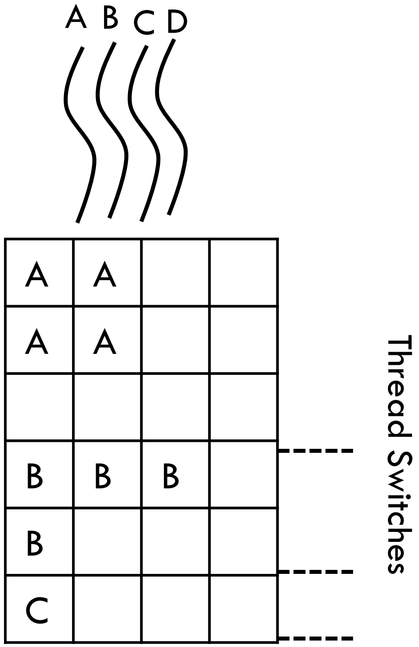
Interleaved multithreading superscalar

Coarse grained multi threading

□ The instructions of a thread are executed successively until an event occurs that may cause delay, such as cache miss. The event induces a switch to another thread.



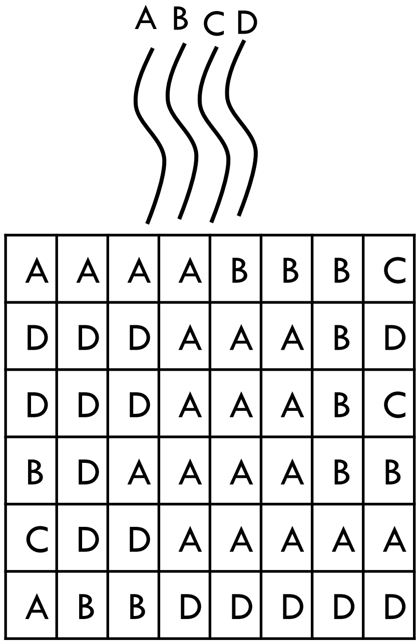
Blocked multithreading scalar



Blocked multithreading superscalar

Simultaneous Multithreading

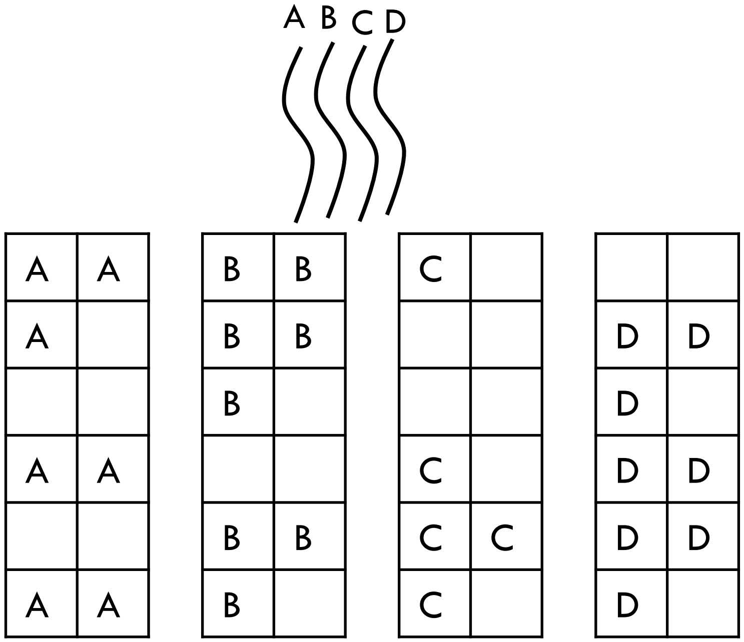
- Instructions are simultaneously issued from multiple threads to execution units of a superscalar processor. This combines the wide superscalar instruction issue capability with the use of multithread contexts.



A	A	A	A	B	B	B	C
D	D	D	A	A	A	B	D
D	D	D	A	A	A	B	C
B	D	A	A	A	A	B	B
C	D	D	A	A	A	A	A
A	B	B	D	D	D	D	D

Simultaneous multithreading (SMT)

Chip multiprocessor (Multicore)

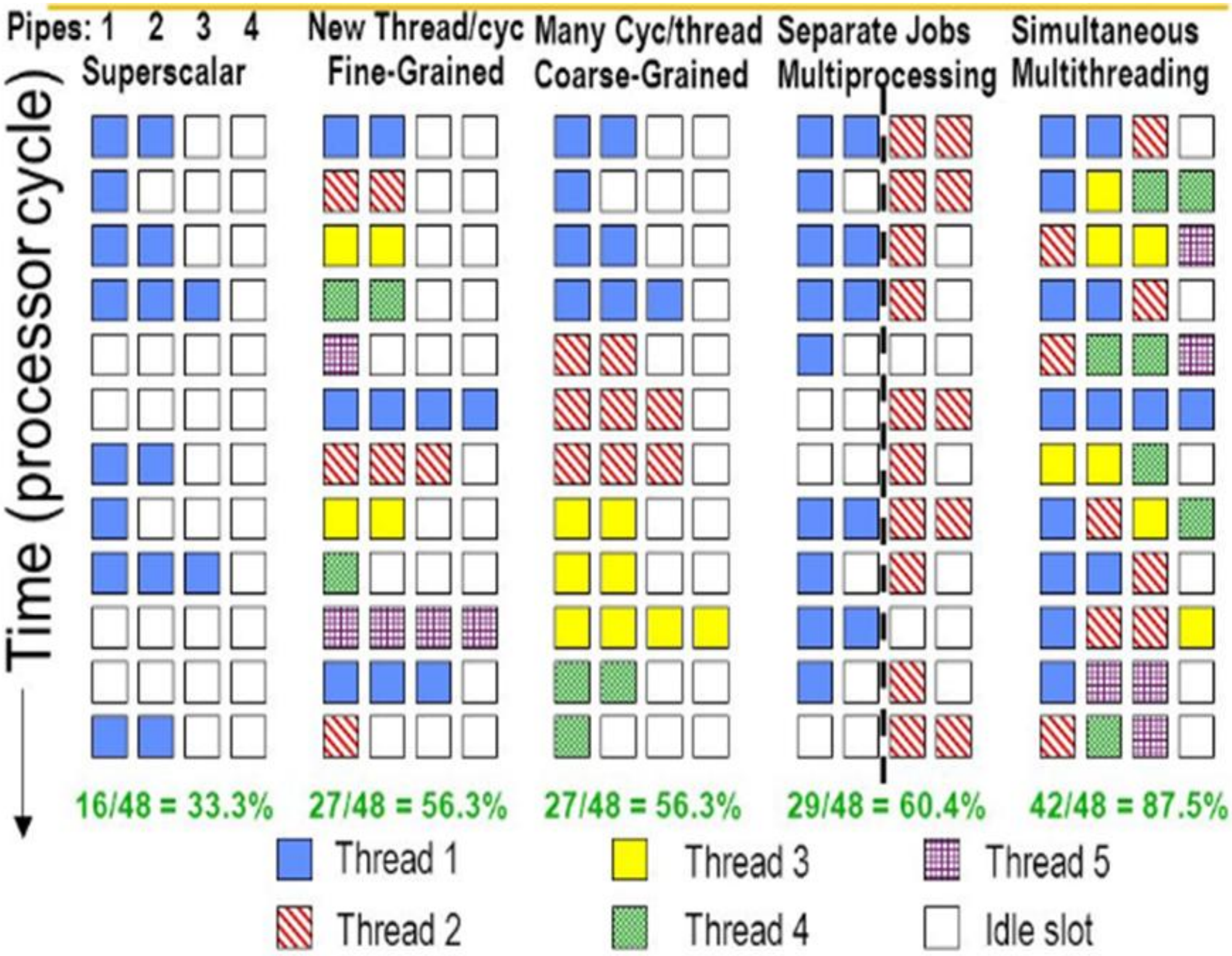


A	A
A	
A	A
A	A

B	B
B	B
B	
B	B
B	

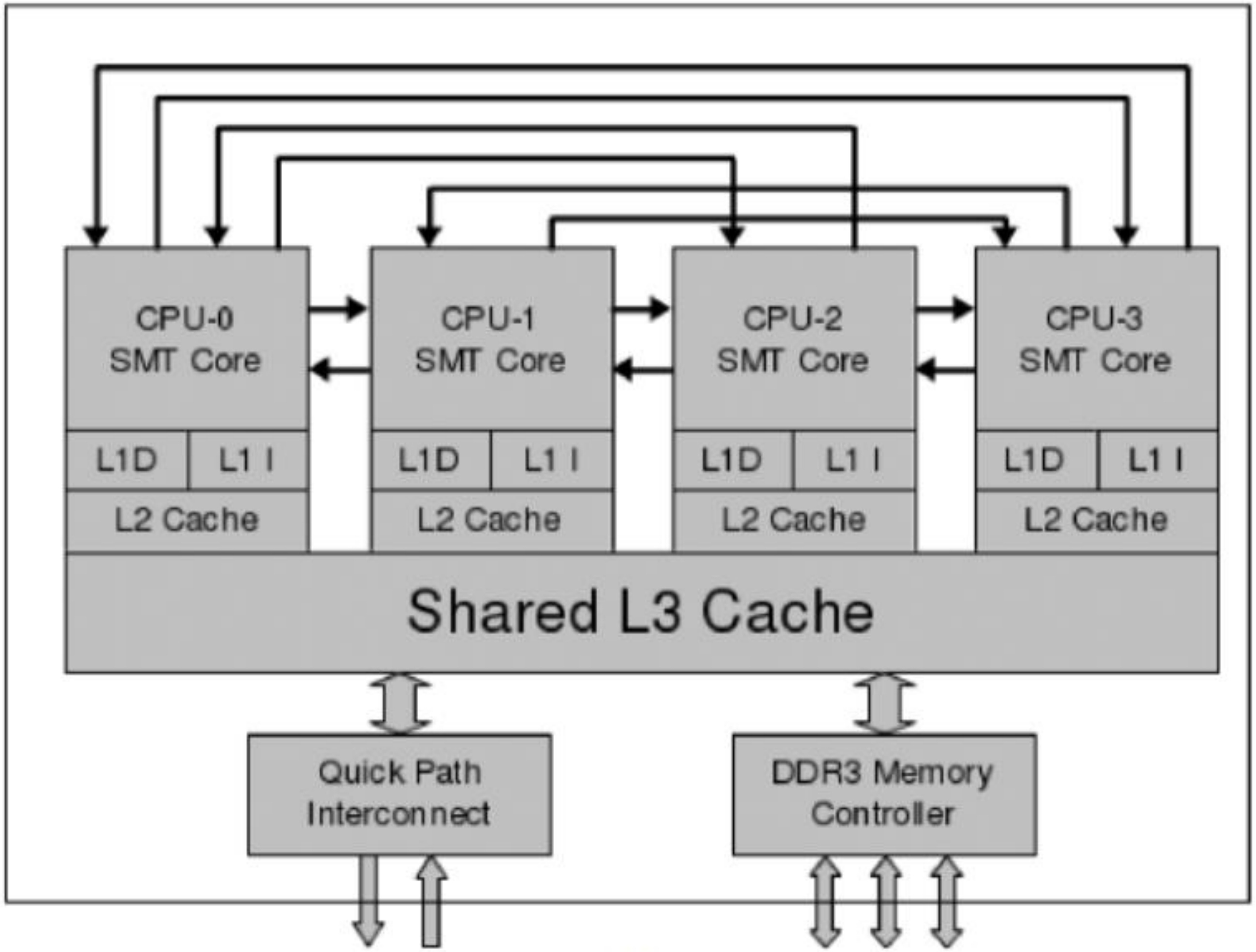
C	
C	
C	C
C	

D	D
D	
D	D
D	D
D	


Pipes: 1 2 3 4
Superscalar
New Thread/cyc
Fine-Grained
Many Cyc/thread
Coarse-Grained
Separate Jobs
Multiprocessing
Simultaneous
Multithreading
Time (processor cycle)
16/48 = 33.3%
27/48 = 56.3%
27/48 = 56.3%
29/48 = 60.4%
42/48 = 87.5%
Thread 1
Thread 2
Thread 3
Thread 4
Thread 5
Idle slot

Intel Core i7

One of the variants of Core i7



Intel Core i7 Processor

Intel Core i7

- Intel Core i7 family
  - https://en.wikipedia.org/wiki/List_of_Intel_Core_i7_processors

- Example: Core i7-7700 Processor
  - 7th generation
  - 3.6 GHz
  - Segment: Desktop
  - **Cores: 4 Threads: 8**
  - L2 Cache: 4x256 KB, L3: 8MB
  - Uses **Hyper-Threading** Technology

Hyper-threading

- **Hyper-threading** (officially called Hyper-Threading Technology or HT Technology and abbreviated as HTT or HT) is Intel's proprietary **simultaneous multithreading (SMT)** implementation used to improve parallelization of computations (doing multiple tasks at once) performed on x86 microprocessors.
- It was introduced on Xeon server processors in February 2002 and on Pentium 4 desktop processors in November 2002. Since then, Intel has included this technology in Itanium, Atom, and **Core 'i' Series** CPUs, among others.

- For each processor core that is physically present, the operating system addresses **two** virtual (logical) cores and shares the workload between them when possible.
- The main function of **hyper-threading** is to increase the number of independent instructions in the pipeline; it takes advantage of **superscalar architecture**, in which multiple instructions operate on separate data in parallel. With HTT, one physical core appears as two processors to the operating system,
  allowing concurrent scheduling of two processes per core.

https://en.wikipedia.org/wiki/Hyper-threading

References

- □ https://www.youtube.com/watch?v=nF3kr5KvUno&list=PLwdnzlV3ogoU0TR333JyxG8T3HDg52S0h
- □ Video lecture by Dr. John Jose , IIT Gowhati, Superscalar Pipeline https://nptel.ac.in/courses/106/103/106103183/
- □ William Stallings, Computer Organization & Architecture, 9th Ed, Pearson