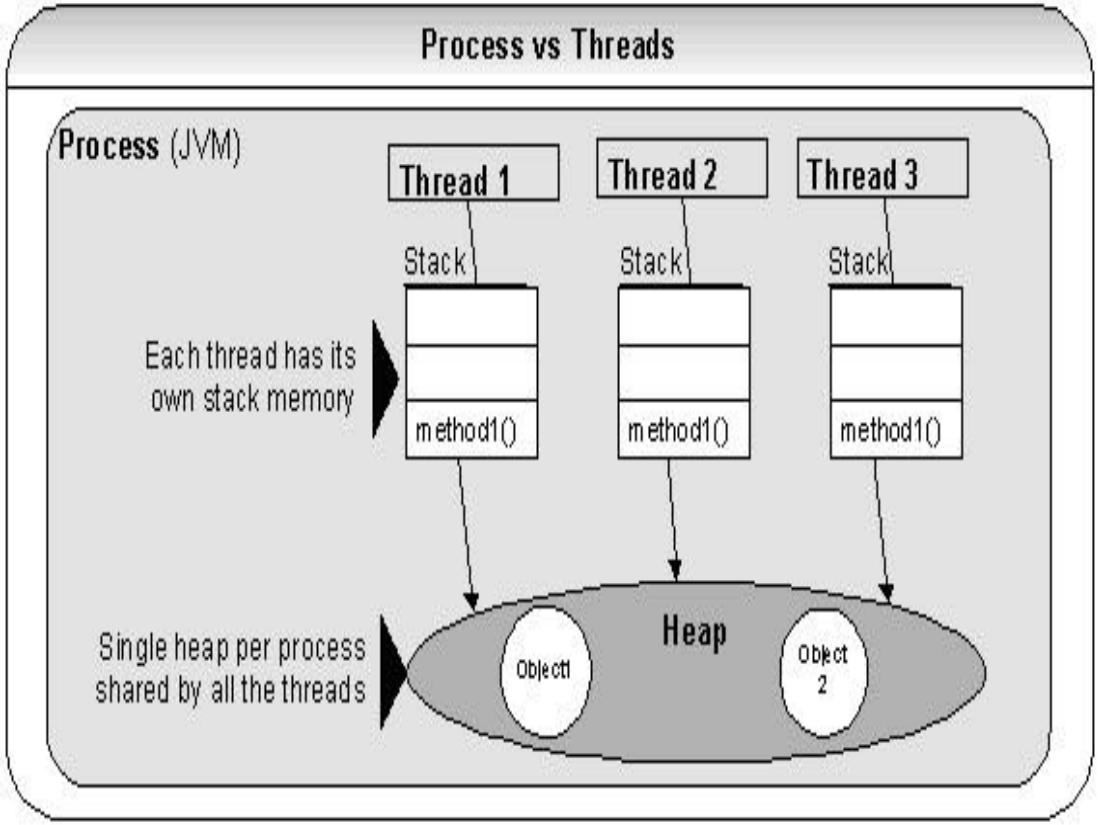
Multicore Processors

Archana Subramanian
Nikhat Farha

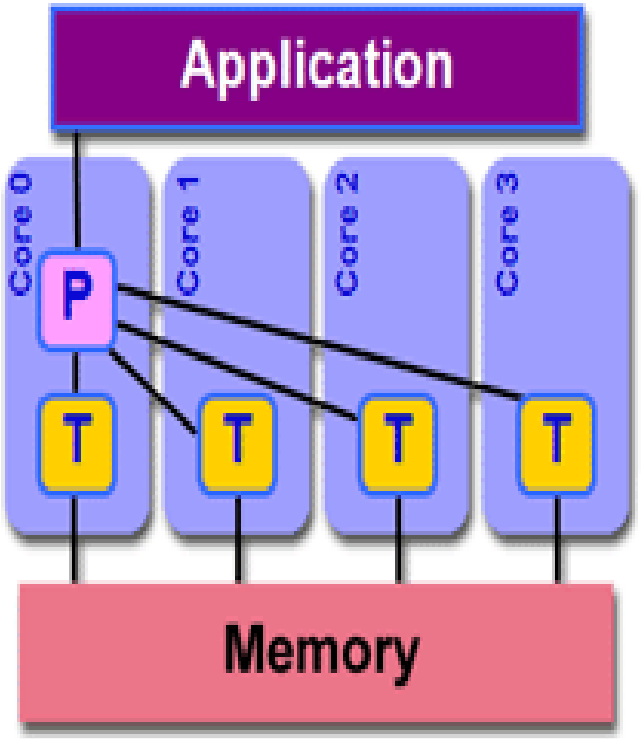
MULTITHREADING

- Application with multiple threads running within a process.
- Ability of a program or operating system to manage its use by more than one user at a time.
- Manage multiple requests without having multiple copies of the same program.



MULTIPROCESSING

- Processes are insulated from each other by the operating system.
- Error in one process cannot bring down another process.
- Operating system keeps track of the tasks and goes from one task to the next without loss of information.



PROCESS VS. THREAD

PROCESS	THREAD
Executable object in an container, an application	Stream of instruction within a process.
Processes do not share memory	Threads share the same memory
Message passing.	Inter thread communication.
Security	Speed

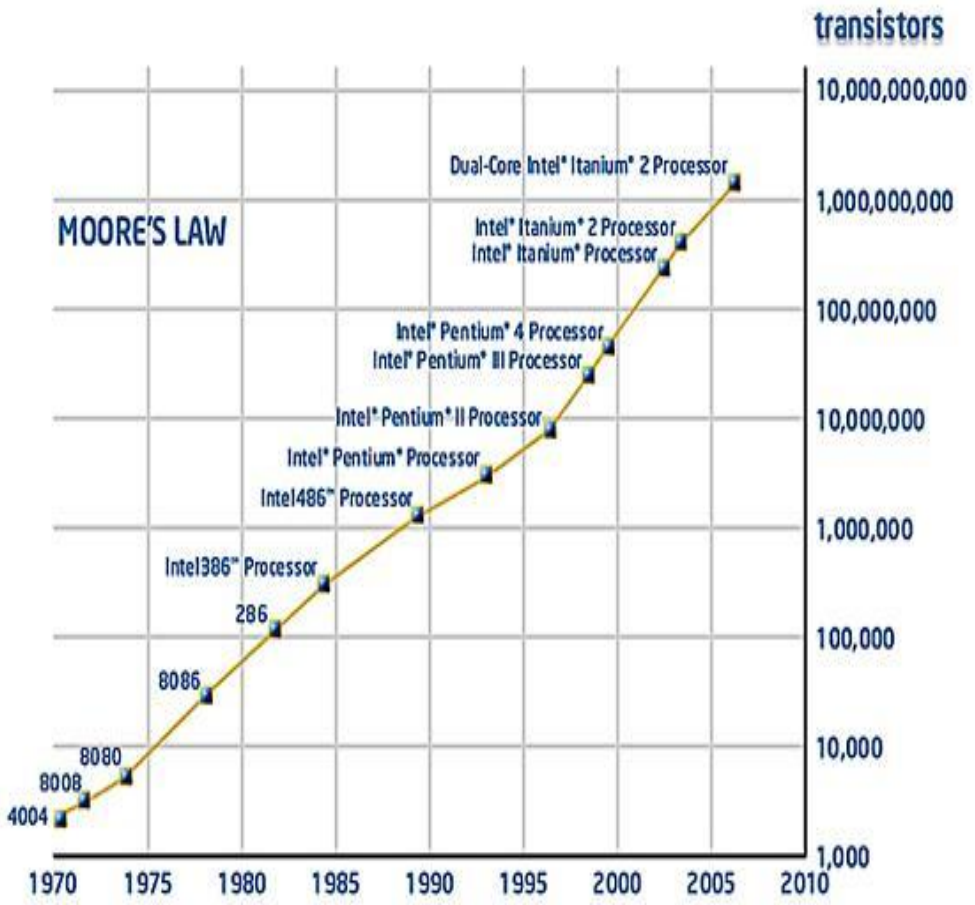
Hybrid design of Multi threading and Multi processing leads to efficient parallel capability of Hardware

MOORE'S LAW

Statement

"The number of transistors on a chip will roughly double each year."

"Computer performance will double every 18 months."



GENERATIONS

1st generation:

- 1971: Intel 4004 (2300 transistors)
- 1974: Intel 8080 (4500 transistors)

➢ Instruction processing was strictly sequential. Instructions were fetched, decoded and executed strictly one at a time

2nd generation:

- 1979: Motorola MC68000 (68000 transistors)

➢ Primitive pipelining with three stages: fetch, decode, execute; only one instruction is in execution at a certain time

GENERATIONS

3rd generation:

• 1984: Motorola MC68020 (240000 transistors)
- ➢ Five stage pipeline; increased parallelism

4th generation:

• 1990: MotorolaMC88110, Intel 80960 - these are RISCs (over 1 Million transistors)
• PowerPC604, Pentium
- ➢ Superscalar architectures; parallel execution based on multiple pipelines and functional units

GENERATIONS

5th generation:

- 1996: P6, PowerPC 620 (over 5 Million transistors)
- MIPS R10000, AMD K5/K6, UltraSparc (not exactly out of order)

➢ Superscalars with out of order execution and sophisticated scheduling and renaming strategies

VLIW

➢ VLIW (very large instruction word) - a compiler schedules the instructions statically (instead of dynamic scheduling as with superscalars).

➢ Compiler is very much architecture dependent ⇒ poor portability

WHY MULTICORE?

- Difficult to make single-core
  -clock frequencies even higher – heat problems
- Deeply pipelined circuits:
  -heat problems, needs special cooling arrangements
- Many new applications are multithreaded.
- General trend in computer architecture (shift towards more parallelism)

MULTICORE TECHNOLOGY

A multi-core processor is a single computing component with two or more independent actual central processing units (called "cores"), which are the units that read and execute program instructions.

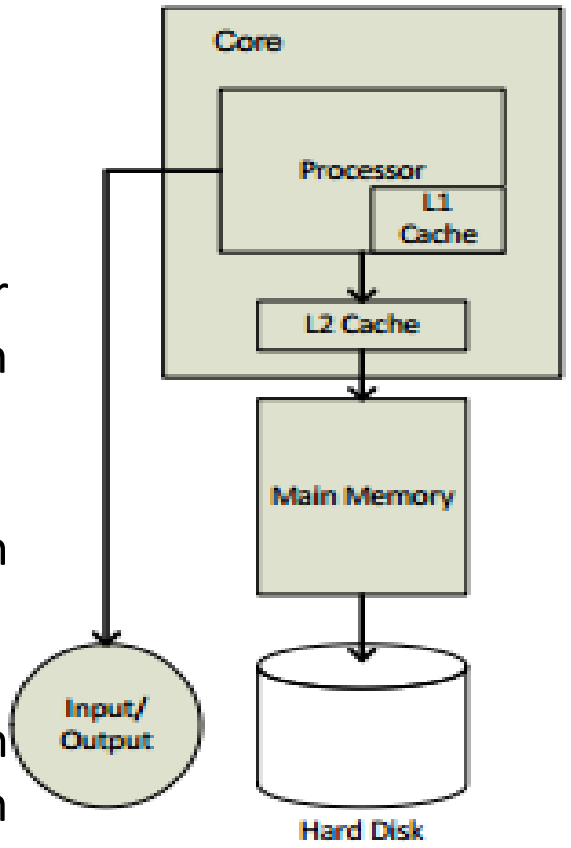
Benefits

- Less Power Consumption
- Simultaneous processing of multiple tasks
- Performance increase

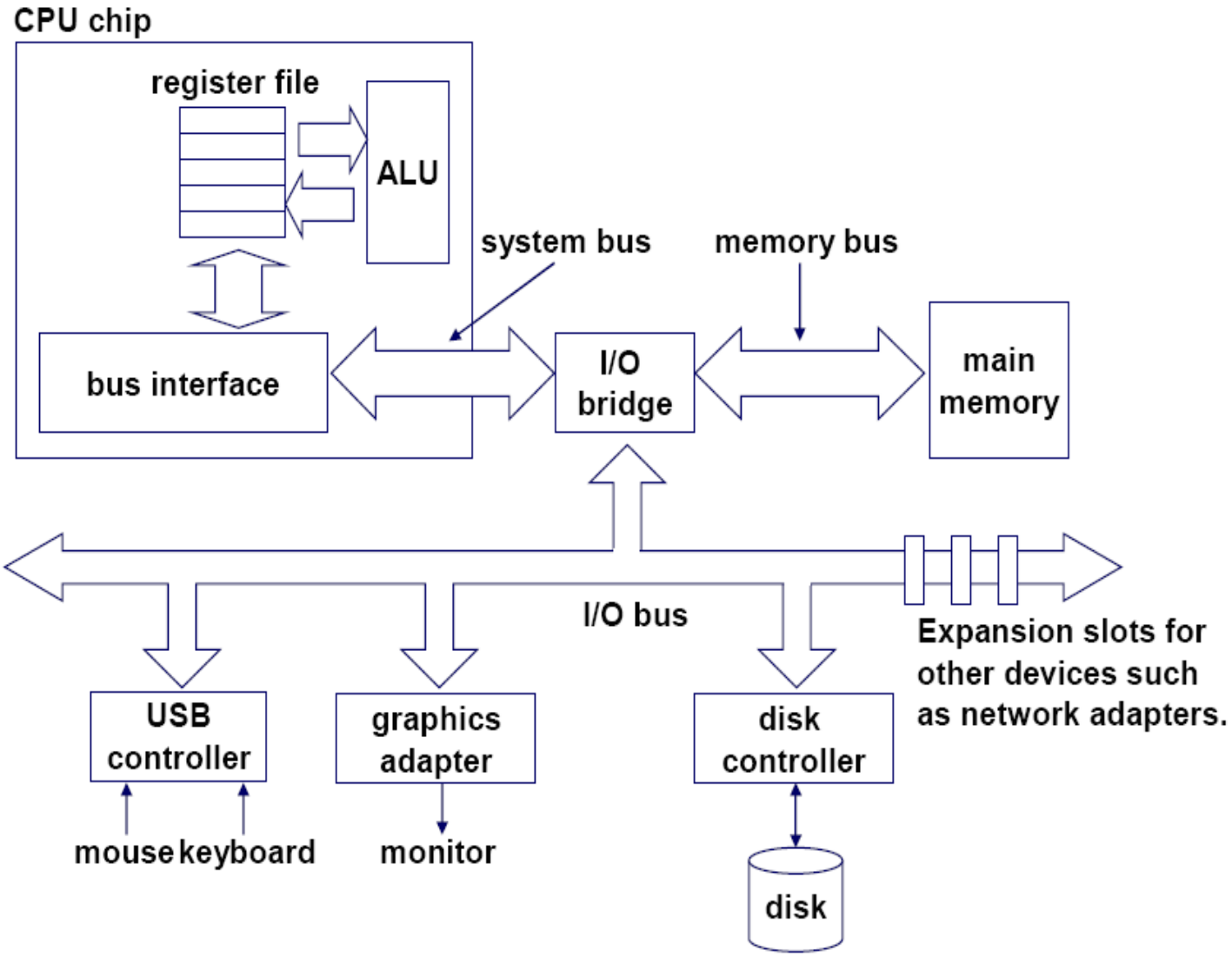
	Single core processor (45nm)	Multi-core processor (45nm)
Vdd	1.0V	1.0V
I/O pins(total)	1280 (ITRS)	3000 (Estimated)
Operating frequency	7.8GHz	4GHz
Chip-package data rate	7.8 Gb/s	4Gb/s
Bandwidth	125GByte/s	1 TeraByte/s
Power	429.78W	107.39W
Total number of pins on chip	3840	9000(Estimated)
Number of pins on the package	2480	4500(Estimated)

SINGLE CORE

- Closest to the processor is Level 1 (L1) cache.
  - very fast memory used to store data frequently used by the processor.
- Level 2 (L2) cache is just off-chip, slower than L1 cache, but still much faster than main memory
- Main memory is very large and slower than cache.
- When data isn't located in cache or main memory the system must retrieve it from the hard disk, which takes exponentially more time than reading from the memory system.

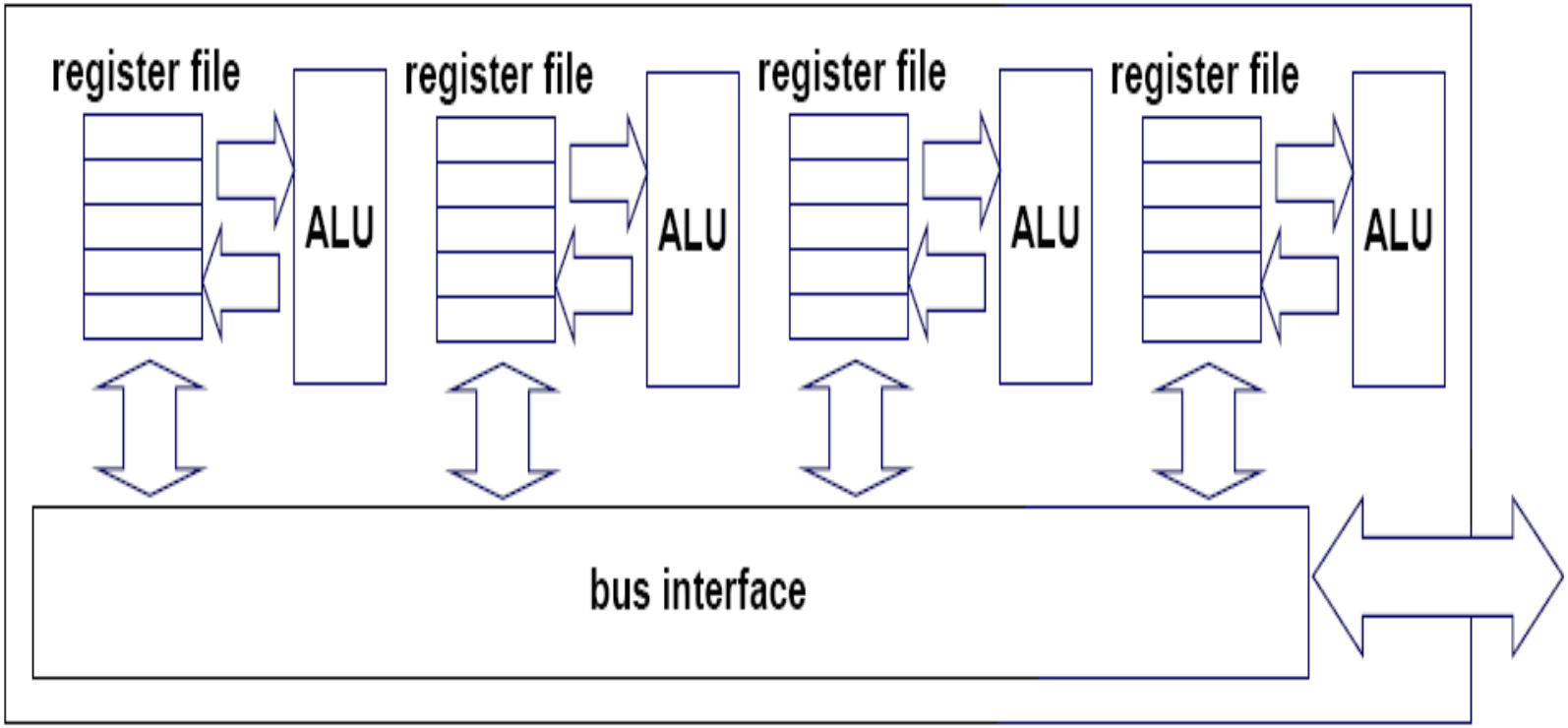


SINGLE CORE COMPUTER

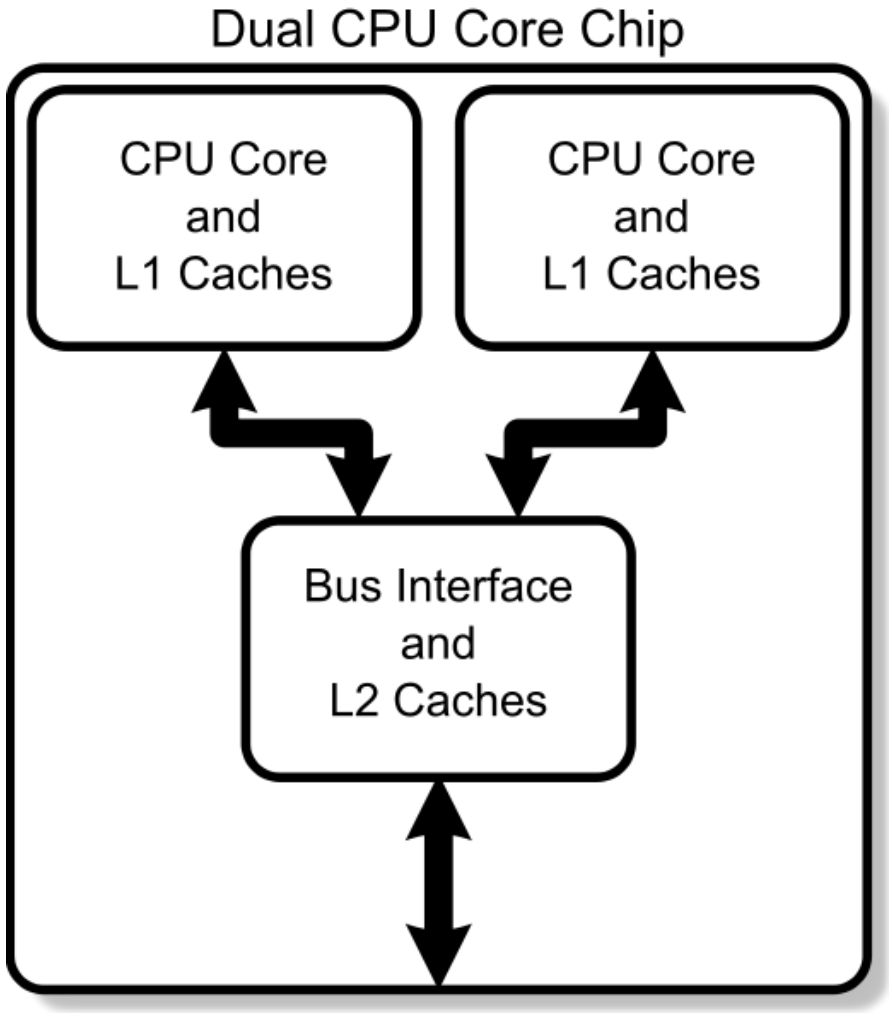


MULTI CORE COMPUTER

Replicate multiple processor cores on a single die

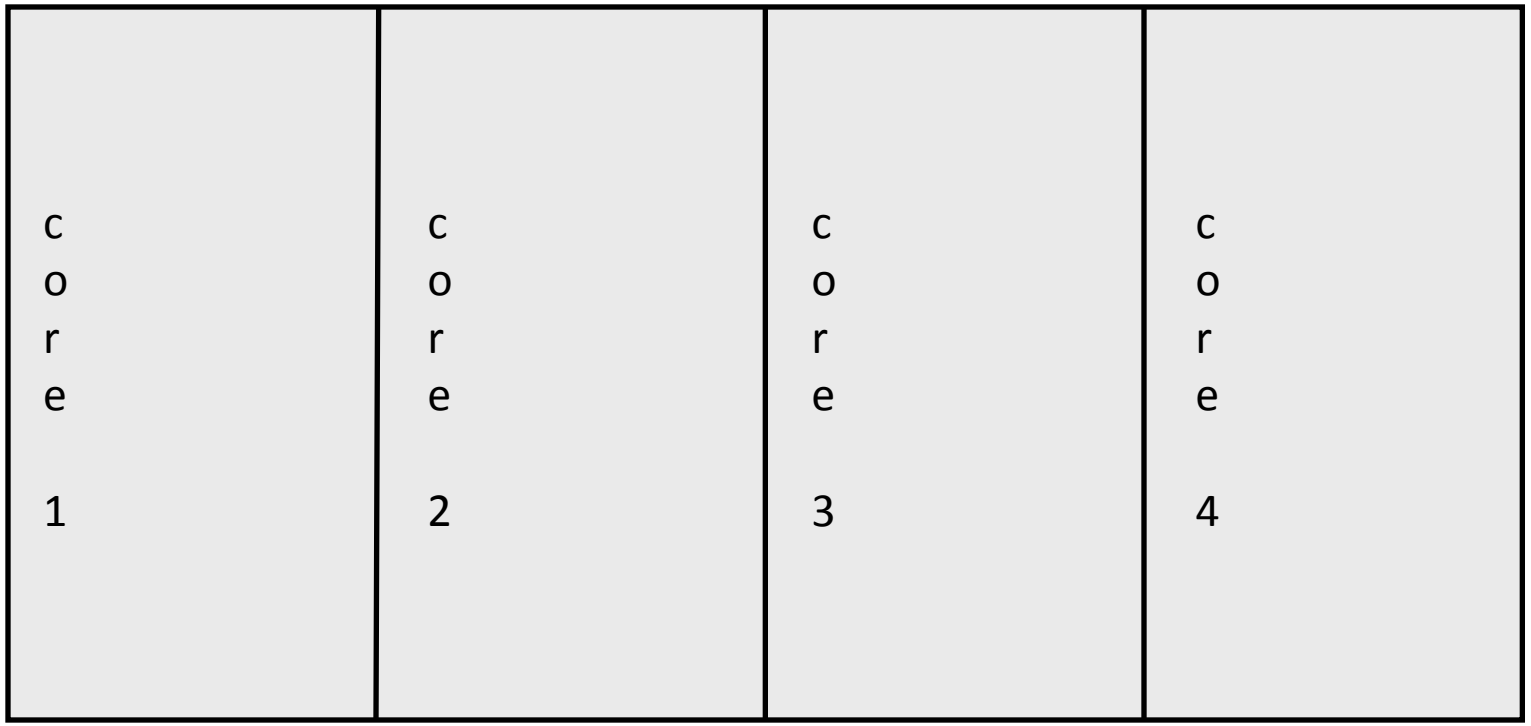


DUAL CORE PROCESSOR

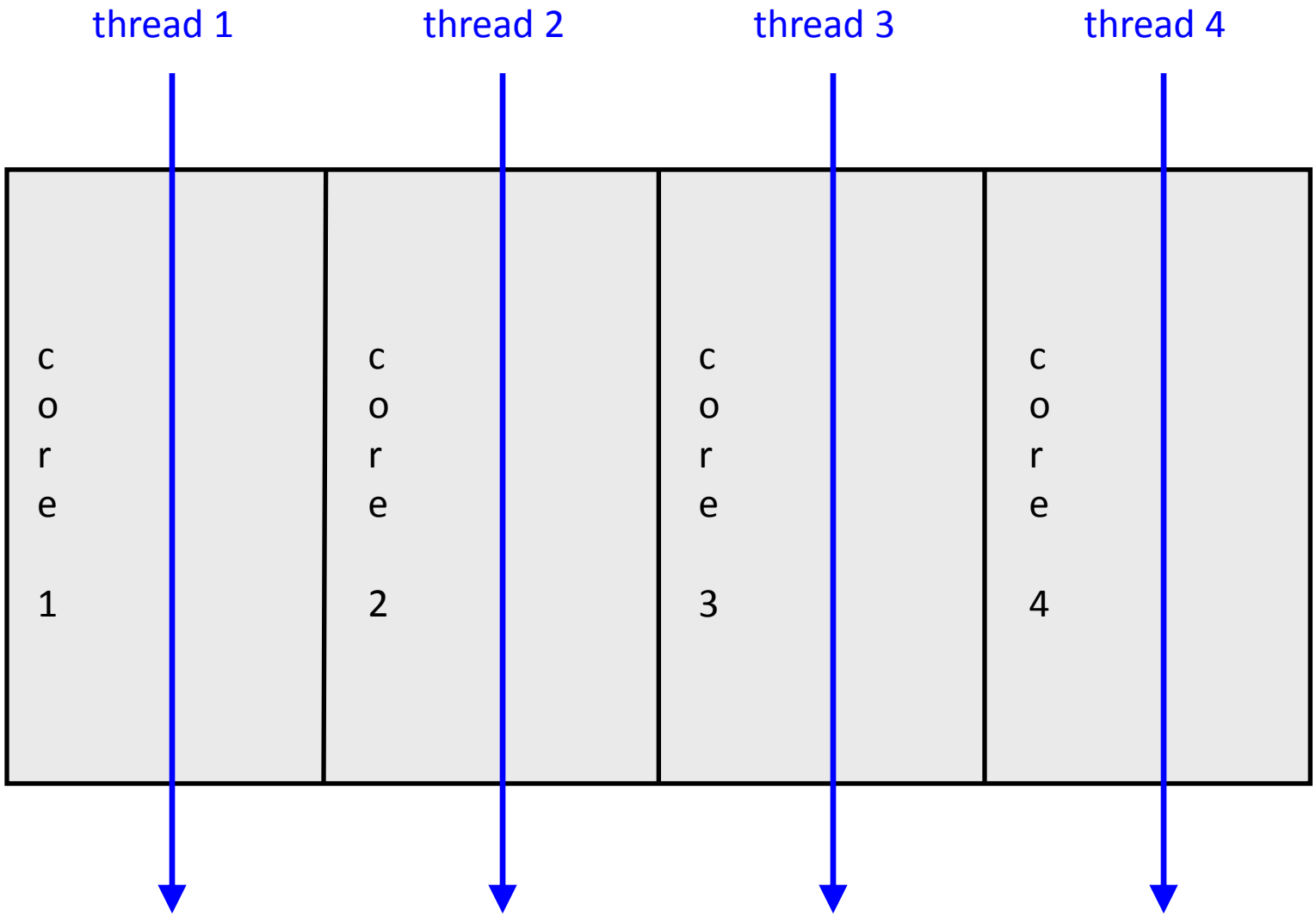


MULTI CORE CHIP

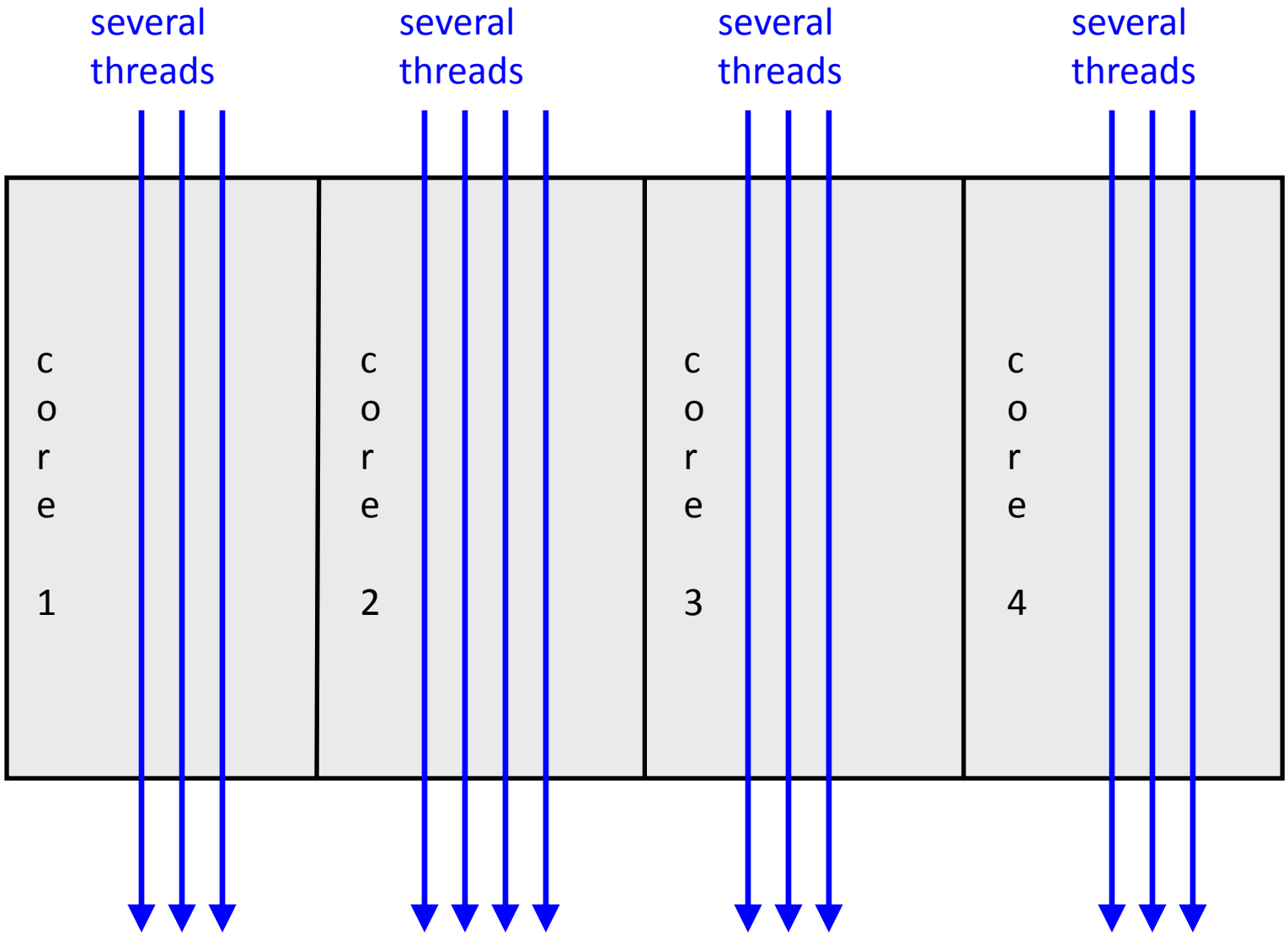
- The cores fit on a single processor socket
- Also called CMP (Chip Multi-Processor)



The cores run in parallel



Within each core, threads are time-sliced (just like on a uniprocessor)



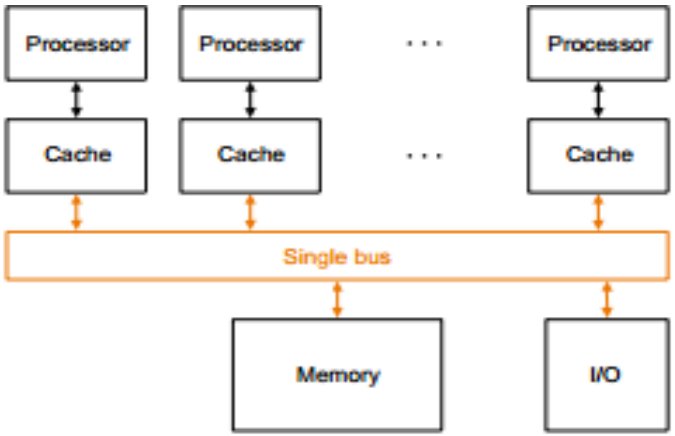
Interaction with the Operating System

- OS perceives each core as a separate processor
- OS scheduler maps threads/processes to different cores
- Most major OS support multi-core today: Windows, Linux, Mac OS X, …

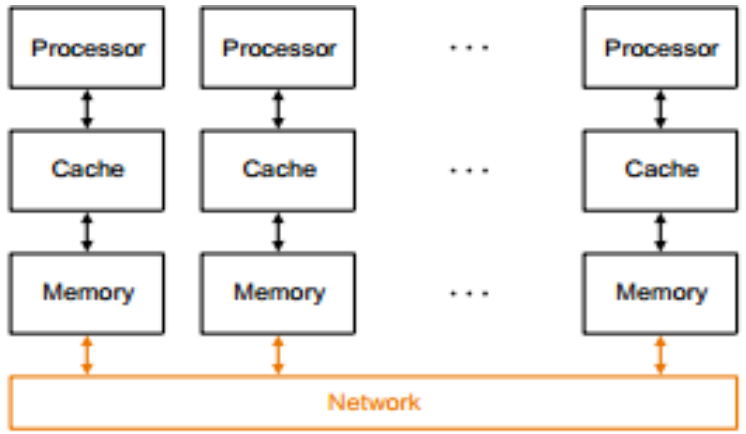
MEMORY MODELS

- **Shared memory**
  In this model, there is one (large) common shared memory for all processors

- **Distributed memory**
  In this model, each processor has its own (small) local memory, and its content is not replicated anywhere else.



(a)



(b)

Multi-core processor is a special kind of a multiprocessor:
All processors are on the same chip

- Multi-core processors are MIMD:
  Different cores execute different threads (Multiple Instructions), operating on different parts of memory (Multiple Data).

- Multi-core is a shared memory multiprocessor:
  All cores share the same memory

MULTICORE ARCHITECTURE

- Intel’s Core 2 Duo
- Advanced Micro Devices(AMD)“ Athlon 64 X2
- Sony-Toshiba IBM’s CELL Processor
- Tilera’s TILE64.

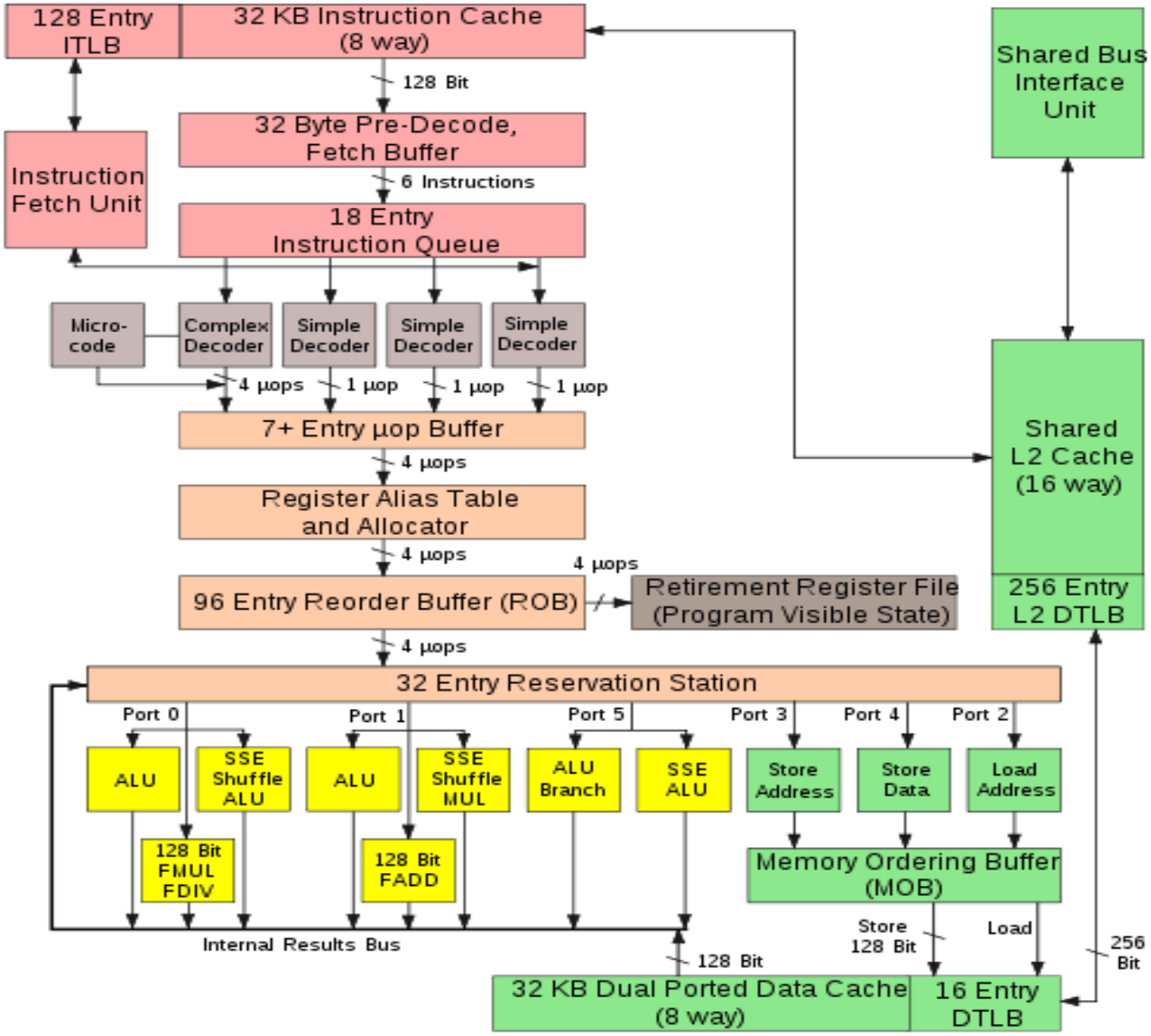
A) INTEL MULTICORE ARCHITECTURE

- It’s as if an Intel Xeon processor were opened up and inside were packaged all the circuitry and logic for two (or more) Intel Xeon processors.

- The Intel® Pentium® 4 and Intel Xeon processors today already use Hyper-Threading Technology (HT Technology) to execute multiple programs simultaneously.

THE INTEL P8

- Intel marketing states that Core is a blend of P-M techniques and NetBurst architecture.
- Core is a descendant of the Pentium Pro, or the P6 architecture.
- Prefetching was inspired by experiences with the Pentium 4. Everything else is an evolution of "Yonah" (Core Duo).
- The main challenge for the CPU designer today is the average memory latency the CPU sees.
- A Pentium 4 3.6 GHz with DDR-400 runs no less than 18 times faster than the base clock of the RAM (200 MHz). Every cycle the memory is being accessed, a minimum of 18 cycles pass on the CPU.



Intel Core 2 Architecture

The Core architecture prefetching is superior to what can be found in the Athlon 64 and Pentium 4:

- two data, one instruction - in each core, plus two prefetchers for the L2-cache.
- to avoid this bottleneck, the prefetch monitor of the Core CPUs always give priority to the demand bandwidth.
- The cache system of the Core CPU is also very impressive. A massive 4 MB L2-cache is shared between the two cores and is accessed in only 12 to 14 cycles. Each core also has a 3 cycle 32 KB Instruction cache and a 32 KB data cache at its disposal.

B) AMD MULTICORE ARCHITECTURE

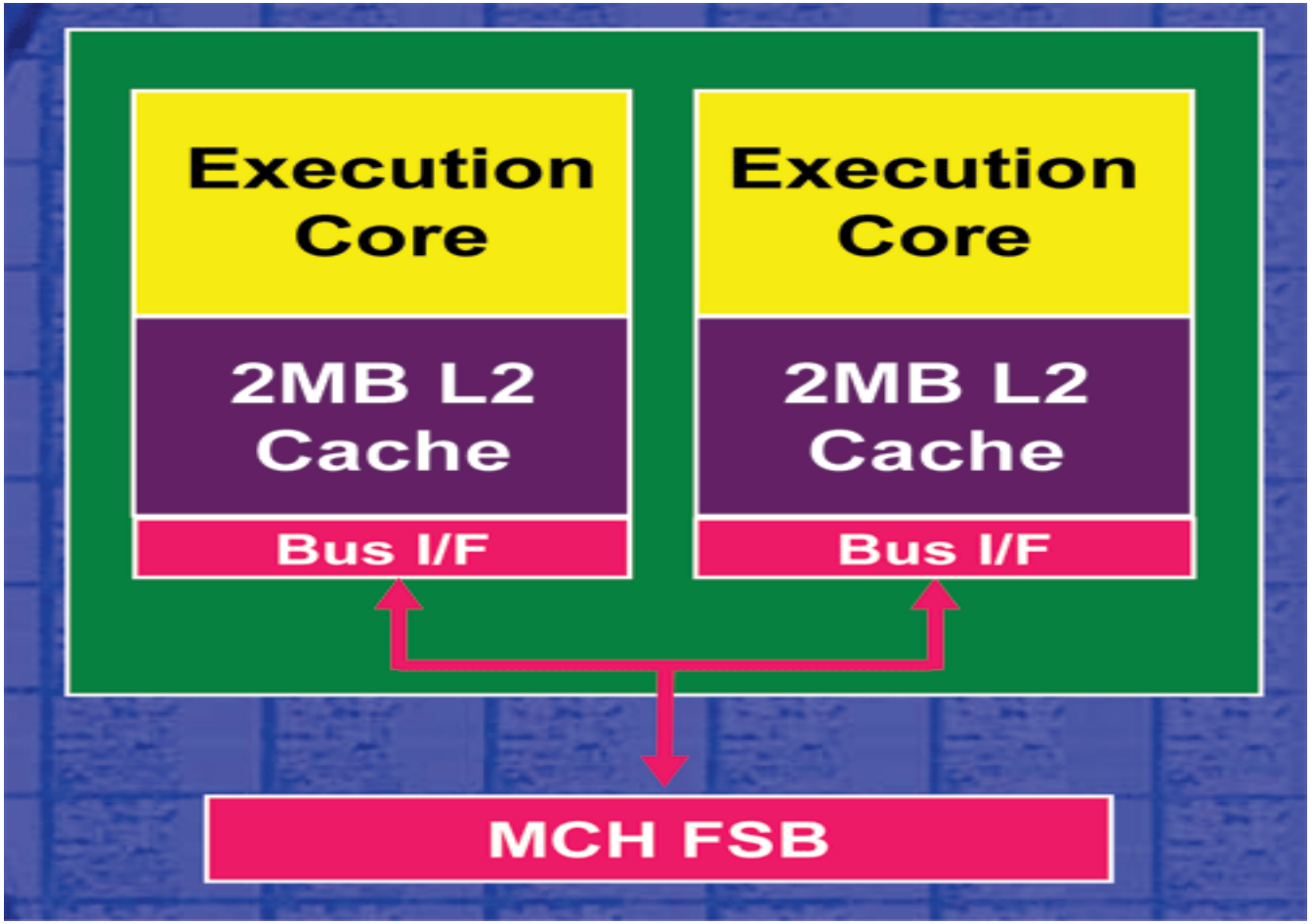
The first desktop dual-core processor from AMD appeared in May 2005 under the Athlon 64 X2 brand name. AMD has several variations of its Athlon 64 X2 processors.

Key Architectural Features AMD Athlon™ X2 Dual-Core Processors (as on AMD ‘s website) :

- AMD64 technology provides full-speed support for x86 code base for uncompromising performance
- 40-bit physical addresses, 48-bit virtual addresses
- Sixteen 64-bit integer registers
- Sixteen 128-bit SSE/SSE2/SSE3 registers
- AMD Digital Media XPress™ provides support for SSE, SSE2, SSE3 and MMX instructions

- A high-bandwidth, low-latency integrated DDR memory controller.
- Hyper Transport™ Technology for high speed I/O communication.
- Large high performance on-chip cache.
  - 64KB Level 1 instruction cache per core
  - 64KB Level 1 data cache per core
  - Up to 1MB Level 2 cache per core
- AMD Power Now!™ Technology (Cool ’n’ Quiet™ Technology) for quieter operation and reduced power requirements.

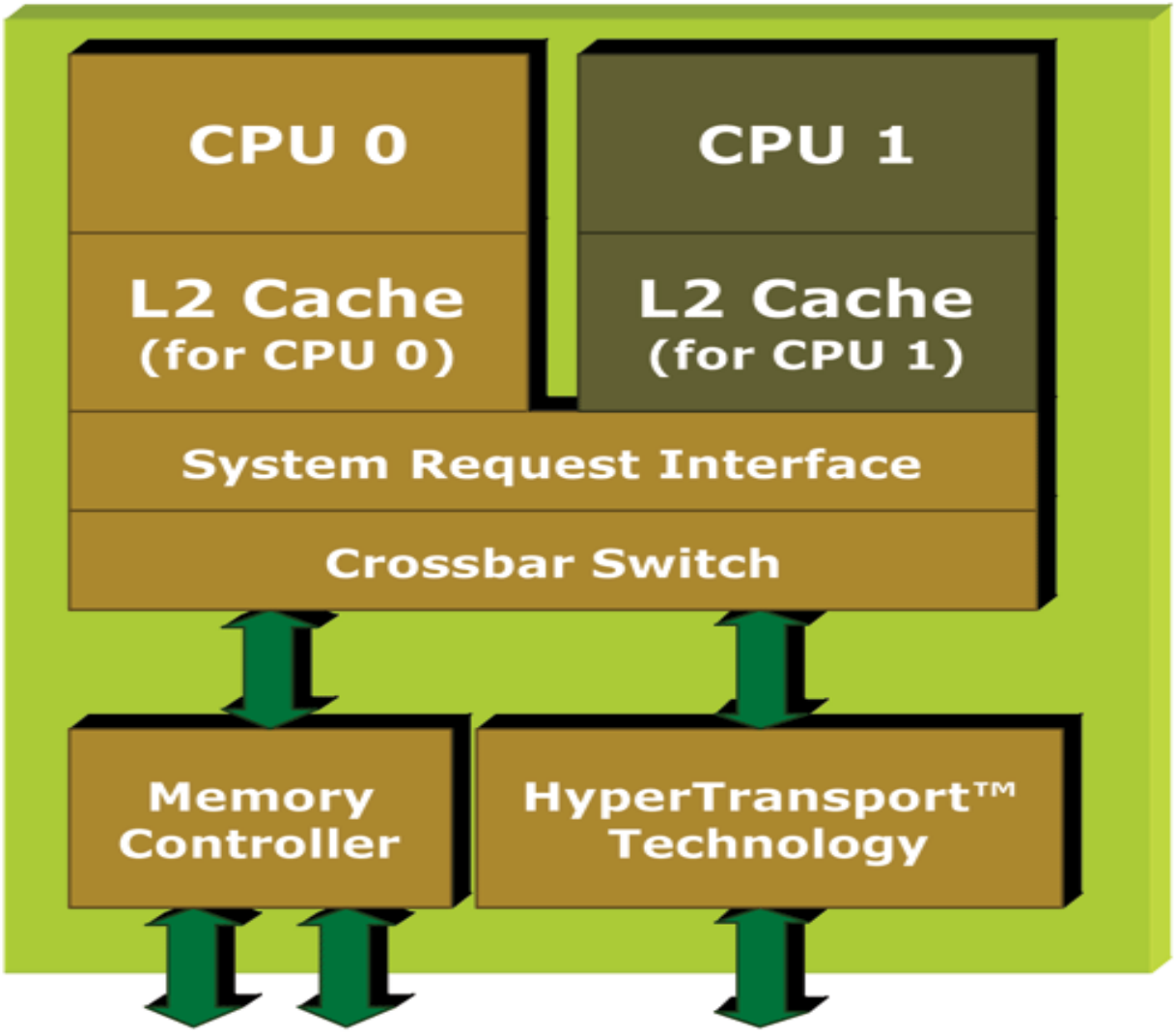
A LOOK AT AMD'S DUAL CORE ARCHITECTURE:



Intel -- Two Pentium 4 cores literally glued together. The two cores share the same voltage and must run in the same power state and all communication between cores must go over the external FSB.

AMD's architecture -- is much more sophisticated.

Instead of having all communication between the cores go over an external FSB, each core will put its request on the System Request Queue (SRQ) and when resources are available, the request will be sent to the appropriate execution core - all without leaving the confines of the CPU's die.



AMD Athlon™ 64 X2
Dual-Core Processor Design

- The one limitation that both AMD and Intel have is bandwidth.

- The benefit is that AMD's dual core CPUs will work in almost all Socket-940 and Socket-939 motherboards. But the downside is that the memory bus remains unchanged at 128-bits wide and supports a maximum memory speed of DDR400.

- While single core Athlon 64 and Opteron CPUs get a full 6.4GB/s of memory bandwidth, today's dual core CPUs are given the same memory bandwidth to share among two cores instead of one.

- Intel is actually in a better position from a memory bandwidth standpoint. At this point, their chipsets provide more memory bandwidth than what a single core needs with their dual channel DDR2-667 controller. The problem is that the Intel dual core CPUs still run on a 64-bit wide 800MHz FSB, which makes Intel's problem more of a FSB bandwidth limitation than a memory bandwidth limitation.

Comparison and Conclusion

Cache configuration	AMD K7 ("Barton")	AMD K8 ("Hammer")	Core	Intel NetBurst "Northwood"	Intel NetBurst "Prescott"	P-M Dothan	P6 ("Coppermine")
L1-cache (Instructions)	64 KB	64 KB	32 KB	8 KB	16 KB	32 KB	16 KB
L1-cache (Data)	64 KB	64 KB	32 KB	8 - 16 KB*	8-16 KB*	32 KB	16 KB
L1-cache latency (load to use)	3	3	3	2	4	3	3
L1 Associativity	2-way	2-way	8-way	4-way D, 8-way I	8-way	8-way	4-way
L1 I TLB (Entries)	24	32	128	128	128	128	32
L1 D TLB (Entries)	24	32	256	8	8	128	64
L2-cache (maximum)	512 KB	1 MB	4 MB	512 KB	2 MB	2 MB	256 KB
L2 Associativity	16-way	16-way	16-way	8-way	8-way	8-way	8-way
L2-cache Width	64-bit	128-bit	256-bit	256-bit	256-bit	256-bit	256-bit
L2-cache Latency load to use (+L1-latency)	16	12	11-20	11-20 (*)	27	12	7
L2 TLB (Entries)	256	512	n/a	64	64	n/a	n/a
Max. Memory Bandwidth to CPU	3.2 GB/s	6.4 GB/s	10.6 GB/s	4.2 GB/s	8.5 GB/s	4.2 GB/s	1.06 GB/s

CORE VERSUS HAMMER: MEMORY SUBSYSTEM

- Hammer has 2 slight advantages over Core
  - The first is its bigger 2 x 64 KB L1 cache.
  - Die memory controller, which lowers the latency to the memory considerably.

Otherwise, the Core CPUs have much bigger caches and much smarter prefetching than the competition. The ***Core architecture's L1 cache delivers about twice as much*** *bandwidth while it's L2-cache is about 2.5 times faster than the Athlon 64/Opteron one.*

Server version of Core ("Woodcrest”):

- It is very hard to find a lot of ILP in server applications.

- The only weakness of Core is that there is no multi-threading in each Core. This is a logical result of the design goal of Core, an architecture which is an all-around compromise for the server, desktop and mobile markets.

POSSIBLE IMPROVEMENTS IN AMD ARCHITECTURE

- To enhance the SSE/SIMD power by increasing the wideness of each unit or by implementing more of them in the out of order FP pipeline.

- To sustain the extra (SIMD) FP power, AMD should definitely improve the bandwidth of the two caches further. The K7 had a pretty slow L2-cache, and the K8 doubled the amount of bandwidth that the L2 could deliver for example.

PARALLEL PROGRAMMING USING MULTI CORE

What applications are good candidates to be moved from serial to multithreaded to experience performance gains on multi-core systems?

✧Programs where threading was already implemented:

- Video encoding, 3D rendering, video/photo editing and high performance computing/workstation applications
- Gamming applications

Question arises !!

Besides using threaded applications, when else might end users experience a performance gain on multi-core systems?"

MULTICORE CHALLENGES

Is there any downside to multi-core architecture? Is it possible that my application will run slower?

- Multi-core platforms provide the next generation of performance, cost-efficiency and business value.

- Developers should encourage their customers to look for different usage or deployment models, make use of greater multi-tasking, or look for additional solution features or functionality that utilize the additional cores.

FINALLY

Multi-core processors are well-positioned to bring performance benefits to emerging software usage models such as virtualization, service-oriented architecture (SOA), and Web services, which require concurrent execution of multiple processes.