Computer Architecture

V semester

Chapter 10: Multicore Computers

3 hours

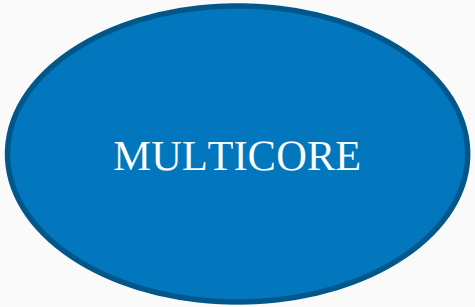
Compiled by :

Ankit Bhattarai, Assistant Professor
Email: ankitbhattarai@cosmoscollege.edu.np

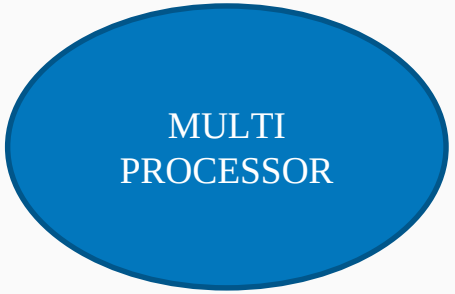
Topics to be discussed

- Hardware Performance Issues
- Software Performance Issues
- Multicore Organization
- Dual Core and Quad Core Processors
- Power Efficient Processors

Multicore Vs Multiprocessor



- A single CPU or processor with two or more independent processing units called cores that are capable of reading and executing program instructions.
- Not so reliable



- It consists of two or more CPUs that allows simultaneous processing of programs.
- Reliable in case of failure.

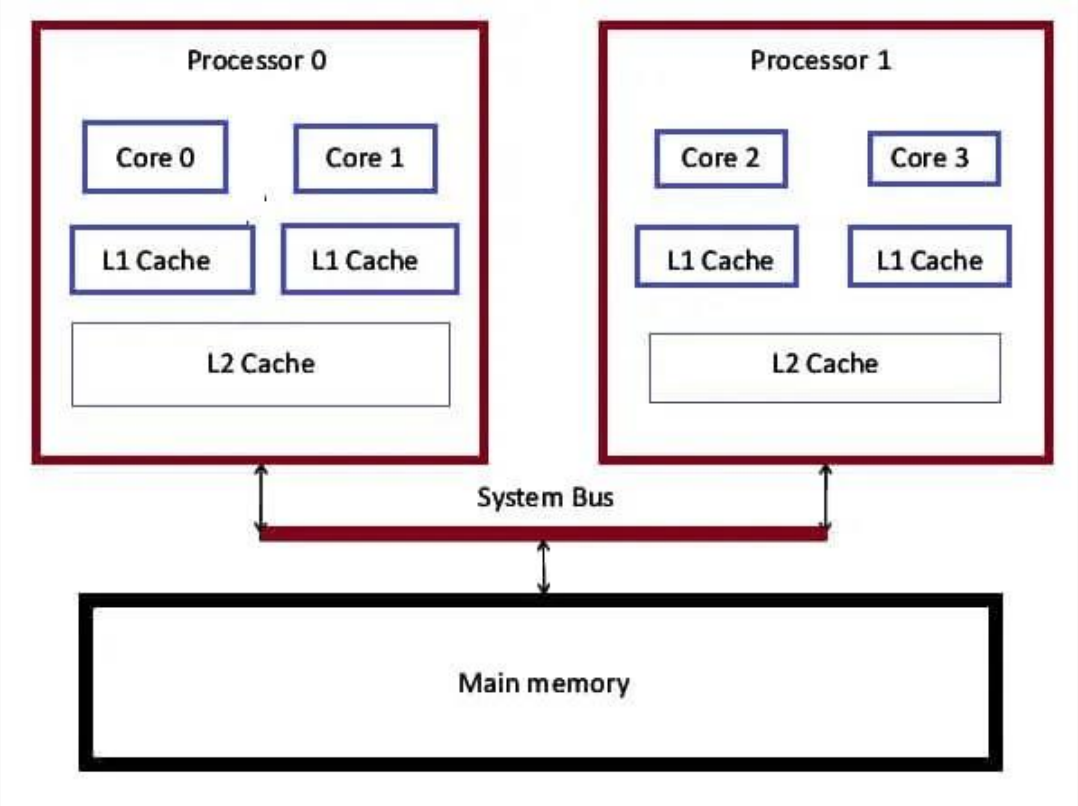
Multicore

- Multicore refers to an architecture in which a single physical processor incorporates the core logic of more than one processor.
- A single integrated circuit is used to package or hold these processors.
- These single integrated circuits are known as a *die*.
- Multicore architecture places multiple processor cores and bundles them as a single physical processor.
- The objective is to create a system that can complete more tasks at the same time, thereby gaining better overall system performance.

Multicore

- Processors were originally developed with only one core.
- Multi-core processors were developed in the early 2000s by Intel, AMD and others.
- Multicore processors may have two cores (dual-core CPUs, for example AMD Phenom II X2 and Intel Core Duo), four cores (quad-core CPUs, for example AMD Phenom II X4, Intel's i5 and i7 processors), six cores (hexa-core CPUs, for example AMD Phenom II X6 and Intel Core i7 Extreme Edition 980X), eight cores (octo-core CPUs, for example Intel Xeon E7-2820 and AMD FX-8350), ten cores (for example, Intel Xeon E7-2850), or more.
- Some of the processors of 12th generation may consists up to 14 cores.

Multicore : Architecture



Multicore

- A multi-core processor implements multiprocessing in a single physical package.
- Designers may couple cores in a multi-core device tightly or loosely. For example, ***cores may or may not share caches***, and they may implement message passing or shared memory inter-core communication methods.
- Common network topologies to interconnect cores include bus, ring, two-dimensional mesh, and crossbar.
- Homogeneous multi-core systems include only identical cores, heterogeneous multi-core systems have cores that are not identical.

Hardware Performance Issues : Increase in parallelism

The organizational changes in processor design have primarily been focused on exploiting Instruction Level Parallelism, so that more work is done in each clock cycle.

- **Pipelining:** Individual instructions are executed through a pipeline of stages so that while one instruction is executing in one stage of the pipeline, another instruction is executing in another stage of the pipeline.

- **Superscalar:** Multiple pipelines are constructed by replicating execution resources. This enables parallel execution of instructions in parallel pipelines, so long as hazards are avoided.

- **Simultaneous multithreading (SMT):** Register banks are replicated so that multiple threads can share the use of pipeline resources.

Hardware Performance Issues : Increase in parallelism

- With each of these innovations, designers have over the years attempted to increase the performance of the system by adding complexity.
- In the case of pipelining, simple three-stage pipelines were replaced by pipelines with five stages. Intel's Pentium 4 "Prescott" core had 31 stages for some instructions.
- There is a ***practical limit*** to how far this trend can be taken, because with more stages, there is the need for more logic, more interconnections, and more control signals

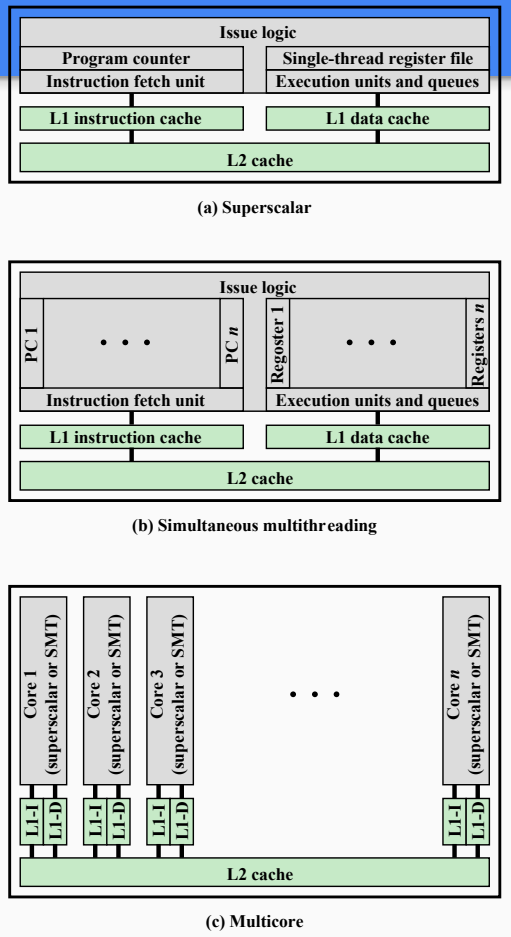


Figure 10.1

Hardware Performance Issues: Alternate Chip Organization

- With each of these innovations, designers have over the years attempted to increase the performance of the system by adding complexity.
- In the case of pipelining, simple three-stage pipelines were replaced by pipelines with five stages. Intel's Pentium 4 "Prescott" core had 31 stages for some instructions.
- There is a ***practical limit*** to how far this trend can be taken, because with more stages, there is the need for more logic, more interconnections, and more control signals

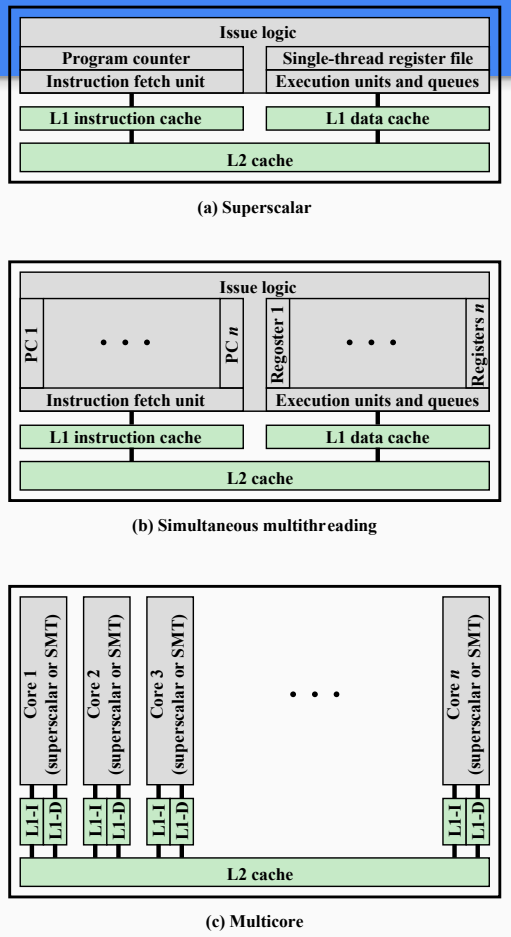


Figure 10.1 Alternate Chip Organization

Hardware Performance Issues

- With superscalar organization, increased performance can be achieved by increasing the number of parallel pipelines. Again, there are diminishing returns as the number of pipelines increases. More logic is required to manage hazards and to stage instruction resources. Eventually, a single thread of execution reaches the point where hazards and resource dependencies prevent the full use of the multiple pipelines available. Also, compiled binary code rarely exposes enough ILP to take advantage of more than about six parallel pipelines.

- This same point of diminishing returns is reached with SMT, as the complexity of managing multiple threads over a set of pipelines limits the ***number of threads and number of pipelines that can be effectively utilized.*** SMT's advantage lies in the fact that two (or more) program streams can be searched for available ILP.

Hardware Performance Issues: Power Consumption

- To maintain the trend of higher performance as the number of transistors per chip rises, designers have resorted to more elaborate processor designs (pipelining, superscalar, SMT) and to high clock frequencies.
- Unfortunately, power requirements have grown exponentially as chip density and clock frequency have risen. This was shown in Figure 10.2.

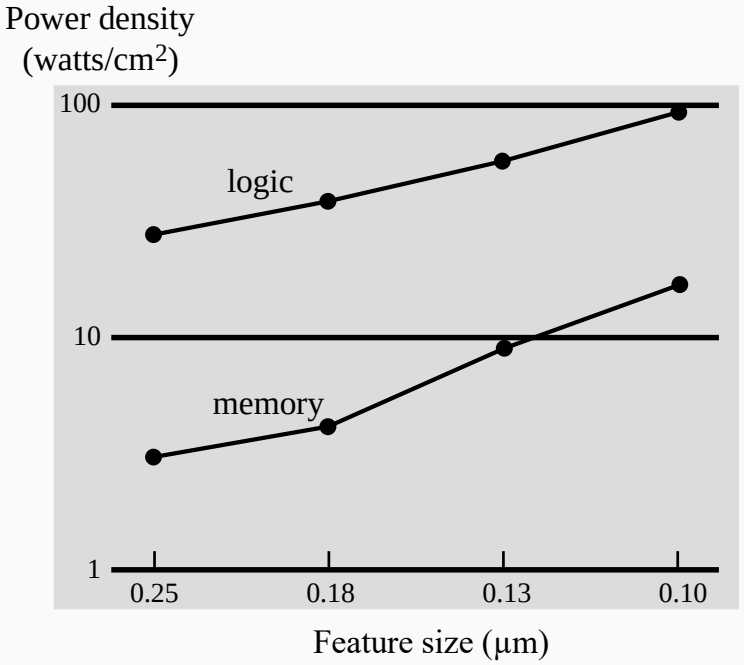


Figure 10.2 Power & memory scenario

Hardware Performance Issues: Power Consumption

- One way to control power density is ***to use more of the chip area for cache memory***.
- Memory transistors are smaller and have a power density an order of magnitude lower than that of logic.
- As chip transistor density has increased, the percentage of chip area devoted to memory has grown, and is now often half the chip area. Even so, there is still a considerable amount of chip area devoted to processing logic.

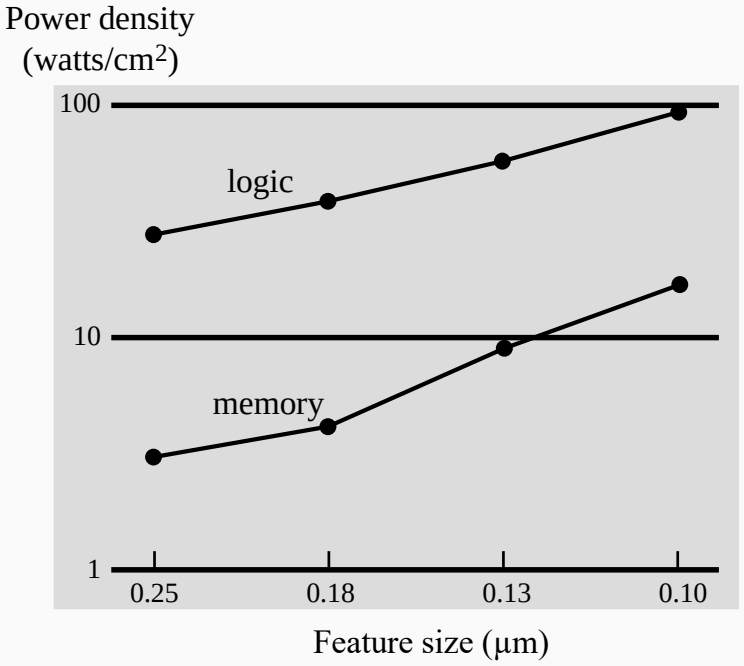


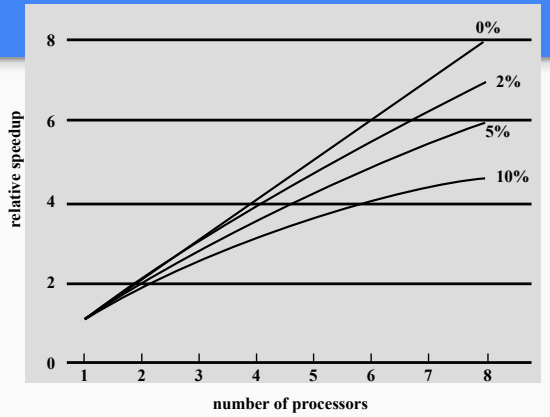
Figure 10.2 Power & memory scenario

Hardware Performance Issues: Power Consumption

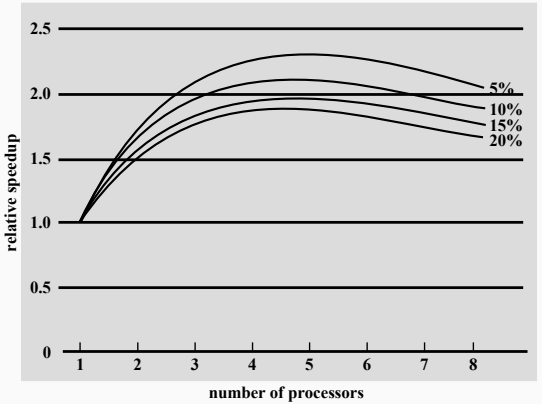
- Power consumption trend is increasing in every year. By 2015 we can expect processors chips with 100 billion transistors on a 30mm2 die.
- Assuming 50-60% of the chip devotes to memory supporting 100 MB of cache and leave about over 1 billion transistors available for logic.
- Pollack's rule states that performance is roughly proportional to square root of increase in complexity i.e. if we double the logic in a processor core, then it delivers only 40% more performance.
- Multicore has potential for near-linear improvement with the increase in number of cores. Power consideration provides another motivation for multicore organization as the chip has such a huge amount of cache memory.

Software Performance Issues: Software on multicore

- The law assumes a program in which a fraction $(1 - f)$ of the execution time involves code that is inherently serial and a fraction f that involves code that is infinitely parallelizable with no scheduling overhead.
- This law appears to make the prospect of a multicore organization attractive. But as Figure shows, even a small amount of serial code has a noticeable impact.
- If only 10% of the code is inherently serial (f = 0.9), running the program on a multi- core system with 8 processors yields a performance gain of only a factor of 4.7.



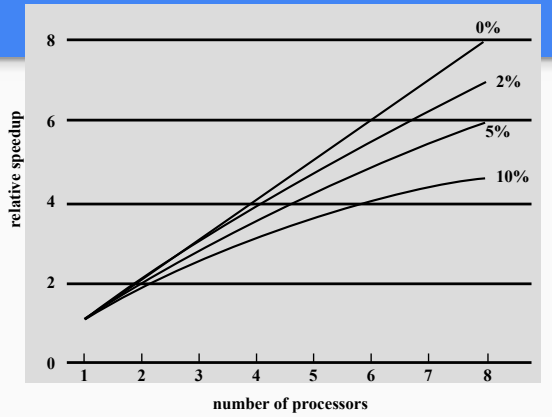
(a) Speedup with 0%, 2%, 5%, and 10% sequential portions



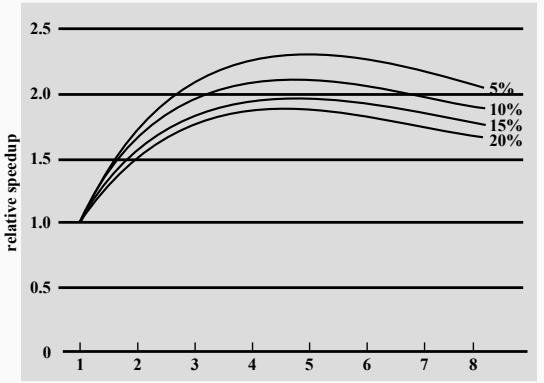
(b) Speedup with overheads

Software Performance Issues: Software on multicore

- In addition, software typically incurs overhead as a result of communication and distribution of work among multiple processors and as a result of cache coherence overhead.
- This results in a curve where performance peaks and then begins to degrade because of the increased burden of the overhead of using multiple processors (e.g., coordination and OS management
- Great attention was paid to reducing the serial fraction within *hardware architectures, operating systems, middleware, and the database application software*.



(a) Speedup with 0%, 2%, 5%, and 10% sequential portions



(b) Speedup with overheads

Dual Core Vs Quad Core Processors

Dual Core	Quad Core
CPU having two core processors.	CPU having four core processors.
Not as fast in terms of speed.	Considered to perform quicker.
Consume less power/ energy since they have only two core processors.	Consume more power/energy since they have four core processors.
Cannot perform many tasks simultaneously.	Can perform many tasks simultaneously.
Do not heat the devices.	It may heat up the devices.
Dual-core processors lack a good graphic.	Quad-core processors have superior and high-quality graphics.

Power efficient processor

- Power efficient processors are more energy-efficient and use less power. That means they can keep going as long as you do on the road—and when they are used in servers, they can help you reduce the space and energy needed to protect and store your data.
- Power efficient processor delivers efficient intelligence into smaller spaces at the network edge.
- Used in a variety of light scale-out workloads that require

✓ very low power

✓ high density

✓ high I/O integration including network routers, switches, storage, security appliances, dynamic web serving, and more.

Example: Intel Atom Processor.

THANK YOU

Any Queries ?

ankitbhattarai@cosmoscollege.edu.np