UNIT I

1. What is called Instruction Level Parallelism?

The technique which is used to overlap the execution of instructions and improve performance is called instruction level parallelism.

2. What are the approaches to exploit ILP?

The two separable approaches to exploit ILP are

- Dynamic, hardware intensive approach
- Static, compiler intensive approach

3. What is pipelining?

Pipelining is an implementation technique whereby multiple instructions are overlapped in execution when they are independent of one another.

4. Write down the formula to calculate the pipeline CPI?

The value of the CPI (Cycles Per Instruction) for the pipelined processor is the sum of the base CPI and all contributions from stalls.

Pipeline CPI = Ideal pipeline CPI
+ Structural stalls
+ Data hazard stalls
+ Control stalls

5. What is called Loop – Level parallelism?

Loop – level parallelism is a way to increase the amount of parallelism available among instructions (ILP) is to exploit parallelism among iterations of a loop.

6. List out the types of dependences.

They are three different types of dependences:

1. Data dependences
2. Name dependences
3. Control dependences

7. When is an instructions said to be dependent?

If two instructions are data dependent, they cannot execute simultaneously or be completely overlapped.

8. Brief on Data dependence?

Data dependence is also called true data dependences. An instruction *j* is data dependent on instruction *i* if either of the following holds.

- Instruction *i* produces a result that may be used by instruction *j*.
    - Instruction *i*
    - Instruction *j*

    or
- Instruction *j* is data dependent on instruction *k*, and instruction *k* is data dependent on instruction *i*
    - Instruction *i*
    - Instruction *k*
    - Instruction *j*

9. What is data hazard?

A hazard is created whenever there is a dependence between instructions, and they are close enough that the overlap caused by pipelining, or other reordering of instructions, would change the order of access to the operand involved in the dependence.

10. What is Control Dependences?

A control dependence determines the ordering of an instruction, *i*, with respect to a branch instruction so that the instruction *i* is executed in correct program order and only when it should be.

11. What are the properties used for preserving Control dependence?

Control dependence is preserved by two properties in a simple pipeline

1. Instructions execute in program order
2. Detection of control or branch hazards.

12. Define dynamic scheduling

Dynamic scheduling is a technique in which the hardware rearranges the instruction execution to reduce the stalls while maintaining data flow and exception behavior.

13. List the advantages of dynamic scheduling?

- It enables handling some cases when dependences are unknown at compile time.
- It simplifies the compiler

- Allows code that was compiled with one pipeline in mind to run efficiently on a different pipeline.
- Uses speculation technique to improve the performance.

14. Why does the imprecise exception occur?

Imprecise exceptions can occur because of two possibilities.

1. The pipeline may have *already completed* instructions that are *later* in program order than the instruction causing the exception.
2. The pipeline may have *not yet completed* some instructions that are *earlier* in program order than the instruction causing the exception

15. List out the fields in each ROB entry?

Each entry in the ROB contains four fields.

1. The instruction type
2. The destination field
3. The value field
4. The ready field

16. What is Loop Unrolling?

A simple scheme for impressing the number of instructions relative to the branch and overhead instructions is *loop controlling.*

17. What are the benefits of Speculating through multiple branches?

The three different situations that can benefit from speculating on multiple branched simultaneously.

1. Very high branch frequency
2. Significant clustering of branches
3. Long delays in functional units.

18. Why do we need branch prediction?

- Increase the number of instructions available for the scheduler to issued. Increases instruction level parallelism (ILP)
- Allows useful work to be completed while waiting for the branch to resolve.

19. What are the basic ideas Pipeline Scheduling?

- To keep pipeline full – Find sequences of unrelated instructions that can be overlapped in the pipeline.
- To avoid pipeline stall – Separate dependent instructions by a distance in clock cycles equal to the pipeline latency of that source instructions.

20. What are the strategies of Branch Predictor?

- Static branch predictors
- Dynamic branch predictors.

UNIT II

1. Define VLIW?

VLIW is a technique for instruction-level parallelism by executing instructions without dependencies (known at compile-time) in parallel, the compiler analysis the program and detects operations to be executed in parallel; such operations are packed into one" large" instruction.

2. What are the responsibilities of VLIW compiler?

The responsibilities of VLIW Compiler are

- Schedules to maximize parallel execution.
- Guarantees intra-instruction parallelism
- Schedules to avoid data hazards (no interlocks)
    - Typically separates operations with explicit NOPs

3. List out the advantages of VLIW Processors

- Simple hardware
    - Number of FUs can be increased without needing additional sophisticated hardware to detect parallelism, like in superscalar's
- Good compilers can detect parallelism based on global analysis of the whole program (no window of execution problem)

4. Define EPIC.

Explicit Parallel Instruction Computing, an architecture frame work proposed by HP. The EPIC architecture is based on VLIW, but was designed to overcome the key limitations of VLIW (in particular, hardware dependence) while simultaneously giving more flexibility to compiler writers.

5. Give the general format of an EPIC bundle.

The general format of an EPIC bundle is shown below.

Tmpl	**Instruction 1**	**Instruction 2**	**Instruction 3**
5	41	41	41
EPIC Bundle			

6. What is called Loop- carried dependence?

Data dependence between different loop iterations (data produced in earlier iteration used in a later one) is called a Loop – carried dependence.

7. When is Loop-level said to be parallel?

A loop is parallel if it can be written without a cycle in the dependences, since the absence of a cycle means that the dependences give a partial ordering on the statements.

8. What is called a recurrence?

When a variable is defined based on value of that variable in an earlier iteration often the one immediately preceding is called as a recurrence.

9. Define dependence analysis algorithm?

Dependence analysis algorithm is algorithm used to detect the dependence by the compiler based on the assumptions that :

- Array indices are affine and
- There exist Greatest Common Divisor (GCD) of the two affine indices

10. When an array index is said to be index?

An array idex is affine if it can be written in the form of an expression

$$axi + b$$

Here, a and b are constant; and

i is the loop index variable

11. What are the components of software pipeline loop?

A software pipeline consists of a loop body, start – up code and clean – up code.

- Additional start-up code to execute code left out from the first original loop iterations.
- Additional finish-code to execute instructions left out from the last original loop iterations.

12. What is Trace scheduling?

Trace scheduling Trace scheduling is a way to organize the process of global code motion it simplifies instruction scheduling by incurring the costs of possible code motion on the less critical paths.

13. Give the limitation of Predicated Instructions?

Annulled instructions take processor resources

- Predicated instructions are not efficient for multiple branches
- Implementing conditional/ predicated instructions has some hardware cost

14. Define IA -64 Processor?

The IA-64 is a RISC-style, register-register instruction set with the features designed to support compiler-based exploitation of ILP.

15. What are the components of IA-64 Register Model?

The components of IA-64 Register Model are

- 128 64-bit General – Purpose Registers
- 128 82-bit Floating-Point Registers, which provide two extra exponent bits over the standard 80-bit IEEE format.
- 64 1- bit predicate registers
- 8 64- bit branch registers, which are used for indirect branches.
- A variety of registers used for system control, memory mapping, performance counters, and communication with the OS

16. What are the parts of CFM pointer?

CFM pointer consists of two parts

1. Local Area
   - Used for local storage
2. Output Area
   - Used to pass values to any called procedure.

17. What are the types of registers and its purpose?

The different types of registers are listed below

1. *Integer Registers*- Used to hold integer data
2. *Floating Point Registers*-Used to hold floating point data
3. *Predicate registers* – Used to hold predicates which control the execution of predicated instruction.
4. *Branch registers*- Used to hold branch destination addresses for indirect branches.

18. What are the benefits of register rotation?

- Makes it easy to allocate registers in software pipelined loops.
- When combined with predication, it can reduce the code expansion incurred by using software pipelining.
- Makes this technique usable for loops with small number of iterations.

19. What is Register Stack Mechanism?

Register stack mechanism is a technique used by integer registers to accelerate procedure calls.

20. What is the use of EFM pointer?

CFM pointer points to the set of registers to be used by a given procedure.

UNIT III

1. What is Parallel Computers?

A parallel computer is a collection of processing elements that cooperate and communicate to evolve large problems fast.

2. List out the categories of Flynn's Taxonomy of Parallel Machines

1. Single instruction stream, single data stream (SISD)
2. Single instruction stream, multiple data stream (SIMD)
3. Multiple instruction stream, Single data stream, (MISD)
4. Multiple instruction stream, Multiple data stream, (MIMD)

3. Define SMP?

Symmetric (shared – memory) multiprocessor (SMP) is a multiprocessor with a single main memory that has a symmetric relationship to all processors.

4. What is Distributed –memory multiprocessor?

Distributed-memory multiprocessor consists of multiprocessors with physically distributed memory. To support larger processor counts, Memory must be distributed among the processors rather than centralized.

5. List out the components of distributed-memory multiprocessor

The basic architecture of a distributed-memory multiprocessor consists of

- Individual nodes containing a processor
- Memory
- I/O
- An interface to an interconnection network that connects all the nodes.

6. What are the benefits of distributed – memory multiprocessor?

Distributing the memory among the nodes has two major benefits.

- It is a cost-effective way to scale the memory bandwidth
- It reduces the latency for access to the local memory

7. What are the advantages of distributed-memory multiprocessor?

- Cost effective way to scale memory bandwidth
- Lower memory latency for local memory access

8. What are the drawbacks of distributed memory multiprocessor?

- Longer communication latency for communicating data between processors
- Software model more complex

9. What is Distributed Shared Memory?

Distributed Shared – Memory (DSM) architecture is a multiprocessor with a shared address space in which communication occurs through a shared address space.

10. What is Message Passing Multiprocessors?

Message passing multiprocessor is a multiprocessor with multiple address space in which communication of data occurs by especially passing messages among the processors.

11. What is multiple address space?

Multiple address space is address space which has the same physical address on two different [processors refers to two different locations in two different memories.

12. Define Synchronous Message Passing?

Synchronous Message Passing is a way in which the initiating processors sends a request and waits until the reply is returned before continuing.

- Send the message that request action or deliver data as Remote Procedure Call (RPC)
- When the destination processor receives the message, it performs the operation or access on behalf of the remote processor.
- Return the result with a reply message.

13. Define Asynchronous Message Passing?

Asynchronous Message Passing is a way in which the initiating processor sends the request message continuously without waiting for the reply message.

14. List out the performance Metrics for Communication Mechanisms?

The three performance metrics are critical in any communication mechanism:

1. Communication bandwidth
2. Communication latency

3. Communication latency hiding

15. Write down the formula to calculate Communication latency?

Communication latency = Sender overhead+
Time of flight+
Transmission time+
Receiver overhead.

16. State Amdahl's Law?

State Amdahl's Law is used to measure the performance of the multiprocessor.

$$Speedup_{overall} = \frac{1}{(1 - Fraction_{enhanced}) + \frac{fraction_{enhanced}}{speedup_{enhanced}}}$$

17. List out the challenges of Parallel Processing?

Two important hurdles which make parallel processing challenging are

1. Limited parallelism available in programs
   a. The problem of inadequate application parallelism must be attacked primarily in software with new algorithms that can have better parallel performance.
2. Large latency of remote access in a parallel processor.
   a. Reducing the impact of long remote latency can be aatacked both by the architecture and by the programmer.

18. What is Symmetric shared – memory Architecture?

Symmetric Shared – Memory Architecture is a small-scale multiprocessor where several processors shared a single physical memory connected by a shared bus. It is called so because each processor has the same relationship to one single shared memory.

19. Define Coherence?

Coherence defines the behavior of reads and writes to the same memory location. Coherence defines what values can be returned by a read

20. Define consistency?

Consistency defines the behavior of reads and writes with respect to accesses to other memory location. It determines when a written value will be returned by a read.

UNIT IV

1. What is Snooping?

Every cache with copy of data also has copy of sharing status of block, but no centralized state is kept

2. What are Spin locks?

Spin locks are the locks that a processor continuously tries to acquire, spinning around a loop until it succeeds.

3. What is Memory Hierarchy?

A memory hierarchy is organized into several levels in which each level is smaller, faster and more expensive per byte than the next lower level.

4. What is the Goal of memory Hierarchy?

The Goal is to provide a memory system with

- Cost almost as low as the cheapest level of memory and
- Speed almost as fast as the fastest level.

5. Define Cache?

Cache is the name given to the first level of the memory hierarchy encountered once the address leaves the CPU.

Examples

- File caches
- Name caches

6. Define Cache bit?

When the CPU find a requested data item in the cache, it is called a cache bit.

- *Hit Rate:* the fraction of cache access found in the cache
- *Hit time:* Time to access the upper level which consists of RAM access time + Time to determine hit/miss

7. Define Cache Miss?

When the CPU does not find a data item it needs in the cache, a cache miss occurs.

- Miss Rate = 1-(Hit Rate)
- Miss penalty: Time to replace a block in cache+ Time to deliver the block to the processor

8. What are the factors on which the cache miss depends on?

The time required for the cache miss depends on both

- Latency
- Bandwidth

9. What is called Principle of Locality?

Program access a relatively small portion of the address space at any instant of time is called Principle of Locality.

10. What is called pages?

The address space is usually broken into fixed-size blocks, called pages. Each page resides either in main memory or on disk.

11. How does page fault occur?

When the CPU references an item within a page that is not present in the cache or main memory, a page fault occurs, and the entire page is moved from the disk to main memory.

12. What is called Memory stall cycles?

The number of cycles during which the CPU is stalled waiting for a memory access is called *memory stall cycles.*

13. What is called the miss penalty?

The number of memory stall cycles depends on both the number of misses and the cost per miss, which is called the *miss penalty.*

14. Write down the formula for calculating Average memory access time?

Average memory access time =Hit time +Miss rate x Miss Penalty

Where Hit time is the time to hit in the cache. This formula can help us decide between split caches and a unified cache

15. What are the techniques to reduce the miss penalty?

- Multi-Level Caches
- Critical Word First and Early Restart
- Giving Priority to Read Misses over Writes
- Merging Write Buffer
- Victim Caches

16. What are the techniques to reduce the miss rate?

- Larger block size
- Larger caches
- Higher associatively
- Way prediction and pseudo associative caches
- Compiler optimizations

17. What are the techniques to reduce hit time?

The four techniques to reduce the hit time are

1. Small and simple cache: direct mapped
2. Avoid address translation during indexing of the cache
3. Pipelined cache access
4. Trace cache

18. What is Sector?

Each track in turn is divided into sectors that contain the information. A sector is the smallest unit that can be read or written.

19. What is called Seek?

To read or write sector, the disk controller sends a command to move the arm over the proper track. This operation is called *seek.*

20. What is called Seek Time?

The time to move the arm to desired track is called seek time.

UNIT V

1. What is Software multithreading?

Software multithreading is a piece of software that is aware of more than one core/processor, and can use these to be able to simultaneously complete multiple tasks.

2. What is Hardware multithreading?

Hardware multithreading is a multithreading that allow multiple threads to share the functional units of a single processor in an overlapping fashion.

3. Difference between Software and Hardware multithreading?

- Multithreading (computer architecture, multithreading in hardware
- Thread (Computer science), multithreading in software

4. List out the approaches in Hardware Multithreading?

The two main approaches in hardware multithreading are

- Fine-grain multithreading
- Course-grain multithreading

5. Define Simultaneous Multithreading (SMT)?

SMT is a variation on multithreading that uses resources of a multiple-issue, dynamically scheduled processors to exploit TLP at the same time it exploits ILP ie., convert thread level parallelism into more ILP.

6. Compare the SMT processor with the base superscalar processor?

The SMT processor are compared to the base superscalar processor in several key measures

- Utilization of functional units
- Utilization of fetch units
- Accuracy of branch predictor
- Hit rates of primary caches
- Hit rates of secondary caches

7. List the factor that limits the issue slot usage?

The issue slot usage is limited by the following factors

- Imbalance in the resource needs
- Resource availability over multiple threads.
- Number of active threads considered
- Finite limitations of buffer
- Ability of fetch enough instruction from multiple threads
- Practical limitations of what instructions combination can issue from one thread and multiple threads.

8. How the Performance of SMT are improved?

The key to maximize SMT performance is to share the following

- Issue slots
- Functional Units
- Renaming registers

9. What is CMP?

Chip-level multiprocessing (CMP or multicore): integrates two or more processors into one chip, each executing threads independently

- Every functional unit of a processor is duplicated.

10. What is Chip Multithreading?

Chip multithreading is the capability of a processor to process multiple software threads and supports simultaneous hardware threads of execution.

Chip Multithreading= Chip Multi processing+ Hardware Multithreading

11. Define Multi-core Microprocessor?

A multi-core microprocessor is one that combines two or more separate processors in one package, often a single chip.

12. What are the performance measure of the Intel micro architecture?

- Dynamic scalability
- Design and performance scalability
- Intelligent performance on-demand
- Increased performance on highly-threaded applications

- Scalable shared memory
- Multi-level shared cache.

13. What is Heterogeneous Multi-core Processors?

Heterogeneous Multi-Core Processor is a processor in which multiple cores of different types are implemented in one CPU.

14. List out the advantage of Heterogeneous Multi-Core Processors?

- Massive parallelism today
- Specialization of hardware for different tasks.

15. List out the disadvantage of Heterogeneous Multi-Core Processors?

- Developer productivity
- Portability
- Manageability

16. What is IBM cell processor?

The IBM cell processor is a heterogeneous multi-core processor comprised of control-intensive processor and compute- intensive SIMD processor cores, each with its own distinguishing features.

17. List the components used in IBM cell Architecture?

- Power Processing Elements (PPE)
- Synergistic Processor Elements(SPE)
- I/O Controller
- Element Interconnect Bus (EIB)

18. What are the Components of SPE?

The SPE is made out of two main units

1. Synergistic Processor Unit (SPU)
2. Memory Flow Controller (MFC)

19. What is function of Synergistic Processor Unit (SPU)?

The Synergistic Processor Unit (SPU) deals with instruction control and execution. It includes various components:

- A register file of 128 registers of 128 bits each
- A unified instruction and data 256- KB Local Store(LS)
- A channel- and- DMA interface.
- As usual, an instruction-control unit, a load and store unit, two fixed- point units, a floating-point unit.

20. What is Memory Flow Controller (MFC)?

The Memory Flow Controller (MFC) is actually the interface between the SPU and the rest of CELL chip. Actually, the MFC interfaces the SPU with the EIB.