CS2354 ACA-2 MARKS & 16 MARKS

UNIT-I

PART-A

1. Explain the concept of pipelining.

Pipelining is an implementation technique whereby multiple instructions are overlapped in execution. It takes advantage of parallelism that exists among actions needed to execute an instruction.

2. What is a Hazard? Mention the different hazards in pipeline.

Hazards are situations that prevent the next instruction in the instruction stream from executing during its designated clock cycle. Hazards reduce the overall performance from the ideal speedup gained by pipelining. The three classes of hazards are,

i)   Structural hazard
ii)   Data hazard
iii)   Control hazard

3. List the various dependences.

Data dependence
Name dependence
Control dependence

4. What is Instruction Level Parallelism? (or) What is ILP?

Pipelining is used to overlap the execution of instructions and improve performance. This potential overlap among instructions is called instruction level parallelism (ILP) since instruction can be evaluated parallel.

5. Give an example of control dependence.

If p1 { s1;}
If p2 {s2;}

S1 is control dependence on p1, and s2 is control dependence on p2.

6. Write the concept behind using reservation station.

Reservation station fetches and buffers an operand as soon as available, eliminating the need to get the operand from a register.

7. Explain the idea behind dynamic scheduling? Also write advantages of dynamic scheduling.

In dynamic scheduling the hardware rearranges the instruction to reduce the stalls while maintaining dataflow and exception behavior.

Advantages: it enables handling some cases when dependence are unknown at compile time and it simplifies the compiler. It allows code that was compiled with one pipeline in mind and run efficiently on a different pipeline.

8. What are the possibilities of imprecise exception?

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

9. What are branch target buffers?

To reduce the branch penalty we need to know from what address to fetch by end of Instruction fetch. A branch prediction cache that stores the predicted address for the next instruction after a branch is called a branch-target buffer or branch target cache.

10. Mention the idea behind hardware-based speculation?

It combines three key ideas:

Dynamic branch prediction to choose which instruction to execute,

Speculation to allow the execution of instructions before control dependence are resolved

Dynamic scheduling to deal with the scheduling of different combinations of basic blocks.

11. What are the fields in the ROB?

Instruction type, destination field, value field and ready field.

12. Mention the advantage of using tournament based predictors?

The advantage of tournament predictor is its ability to select the right predictor for right branch.

13. What is loop unrolling?

A simple scheme for increasing the number of instructions relative to the branch and overhead instructions is loop unrolling. Unrolling simply replicates the loop body multiple times, adjusting the loop termination code.

14. What are the basic compiler techniques?

Loop unrolling, Code scheduling.

15. Give an example for data dependence.

```
Loop: L.D       F0, 0(R1)
      ADD.D   F4, F0, F2
      S.D       F4, 0(R1)
      DADDUI  R1, R1, #-8
      BNE       R1, R2, LOOP
```

PART B

1. Explain the data dependences and hazards in detail with examples. (16)

Data Dependences and Hazards

CS2354 ACA-2 MARKS & 16 MARKS

Name dependencies and Hazards
Control dependencies and Hazards

2.**Briefly explain the concept of dynamic scheduling. (8)**

Dynamic scheduling

3.**Explain dynamic scheduling using tomosulo's approach with an example. (or)**

Explain technique used for overcoming the data hazards with an example.

-Diagram

-Use of reservation stations

Components of tomosulo's architecture:

4.**Explain in detail about the hardware based speculation and explain how it**

overcomes the control dependencies. (16)

-**Hardware-based speculation combines three key ideas**:

-**Role of re-order buffer**

Issue

Execute

Write result

Commit

5.**Explain the basic compiler techniques for exposing ilp. (8)**

Basic Pipeline Scheduling and Loop Unrolling

Loop unrolling

6**. Explain static and dynamic branch prediction schemes (16)**

Static Branch Prediction

Correlating Branch Predictors

Tournament Predictors: Adaptively Combining Local and Global Predictors

CS2354 ACA-2 MARKS & 16 MARKS

UNIT-II

PART-A

1. Explain the VLIW approach?

They uses multiple, independent functional units. Rather than attempting to issue multiple, independent instructions to the units, a VLIW packages the multiple operations into one very long instruction.

2. Mention the advantage of using multiple issue processors?

They are less expensive, they have cache based memory system, more parallelism.

3. What are loop carried dependence?

They focus on determining whether data accesses in later iterations are dependent on data values produced in earlier iterations; such a dependence is called loop carried dependence. E.g for(i=1000;i>0;i=i-1)

$$X[i] = x[i] +s;$$

4. Use the G.C.D test to determine whether dependence exists in the following loop.

for(i=1;i<=100;i=i+1)
X[2*i+3]=X[2*i]*5,0;
Solution: a=2, b=3, c=2, d=0
GCD(a,c)=2 and d-b=-3
Since 2 does not divide -3, no dependence is possible.

5. What is software pipelining?

Software pipelining is a technique for reorganizing loops that each iteration in the software pipelined code is made from instruction chosen from different iterations of the original loop.

6. What is global code scheduling?

Global code scheduling aims to compact code fragment with internal control structure into the shortest possible sequence that preserves the data and control dependence. Finding a shortest possible sequence is finding the shortest sequence for the critical path.

7. What is Trace?

Trace selection tries to find a likely sequence of basic blocks whose operations will be put together into smaller number of instructions, this sequence is called trace.

8. What are the steps involved in Trace scheduling?

Trace selection, Trace compaction, Book Keeping.

9. What is superblock?

Superblocks are formed by a process similar to that used for traces, but are form of extended basic block, which are restricted to a single entry point but allow multiple exits.

10. What is poison bit?

Poison bits are a set of status bits that are attached to the result registers written by the specified instruction when the instruction causes exceptions. The poison bits cause a fault then a normal instruction attempts to use the register.

11. What are the disadvantages of supporting speculation in hardware?

Complexity, additional hardware required.

12. What is an instruction group?

It is a sequence of consecutive instructions with no register data dependences among them. All the instructions in the group could be executed in parallel. An instruction group can be arbitrarily long.

13. What are the limitations of ILP?

The hardware model, limitation on the window size and maximum issue count, the effects of realistic branch and jump prediction, the effects of finite registers, the effect of imperfect alias analysis.

14. What is copy propagation?

Within a basic block, algebraic simplification of expressions and an optimization called copy propagation. Which eliminates operations that copy values, can be used to simplify sequences like the following:

DADDUI R1, R2, #4

DADDUI R1, R1, #4

to DADDUI R1, R2, #8

15. What is CFM?

CS2354 ACA-2 MARKS & 16 MARKS

A special register called the current frame pointer points to the set of registers to be used by a given procedure.

PART B

1. **Explain the VLIW and EPIC processors. (8)**

   ILP in VLIW

   EPIC ARCHITECTURE

   Instruction-level parallelism in EPIC

   Static Multiple Issue : The VLIW Approach

   The Basic VLIW Approach

2. **Explain hardware support for compilers for exposing ilp. (16)**

   Conditional or Predicated Instructions

   Compiler Speculation with Hardware Support

   Hardware Support for Preserving Exception Behavior

   Hardware Support for Memory Reference Speculation

3. **Hardware versus software speculation mechanisms. (8)**

4. **Explain the ia 64 architecture. (10)**

   Specification

   Memory Address Modes

   Branching

   Five execution unit slots

   Register mechanism

   Predication

5.Advanced compiler support for exposing and exploiting ilp (16).

- **Detecting and Enhancing Loop-Level Parallelism**

- **Identifying dependencies**

- **Copy propagation**

- **Software pipelining**

- **Global code scheduling**

- **Trace scheduling**

6.Briefly explain global code scheduling(6)

7.Explain software pipelining (6)

UNIT-III
PART-A

1. **What are multiprocessors? Mention the categories of multiprocessors?**
   Multiprocessors are used to increase performance and improve availability. The different categories are
   SISD(Single Instruction and Single Data stream),
   SIMD(Single Instruction and Multiple Data stream),
   MISD(Multiple Instruction and Single data stream)
   MIMD(Multiple Instruction and Multiple Data stream)

2. **What are threads?**
   These are multiple processors executing a single program and sharing the code and most of their address space. When multiple processors share code and data in the way, they are often called threads.

3. **What is cache coherence problem?**
   Two different processors have two different values for the same location.

4. **What are the ways to maintain cache coherence? OR what are the ways to enforce cache coherence?**
   Directory based protocol, Snooping based protocol.

5. **What are the ways to maintain cache coherence using snooping protocol?**
   Write invalidate protocol, write update or write broadcast protocol.

6. **What is write invalidate and write update protocol?**
   - Write invalidate provide exclusive access to caches. These exclusive caches ensure that no other readable or writable copies of an item exist when the write occurs.
   - Write updates protocol updates all cached copies of a data item when that item is written.

7. **What are the disadvantages of using symmetric shared memor?**
   Compiler mechanisms are very limited, and create large latency for remote memory access fetching multiple words in a single cache block will increase the cost.

8. **Mention the information available in the directory?**
   The directory keeps the state of each block that is cached. It keeps track of which caches have copies of the block.

9. **What are the states of cache block in directory based approach?**
   Shared, Un-cached and Exclusive.

10. **What are the uses of having a bit vector?**
    When a block is shared, the bit vector indicates whether the processor has the copy of the block. When block is in exclusive state, bit vector keep track of the owner of the block.

11. **When do we say that a cache block is exclusive?**
    When exactly one processor has the copy of the cache block, and it has written the block, then the processor is called the owner of the block. During this phase the block will be in exclusive state.

12. **Explain the types of messages that can be send between the processors and directories?**

    Local Node: node where the requests originates
    Home Node: Node where memory location and directory entry of the address resides.
    Remote Note: the copy of the block in the third node called remote node.

13. **What is consistency? And what are the models used for consistency?**
    Consistency says in what order processor must observe the data writes of another processor.
    Models used for Consistency:
    Sequential Consistency Model,
    Relaxed Consistency Model

14. **What is sequential consistency?**

Sequential consistency requires that the result of any execution be the same, as if the memory accesses executed by each processor were kept in order and the accesses among different processors were interleaved.

15. What is relaxed consistency model?
Relaxed consistency model allows reads and writes to be executed out of order.
Three sets of ordering are:
W->R Ordering – Total store ordering
W->W Ordering – Partial ordering
R->W and R->R Ordering – Weak ordering

16. What is coarse grained and fine grained multithreading?
Coarse Grained: it switches only on costly stalls (or) event. Thus it is much less likely to slow down the execution of an individual thread.
Fine Grained: it switches threads between threads on each instruction, causing the execution of multiple threads to be interleaved.

PART B

1. **Explain the symmetric shared memory architecture. Explain the snooping based protocols with neat diagram. (16)**
   Symmetric Shared Memory Architectures:
   Cache Coherence in Multiprocessors:
   Basic Schemes for Enforcing Coherence
   Snooping Protocols
   Write invalidate
   Write update protocol
   State transition diagram

2. **Explain the concept of distributed shared memory and also explain directory based protocols with an example. (or) Explain the numa architecture with neat diagram.**

   Directory-Based Cache-Coherence Protocols: The Basics

3. **Explain the basics of multithreading and it types. (8) (or) Explain how multihreading approach can be used to exploit thread level parallelism within a processor? (8)**

   Fine-grained multithreading
   Coarse-grained multithreading
   Simultaneous Multithreading
   Design Challenges in SMT processors

4. **Explain in detail about the need for consistency models and it types.(8)**

    Sequential consistency

    Relaxed Consistency Models:

5. **Explain the performance of the symmetric shared memory with necessary graphs. (16)**

    Distribution of Execution Time in the Multi-programmed Parallel Make Workload Data Miss Rate VS. Data Cache Size

UNIT-IV

PART-A

1. **What is cache miss and cache hit?**

Cache Miss: When the CPU finds a requested data item in the cache, it is called cache miss.

Cache Hit: When the CPU finds a requested data item is available in the cache, it is called cache hit.

2. **What is write through and write back cache?**

Write through cache: The information is written to both the block in the cache and to the block in the lower level memory.

Write Back Cache: The information is written only to the block in the cache. The modified cache block is written to main memory only when it is replaced.

3. **What is Miss Rate and Miss Penalty?**

Miss Rate is the fraction of cache access that results in a miss.

Miss Penalty depends on the number of misses and clock per miss.

4. **Write the equation of Average memory access time.**

Average memory access time = hit time+miss rate X Miss Penalty

5. What is stripping?

Spreading multiple data over multiple disks is called stripping, which automatically forces accesses to several disks.

6. What is disk mirroring? And write the drawbacks of disk mirroring.

Disks in the configuration are mirrored or copied to another disk. With this arrangement data on the failed disks can be replaced by reading it from the other mirrored disks.

Drawback: writing onto the disk is slower since the disks are not synchronized, seek time will be different.

It imposes 50% space penalty hence expensive.

7. Mention the factors that measure I/O performance ?

Diversity capacity, response time, throughput, interference with CPU execution.

8. What is transaction time?

The sum of entry time, response time and think time is called transaction time.

9. State little law?

Little law relates the average number of tasks in the system. It relates to Average arrival rate of new tasks with the average time to perform a task.

10. What are the steps to design an I/O system?

- Naïve cost – performance design and evaluation

- Availability of naïve design

- Response time

- Realistic cost performance, design and evaluation

- Realistic design for availability and its evaluation

11. Write the classification of buses.

I/O buses – these buses are lengthy and have any types of devices connected to it.

CPU memory buses – They are short and generally of high speed.

12. What is bus master?

Bus master are devices that can initiate the read or write transaction.

Eg. Processor – processor are always has the bus mastership.

13. Mention the advantages of using bus master.

It offers higher bandwidth by using packets, as opposed to holding the bus for full transaction.

14. What is split transaction?

The idea behind this is to split the bus into request and replies, so that the bus can be used in the time between request and the reply.

15. What are the measures of latency in memory technology?

Access Time:

Is the time between when a read is required and when the desired word arrives.

Cycle Time:

Is the minimum time between requests to memory.

PART B

1. Explain in detail about the vaious optimization techniques for improving the cache performance. (16)

> **Cache performance**
> **There are 17 cache optimizations into four categories**:
> **First Miss Penalty Reduction Technique: Multi-Level Caches**
> **Second Miss Penalty Reduction Technique: Critical Word First and Early Restart**
> **Third Miss Penalty Reduction Technique: Giving Priority to Read Misses over Writes**
> **Fourth Miss Penalty Reduction Technique: Merging Write Buffer**
> **Fifth Miss Penalty Reduction Technique: Victim Caches**
> **First Miss Rate Reduction Technique: Larger Block Size**
> **Second Miss Rate Reduction Technique: Larger caches**
> **Miss Rate Reduction Technique: Higher Associativity**
> **Fourth Miss Rate Reduction Technique: Way Prediction and Pseudo-Associative Caches**
> **Fifth Miss Rate Reduction Technique: Compiler Optimizations**

2. What is virtual memory? write techniques for fast address translation

> **Techniques for Fast Address Translation**
>
> Selecting a Page Size

3. Explain various types of storage devices (16)

> **Types of Storage Devices**

Magnetic Disks

Optical Disks

Magnetic Tape

Automated Tape Libraries

Flash Memory

4. **Explain the buses and i/o devices (16)**

Bus Design Decisions

Bus Standards

Interfacing Storage Devices to the CPU

5.**Briefly explain raid (or) redundant arrays of inexpensive disks (8)**

No Redundancy (RAID 0)

Mirroring (RAID 1)

Bit-Interleaved Parity (RAID 3)

Block-Interleaved Parity and Distributed Block-Interleaved Parity (RAID 4 and RAID 5)

6.**Explain the design of i/o systems and its performance. (16)**

— The CPU

— The memory system:

– Internal and external caches, Main Memory

— The underlying interconnection (buses)

UNIT-V
PART-A

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

Hardware multithreading is a hardware that is aware of more than one core/processors and can use these to be able simultaneously completes multiple threads.

2. **Write short note on SMT.**

Simultaneous multithreading is a technique for improving the overall efficiency of superscalar CPU's with hardware multithreading. SMT permits multiple independent threads of execution to better utilize the resources provided by modern processor architectures.

3. **List the design challenges of SMT.**
   - Larger register file needed to hold multiple contexts.
   - Not affecting clock cycle time, especially its,
     - o Instruction issue: more candidate instructions need to be considered.
     - o Instruction completion: choosing which instructions to commit may be challenging.
   - Ensuring that cache and TLP conflicts generated by SMT do not degrade performance.

4. **Define CMP.**

Chip level multiprocessing(CMP or multi-core): integrates two or more independent cores into a single package composed of a single integrated circuit, called a die, or more dies packaged, each executing threads independently.

5. **Define chip multithreading:**
   - Chip multithreading is defined as the combination of chip multiprocessing and the hardware multithreading.

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

6. **Define multicore processors.**

   A multicore design in which a single physical processor contains the core logic of more than one processor.

7. **What is heterogeneous multicore processor?**

   The architecture consists of a chip-level multi-processors with multiple, processor cores. These cores all execute the same instruction set, but include significantly different resources and achieve different performance and energy efficiency on the same application.

8. **List out the advantages of heterogeneous multicore processors.**

   More efficient adaptation to application diversity.

Applications place different demands on different architectures.

A more efficient use of die area for a given thread-level parallelism.

Chip level multiprocessors with a mix of cores.

9. What is IBM Cell processor?

Cell is a microprocessor architecture jointly developed by sony, sony computer entertainment, Toshiba and IBM, an alliance known as "STI". The architectural design and first implementation were carried out at the STI design center in Austin, Texas over four-year period.

10. List out the components of cell processors.

External input and output structures

The main processor called the Power Processing Element(PPE)

Eight fully functional co-processors called the Synergistic Processing elements or SPE's.

A specialized high bandwidth circular data bus connecting the PPE.

What are the components of SPE?

Synergistic processor unit(SPU)

Memory flow controller (MFC)

11. What is memory flow controller?

The memory flow controller is the interface between the SPU and the rest of the cell chip.

12. State fine grained multi-threading.

Switches between threads on each instructions, causing the execution of multiple threads to be interleaved.

Usually done in a round robin fashion, skipping any stalled threads.

CPU must be able to switch threads every clock.

13. List Advantages of coarse grain multithreading.

It is hard to overcome throughput losses from shorter stalls, due to pipeline start-up costs.

Since CPU, issues instruction from 1 thread, when a stall occurs, the pipeline must be emptied or frozen.

New thread must fill pipeline before instruction can complete.

14. Define coarse grain multithreading.

It switches thread only on costly stalls, such as L2 cache misses.

PART B

1.**Briefly explain multicore architecture with neat diagram and also write its applications. (8) (or) chip multiprocessors (or) cmp architecture.**

> **Single-core computer**
> **CMP architecture:**
>
> **CHIP multi threading:**
>
> **Multi-core architectures**

2.**Explain in detail about the various types of software and hardware based multi-threading. (8)**

> **Software and hardware multithreading**
>
> **Multithreading in Hardware**
> **Hardware Multithreading Techniques**
>
> **Cycle-by-cycle interleaving (Fine Grained Multithreading)**
>
> **Block interleaving (Coarse Grain Multithreading)**

3.**Explain the architectural features of ibm cell processors with neat diagram. (or) explain cell processors. (or) explain cell broadband engine. (10)**

> **IBM Cell Processor**
> **Applications of cell processors**
> **Power Processor Element (PPE)**
> **Synergistic Processing Elements**
> **Input/output interfaces**

4.**Explain sun cmp architecture with neat diagram. (or) explain sun niagara processor architecture. (10)**

5.Explain simultanous multi-threading concept for converting thread level parallelism into instruction level parallelism. (8)

        Design Challenges in SMT
        Potential Performance Advantages from SMT