

PROCESSORS MEMORYHER ARCHYARCHY

4

CONTENTS

Processors and memory hierarchy

- Advanced processor technology
 - Design Space of processors
 - Instruction -set Architecture
 - CISC Scalar Processors
 - RISC Scalar Processors
- Super scalar and vector processons
 - Super Scalar Processors
 - Vector and Symbolic Processors

Memory hierarchy technology.

4

ADVANCED D PROCESSOR R TECHNOLOGY

• Major processor families are:-

- □ CISC
- □ RISC
- □ Superscalar
- □ <u>VLIW</u>
- □ <u>Super pipelined</u>
- □ <u>Vector</u>
- Symbolic processors

- Scalar and vector processors are used for numerical computations
- Symbolic processors are used for AI applications

Design space of processors s



DESIGN N SPACE OF PROCESSORS RS

- **CPI** Vs clock speed of various processors
 - Processors designed for following aims have lower clock speeds
 - Multi-core chips
 - Embedded applications
 - Low cost applications
 - Low power consumption

High performance processors are designed tooperate at higher clock speeds

DESTENN SPACEE

• CISC architecture design space

- Clock rate of CISC processors ranges upto few GHz
- **CPI** of CISC instructions varies from 1 to 20
- Eg: Intel Pentium, M68040

• **RISC** <u>architecture</u> design space

CPI of RISC instruction is reduced between 1 and 2 cycles

 \Box Eg:

- SPARC (Scalable Processor Architecture)
- MIPS (Microprocessor without Interlocked Pipeline Stages)

• Superscalar processor design space

- □ Subclass of RISC processor
- Multiple instructions are issued simultaneously during each cycle
- □ Effective **CPI** is lower than scalar RISC
- Clock rate matches with that of scalar RISC

• Design space of VLIW architecture

- □ Use more functional units than superscalar processor
- **CPI** is further lowered
- □ Eg: Intel i860

• Vector supercomputers design space

- □ Use multiple functional units
- Perform concurrent scalar and vector operations
- **CPI** is very low

INSTRUCTION PIPELINES [1]

• **Execution cycle** of an instruction involves 4 phases

- Fetch
- Decode
- E<u>xecute</u>
- Write back
- **Pipeline** receives successive instructions fromone end
- It executes them in a streamlined overlapped fashion

INSTRUCTION PIPELINES[2]

o Pipeline cycle

It is defined as the time required for each phase to complete its operation assuming equal delay in all phases

o Instruction pipeline cycle

□ It is the clock period of the instruction pipeline

o Instruction issue latency

Time required between the issue of two adjacent instructions

o Instruction issue rate

No: of instructions issued per cycle

INSTRUCTION PIPELINES[3]

o Simple operation latency

- □ No: of cycles needed to perform simple operations
- □ Simple operations are:
 - o Integer adds
 - <u>Loads</u>
 - Stores
 - Branches
 - o moves

o Resource conflicts

It is a situation where 2 or more instructions demand use of the same functional unit at the same time





(a) Execution in a base scalar processor



(b) Underpipelined with two cycles per instruction issue



BASIC C SCALAR R PROCESSOR R

• It is a machine with following features:-

1 <u>instruction issued per cycle</u>

□ 1 cycle latency for simple operation

□ 1 cycle latency between instruction issues

INSTRUCTION PIPELINES [5]

- Another **under-pipelined** situation is in which the pipeline cycle time is doubled by combining pipeline stages.
- In <u>this case, the fetch and decode phases are</u> com<u>bined into one pipeline stage, and execute and</u> write <u>back are combined into another stage.</u>
- This will also result in **poor pipeline** utilization.
- The effective CPI rating is l for the ideal pipeline
- The effective CPI rating is 2 for fig b

Δ.

• The effective CPI rating is one half for fig c

INSTRUCTION ON SET ARCHITECTURES RES

• Instruction set of a computer specifies

- Machine instructions that a programmer can use while programming the machine
- Complexity of an instruction depends on:-
 - □ Instruction format
 - □ Da<u>ta format</u>
 - □ Ad<u>dressing modes</u>
 - □ General purpose registers
 - Opcode specifications
 - □ Flow control mechanisms

TYPES S OF INSTRUCTION SET ARCHITECTURES RES

- o 2 types
 - Complex instruction set
 - Reduced instruction set

COMPLEX INSTRUCTION ON SETS:

- CISC contains 120 to 350 instructions
- Instructions use variable instruction and data formats
- Use <u>small set of 8 to 24 general purpose registers</u> GPR
- Executes large no: of memory reference operations
- More than a dozen addressing modes

REDUCED INSTRUCTION SET

- Contains less than 100 instructions
- Have fixed instruction formats (32bits)
- Only 3 to 5 simple addressing modes are used
- Most instructions are register based
- Mem<u>ory access is done by load/store instructions</u> only
- Most instructions execute in one cycle
- Hardwired control is used
- Higher clock rate
- Lower **CPI**
- Higher processor performance

ARCHITECTURALI RAL DISTINCTIONS N S

CISC architecture	RISC architecture
Uses unified cache for holding both instructions and data	Uses separate instruction cache and data cache
same data/instruction	different
Use of micro-programmed control	Use hardwired control
Control memory was needed in early CISC	No need of control memory
High CPI	Less CPI





(b) The RISC architecture with hardwired control and split instruction cache and data cache

CISC SCALAR PROCESSOR

4

INTRODUCTION IN

• Scalar processor executes scalar data

o Simplest scalar processor

executes integer instruction using fixedpoint operands

o Capab<u>le scalar processor</u>

Execute both integer & floating point operations

o Modern scalar processor

Posses both integer unit & floating point unit

• CISC uses pipelined design

Representative CISC

- VAX 8600
- Motorola MC68040
- Intel i486



- Introduced by **Digital Equipment Corporation**
- o Uses Micro programmed control
- o 300 instructions
- o 20 addressing mode
- 16 G<u>PRs</u>
- CPI range from 2 to 20 cycles

4

- CPU has 2 functional units
 - Execution unit
 - □ Floating point unit
- Used for concurrent execution of integer & floating point instructions
- Unified cache used for holding data and instruction

- Pipelining has 6 stages
- Instruction unit
 - □ It pre-fetches & decoded the instructions
 - Handles branching operations
 - Supply operands to 2 functional units in a pipelined manner
- Translation look aside buffer TLB
 - Used in memory control unit for the fast generation of physical address from virtual address



MOTOROLA MG8040

- Consist of 1.2 million transistors
- o 100 instructions
- o 16 GPR
- 18 addressing modes
- Separate instruction & data cache
 - □ 4Kb instruction cache
 - □ 4Kb data cache
- Separate MMU supported by ATC (address translation cache)
 - This is equivalent to TLB

Δ.

• Data format range from 8 to 80 bits





MOTOROLA M(68040[2])

• CPU has 2 units

- Integer unit
- Floating point unit

o Integer unit

- □ Integer unit has 6 stage instruction pipeline
- All instructions are decoded in this unit
- Floating point instructions are forwarded to floating point unit for execution

o Floating point unit

□ It consist of 3 pipeline stages

MOTOROLAA M(68040[3]]

Memory units

- Data memory unit
- Instruction memory unit
- Separate instruction & data buses are used for instruction & data memory units
- □ Both buses are 32 bit wide
- Each of 2 ATC have 64 entries
 - This provide fast translation from virtual to physical addresses

CISC MICROPROCESSOR D R FAMILESE S

• Intel family

- □ Intel 4004 → 4 bit
- □ Intel 8008,8080,8085 → 8 bit
- □ <u>Intel 8086,80186,80268 → 16 bit</u>
- □ I<u>ntel 80386,80486 → 32 bit</u>
- Mot<u>orola family</u>
 - □ MC6800 → 8 bit
 - □ MC68000 → 16 bit
 - □ MC68020 → 32 bit



INTRODUCTION IN

- Generic RISC processors are called as scalar RISC
- They are designed to issue one instruction per cycle
- Similar to **Base scalar processor**

Δ.

• A good <u>compiler is required in RISC processor</u>than in C<u>ISC</u>

REPRESENTATIVE RSC PROCESSORS &

- □ Sun SPARC
- Intel i860
- □ Motorola M88100
- All <u>of these use 32 bit instructions</u>
 SPA<u>RC</u> stands <u>for</u> <u>Scalable</u> <u>Processor</u> <u>Architecture</u>

SUN SPARCE ARCHITECTUREURE

- These processors are implemented by a no: of licensed manufacturers
- All of them implemented floating point unit on a separate coprocessor chip
- Sun <u>SPARC instruction set has 69 basic instructions</u>
- It runs each procedure using 32 bit registers

- 8 of these registers are global registers shared by all procedures
- Remaining 24 are window registers associated with only each procedure

INTEL SEOFROCESSOR S D R ARCHITECTURE U R E

- o 64 bit RISC processor fabricated on a single chip
- It executed 82 instructions
- They included:-
 - 42 RISC integer
 - 24 <u>floating point</u>
 - 10 graphics
 - 6 assembler pseudo operations
- All instructions are executed in one cycle
- Applications
 - Used in floating point accelerators

- Graphics subsystem
- Workstations
- Multiprocessors
- multicomputers
ARCHITECTURE OF MELISOO

• There are 9 functional units

- Bus control unit
- Data cache
- Instruction cache
- <u>MMU</u>
- Integer unit
- Floating point control unit
- Pipelined adder unit
- Pipelined multiplier unit
- Graphics unit
- These units are interconnected using multiple data paths
 - □ Width ranges from 32 to 128 bits



o Bus control unit

• External and internal address buses are 32 bit wide

- External and internal data buses are 64 bits wide
- It coordinated 64 bit data transfer b/w chip and the outside world

o Instruction cache

- □ It is 4Kb cache
- Organized as 2 way set associative memory
- □ Each cache block have a size of 32 bytes
- □ It transferred 64 bits per clock cycle

o Data cache

- □ It is 8 kb cache
- □ 2 way set associative
- □ It transferred 128 bits per clock cycle
- Write back policy was used

Δ.

o MMU

This unit implemented 4Kb paged virtual memory via <u>TLB</u>

o RISC integer unit

- RISC integer ALU is 32 bits
- □ It executed
 - Load
 - Store
 - Integer
 - Bit
 - Control instructions
- □ It fetched instructions for floating point control unit also

o Fl<u>oating point units</u>

- There are 2 floating pointunits
 - Multiplier unit
 - Adder unit
- They can be used separately or simultaneously
- These units are controlled by floating point control unit
- Dual operation floating point instructions use both adder & multiplier units in parallel
 - Eg: add-and-multiply instruction, Subtract- and- multiply instruction

- Both integer unit & floating point control unit execute concurrently
- Floating point unit operates with 32 bit & 64 bit operands

o Graphics unit

- This unit executed integer operations corresponding to 8,
 16 or 32 bit pixel data types
- □ Supported 3D drawing in graphics frame buffer
- □ It <u>also supported:-</u>
 - Color intensity
 - Shading
 - Hidden surface elimination

Δ.

• Merge register

- Used only by vector integers instructions
- This register accumulated results of multiple addition 42 operations

SUPERSCALAR PROCESSORS

INTRODUCTION IN

- A CISC or a RISC scalar processor can be improved with a superscalar or vector architecture.
- Scalar processors are those executing one instruction per cycle.
- Only one instruction is issued per cycle, and only one completion of instruction is expected from the pipeline per cycle.
- In a superscalar processor, multiple instructions are issued per cycle and multiple results are generated per cycle

PIPELINING IN SUPERSCALAR PROCESSOR R

- Superscalar processors were originally developed as an alternative to vector processors
- They exploit **higher degree** of instruction level paral<u>lelism.</u>
- A **superscalar processor** of degree m can issue m instructions per cycle.
- The **base scalar processor**, implemented either in RISC or ClSC, has m = 1.
- Instruction issue degree is limited to 2 to 5

Δ.

• Thus m instructions must be executable in parallel.

PIPELINING IN SUPERSCALARLAR PROCESSORIR



Fig: A superscalar processor of degree m=3

PIPELINING IN SUPERSCALAR PROCESSOR R

- In a superscalar processor; the simple operation latency should require only one cycle, as in the base scalar processor.
- Due to the desire for a higher degree of instructionlevel parallelism in programs, the superscalar processor depends more on an optimizing compiler to exploit parallelism

REPRESENTATIVE SUPERSCALAR PROCESSOR R

- IBM RS/6000
- o DEC alpha 21064
- Intel i960CA

IBM RS/6000 ARCHITECTURE URE

• It has 3 functionalunits

- Branch processor
- Fixed point unit FXU
- Floating point unit FPU
- These could operate in parallel

o Bra<u>nch processor</u>

- This unit could arrange the execution of up to 5 instructions per cycle
- □ The instructions included are:-

- 1 branch instruction \rightarrow executed by branch processor
- 1 fixed point instruction → done by FXU
- 1 condition register instruction \rightarrow done by branch processor
- 1 floating point multiply-add instruction → done by FPU
 - This inst can be counted as 2 floating point operations

IBM RS/6000 ARCHITECTURE URE



- RS/6000 used hardwired control logic
- System used no: of buses
 - 32 bit for FXU
 - 64 <u>bit for FPU</u>
 - 128 bit for I cache and D cache

- Application
 - Numerically intensive scientific & engineering applications
 - Multiuser commercial environments

MEMORY HIERARCHY TECHNOLOGY

HERARCHICALCAL MEMORY TECHNOLOGYY

- In computer architecture the **memory hierarchy** separates computer storage into hierarchy based on **response time**
- Since **Response time**, complexity and capacity are related the level may also be distinguished by their performance and contro<u>lling technology</u>
- Storage devices are organized into a hierarchy of layers
 - Registers
 - Caches
 - Main memory
 - Disk devices
 - Backup storage

MEMORY HIERARCHY Y





PARAMETERS: OF MEMORY TECHNOLOGESES

• Memory technology at each level is characterized by 5 parameters

• Access time (ti)

□ It refers to the round-trip time from CPU to the ith level memory

- o Me<u>mory size (si)</u>
 - □ No: of bytes or words in leveli
- o Cos<u>t per byte ((ci)</u>)
 - □ Cost of ith level memory
 - □ It is estimated by the product of **ci.si**

Δ.

o Bandwidth (bi)

□ It refers to the **rate** at which information is transferred between adjacent levels

o Unit of transfer (xi)

Refers to the grain size for data transfer between levels i and level i+1

ANALYSIS IS OF PARAMETERS R S

- Memory devices at a lower level are :-
 - □ faster to access
 - □ Smaller in size
 - □ <u>More expensive per byte</u>
 - □ Higher bandwidth
 - Smaller units of transfer as compared with higher levels
- ti-1< ti, si-1< si, ci-1 > ci, bi-1> bi, xi-1<xi

MEMORY HERARCHY Y

• Registers and caches

- □ Registers are part of the **processors**
- Compiler does the register assignment
- Register transfer operations are directly controlled by the processor
- Transfer is conducted at processor speed in one clock cycle

o Caches

- Multilevel caches are built either on processor chip or on the processor board
- They are controlled by MMU & are programmer transparent
- □ They are implemented in 1 or **multiple levels**
- □ Speed of processor is faster than memory speed

o Main memory

- □ Also called **primary memory**
- □ Larger than the cache
- Implemented by cost effective RAM chips like DDR SDRAM
- □ Managed by MMU with OS

o Dis<u>k drives</u>s

- □ Disk storage is the highest level of on-line memory
- Holds system programs like OS, compilers, user programs and data sets

• Optical disk & magnetic tape (backup storage)

- □ They are **offline memory**
- Used for archival and backupstorage
- They hold copies of present & past userprograms, processed results & files

Memory level Characteristics	Level 0 CPU Registers	Level 1 Cache	Leve 2 Main Memory	Level 3 Disk Storage	Level 4 Tape Storage
Device technology	ECL	256K-bit SRAM	4M-bit DRAM	1-Gbyte magnetic disk unit	5-Gbyte magnetic tape unit
Access time, t_i	10 ns	25-40 ns	60–100 ns	12–20 ms	2-20 min (search time)
Capacity, s _i (in bytes)	512 bytes	128 Kbytes	512 Mbytes	60–228 Gbytes	512 Gbytes- 2 Tbytes
Cost, c _i (in cents/KB)	18,000	72	5.6	0.23	0.01
Bandwidth, b_i (in MB/s)	400-800	250-400	80-133	3–5	0.18-0.23
Unit of transfer, x_i	4-8 bytes per word	32 bytes per block	0.5-1 Kbytes per page	5-512 Kbytes per file	Backup
Allocation management	Compiler assignment	Hardware control	Operating system	Operating system/user	Operating system/user

 Table 4.7 Memory Characteristics of a Typical Mainframe Computer in 1993

PROPERTIES OF MEMORY HIERARCHY

INTRODUCTION IN

- Information stored in memory hierarchy (M1,M2,M3...Mn) satisfies following properties:-
 - I. Inclusion property
 - 2. Coherence property
 - 3. Locality property
- Cache is considered as innermost level M1
- This communicates with CPU registers directly
- Outer most level **Mn** contains all the information words stored

1. INCLUSION: N PROPERTYTY

- This property implies that M1 subset of M2 subset of M3 subset ofMn
- Inclusion relationship implies that all information items are originally stored in outermost level Mn
- All <u>the information will be found in **highest** level</u> which <u>is backup storage</u>
 - During processing, subsets of Mn is copied to Mn-1
 - \Box Subset of Mn-1 are copied to Mn-2 and so on
 - □ If an information is found in Mi, then copies of the same word is found in all upper levels, Mi+1, Mi+2....Mn
 - □ However word in Mi+1 may not be found in Mi

Δ.

■ Word misss in Mi→ word is missing from all its lower levels

UNITSS OF DATA TRANSFER AT EACH LEVEL

• Unit of data transfer between **CPU** & **Cache** words

- Word \rightarrow 4 or 8 bytes
- Cache(M1) is divided into cache blocks
 - \Box Each block \rightarrow 32bytes
 - Unit <u>of transfer between Cache & Main memory</u> blocks
- Main memory is divided into pages
 - □ Each <u>page → 4 Kb</u>

Unit of data transfer between Main memory & Disks > pages

- Disk memory is organized as segments
 - □ Size of segment varies with user needs

Δ.

Unit of data transfer between Disk & Backup storage segments



4.18 The inclusion property and data transfers between adjacent levels of a memory hierarchy

2.COHERENCE: E PROPERTY Y

- This property states that, copies of same information item at successive memory levels must be consistent
- If a <u>word is modified in cache, the copies of</u> that wor<u>d must be updated immediately or eventually</u> at all <u>higher levels</u>
- 2 strategies for maintaining coherence
 - □ Write through (WT)

Δ.

□ Write back (WB)

o Write through

This strategy demands immediate update in Mi+1, if a word is modified in Mi for i=1,2...n-1

o Write back

□ It <u>delays the update in M1+1 until the word being</u> modified in Mi is replaced or removed from Mi

Write-through vs. write-back

- Write-though
 - Every write from every processor goes to shared bus and memory
 - Easy protocol
 - Write-through unpopular for SMPs
- Write-back
 - absorb most writes as cache hits
 - Write hits don't go on bus
 - But now how do we ensure write propagation and serialization?
 - Need more sophisticated protocols



9

3. LOCALITY TY OF REFERENCES ES

- Locality of reference refers to a phenomenon in which a computer program tends to access same set of memory locations for a particular time period.
- This property of locality of reference is mainly shown by loops and subroutine calls in a program.

LOCALITY OF REFERENCES ES

- →In case of Loops in program CPU repeatedly refers to set of instructions that constitute the loop.
- →In case of subroutine calls, everytime set of instructions are fetched from memory.
- →References to data items also get localized that means same data item is referenced again and again.

LOCALITY OF REFERENCES ES






3. LOCALITY TY OF REFERENCES ES

- Memory hierarchy is developed based on locality of reference
- Memory access tends to cluster in certain regions in time, space and ordering
 - Program spend 90% of execution time only in 10% of code

• Eg: innermost loop of a nested loopoperation

• 3 dimensions of locality property

- [a] Temporal locality
- [b] Spatial locality

A. TEMPORAL LOCALITY Y

- Recently referenced items are likely to be referenced again in near future
- This is caused due to program constructs like:-
 - Iterative loops

- Subroutines
- When a loop is encountered or a subroutine is called, a small code segment is referenced repeatedly
- Temporal locality tends to cluster the access in recently used areas



Main memory

Cache memory

4 -

SPATIALA L LOCALITY Y

• Instruction or data near to main memory location may be needed soon

□ Eg: operations on arrays

• It involves access of certain clustered area in the address space



MEMORY CAPACITY TY PLANNING G

- Performance of a memory hierarchy is determined by effective access time Teff
- Teff depends on
 - □ hit <u>ratio</u>
 - □ access frequencies of successive levels

HIT RATO

- It is a concept defined for any 2 adjacent level of memory hierarchy
- When an item is found in Mi→ hit
- Otherwise it is a miss
- Hit <u>ratio Hi at Mi is the probability that an</u> item is found at Mi
- Miss ratio is defined as 1-Hi

ACCESS S FREQUENCY: Y

• Access frequency to Mi is defined as

- □ the probability of successfully accessing Mi, when there are i-1 misses at lower levels and a hit at Mi
- fi= (i-h1)(1-h2).....(1-hi-1)hi

- Access frequency decreases from lower to high levels
- f1>>f2>>f3>>....>>fn
- This implies that inner levels of memory are accessed more often than outer levels

EFFECTIVE VE ACCESS TIME

- Our aim is to achieve high hit ratio at Mi
- Coz when a Miss occur, penalty must be paid to access higher level of memory
 - □ <u>Misses in cache</u> → <u>block misses</u>

- □ M<u>isses in main memory → page fault</u>
- Time <u>penalty for page fault is longer than block</u> miss
 - \Box Because t1<t2<t3
- Cache miss is 2 to 4 times costly than cache hit
- Page fault is 1000 to 10000 times costly than page hit

• Teff=
$$\sum_{i=1}^{n}$$
 fi.ti

 $=h1t1+(1-h1)h2t2+(1-h1)(1-h2)h3t3+....+(1-h1)(1-h2)....(1-h_{n-1})tn$

Teff depends on program behavior and memory choices

HERARCHYHY OPTIMIZATION IN

• Total cost of memory hierarchy is C_{total} • $C_{total} = \sum_{i=1}^{n} ci.si$

• This implies that cost is estimated over n levels

o ci and <u>si depends on ti at each levels</u>

PROBLEM M

• Consider the design of a 3 level memory hierarchy with following features

Memory level	Access time	Capacity	Cost/kbyte
Cache	t1= 25 ns	s1= 512 Kbytes	c1=\$1.25
Main memory	t2= unknown	s2= 32 Mbytes	c2= \$.2
Disk a <u>rray</u>	<u>t3= 4 ms</u>	<u>s3= unknown</u>	<u>c3= \$0.0002</u>

- Aim is to achieve effective memory accesstime Teff=850 ns
- Cache hit ratio h1=0.98
- Main memory Hit ratioh2=0.99
- Total cost is upper bounded by \$15000
- Calculate the unknown specifications based on the given conditions

• Ctotal = c1.s1+c2.s2+c3.s3 <=1500 15000=1.25*512+.2*32000+.0002s3 15000 = 640+6400+.0002s3 .0002s3 = 15000-640-6400 s3=7960/.0002 $= 39800000x \ 10^{-6}$ = 39.8

- Teff= h1t1+(1-h1)h2t2+(1-h1)(1-h2)h3t3
- Teff= $h1t1+(1-h1)h2t2+(1-h1)(1-h2)h3t3 \le 850 \text{ ns}$
- \circ 850x10⁻⁹= .98*25x10⁻⁹+.02*.99*t2+.02*.01*1*4*10⁻³
- $\circ 850 \underline{x10^{-9}} = 24.5 \underline{x10^{-9}} + .0198 \underline{t2} + .0008 \underline{x10^{-3}}$
- $\circ .0198 \underline{t2} = 850 \underline{x10} 9 24.5 \underline{x10} 9 .0008 \underline{x10} 3$
 - $= 825.5 \times 10^{-9} .0008 \times 10^{-3}$
 - $=825.5 \times 10^{-9} 800 \times 10^{-9}$
 - $= 25.5 \times 10^{-9}$
 - $= 25.5 \times 10^{-9} / .0198$
 - $t2 = 1287 \text{ x} 10^{-9}$