Syllabus

410241 : High Performance Computing

Teaching Scheme :	Credit : 04	Examination Scheme :
TH : 04 Hours/Week		In-Sem (Paper) : 30 Marks
		End-Sem (Paper) : 70 Marks

Prerequisite Courses :

210253-Microprocessor, 210244 - Computer Organization and Architecture, 210254-Principles of Programming Languages, 310251- Systems Programming and Operating System

Course Objectives :

- To study parallel computing hardware and programming models
- To be conversant with performance analysis and modeling of parallel programs
- To understand the options available to parallelize the programs
- To know the operating system requirements to qualify in handling the parallelization

Course Outcomes :

On completion of the course, student will be able to

- Describe different parallel architectures, inter-connect networks, programming models
- Develop an efficient parallel algorithm to solve given problem
- Analyze and measure performance of modern parallel computing systems
- Build the logic to parallelize the programming task

Course Contents

Unit I : Introduction

(09 Hours)

Motivating Parallelism, Scope of Parallel Computing, Parallel Programming Platforms : Implicit Parallelism, Trends in Microprocessor and Architectures, Limitations of Memory, System Performance, Dichotomy of Parallel Computing Platforms, Physical Organization of Parallel Platforms, Communication Costs in Parallel Machines, Scalable design principles, Architectures : N-wide superscalar architectures, Multi-core architecture.

(Refer Chapter 1)

Unit II : Parallel Programming

(09 Hours)

Principles of Parallel Algorithm Design : Preliminaries, Decomposition Techniques, Characteristics of Tasks and Interactions, Mapping Techniques for Load Balancing, Methods for Containing Interaction Overheads, Parallel Algorithm Models, The Age of Parallel Processing, the Rise of GPU Computing, A Brief History of GPUs, Early GPU.

(Refer Chapter 2)

Unit III : Basic Communication

(09 Hours)

Operations- One-to-All Broadcast and All-to-One Reduction, All-to-All Broadcast and Reduction, All-Reduce and Prefix-Sum Operations, Scatter and Gather, All-to-All Personalized Communication, Circular Shift, Improving the Speed of Some Communication Operations.

(Refer Chapter 3)

Unit IV : Analytical Models of Parallel Programs

(9 Hours)

Analytical Models : Sources of overhead in Parallel Programs, Performance Metrics for Parallel Systems, and The effect of Granularity on Performance, Scalability of Parallel Systems, Minimum execution time and minimum cost, optimal execution time. Dense Matrix Algorithms : Matrix-Vector Multiplication, Matrix-Matrix Multiplication.

(Refer Chapter 4)

Unit V : Parallel Algorithms- Sorting and Graph

(09 Hours)

Issues in Sorting on Parallel Computers, Bubble Sort and its Variants, Parallelizing Quick sort, All-Pairs Shortest Paths, Algorithm for sparse graph, Parallel Depth-First Search, Parallel Best-First Search.

(Refer Chapter 5)

Unit VI : CUDA Architecture

(09 Hours)

CUDA Architecture, Using the CUDA Architecture, Applications of CUDA Introduction to CUDA C-Write and launch CUDA C kernels, Manage GPU memory, Manage communication and synchronization, Parallel programming in CUDA- C.

(Refer Chapter 6)

❑❑❑

UNIT I

Chapter 1 : Parallel Processing Concepts 1-1 to 1-44

Syllabus : Motivating Parallelism, Scope of Parallel Computing, Parallel Programming Platforms : Implicit Parallelism, Trends in Microprocessor and Architectures, Limitations of Memory, System Performance, Dichotomy of Parallel Computing Platforms, Physical Organization of Parallel Platforms, Communication Costs in Parallel Machines, Scalable design principles, Architectures : N-wide superscalar architectures, Multi-core architecture.

1.1 Introduction to Parallel Computing 1-1
✓ **Syllabus Topic :** Motivating Parallelism 1-1
1.2 Motivation for Parallelism 1-1
1.2.1 Increased Number of Transistors in Integrated Chips Enhances Computing Power 1-2
1.2.2 Improvements in Storage Technology (Memory/Disk) 1-2
1.2.3 Improved Networking Devices and Systems for Data Communications 1-3
✓ **Syllabus Topic :** Scope of Parallel Computing 1-3
1.3 Scope of Parallel Computing 1-3
1.3.1 Applications of Parallel Computing in Engineering 1-3
1.3.2 Scientific Applications using Parallel Computing 1-4
1.3.3 Commercial Applications Based on Parallel Computing 1-4
1.3.4 Parallel Computing Applications in Computer Systems 1-4
✓ **Syllabus Topic :** Parallel Programming Platforms 1-4
1.4 Parallel Programming Platforms 1-4
✓ **Syllabus Topic :** Implicit Parallelism - Trends in Microprocessor and Architectures 1-5
1.4.1 Implicit Parallelism : Trends in Microprocessor and Architectures 1-5
1.4.1(A) Pipelining and Superscalar Execution 1-5
1.4.1(B) Very Long Instruction Word Processors 1-7
1.4.1(C) Basic Working Principle of VLIW Processor 1-8
1.4.1(D) Advantages of VLIW Processor 1-8
1.4.1(E) Disadvantages of VLIW Processor 1-9
✓ **Syllabus Topic :** Limitations of Memory System Performance 1-9
1.5 Limitations of Memory System Performance 1-9
1.5.1 Use of Caches for Improvement of Latency 1-10
1.5.2 Memory Bandwidth 1-10
1.5.3 Memory Latency Hiding Techniques 1-10
1.5.4 Effects of Multithreading and Prefetching (Tradeoffs of Multithreading and Prefetching) 1-11
✓ **Syllabus Topic :** Dichotomy of Parallel Computing Platforms 1-11
1.6 Dichotomy of Parallel Computing Platforms 1-11
1.6.1 Control Structure of Parallel Platforms 1-11
1.6.2 Communication Model of Parallel Platforms 1-13
1.6.2(A) Shared-Address-Space 1-13
✓ **Syllabus Topic :** Physical Organization of Parallel Platforms 1-14
1.7 Physical Organization of Parallel Platforms 1-14
1.7.1 Architecture of an Ideal Parallel Computer 1-14
1.7.2 Interconnection Networks for Parallel Computers 1-15
1.7.3 Network Topologies 1-16
✓ **Syllabus Topic :** Communication Costs in Parallel Machines 1-18
1.8 Communication Costs in Parallel Machines 1-18
1.8.1 Message Passing Costs 1-18
1.8.2 Communication Costs in Shared-Address-Space Machines 1-20
✓ **Syllabus Topic :** N-wide Superscalar Architectures, Multi-core Architecture 1-20
1.9 Architectures : N-wide Superscalar Architectures, Multi-core, Multi-Threaded 1-20
1.9.1 Multithreaded Architecture 1-20
1.9.2 Multi - Core Architecture 1-20
1.9.3 N-wide Superscalar Architectures 1-21
1.9.3(A) Dynamically-scheduled Superscalar Processor 1-23
1.10 Cache Coherence in Multiprocessor Systems 1-24
1.10.1 Hardware-based Protocols 1-25
1.11 Levels of Parallelism 1-30
1.11.1 Instruction - Level Parallelism 1-30
1.11.2 Thread - Level Parallelism or Task - Level Parallelism 1-30
1.11.3 Transaction - Level Parallelism 1-30
1.11.4 Memory 1-30
1.11.5 Function - Level Parallelism 1-31
1.12 Models 1-31
1.12.1 SIMD (Single Instruction Stream Multiple Data Stream) Architecture 1-32
1.12.2 MIMD (Multiple Instruction Stream Multiple Data Stream) Architecture 1-32
1.12.3 SIMT 1-33
1.12.3(A) Micro Architecture of SIMT 1-34

1.12.3(B) Mapping SIMT Workloads to SIMT Cores 1-34

1.12.3(C) SIMT Programming Model 1-35

1.12.4 SPMD (Single Program Multiple Data) 1-35

1.12.5 Dataflow Model 1-36

1.12.5(A) Dataflow Graph 1-37

1.12.5(B) Dataflow Computations 1-38

1.12.5(C) Types of Dataflow Execution Model 1-39

1.12.5(D) Dataflow Machine Architecture 1-40

1.12.5(E) Types of Dataflow Machines 1-41

1.12.5(F) Drawbacks in Dataflow Computing 1-42

1.12.6 Demand-driven Computation 1-43

UNIT II

Chapter 2 : Parallel Programming 2-1 to 2-25

Syllabus : Principles of Parallel Algorithm Design : Preliminaries, Decomposition Techniques, Characteristics of Tasks and Interactions, Mapping Techniques for Load Balancing, Methods for Containing Interaction Overheads, Parallel Algorithm Models, The Age of Parallel Processing, the Rise of GPU Computing, A Brief History of GPUs, Early GPU.

✓ **Syllabus Topic :** Principles of Parallel Algorithm Design - Preliminaries 2-1

2.1 Principles of Parallel Algorithm Design 2-1

2.1.1 Preliminaries 2-1

2.1.2 Decomposition, Tasks, and Dependency Graphs 2-1

2.1.3 Granularity, Concurrency and Task-Interaction 2-3

2.1.3(A) Granularity 2-3

2.1.3(B) Concurrency 2-4

2.1.3(C) Task-Interaction 2-4

2.1.4 Processes and Mapping 2-5

2.1.4(A) Processes 2-5

2.1.4(B) Mapping 2-5

2.1.5 Processes Versus Processors 2-5

✓ **Syllabus Topic :** Decomposition Techniques 2-5

2.2 Decomposition Techniques 2-5

2.2.1 Data-Decomposition 2-6

2.2.2 Recursive Decomposition 2-6

2.2.3 Exploratory Decomposition 2-7

2.2.4 Speculative Decomposition 2-9

✓ **Syllabus Topic :** Characteristics of Tasks and Interactions 2-11

2.3 Characteristics of Tasks and Interactions 2-11

2.3.1 Tasks Characteristics 2-11

2.3.2 Characteristics of Inter-Task Interactions 2-12

2.3.3 Static Versus Dynamic Pattern 2-13

2.3.4 Regular Versus Irregular 2-13

2.3.5 Read-Only Versus Read-Write 2-13

2.3.6 One-way Versus Two-way 2-13

✓ **Syllabus Topic :** Mapping Techniques for Load Balancing 2-14

2.4 Mapping Techniques for Load Balancing 2-14

2.4.1 Mapping Techniques 2-14

2.4.2 Schemes for Static Mapping 2-15

2.4.2(A) Mapping Based on Data Partitioning 2-15

2.4.2(B) Block Distributions 2-15

2.4.2(C) Cyclic Distributions 2-17

2.4.2(D) Block - Cyclic Distribution 2-17

2.4.3 Randomized Block Distributions 2-18

2.4.4 Hierarchical Mappings 2-18

2.4.5 Schemes for Dynamic Mapping 2-19

✓ **Syllabus Topic :** Methods for Containing Interaction Overheads 2-20

2.5 Methods for Containing Interaction Overheads 2-20

2.5.1 Maximizing Data Locality 2-20

2.5.2 Overlapping Computations with Interactions 2-20

2.5.3 Replicating Data or Computations 2-21

2.5.4 Overlapping Interactions with Other Interactions 2-21

✓ **Syllabus Topic :** Parallel Algorithm Models 2-21

2.6 Parallel Algorithm Models 2-21

2.6.1 The Data-Parallel Model 2-21

2.6.2 The Task Graph Model 2-22

2.6.3 The Work Pool Model (or Task Pool Model) 2-22

2.6.4 The Master Slave Model (Manager - Worker Model) 2-22

2.6.5 The Pipeline or Producer Consumer Model 2-23

2.6.6 Hybrid Models 2-23

✓ **Syllabus Topic :** The Age of Parallel Processing 2-23

2.7 The Age of Parallel Processing 2-23

2.7.1 Central Processing Units and GPUs........................2-23
✓ **Syllabus Topic :** Brief History of GPUs, Early GPU 2-24
2.7.2 The Rise of GPU Computing 2-24

UNIT III

Chapter 3 : Basic Communication 3-1 to 3-21

Syllabus : Operations - One-to-All Broadcast and All-to-One Reduction, All-to-All Broadcast and Reduction, All-Reduce and Prefix-Sum Operations, Scatter and Gather, All-to-All Personalized Communication, Circular Shift, Improving the Speed of Some Communication Operations.

✓ **Syllabus Topic :** Operations - One-to-All Broadcast and All-to-one Reduction 3-1
3.1 One-to-All Broadcast and All-to-One Reduction...........3-1
3.1.1 Ring and Linear Array...3-2
3.1.2 Mesh 3-3
3.1.3 Hypercube..3-3
3.1.4 Balanced Binary Tree 3-4
3.1.5 Broadcast and Reduction Algorithms..........................3-4
3.1.6 Cost Analysis for One-to-Many Broadcast and Many-to-One Reduction on Hypercube................3-8
✓ **Syllabus Topic :** All-to-All Broadcast and Reduction 3-8
3.2 All-to-All Broadcast and Reduction 3-8
3.2.1 All - to - all Broadcast and Reduction on a Ring 3-9
3.2.2 All - to - all Broadcast on a mesh...............................3-11
3.2.3 All - to - all Broadcast on a Hypercube 3-12
✓ **Syllabus Topic :** All-Reduce and Prefix-Sum Operations 3-13
3.3 All-Reduce and Prefix-Sum Operations 3-13
3.3.1 Prefix-Sum Algorithm 3-14
✓ **Syllabus Topic :** Scatter and Gather.......................3-14
3.4 Scatter and Gather..3-14
✓ **Syllabus Topic :** All-to-All Personalized Communication 3-15
3.5 All-to-All Personalized Communication......................3-15
3.5.1 Total Exchange on a Ring 3-16
3.5.2 Total Exchange on a Mesh...3-16
3.5.3 Total Exchange on a Hypercube................................3-17
3.5.4 All-to-All Personalized Communication on a Hypercube: Optimal Algorithm 3-18
✓ **Syllabus Topic :** Circular Shift..................................3-20
3.6 Circular Shift..3-20
3.6.1 Circular Shift on a Mesh 3-20
3.6.2 Circular Shift on a Hypercube....................................3-20
✓ **Syllabus Topic :** Improving the Speed of Some Communication Operations 3-21
3.7 Improving the Speed of Some Communication Operations 3-21
3.7.1 One-to-all Broadcast 3-21
3.7.2 All-to-one reduction 3-21
3.7.3 All Reduce...3-21

UNIT IV

Chapter 4 : Analytical Modelling of Parallel Programs 4-1 to 4-22

Syllabus : Analytical Models: Sources of overhead in Parallel Programs, Performance Metrics for Parallel Systems, and The effect of Granularity on Performance, Scalability of Parallel Systems, Minimum execution time and minimum cost, optimal execution time. Dense Matrix Algorithms: Matrix-Vector Multiplication, Matrix-Matrix Multiplication.

✓ **Syllabus Topic :** Sources of Overhead in Parallel Programs 4-1
4.1 Sources of Overhead in Parallel Programs.................4-1
✓ **Syllabus Topic :** Performance Metrics for Parallel Systems 4-2
4.2 Performance Metrics for Parallel Systems..................4-2
✓ **Syllabus Topic :** Effect of Granularity on Performance..4-6
4.3 Effect of Granularity on Performance 4-6
✓ **Syllabus Topic :** Scalability of Parallel Systems.........4-7
4.4 Scalability of Parallel Systems....................................4-7
✓ **Syllabus Topic :** Minimum Execution Time and Minimum Cost-Optimal Execution Time 4-10
4.5 Minimum Execution Time and Minimum Cost-Optimal Execution Time.................................4-10
✓ **Syllabus Topic :** Dense Matrix Algorithms 4-11
4.6 Dense Matrix Algorithms 4-11
✓ **Syllabus Topic :** Matrix-Vector Multiplication...........4-11
4.6.1 Matrix-Vector Multiplication.......................................4-11
4.6.2 Row-wise 1-D Partitioning 4-12
4.6.3 2-D Partitioning..4-14
4.6.4 Using Fewer than n^2 Processes 4-16

✓ **Syllabus Topic :** Matrix Multiplication 4-17

4.7 Matrix Multiplication 4-17

4.7.1 Sequential Algorithm for Matrix Multiplication 4-17

4.7.2 Partitioning into Submatrices 4-17

4.7.3 The Steps of the Parallel Algorithm for Dense Matrix Multiplication 4-18

4.7.4 A Simple Parallel Algorithm 4-19

4.7.5 Cannon's Algorithm 4-20

UNIT V

Chapter 5 : Parallel Algorithms- Sorting and Graph 5-1 to 5-21

Syllabus : Issues in Sorting on Parallel Computers, Bubble Sort and its Variants, Parallelizing Quick sort, All-Pairs Shortest Paths, Algorithm for sparse graph, Parallel Depth-First Search, Parallel Best-First Search.

✓ **Syllabus Topic :** Issues in Sorting on Parallel Computers 5-1

5.1 Issues in Sorting on Parallel Computers 5-1

5.1.1 Where the Input and Output Sequences are Stored ? 5-1

5.1.2 How Comparisons are Performed ? 5-1

5.1.2(A) One Element Per Process 5-2

5.1.2(B) More than One Element Per Process 5-2

✓ **Syllabus Topic :** Bubble Sort and its Variants 5-2

5.2 Bubble Sort and its Variants 5-2

5.2.1 Odd-Even Transposition 5-3

5.2.1(A) Parallel Formulation 5-3

5.2.1(B) Shell sort 5-4

5.2.1(B)1 General Scheme of Shell Sort Method 5-5

5.2.1(B)2 Shell Sort Parallel Algorithm 5-5

5.3 Quick Sort 5-6

✓ **Syllabus Topic :** Parallelizing Quick Sort 5-7

5.3.1 The Parallel Quick Sort Algorithm 5-7

✓ **Syllabus Topic :** All-Pairs Shortest Paths 5-8

5.4 All-Pairs Shortest Paths 5-8

5.4.1 Dijkstra's Algorithm 5-8

5.4.2 Floyd Algorithm 5-9

✓ **Syllabus Topic :** Algorithms for Sparse Graphs 5-11

5.5 Algorithms for Sparse Graphs 5-11

5.5.1 Finding a Maximal Independent Set 5-12

5.5.2 Parallel Formulation of Luby's Algorithm in Shared-Address-Space Computer 5-14

5.5.2(A) Single-Source Shortest Paths 5-14

5.5.3 Parallel Johnson's Algorithm with Centralized Queue 5-15

5.5.4 Parallel Johnson's Algorithm with Distributed Queue 5-16

✓ **Syllabus Topic :** Parallel Depth First Search (DFS) 5-17

5.6 Parallel Depth First Search (DFS) 5-17

5.7 Best-First Search 5-18

✓ **Syllabus Topic :** Parallel Best-First Search 5-19

5.7.1 Parallel Best-First Search 5-19

UNIT VI

Chapter 6 : CUDA Architecture 6-1 to 6-23

Syllabus : CUDA Architecture, Using the CUDA Architecture, Applications of CUDA Introduction to CUDA C-Write and launch CUDA C kernels, Manage GPU memory, Manage communication and synchronization, Parallel programming in CUDA-C.

✓ **Syllabus Topic :** CUDA Architecture 6-1

6.1 CUDA Architecture 6-1

✓ **Syllabus Topic :** Using the CUDA Architecture 6-2

6.2 Using the CUDA Architecture 6-2

✓ **Syllabus Topic :** Applications of CUDA 6-2

6.3 Applications of CUDA 6-2

6.4 Heterogeneous Architecture 6-4

6.4.1 Paradigm of Heterogeneous Computing 6-4

6.4.1(A) Processor Architecture 6-5

6.4.2 CUDA Programming Models for High Performance Computing Architectures 6-7

6.4.3 GPU Programming Model 6-7

6.5 Memory Hierarchy 6-8

6.5.1 CUDA Memory and Cache Architecture 6-10

6.6 Introduction to CUDA C 6-10

6.6.1 CUDA Kernels 6-12

6.6.1(A) Hello World Program in CUDA 6-13

6.6.1(B) Parameter Passing in Kernel 6-13

6.6.2 Organizing Threads 6-14

✓ **Syllabus Topic :** Launching a CUDA Kernel 6-15

6.6.3 Launching a CUDA Kernel 6-15

✓ **Syllabus Topic :** Writing a CUDA Kernel 6-16

6.6.4 Writing a CUDA Kernel 6-16

✓ **Syllabus Topic :** Manage Communication and Synchronization 6-17

6.6.5 Manage Communication and Synchronization 6-17

✓ **Syllabus Topic :** Manage GPU Memory 6-17

6.6.6 Manage GPU Memory 6-17

✓ **Syllabus Topic :** Parallel Programming in CUDA- C 6-18

6.7 Parallel Programming in CUDA- C 6-18

6.7.1 Parallel Programming Pattern in CUDA 6-18

6.7.2 A Simple Example 6-18

6.7.3 Vector Addition in Parallel 6-20

6.7.4 Vector Addition using Parallel Threads 6-21

6.7.5 Parallel Dot Product Using CUDA 6-22

❑❑❑