Peruvian Computing Society (SPC)
School of Computer Science
Sillabus 2023-I

1. COURSE

CS3P1. Parallel and Distributed Computing (Mandatory)

2. GENERAL INFORMATION

2.1 Course	:	CS3P1. Parallel and Distributed Computing
2.2 Semester	:	8^{vo} Semestre.
2.3 Credits	:	4
2.4 Horas	:	2 HT; 4 HP;
2.5 Duration of the period	:	16 weeks
2.6 Type of course	:	Mandatory
2.7 Learning modality	:	Blended
2.8 Prerrequisites	:	

- CS212. Analysis and Design of Algorithms. (5^{th} Sem)
- CS231. Networking and Communication. (7^{th} Sem)
- CS212. Analysis and Design of Algorithms. (5^{th} Sem)
- CS231. Networking and Communication. (7^{th} Sem)

3. PROFESSORS

Meetings after coordination with the professor

4. INTRODUCTION TO THE COURSE

The last decade has brought explosive growth in computing with multiprocessors, including Multi-core processors and distributed data centers. As a result, computing parallel and distributed has become a widely elective subject to be one of the main components in the mesh studies in computer science undergraduate. Both parallel and distributed computing the simultaneous execution of multiple processes, whose operations have the potential to intercalar in a complex way. Parallel and distributed computing builds on foundations in many areas, including understanding the fundamental concepts of systems, such as: concurrency and parallel execution, consistency in state / memory manipulation, and latency. The communication and coordination between processes has its foundations in the passage of messages and models of shared memory of computing and algorithmic concepts like atomicity, consensus and conditional waiting.Achieving acceleration in practice requires an understanding of parallel algorithms, strategies for decomposition problem, systems architecture, implementation strategies and analysis of performance. Distributed systems highlight the problems of security and tolerance to Failures, emphasize the maintenance of the replicated state and introduce additional problems in the field of computer networks.

5. GOALS

- That the student is able to create parallel applications of medium complexity by efficiently leveraging machines with multiple cores.
- That the student is able to compare sequential and parallel applications.
- That the student is able to convert, when the situation warrants, sequential applications to parallel efficiently

6. COMPETENCES

1) Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions. (**Usage**)

6) Apply computer science theory and software development fundamentals to produce computing-based solutions. (**Usage**)

7. TOPICS

<table>
<tr><td colspan="2">Unit 1: Parallelism Fundamentals (18)</td></tr>
<tr><td colspan="2">Competences Expected:</td></tr>
<tr><td>Topics</td><td>Learning Outcomes</td></tr>
<tr><td>
• Multiple simultaneous computations
• Goals of parallelism (e.g., throughput) versus concurrency (e.g., controlling access to shared resources)
• Parallelism, communication, and coordination
  – Parallelism, communication, and coordination
  – Need for synchronization
• Programming errors not found in sequential programming
  – Data races (simultaneous read/write or write/write of shared state)
  – Higher-level races (interleavings violating program intention, undesired non-determinism)
  – Lack of liveness/progress (deadlock, starvation)
</td><td>
• Distinguish using computational resources for a faster answer from managing efficient access to a shared resource [Familiarity]
• Distinguish multiple sufficient programming constructs for synchronization that may be inter-implementable but have complementary advantages [Familiarity]
• Distinguish data races from higher level races [Familiarity]
</td></tr>
<tr><td colspan="2">Readings : [Pac11], [Mat14], [quinnz], [Geo10]</td></tr>
</table>

<table>
<tr><td colspan="2">Unit 2: Parallel Architecture (12)</td></tr>
<tr><td colspan="2">Competences Expected:</td></tr>
<tr><td>Topics</td><td>Learning Outcomes</td></tr>
<tr><td>
• Multicore processors
• Shared vs distributed memory
• Symmetric multiprocessing (SMP)
• SIMD, vector processing
• GPU, co-processing
• Flynn's taxonomy
• Instruction level support for parallel programming
  – Atomic instructions such as Compare and Set
• Memory issues
  – Multiprocessor caches and cache coherence
  – Non-uniform memory access (NUMA)
• Topologies
  – Interconnects
  – Clusters
  – Resource sharing (e.g., buses and interconnects)
</td><td>
• Explain the differences between shared and distributed memory [Assessment]
• Describe the SMP architecture and note its key features [Assessment]
• Characterize the kinds of tasks that are a natural match for SIMD machines [Usage]
• Describe the advantages and limitations of GPUs vs CPUs [Usage]
• Explain the features of each classification in Flynn's taxonomy [Usage]
• Describe the challenges in maintaining cache coherence [Familiarity]
• Describe the key performance challenges in different memory and distributed system topologies [Familiarity]
</td></tr>
<tr><td colspan="2">Readings : [Pac11], [KH13], [SK10], [Geo10]</td></tr>
</table>

<table>
<tr><td colspan="2">Unit 3: Parallel Decomposition (18)</td></tr>
<tr><td colspan="2">Competences Expected:</td></tr>
<tr><td>Topics</td><td>Learning Outcomes</td></tr>
<tr><td>
Need for communication and coordination/synchronization
Independence and partitioning
Basic knowledge of parallel decomposition concept
Task-based decomposition
Implementation strategies such as threads
Data-parallel decomposition
Strategies such as SIMD and MapReduce
Actors and reactive processes (e.g., request handlers)
</td><td>
Explain why synchronization is necessary in a specific parallel program [Usage]
Identify opportunities to partition a serial program into independent parallel modules [Familiarity]
Write a correct and scalable parallel algorithm [Usage]
Parallelize an algorithm by applying task-based decomposition [Usage]
Parallelize an algorithm by applying data-parallel decomposition [Usage]
Write a program using actors and/or reactive processes [Usage]
</td></tr>
<tr><td colspan="2">Readings : [Pac11], [Mat14], [Qui03], [Geo10]</td></tr>
</table>

<table>
<tr><td colspan="2">Unit 4: Communication and Coordination (18)</td></tr>
<tr><td colspan="2">Competences Expected:</td></tr>
<tr><td>Topics</td><td>Learning Outcomes</td></tr>
<tr><td>

- Shared Memory
- Consistency, and its role in programming language guarantees for data-race-free programs
- Message passing
  - Point-to-point versus multicast (or event-based) messages
  - Blocking versus non-blocking styles for sending and receiving messages
  - Message buffering (cross-reference PF/Fundamental Data Structures/Queues)
- Atomicity
  - Specifying and testing atomicity and safety requirements
  - Granularity of atomic accesses and updates, and the use of constructs such as critical sections or transactions to describe them
  - Mutual Exclusion using locks, semaphores, monitors, or related constructs
    * Potential for liveness failures and deadlock (causes, conditions, prevention)
  - Composition
    * Composing larger granularity atomic actions using synchronization
    * Transactions, including optimistic and conservative approaches
- Consensus
  - (Cyclic) barriers, counters, or related constructs
- Conditional actions
  - Conditional waiting (e.g., using condition variables)

</td><td>

- Use mutual exclusion to avoid a given race condition [Usage]
- Give an example of an ordering of accesses among concurrent activities (eg, program with a data race) that is not sequentially consistent [Familiarity]
- Give an example of a scenario in which blocking message sends can deadlock [Usage]
- Explain when and why multicast or event-based messaging can be preferable to alternatives [Familiarity]
- Write a program that correctly terminates when all of a set of concurrent tasks have completed [Usage]
- Give an example of a scenario in which an attempted optimistic update may never complete [Familiarity]
- Use semaphores or condition variables to block threads until a necessary precondition holds [Usage]

</td></tr>
<tr><td colspan="2">Readings : [Pac11], [Mat14], [Qui03], [Geo10]</td></tr>
</table>

Unit 5: Parallel Algorithms, Analysis, and Programming (18)	
Competences Expected:	
Topics	**Learning Outcomes**
• Critical paths, work and span, and the relation to Amdahl's law • Speed-up and scalability • Naturally (embarrassingly) parallel algorithms • Parallel algorithmic patterns (divide-and-conquer, map and reduce, master-workers, others) – Specific algorithms (e.g., parallel MergeSort) • Parallel graph algorithms (e.g., parallel shortest path, parallel spanning tree) (cross-reference AL/Algorithmic Strategies/Divide-and-conquer) • Parallel matrix computations • Producer-consumer and pipelined algorithms • Examples of non-scalable parallel algorithms	• Define "critical path", "work", and "span" [Familiarity] • Compute the work and span, and determine the critical path with respect to a parallel execution diagram [Usage] • Define "speed-up" and explain the notion of an algorithm's scalability in this regard [Familiarity] • Identify independent tasks in a program that may be parallelized [Usage] • Characterize features of a workload that allow or prevent it from being naturally parallelized [Familiarity] • Implement a parallel divide-and-conquer (and/or graph algorithm) and empirically measure its performance relative to its sequential analog [Usage] • Decompose a problem (eg, counting the number of occurrences of some word in a document) via map and reduce operations [Usage] • Provide an example of a problem that fits the producer-consumer paradigm [Usage] • Give examples of problems where pipelining would be an effective means of parallelization [Usage] • Implement a parallel matrix algorithm [Usage] • Identify issues that arise in producer-consumer algorithms and mechanisms that may be used for addressing them [Usage]
Readings : [Mat14], [Qui03], [Geo10]	

Unit 6: Parallel Performance (18)	
Competences Expected:	
Topics	**Learning Outcomes**

• Load balancing
• Performance measurement
• Scheduling and contention (cross-reference OS/Scheduling and Dispatch)
• Evaluating communication overhead
• Data management
  – Non-uniform communication costs due to proximity (cross-reference SF/Proximity)
  – Cache effects (e.g., false sharing)
  – Maintaining spatial locality
• Power usage and management
	• Detect and correct a load imbalance [Usage]
	• Calculate the implications of Amdahl's law for a particular parallel algorithm (cross-reference SF/Evaluation for Amdahl's Law) [Usage]
	• Describe how data distribution/layout can affect an algorithm's communication costs [Familiarity]
	• Detect and correct an instance of false sharing [Usage]
	• Explain the impact of scheduling on parallel performance [Familiarity]
	• Explain performance impacts of data locality [Familiarity]
	• Explain the impact and trade-off related to power usage on parallel performance [Familiarity]

Readings : [Pac11], [Mat14], [KH13], [SK10], [Geo10]	

8. WORKPLAN

8.1 Methodology

Individual and team participation is encouraged to present their ideas, motivating them with additional points in the different stages of the course evaluation.

8.2 Theory Sessions

The theory sessions are held in master classes with activities including active learning and roleplay to allow students to internalize the concepts.

8.3 Practical Sessions

The practical sessions are held in class where a series of exercises and/or practical concepts are developed through problem solving, problem solving, specific exercises and/or in application contexts.

9. EVALUATION SYSTEM

********* EVALUATION MISSING ********

10. BASIC BIBLIOGRAPHY

[Geo10] Gerhard Wellein Georg Hager. *Introduction to High Performance Computing for Scientists and Engineers (Chapman & HallCRC Computational Science)*. Ed. by CRC Press. 1st. 2010. ISBN: 978-1439811924.

[KH13] David B. Kirk and Wen-mei W. Hwu. *Programming Massively Parallel Processors: A Hands-on Approach*. 2nd. Morgan Kaufmann, 2013. ISBN: 978-0-12-415992-1.

[Mat14] Norm Matloff. *Programming on Parallel Machines*. University of California, Davis, 2014. URL: `http://heather.cs.ucdavi`

[Pac11] Peter S. Pacheco. *An Introduction to Parallel Programming*. 1st. Morgan Kaufmann, 2011. ISBN: 978-0-12-374260-5.

[Qui03] Michael J. Quinn. *Parallel Programming in C with MPI and OpenMP*. 1st. McGraw-Hill Education Group, 2003. ISBN: 0071232656.

[SK10] Jason Sanders and Edward Kandrot. *CUDA by Example: An Introduction to General-Purpose GPU Programming*. 1st. Addison-Wesley Professional, 2010. ISBN: 0131387685, 9780131387683.