The Hong Kong Polytechnic University

Subject Description Form

Subject Code	COMP5585
Subject Title	Cloud Computing and Hyperscale AI Infrastructure
Credit Value	3
Level	5
Pre-requisite/ Co-requisite/ Exclusion	Knowledge of Computer Organization and Operating Systems is preferable.
Objectives	This is a fundamental course that provides students with the foundations of cloud computing and hyperscale AI infrastructure. It covers the principles and concepts, the technical underpinnings and supporting technologies, and the best practices and applications. The objectives of this subject include: 1. To provide students with a broad view of the theoretical and technological aspects that have led to the evolution of cloud computing and hyperscale AI infrastructure. 2. To teach students how cloud computing and hyperscale AI infrastructure support emerging big data and AI applications that are essential to the industry, and the requirement of working with cloud computing environments & hyperscale AI infrastructure and develop cloud-based applications. 3. To equip students with the underlying technologies of cloud computing, including (1) basic knowledge of cloud computing, (2) data center networking, (3) virtualization, (4) large-scale distributed computing framework. 4. To equip students with the underlying technologies of hyperscale AI infrastructure, including (1) basic knowledge of large-scale AI model training, fine-tuning, and inference, (2) networking stacks in Artificial Intelligence Data Center, and communication libraries for large-scale AI infrastructure, such as MPI, NCCL, etc, (3) distributed strategies for model training, fine-tuning, and inference. (4) emerging system performance optimization technologies for model training, fine-tuning, and inference. 5. To equip students with the knowledge and skills for the planning, design and programming of cloud systems and AI infrastructure for the real-world applications.

Intended Learning Outcomes	Upon completion of the subject, students will be able to: *Professional/academic knowledge and skills* 1. Understand and appreciate the technological impact of cloud computing and hyperscale AI infrastructure for future enterprises, and the technologies underpinning it. 2. Apply systematic and principled practices to designing, implementing and deploying cloud computing and hyperscale AI infrastructure. 3. Review and assess the opportunities, costs, and system performance optimization for how the cloud computing support emerging AI applications. *Attributes for all-roundedness* 1. Systematic and incremental approach to resolving practical enterprise computing problems and challenges. 2. Learn to work effectively as a team member. 3. Write technical reports and present solutions.
Subject Synopsis/ Indicative Syllabus	**Topics:** **1. Overview** The evolution of cloud computing paradigms; Motivations and benefits of cloud computing; Definitions and principles of cloud computing; Applications of cloud computing; The overview of hyperscale AI infrastructure. **2. Part I: Cloud Architecture and Service Models** Cloud architecture and major components, Physical infrastructure; Service models; Service provisioning; Representative providers and platforms (Amazon, Microsoft, IBM Google, Alibaba, etc); AWS (EC2, S3, etc.) **3. Part I: Virtualization Techniques** CPU virtualization; Memory virtualization; I/O virtualization; Network virtualization; Xen; Docker (Container); Kubernetes. **4. Part I: Large-Scale Distributed Computing Platform** Basic knowledge of large-scale distributed computing platforms, such as Hadoop, Spark, Ray, etc; **5. Part II: Networking Stacks and Communication libraries in Artificial Intelligence Data Center** Networking architecture in emerging AI Data Center; Communication libraries like MPI and NCCL, etc. **6. Part II: Large-Scale Model Training and Fine-Tuning** Basic knowledge about distributed large-scale model training and fine-tuning; Various distributed training strategies; Various model training frameworks (Pytorch, Tensorflow, etc). **7. Part II: Large-Scale Model Inference and Serving.**

<table>
<tr><td></td><td>Basic knowledge about distributed large-scale model inference and serving; Various distributed inference strategies; Various model inference frameworks, such as VLLM, etc.
8. Part II: Emerging Optimization Technologies for Model Training, Fine-Tuning, and Inference
Model Compression; Data management; Strategy design, etc.</td></tr>
<tr><td>Teaching/Learning Methodology</td><td>The course is comprised of lectures, tutorials and laboratory exercises. During lecture, students are taught the important concepts and principles that drive the development of cloud computing and hyperscale AI infrastructure. In the lecture, students are encouraged to actively participate in mini-discussions and questions that are designed to reinforce their understanding of concepts taught.

During tutorials, students will be presented with real and practical scenarios of enterprise case studies. In particular, they will be given the unique opportunities to study, analyze and propose solutions that leverage cloud computing and hyperscale AI infrastructure concepts. Small group discussions will be encouraged, and students will need to present their results and solutions in the form of reports and presentations.

To reinforce practical aspects of their training, simple lab exercises will be conducted to expose students to the state-of-the-art tools and development environment that uses cloud computing and AI infrastructure as the underlying architecture to provide enterprise solutions.</td></tr>
<tr><td>Assessment Methods in Alignment with Intended Learning Outcomes</td><td>
<table>
<tr><td rowspan="2">Specific assessment methods/tasks</td><td rowspan="2">% weighting</td><td colspan="6">Intended subject learning outcomes to be assessed (Please tick as appropriate)</td></tr>
<tr><td>a</td><td>b</td><td>c</td><td>d</td><td>e</td><td>f</td></tr>
<tr><td>1. Continuous Assessments</td><td>30%</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td></tr>
<tr><td>2. Final Examination</td><td>70%</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td></td><td></td></tr>
<tr><td>Total</td><td>100 %</td><td colspan="6"></td></tr>
</table>
Explanation of the appropriateness of the assessment methods in assessing the intended learning outcomes:

Students taking the subject will be assessed by performance in two parts: continuous assessments and examination. Continuous assessment may include in-class discussions / quizzes, assignments, and tests and / or exams.</td></tr>
</table>

<table>
<tr><td></td><td colspan="2">The in-class discussions and quizzes engage students to actively participate in learning during lectures and tutorials. Students are to collaboratively work together to apply what they have learned in the class to solve practical problems. Assignment may include group projects that are designed to help students to work together in a small group to solve practical case studies and examples by applying concepts that are taught in the class. The results are to be presented in the form of reports and / or presentations. Tests and assignments are designed to help students reinforced their understanding of concepts and principles that are taught in the class. They are conducted to assess independent problem solving and critical thinking skills.</td></tr>
<tr><td rowspan="5">Student Study Effort Expected</td><td>Class contact:</td><td></td></tr>
<tr><td>▪ Lectures, Tutorials / Labs</td><td>39 Hrs.</td></tr>
<tr><td>Other student study effort:</td><td></td></tr>
<tr><td>▪ Assignment, Projects, Reading and Exam</td><td>66 Hrs.</td></tr>
<tr><td>Total student study effort</td><td>105 Hrs.</td></tr>
<tr><td>Reading List and References</td><td colspan="2"><u>Reference Books for Cloud Computing:</u>
1. “Essentials of Cloud Computing”, by Chellammal Surianarayanan and Pethuru Raj Chelliah. Publisher: Springer, edition: 2019.
2. “Cloud Computing Solutions Architect: A Hands-On Approach”, by Arshdeep Bahga and Vijay Madisetti. Arshdeep Bahga & Vijay Madisetti, 2019.
3. Articles from web, technical journals, and conference proceedings will be handed out or posted on L@PU Blackboard when needed.
<u>Reading List for AI Infrastructure:</u>
1. Model Training and Fine-Tuning:
(1) PipeDream: Generalized Pipeline Parallelism for DNN Training (SOSP'19)
(2) Alpa: Automating Inter- and Intra-Operator Parallelism for Distributed Deep Learning (OSDI'22)
2. Model Inference:
(1) Orca: A Distributed Serving System for Transformer-Based Generative Models (OSDI'22)
(2) Efficient Memory Management for Large Language Model Serving with PagedAttention (SOSP'23)
(3) https://github.com/xlite-dev/Awesome-LLM-Inference?tab=readme-ov-file#Mixture_of_Experts_LLM_Inference</td></tr>
</table>