

APPLIED ACCELERATED ARTIFICIAL INTELLIGENCE

MULTI-FACULTY

INTENDED AUDIENCE: Ph.D Scholars (any stream of Science or Engineering); Post Graduate Students (any stream of Science or Engineering); 3rd and 4th year UG Students (any stream of Engineering); Non-Students and Working Professionals

PREREQUISITES: Prior knowledge of Computer Organization, High-Performance; Computing, Machine Learning and Deep learning is desirable

INDUSTRY SUPPORT: Companies working in the domains of Machine Learning and Artificial Intelligence

COURSE OUTLINE:

This course will cover the fundamentals of the compute capabilities and the system software required for implementing artificial intelligence (AI) based solutions with accelerated computing on industrial use cases such as the one in the domains of Smart City. The course will discuss end to end deployments of industrial use cases like using large language models with demonstration, and hence will help participants use state-of-the-art AI SDKs effectively to solve complex problems.

ABOUT INSTRUCTOR:

Prof. Satyajit Das is an Assistant Professor in the Department of Computer Science and Engineering, IIT Palakkad. He received his joint Ph.D. degree from University of South Brittany (UBS), France, and University of Bologna (UniBo), Italy. Prior to joining IIT Palakkad, he was a postdoctoral fellow at LabSTICC, UBS. His research spans the areas of architecture, methods, and tools for embedded systems, including CGRAs, custom processors, multi-cores, high-level synthesis, and compilers. The main focus of Dr. Das's research is to implement highly energy efficient solutions for digital architectures in the domain of heterogeneous and reconfigurable multi-core System on Chips (SoCs). This includes architectures, design implementation strategies, runtime, and compilation support.

Prof. Satyadhyam Chickerur received his B.E degree in E&C, M.Tech in CSE and PhD in Computer and Information Sciences. He is presently Professor at School of Computer Science and Engineering and head of Centre for High Performance Computing at KLE Technological University, Hubli. He has served as faculty in various engineering colleges in India. He is a member of ISTE, IEEE and ACM. He was the Execom member of IEEE signal processing society, Bangalore chapter (2007-2009). He was a Member of Intel - IISC - VTU multicore Curriculum Development committee. He was one of the judges and problem setter for ACM ICPC programming contest of the Asia Regional's in the year 2007 and 2008. He has received various grants from industry and other Organisations for research as well.

Prof. Bharatkumar Sharma obtained a master's degree in information technology from the Indian Institute of Information Technology, Bangalore. He has around 10 years of development and research experience in the domains of software architecture and distributed and parallel computing. He is currently working with NVIDIA as a senior solutions architect, South Asia.

Prof. Tosin Adesuyi currently works at the Department of Computer and Software Engineering, Kumoh National Institute of Technology. Prof. Tosin does research in Deep Learning, Data Science, Data Mining and Computer Security. Their current project is Privacy in deep neural networks '

COURSE PLAN:

Week 1: Introduction to AI Systems Hardware part 1, Introduction to AI Systems Hardware part 2, Introduction to AI Accelerators, GPUs, Introduction to Operating Systems, Virtualization, Cloud part 1, Introduction to Operating Systems, Virtualization, Cloud part 2

Week 2: Introduction to Containers and IDE Dockers part1, Introduction to Containers and IDE Dockers part 2, Scheduling and Resource Management part 1, Scheduling and Resource Management part 2, DeepOps: Deep Dive into Kubernetes with deployment of various AI based Services Part 1, DeepOps: Deep Dive into Kubernetes with deployment of various AI based Services Part 2

Week 3: DeepOps: Deep Dive into Kubernetes with deployment of various AI based Services Session II part1, DeepOps: Deep Dive into Kubernetes with deployment of various AI based Services Session II part 2, Fundamentals of Distributed AI Computing Session 1 Part 1, Fundamentals of Distributed AI Computing Session 1 Part 2, Fundamentals of Distributed AI Computing Session 2 Part 1, Fundamentals of Distributed AI Computing Session 2 Part 2

Week 4: Introduction to Pytorch part 1, Introduction to Pytorch part 2, Introduction to Pytorch part 3, Introduction to Pytorch part 4, Profiling with DLProf Pytorch Catalyst part 1, Profiling with DLProf Pytorch Catalyst part 2

Week 5: Introduction to TensorFlow part 1,Introduction to TensorFlow part 2,Accelerated TensorFlow,Accelerated TensorFlow,Accelerated TensorFlow - XLA Approach,Accelerated TensorFlow - XLA Approach

Week 6: Optimizing Deep learning Training: Automatic Mixed Precision part 1,Optimizing Deep learning Training: Automatic Mixed Precision part 2,Optimizing Deep learning Training: Transfer Learning part 1,Optimizing Deep learning Training: Transfer Learning part 2

Week 7: Distributed Deep Learning using Tensorflow and Horovod ,Challenges with Distributed Deep Learning Training Convergence , Fundamentals of Accelerating Deployment part 1,Fundamentals of Accelerating Deployment part 2

Week 8: Accelerating neural network inference in PyTorch and TensorFlow part 1,Accelerating neural network inference in PyTorch and TensorFlow part 2,Accelerated Data Analytics part 1,Accelerated Data Analytics part 2,Accelerated Data Analytics part 3,Accelerated Data Analytics part 4,Accelerated Machine Learning

Week 9:Scale Out with DASK,Web visualizations to GPU accelerated crossfiltering part 1,Web visualizations to GPU accelerated crossfiltering part 2, Accelerated ETL Pipeline with SPARK part 1, Accelerated ETL Pipeline with SPARK part 2

Week 10: Introduction to NLP part 1 ,Introdcution to NLP part 2

Week 11: Applied AI: Smart City (Intelligent Video Analytics) Session 1 part 1,Applied AI: Smart City (Intelligent Video Analytics) Session 1 part 2,Applied AI: Smart City (Intelligent Video Analytics) Session 2 Deepstream part 1,Applied AI: Smart City (Intelligent Video Analytics) Session 2 Deepstream part 2

Week 12: Text Classification,Machine Translation,Introduction to Large Language Models and Transformer Architecture,Pre-training and Fine-tuning and Bias and fairness in large language models