Indian Institute of Technology Kanpur
Department of Computer Science and Engineering

Proposal For a Course

Course Title: System Level Data Formats and Representations for AI
Course No.: CS6XX
Credits: 3-0-0-9
Prerequisites: CS220
Who can take the course: PhD, Masters, 3rd and 4th year UG Students
Proposer: Urbi Chatterjee
Other Interested Faculty: Debapriya Basu Roy
Departments that may be interested: CSE, EE, DIS

Course Objective: Artificial Intelligence workloads have fundamentally transformed the way computing systems are designed, optimized, and deployed. Traditional numeric representations such as IEEE FP32, originally developed for scientific computing, are increasingly inefficient for modern AI systems that demand high throughput, low energy consumption, reduced memory footprint, and scalable training across distributed platforms. This course aims to provide a comprehensive system-level understanding of data formats and numerical representations that enable efficient AI computation across hardware and software layers.

The primary objective of this course is to equip students with the theoretical foundations and practical insights required to analyze, design, and evaluate data formats tailored for AI workloads. The course begins by examining the evolution of number representations, from fixed-point and floating-point to emerging AI-specific formats, while grounding discussions in the numerical characteristics of neural networks, including weight, activation, and gradient distributions. Students will develop a deep understanding of IEEE floating-point limitations and explore alternatives such as FP16, BFloat16, FP8, mixed-precision and trans-precision training techniques, block floating point, and micro-scaling formats.

Beyond conventional formats, the course introduces advanced and emerging representations including quantization schemes (uniform, non-uniform, ultra-low precision), posit and universal number systems, logarithmic and stochastic formats, and sparse tensor representations. Emphasis is placed on hardware–software co-design, accelerator architectures, approximate computing, distributed training representations, and memory-efficient model storage strategies.

By the end of the course, students will be able to critically evaluate trade-offs between accuracy, stability, energy efficiency, and hardware complexity; design representation-aware AI pipelines; and understand how modern accelerators integrate numeric innovations. The course prepares students to contribute to next-generation AI hardware, large-scale model training, and efficient edge intelligence systems.

Course Contents:

Lecture No.	**Topic / Key Focus**
1	Evolution of data representation: unsigned/signed, fixed-point, config-point, floating-point, AI formats; motivation for new formats
2	Numerical characteristics of neural networks: distributions of weights, activations, gradients
3	AI hardware perspective: GPUs, TPUs, accelerators, tensor cores and numeric precision
4	Fixed-point and Floating-Point arithmetic fundamentals and scaling
5	Rounding errors, ULP, numerical stability in large computations

Lecture No.	Topic / Key Focus
6	Why FP32 is inefficient for AI: memory, energy, compute overhead, CORDIC alternative
7	FP16 format and its role in deep learning acceleration
8	BFloat16 format and training stability for large models
9	FP8 formats (E4M3, E5M2) and trade-offs
10	Mixed precision training pipelines
11	Trans-precision and hybrid training techniques
12	Motivation for micro-scaling formats in modern AI training
13	Block floating-point representation and shared exponents
14	Architecture of MX formats (MXFP, MXINT)
15	Hardware implementation challenges of micro-scaling
16	Case study: micro-scaling in large model training
17	Fundamentals of quantization in neural networks
18	Uniform quantization and INT8 inference
19	Non-uniform and logarithmic quantization
20	Post-training quantization methods
21	Quantization-aware training techniques
22	Ultra-low precision models (4-bit / 3-bit / 2-bit AI models)
23	Introduction to posit number systems
24	Posit vs IEEE floating-point comparison
25	Hardware architecture of posit arithmetic
26	Logarithmic number systems
27	Emerging universal number systems
28	Stochastic rounding principles
29	Probabilistic computing methods
30	Hardware support for stochastic arithmetic
31	Impact of stochastic formats on AI training
32	Sparsity in neural networks and pruning
33	Sparse tensor formats (CSR, CSC, block sparse)
34	Format-specific model compression
35	Sparse AI accelerator architectures
36	Representation- and hardware-aware pruning
37	Numerical challenges in large language models
38	Memory-efficient training strategies, Distributed training data representations
39	Efficient checkpointing and model storage
40	Hardware–software co-design principles for AI, Approximate computing for AI

Books:

1.Jean-Michel Muller, Nicolas Brisebarre, Florent de Dinechin, Claude-Pierre Jeannerod, Vincent Lefèvre, Guillaume Melquiond, Nathalie Revol, Damien Stehlé, and Serge Torres. 2009. Handbook of Floating-Point Arithmetic (1st. ed.). Birkhäuser Basel.

2. Michael L. Overton. 2001. Numerical computing with IEEE floating point arithmetic. Society for Industrial and Applied Mathematics, USA.

References:

1. David Mallasén, Raul Murillo, Guillermo Botella, and Alberto Antonio Del Barrio. 2025. Navigating Posit Arithmetic: A Comprehensive Survey of Principles, Hardware, and Applications. ACM Comput. Surv. 58, 5, Article 131 (April 2026), 36 pages. https://doi.org/10.1145/3772284

Urbi Chatterjee

Proposer: Urbi Chatterjee
Dated: 01/03/2026

DPGC Convener

Chairman SPGC

DOAA