CSC290/420 Machine Learning Systems for Efficient AI
Evolution

Sreepathi Pai

November 18, 2026

URCS

Outline

Programs

Machines

0 A.C.

Current Bottlenecks

AI Bubble Bursts?

Outline

Programs

Machines

0 A.C.

Current Bottlenecks

AI Bubble Bursts?

High-level programming models

- Dataflow Graphs
  - TensorFlow
  - PyTorch
  - etc.
- Unlikely to change.
- Some of the operators will become more specalized
  - Attention, for example

Low-level programming models for Operators

- Assembly language
- C/C++ and SIMD
- CUDA and other GPU programming models
- Device-specific programming models

High-Level Low-Level Programming Models

- Generate low-level code
- Loop Based languages
- Tensor languages
  - TVM, etc.
- Tile-based languages
- Asynchronous Data Flow
  - AMD NPU
  - AWS Trainium(?)
  - (also Newer GPUs)
- ?

Outline

Programs

Machines

0 A.C.

Current Bottlenecks

AI Bubble Bursts?

Edge Compute

- AI on “edge” devices
- AI on microcontrollers
- Neural Processing Units (NPUs)
    - Intel, AMD, etc.

Beyond GPUs

- Training
  - TPUs
  - Trainium
- Inference
  - CPU + SIMD + Matrix Multiply Accelerators
- Beyond NVIDIA GPUs: AMD, Qualcomm, etc.
- Microsoft Maia, OpenAI Jalapeño, Meta MTIA, etc.

Implications of Heterogeneity

- "Write Once, Run Anywhere"
  - Old Java Slogan
- The "Moat"
  - "CUDA Moat"
  - After agents: Architecting for Hardware Diversity: Meta's End-to-End Co-Design for High-Performance AI Training on AMD GPUs

Outline

Programs

Machines

0 A.C.

Current Bottlenecks

AI Bubble Bursts?

After Claude

- The version of Claude released Jan 2026 was significantly better
  - Not AGI
- Could write code
  - Not very good at it, but better than most
- "Agentic AI"
  - AI "conversing" with each other and other systems

Would AI make this class obsolete?

- KernelBench: can LLMs write efficient GPU kernels?
  - Milestones: 100% correct, then "speed-of-light" kernels
- Meta, OpenAI, and others report kernels written for their own hardware
  - Much faster than humans writing code
  - How GPT-5.6 fuses frontier intelligence with frontier efficiency
- AI directly writing assembly language?
  - AI as a Compiler: Compiling Triton Kernels without the Triton Compiler
- Coming soon: "Recursive Self Improvement"?

Outline

Programs

Machines

0 A.C.

Current Bottlenecks

AI Bubble Bursts?

Reinforcement Learning

- Deep Reinforcement Learning is an unsupervised learning technique
- Can be used to train LLMs
    - "RLHF" – reinforcement learning from human feedback
- DeepSeek
    - RL from a "reasoning model"

Federated Learning

- Motivated largely by privacy
- “Data doesn’t leave your phone”

Homomorphic Encryption

- Mathematical operations on encrypted, private data
- Computation never sees the data
- Paper: Orion: A Fully Homomorphic Encryption Framework for Deep Learning

Small Language Models

- Pre-training with fewer parameters
- Model compression of a larger model
- “Distillation”

Outline

Programs

Machines

0 A.C.

Current Bottlenecks

AI Bubble Bursts?

Classifying AI Applications

- Sensing
  - Computer vision
  - Text-to-speech
- "Intelligent"
  - Chatbots
  - Proofs
  - Code writing

2027:

- Assume
  - OpenAI is no more, and took down the whole industry
  - Anthropic's IPO was below expectations
- A brutal AI winter is upon us.
- PyTorch development has been abandoned

Characteristics of AI Applications

- Lots of compute
- Large amounts of Data
- Significant Communication

Computational Science

- Simulate physical laws
  - i.e., solve equations numerically
- Now with AI-built models
  - Weather prediction
- "Foundation Models for Science"
  - Genomics
  - Chemistry

Big Data

- Big data analysis (e.g., bioinformatics)

Conclusion

- Most techniques introduced in this class can be used for high performance