The City College of New York

CSC 36000:
Modern Distributed Computing *with AI Agents*

By Saptarashmi Bandyopadhyay
Email: sbandyopadhyay@ccny.cuny.edu
Assistant Professor of Computer Science
City College of New York and Graduate Center at City University of New York

Today's Lecture

Reliable and Resilient Distributed AI

- Consistency-Availability Trade-off
- Replication for High-Availability AI

Optimizing Performance and Efficiency

- The Energy Footprint
- Quantization
- Pruning

Live Coding Demonstration

Reliable and Resilient Distributed AI

Consistency

Consistency Models

- We need a way to make sure changes that happen in *one* node are reflected across *all* nodes
- This is accomplished using a *consistency model*: a fundamental contract that governs how and when changes become visible across all nodes of a distributed system
- A consistency model is core to the correctness, performance, and availability of a system, especially in Distributed AI systems
- There is a broad spectrum of consistency models ranging from strict immediate consistency to relaxed eventual consistency

The Spectrum of Consistency

There are two prominent models for consistency:

- **Strong Consistency:** Once a write operation completes, *any* subsequent read, regardless of the node it's directed to, will return the value of this write or a subsequent one
    - Unified and Up-to-date but introduces latency
- **Eventual Consistency:** Allows for *temporary* inconsistency. If no new updates are made to a data item, all its replicas will eventually converge to the same state
    - Can be inconsistent at times but allows for lower latency and more availability

Various hybrid models exist in between, such as **Sequential Consistency** (all operations happen in-order) and **Causal Consistency** (related operations in-order, others can vary)

Real-world Use Case: AI Fraud Detection

- The choice of consistency model is extremely important for many applications
- For some, the *cost of staleness*, i.e. the downsides of using slightly old data can be steep
- What consistency model would you use for an AI system to detect fraud in banking?

Real-world Use Case: AI Recommender Systems

- For some applications like TikTok, user interaction data such as the number of likes powers personalized recommendations
- Which is more important, showing the exact number of likes or immediately being able to show a video?
- Would this be a low or high cost of staleness?
- Also it crashes your phone!
  - How much cookies should distributed apps like this store?!

Questions?