AI Metropolis: Scaling LLM Multi-Agent Simulation with Out-of-order Execution

Zhiqiang Xie†

Joint work with

Hao Kang*, Ying Sheng †, Tushar Krishna*, Kayvon Fatahalian† and Christos Kozyrakis †

† Stanford University *Georgia Institute of Technology

From LLM Agents to Agent Society

Please do a portrait of an LLM agent

Image created

From LLM Agents to Agent Society

Please do a portrait of an LLM agent

Image created

From LLM Agents to Agent Society

Please do a portrait of an LLM agent

Image created

OASIS

Automation or Simulation?

The Biggest Potent

Multi-agent Systen

CAMEL-AI

Project Sid: 1,000 AI Agents Living in Minecraft

Explore Project Sid where 1,000 AI Agents test their wits in the immersive world of Minecraft. Watch AI evolve in a gaming realm!

Case Studies News September 8, 2024

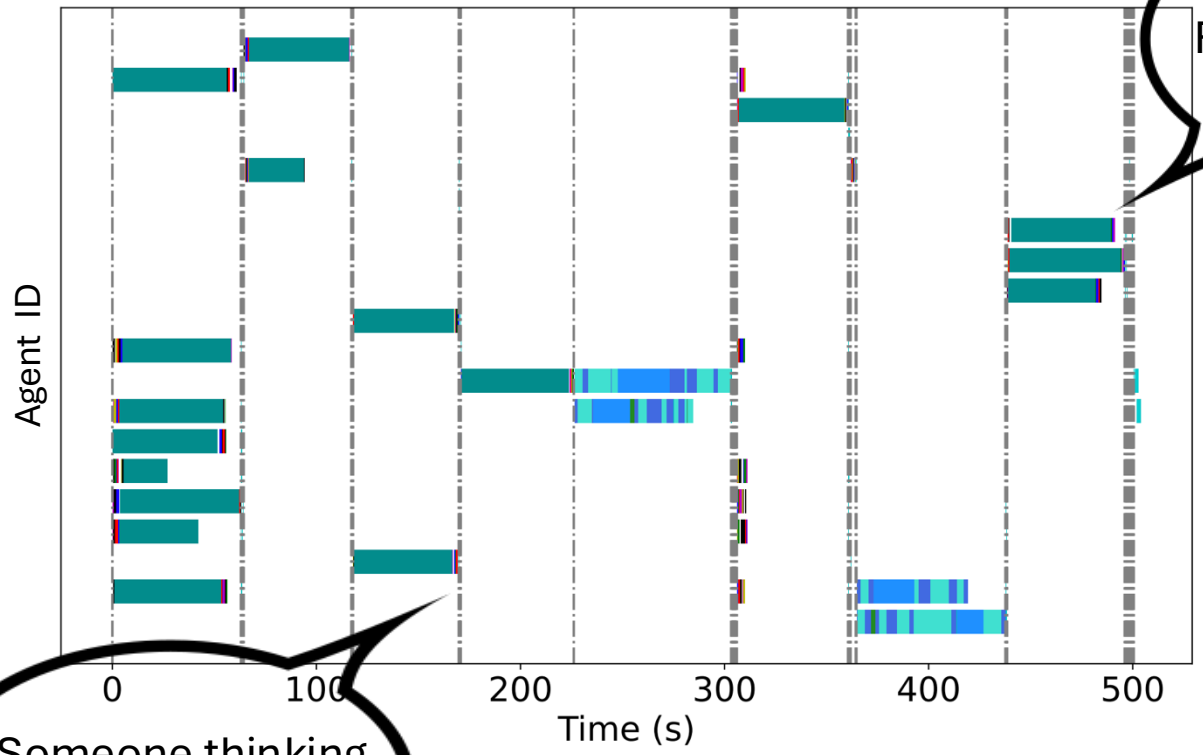
How is the Simulated World Constructed

Input: target_step, agents, world
Initialize: step ← 0
while step < target_step **do**
  actions ← [ ]
  for all agent **in** agents **do**
    actions.append(agent.***proceed***(world))
  end for
  world.***step***(actions)
  step ← step + 1
end while

Step function, a paradigm we long loved since training RL agents

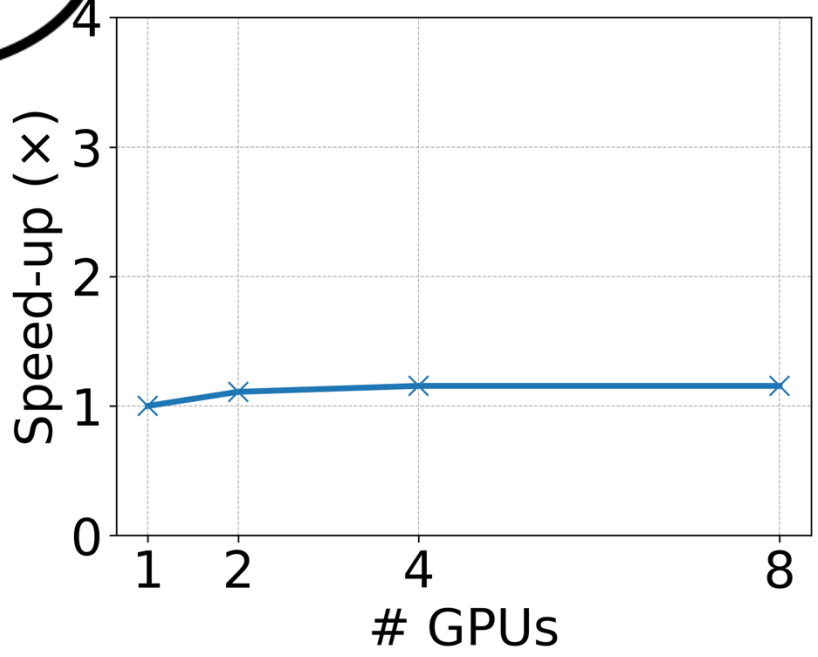
It does not SCALE!

Trace snippet of agent activities



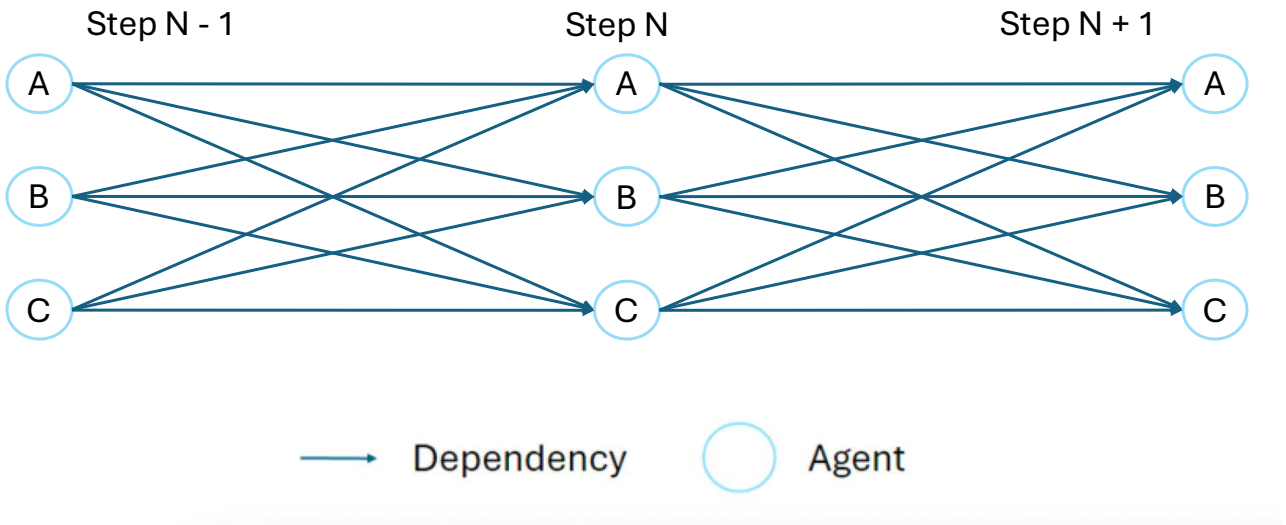
Folks chatting

Someone thinking about what to eat



GPU is not all your need : (

Causality and Dependency



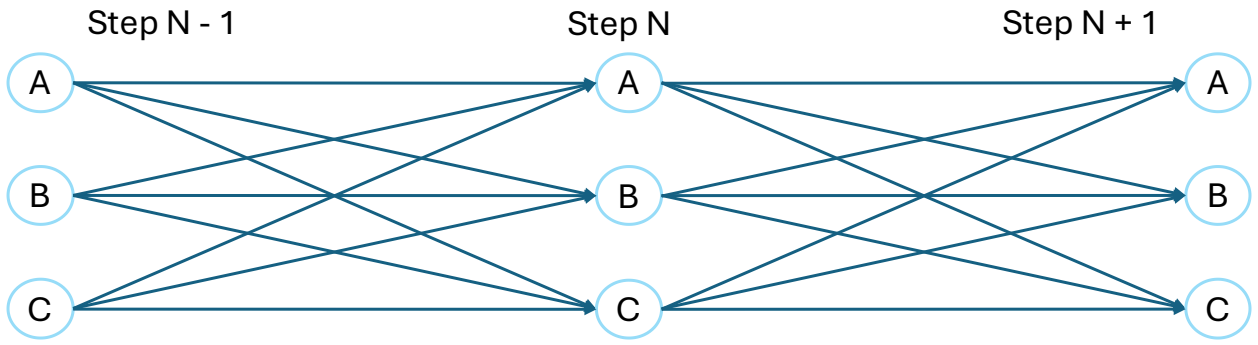
Global Synchronization

introduces excessive **Dependency**

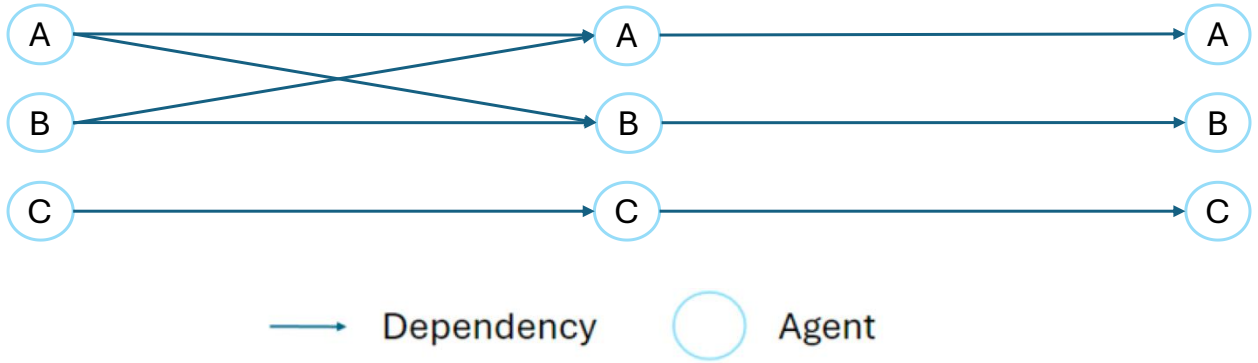
to enforce **Causality**

which limits **Parallelism**

Causality and Dependency



Real causality dependency might be:



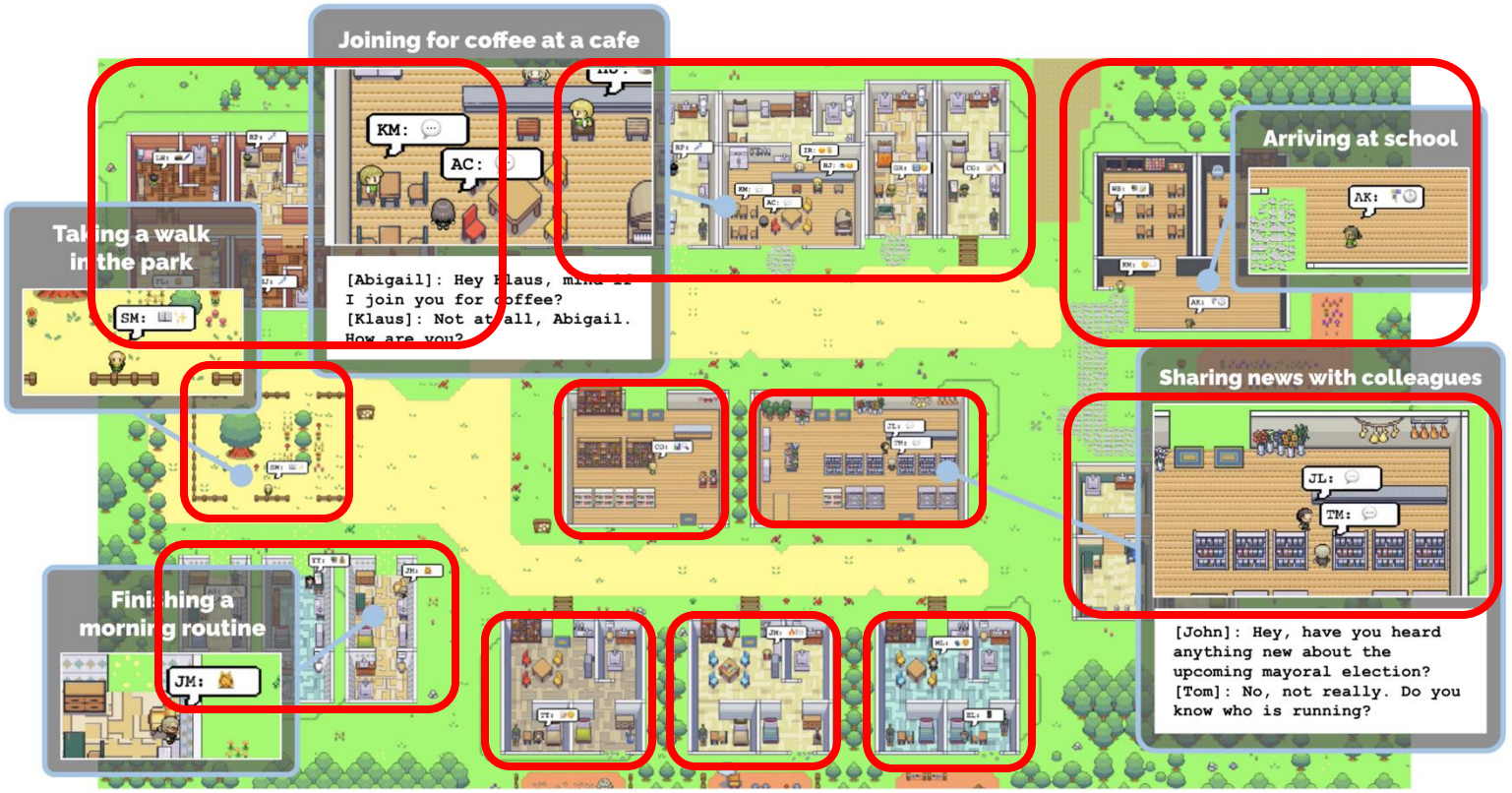
Global Synchronization

introduces excessive **Dependency**

to enforce **Causality**

which limits **Parallelism**

AI Metropolis



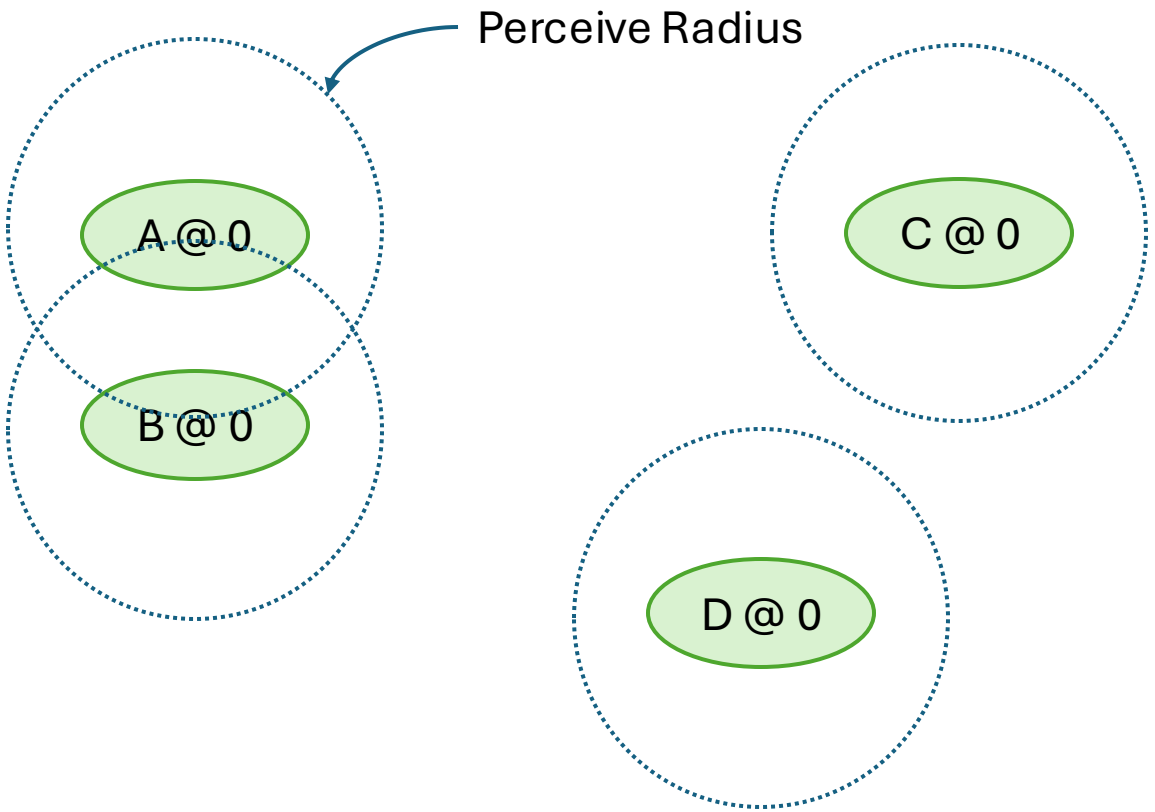
Smart **Dependency Tracking**

enables **Out-of-order execution**

which guarantees **Causality**

and release more **Parallelism**

Spatiotemporal Dependency Graph



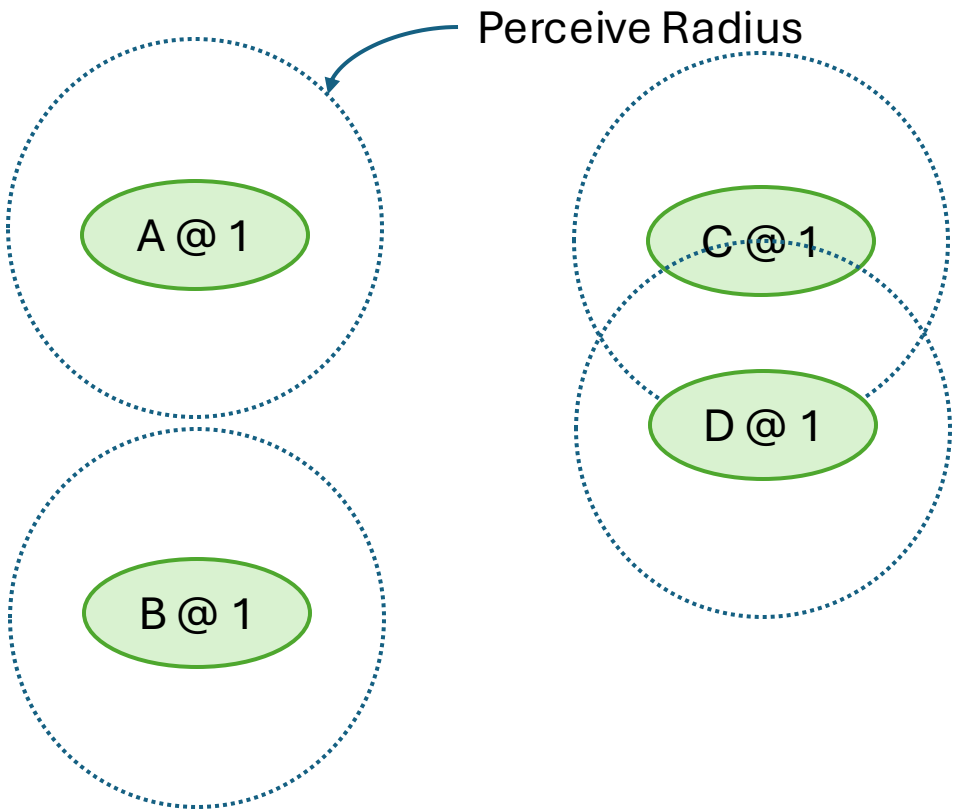
Events of Time 0:

- A: perceive B and act (8s)
- B: perceive A and act (8s)
- C: perceive nothing and act (1s)
- D: perceive nothing and act (1s)

Wall time: 8s

Average parallelism: 2.25

Spatiotemporal Dependency Graph



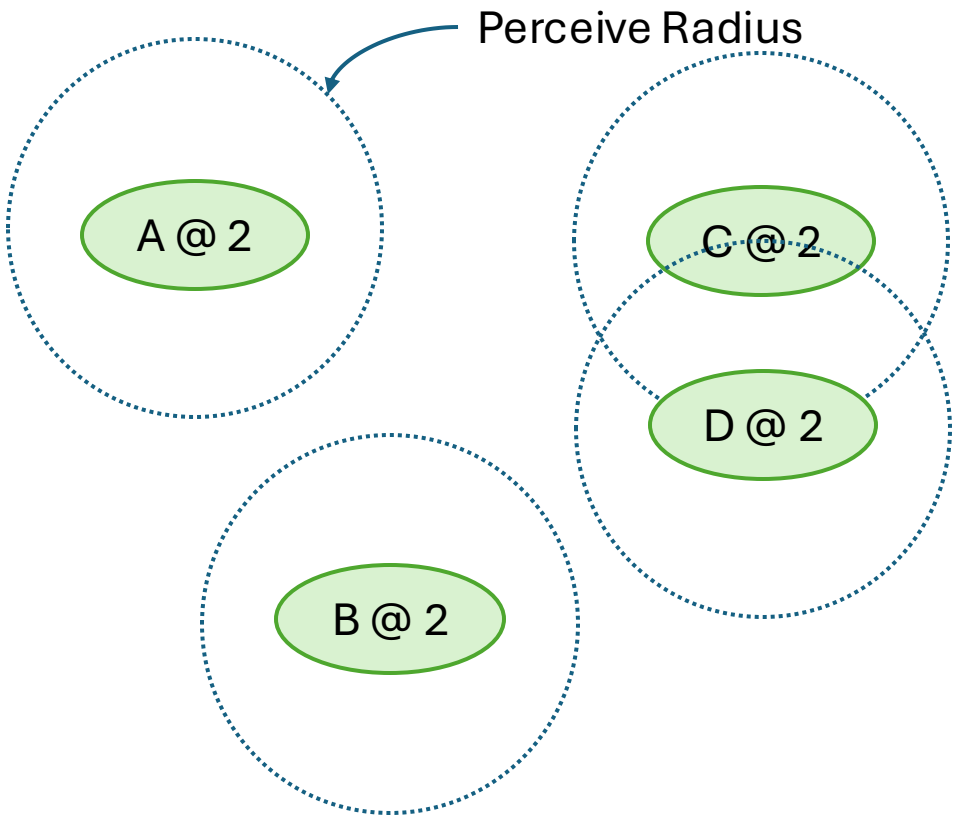
Events of Time 1:

- A: perceive nothing and act (1s)
- B: perceive nothing and act (1s)
- C: perceive D and act (10s)
- D: perceive C and act (10s)

Wall time: 10s

Average parallelism: 2.2

Spatiotemporal Dependency Graph



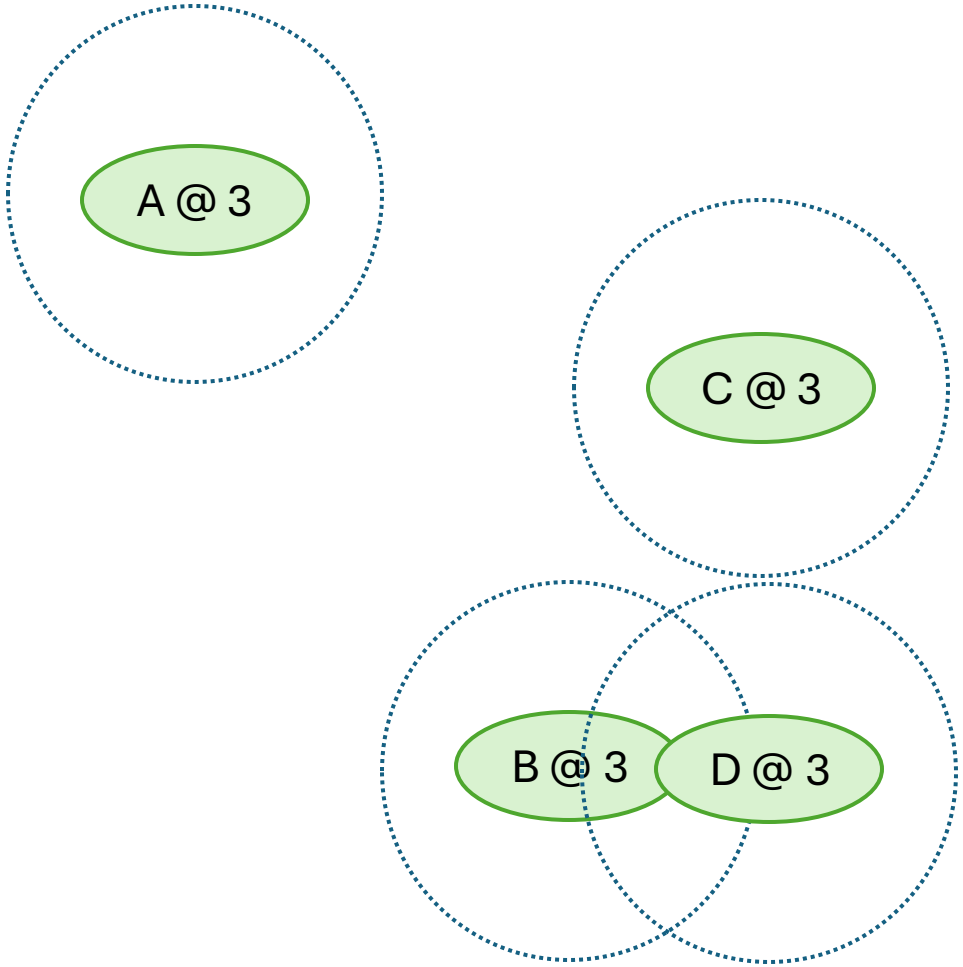
Events of Time 2:

- A: perceive nothing and act (1s)
- B: perceive nothing and act (1s)
- C: perceive D and act (2s)
- D: perceive C and act (2s)

Wall time: 2s

Average parallelism: 3

Spatiotemporal Dependency Graph



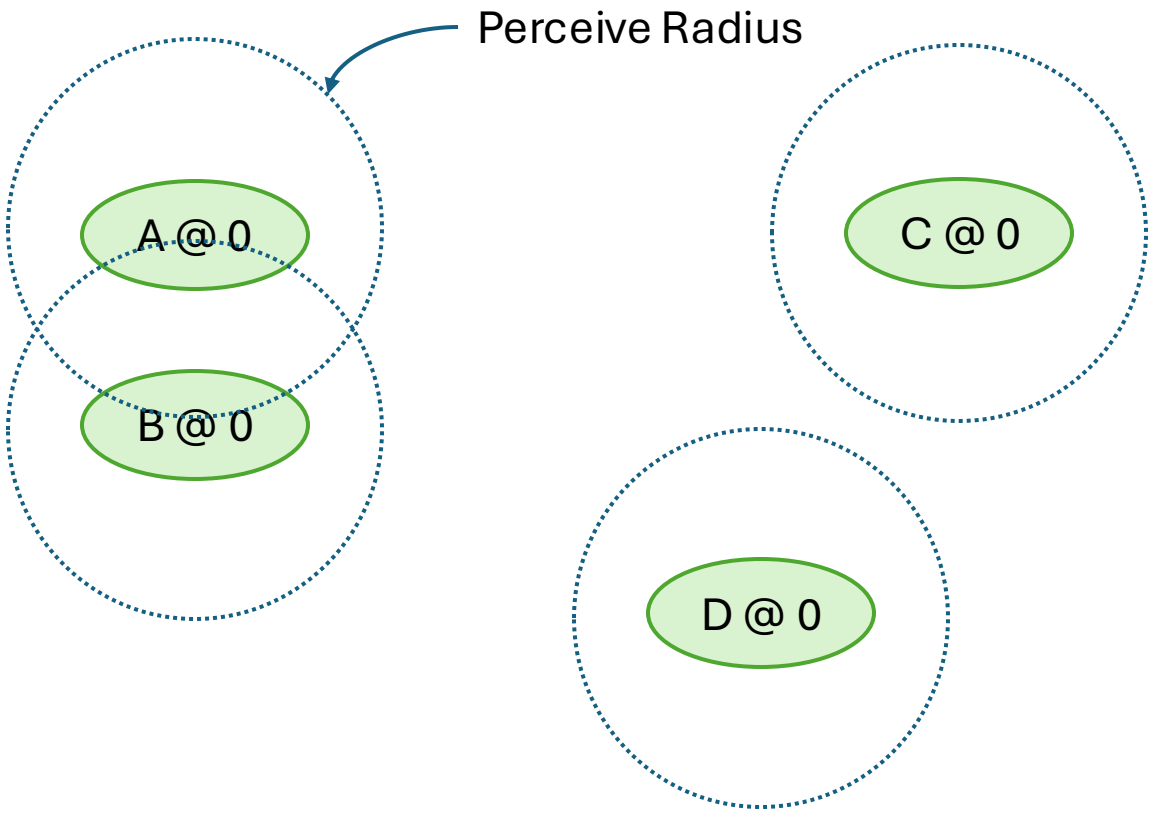
Events of Time 3:

- A: perceive nothing and act (1s)
- B: perceive D and act (2s)
- C: perceive nothing and act (1s)
- D: perceive B and act (2s)

Wall time: 2s

Average parallelism: 3

Spatiotemporal Dependency Graph



Events:

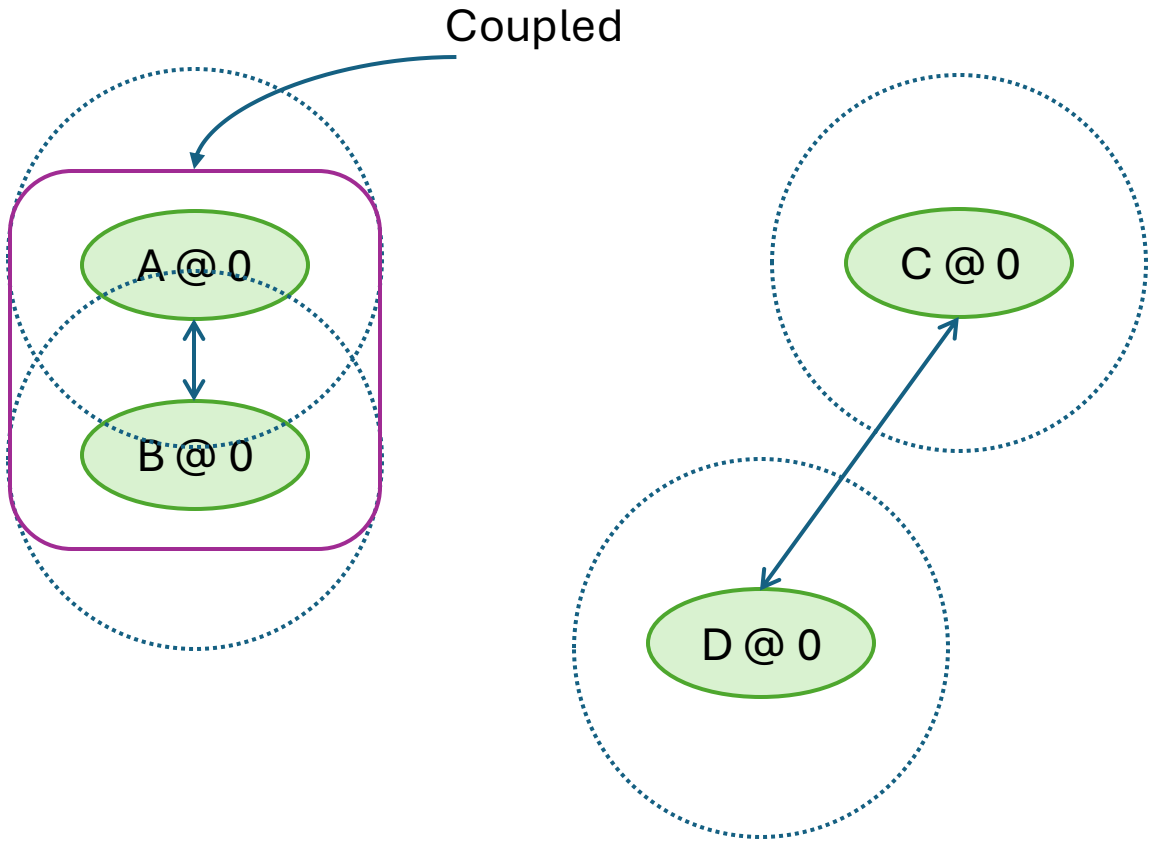
- Time 0 (0-8s): A <> B, C, D
- Time 1 (8-18s): C<> D, A, B
- Time 2 (18-20s): C<> D, A, B
- Time 3 (20-22s): B <> D, A, C

Wall time: 22s

Average parallelism: 2.36

Causal Violation: Agent observes an **out-of-sequence event** or misses an **expected event**.

Spatiotemporal Dependency Graph

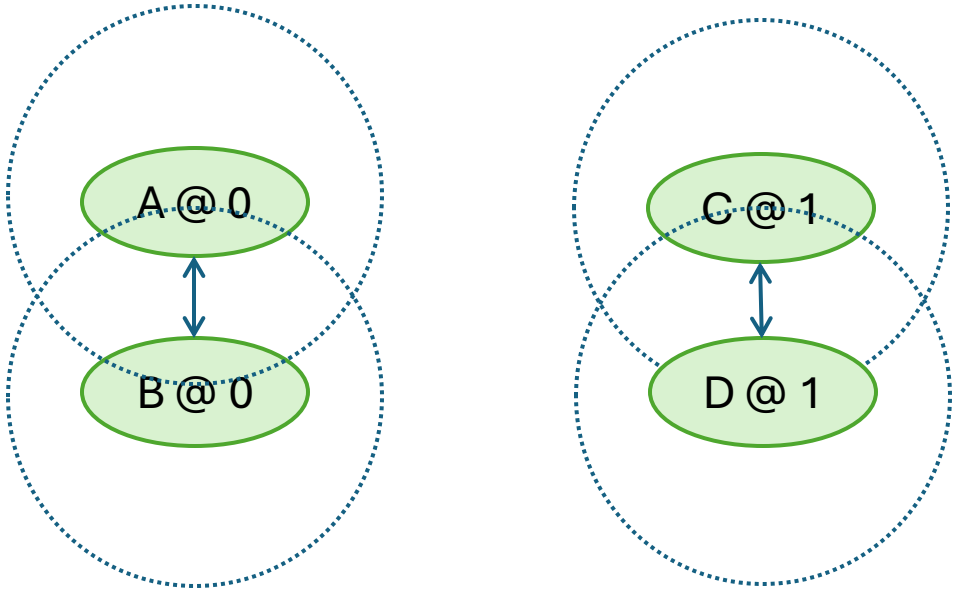


Events:

- Time 0: **A <> B** (0-8s), **C** (0-1s), **D** (0-1s)
- Time 1: C<> D, A, B
- Time 2: C<> D, A, B
- Time 3: B <> D, A, C

Causal Violation: Agent observes an **out-of-sequence event** or misses an **expected event.**

Spatiotemporal Dependency Graph

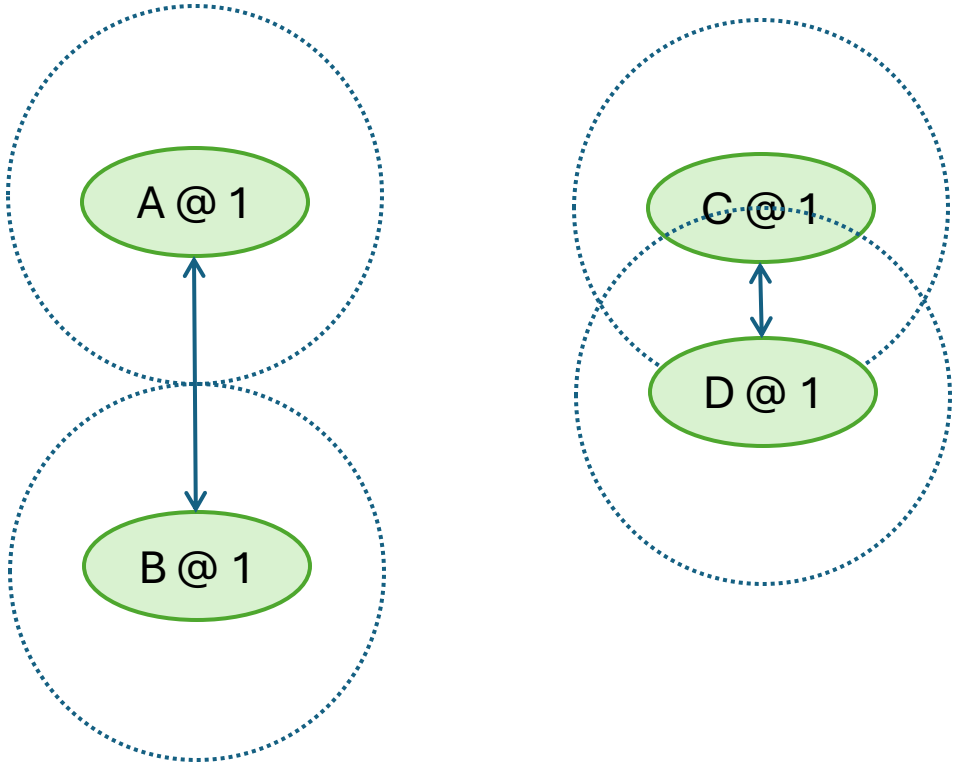


Events:

- Time 0: **A <> B** (0-8s), C (0-1s), D (0-1s)
- Time 1: **C<> D** (1-11s), A, B
- Time 2: C<> D, A, B
- Time 3: B <> D, A, C

Causal Violation: Agent observes an **out-of-sequence event** or misses an **expected event.**

Spatiotemporal Dependency Graph

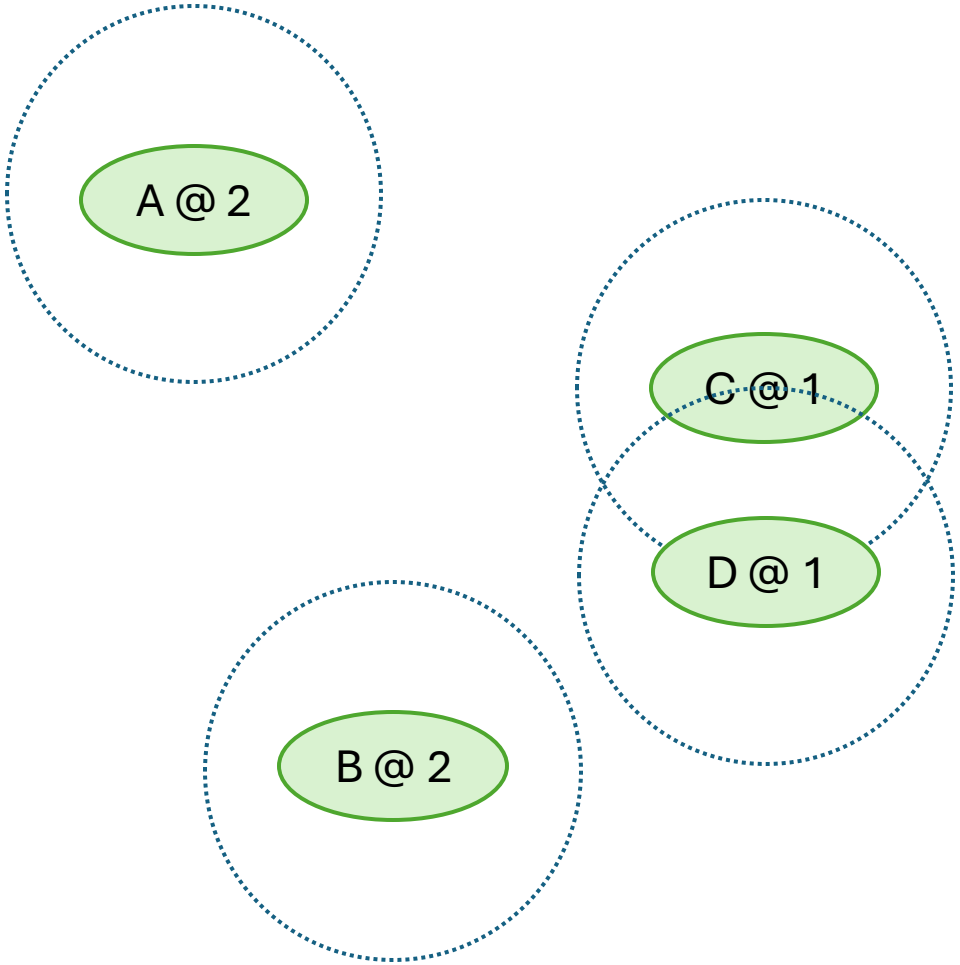


Events:

- Time 0: A <> B (0-8s), C, D (0-1s)
- Time 1: **C<> D** (1-11s), **A, B** (8-9s)
- Time 2: C<> D, A, B
- Time 3: B <> D, A, C

Causal Violation: Agent observes an **out-of-sequence event** or misses an **expected event.**

Spatiotemporal Dependency Graph

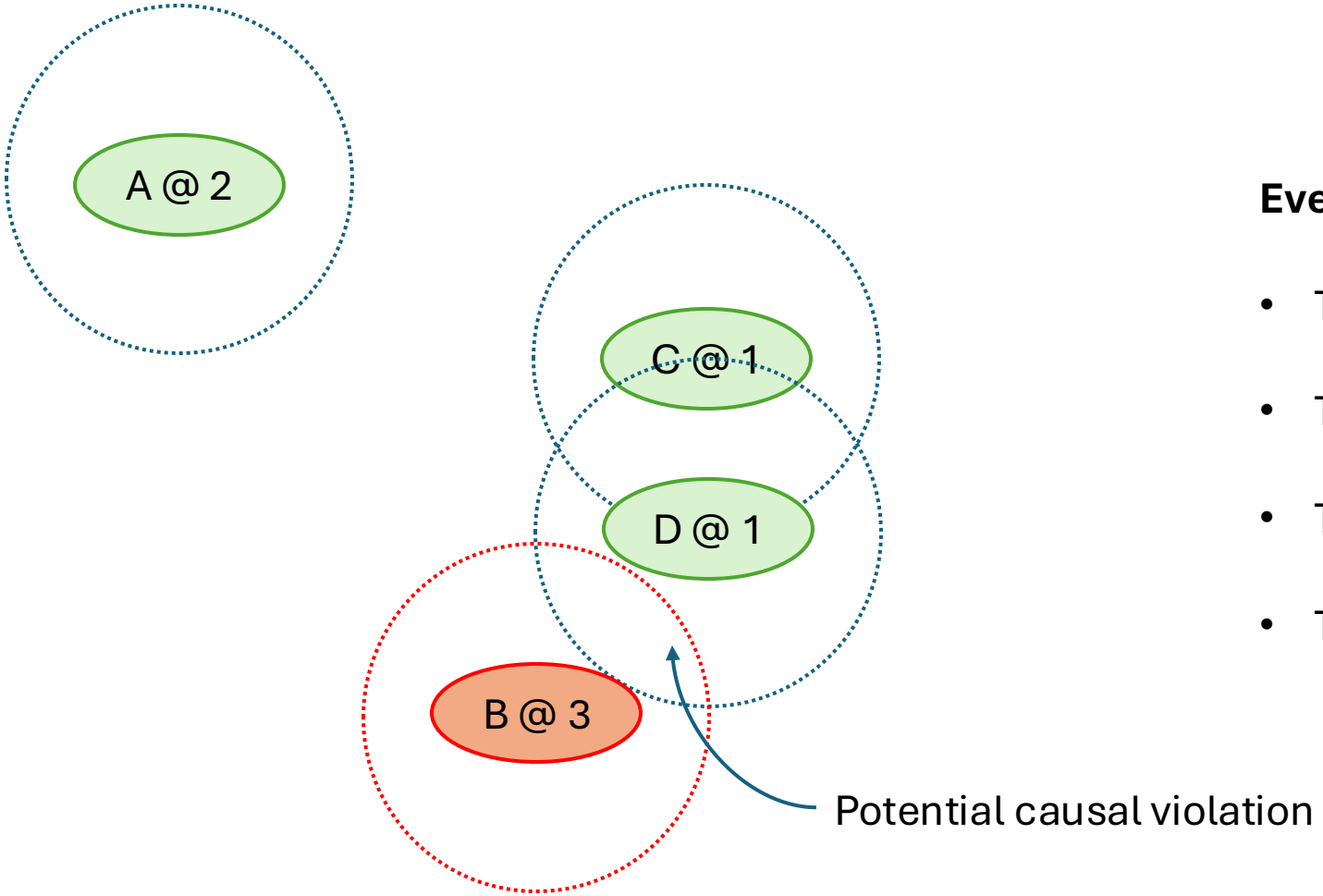


Events:

- Time 0: A <> B (0-8s), C, D (0-1s)
- Time 1: **C<> D** (1-11s), A, B (8-9s)
- Time 2: C<> D, **A, B** (9-10s)
- Time 3: B <> D

Causal Violation: Agent observes an **out-of-sequence event** or misses an **expected event.**

Spatiotemporal Dependency Graph

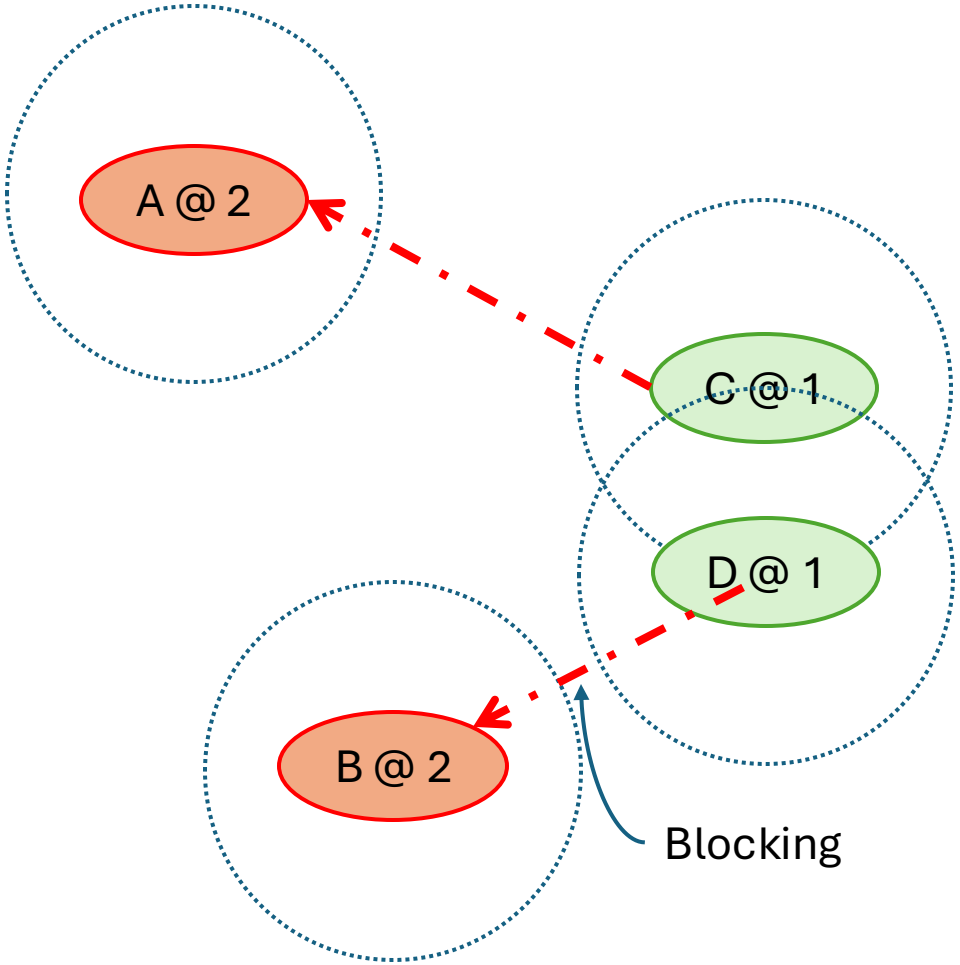


Events:

- Time 0: A <> B (0-8s), C, D (0-1s)
- Time 1: **C<> D (1-11s),** A, B (8-9s)
- Time 2: C<> D, A, B (9-10s)
- Time 3: **B** <> D, **A**, C

Causal Violation: Agent observes an **out-of-sequence event** or misses an **expected event.**

Spatiotemporal Dependency Graph

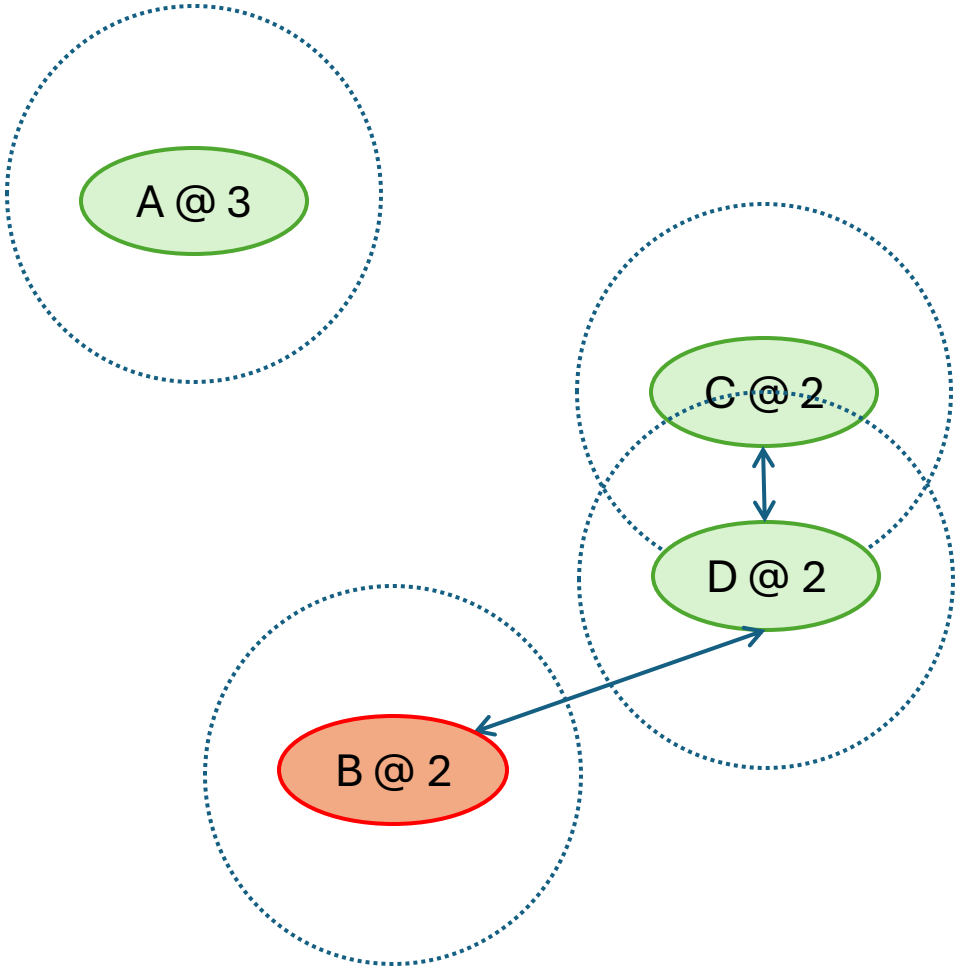


Events:

- Time 0: A <> B (0-8s), C, D (0-1s)
- Time 1: **C<> D (1-11s),** A, B (8-9s)
- Time 2: C<> D, A, B (9-10s)
- Time 3: **B** <> D, **A**, C

Causal Violation: Agent observes an **out-of-sequence event** or misses an **expected event.**

Spatiotemporal Dependency Graph

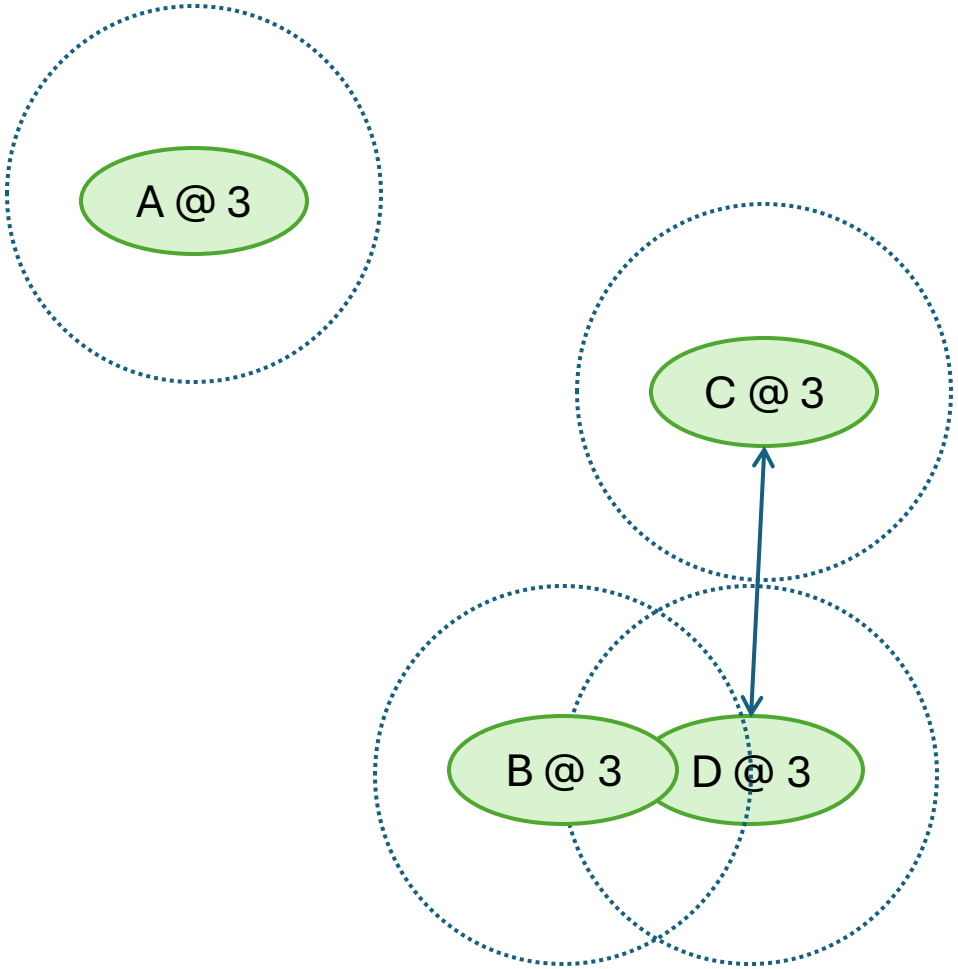


Events:

- Time 0: A <> B (0-8s), C, D (0-1s)
- Time 1: C<> D (1-11s), A, B (8-9s)
- Time 2: **C<> D (11-13s)**, A, B (9-10s)
- Time 3: **B** <> D, **A (11-12s)**, C

Causal Violation: Agent observes an **out-of-sequence event** or misses an **expected event.**

Spatiotemporal Dependency Graph



Events:

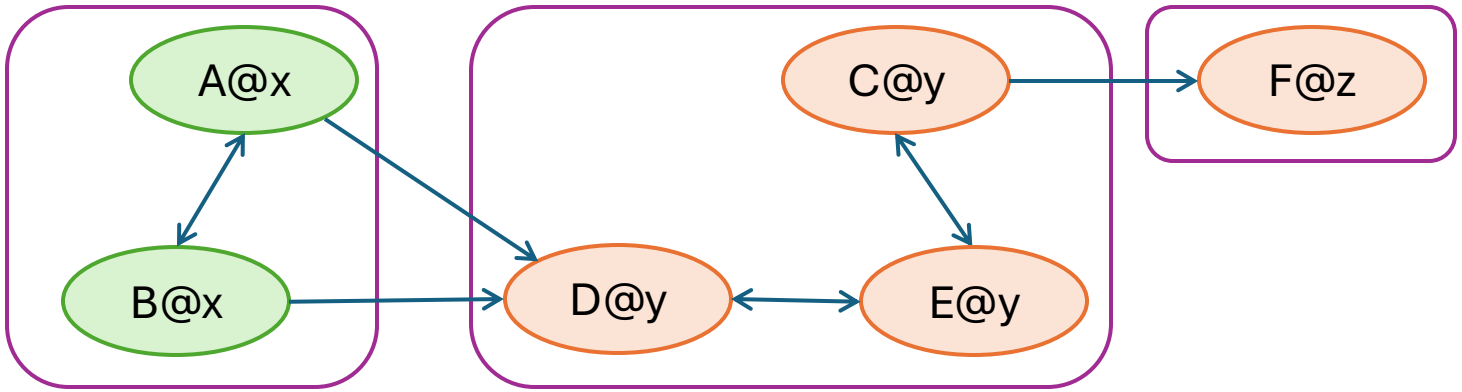
- Time 0: A <> B (0-8s), C, D (0-1s)
- Time 1: C<> D (1-11s), A, B (8-9s)
- Time 2: C<> D (11-13s), A, B (9-10s)
- Time 3: **B <> D(13s-15s)**, A (11-12s), **C (13-14s)**

Wall time: 15s

Average parallelism: 3.47

Causal Violation: Agent observes an **out-of-sequence event** or misses an **expected event.**

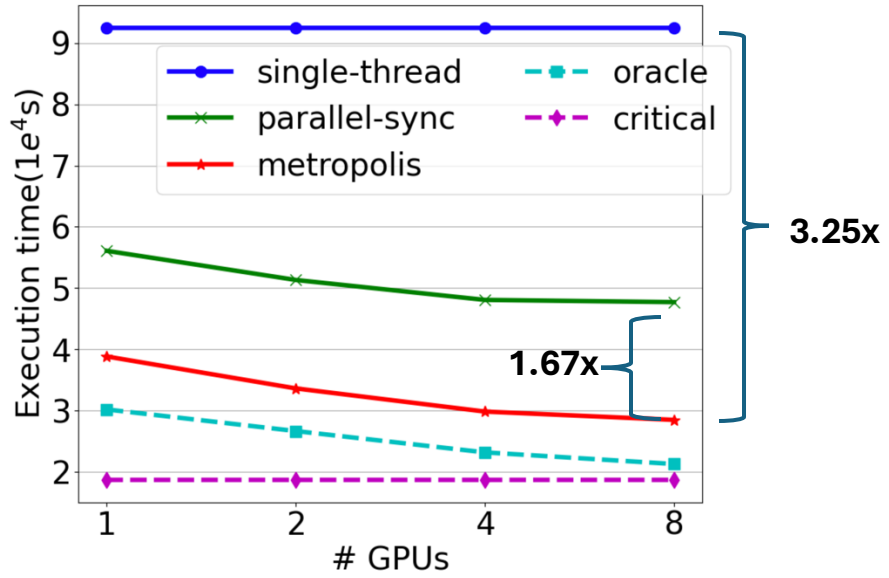
Spatiotemporal Dependency Graph



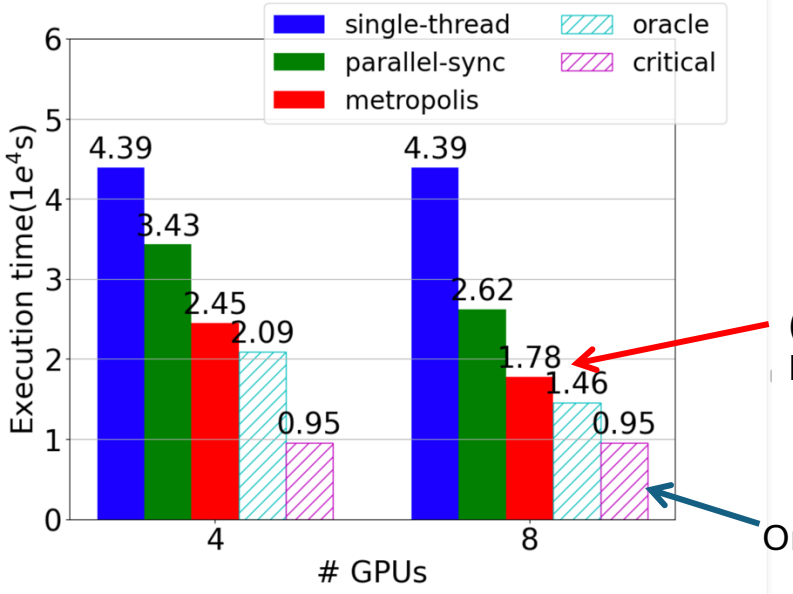
See paper for and more details:

- Proofs
- Efficient graph update
- Scalable implementation
- Priority scheduling

Full Day 25 Agent Simulation



(a) Simulation using Llama-3-8b-instruct on NVIDIA L4 GPU



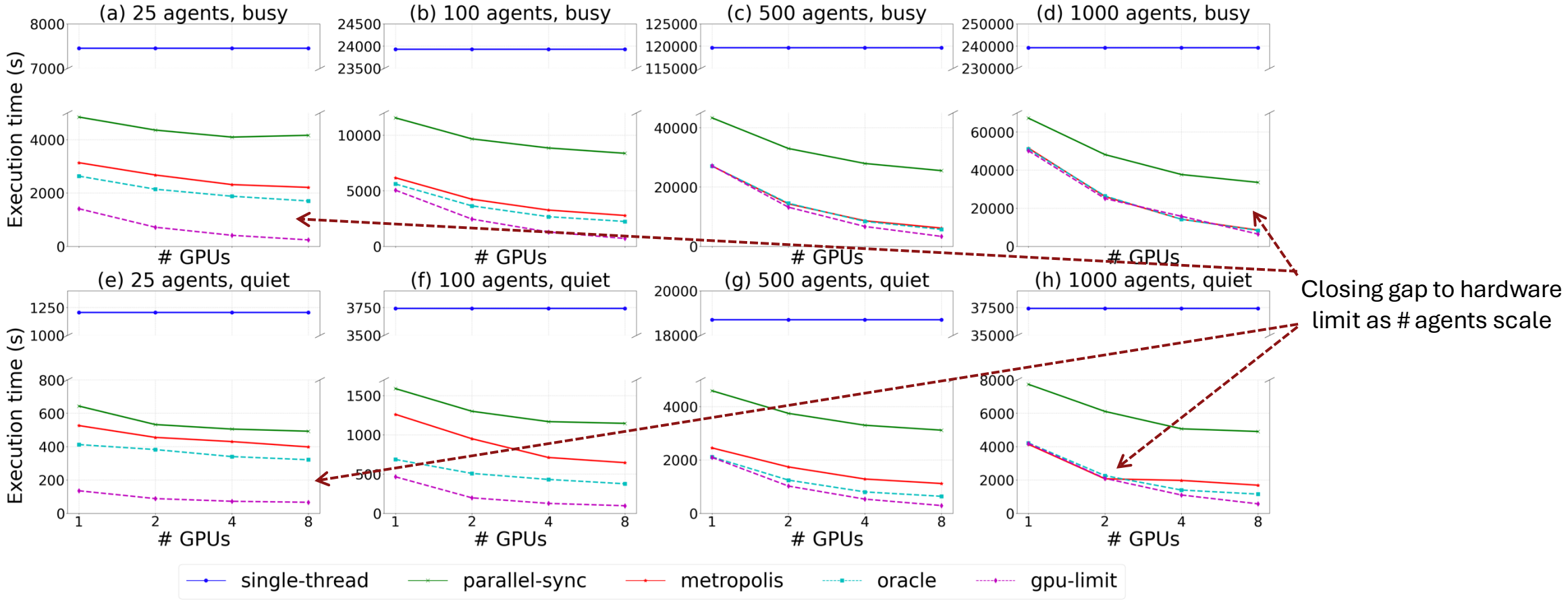
(b) Simulation using Llama-3-70b-instruct on NVIDIA A100 GPUs

Similar scalability as oracle (optimal scheduling, knowing all future events)

Only compute the tasks on longest causality path

- Up to **3.25x** faster than the original implementation (single-thread), **1.67x** than parallel implementation (parallel-sync)
- **74.7%** of oracle performance on 8 GPUs, **82.9%** on single

Scaling Up to 1000 Agents



- Up to **19.5x** faster than the original implementation (single-thread), **4.25x** than parallel implementation (parallel-sync) as # agents scale

To build an AI Society

Environments: Madrona, NVIDIA OMNIVERSE, UNREAL ENGINE

Scheduler: AI METROPOLIS

Serving Engines: SGL, TensorRT, vLLM

Why do we need a scheduler?

- Shared states in the environment require synchronization for **causality**
- Human-like interaction expect **faster** and more **predictable** response, contrary to LLM inference
- It utilize application-specific **dependency** and **priority** information to achieve better efficiency and user satisfaction

Prototype and traces will release soon!

Tentative Logo, acknowledgement to ChatGPT