Generalizing Bulk-Synchronous Parallel Model for Data Science: From Data to Threads and Agent-Based Simulations

Zilu Tian, Peter Lindner, Christoph Koch, Markus Nissl, Val Tannen

DATA, EPFL  University of Pennsylvania

Why agent-based simulations matter?

- Epidemics
  - London's Imperial College predicts utilize to be from consuming excess in UK and
  - Behind the Virus Report That Jailed the UK, the UK in Action

- Economics
  - The economy needs agent-based modelling
  - What is the average wealth of the population if we increase the initial wealth by 10%, 20% and 50%, respectively?

What are agent-based simulations, really?

- Depend on who you ask!
  - A recent survey in 2020 listed
    - 36 general-purpose frameworks
    - 100+ specialized frameworks
  - Different assumptions about agents
    - NetLogo considers turtles as agents, along with patches and links
    - DMASON assumes each agent belongs to a temporal region
    - Repast Symphony assumes that agents actions are scheduled
  - Different assumptions about interaction
    - NetLogo assumes spatial-based interaction
    - DMASON is based on publish-subscribe paradigm
    - Repast Symphony allows instant changes to other agents’ states
  - The lack of formal models causes high heterogeneity
    - Increase users’ learning curves
    - Decrease cross-platform result verification
    - Hard to select the right tool
    - Limit performance optimizations to framework-dependent
  - Generally speaking, frameworks are round-based or asynchronous
    - whether agents proceed in lockstep
    - But frameworks have different flavors of “round-based” or “asynchronous”

Challenges of agent-based simulations

- Agent-based simulations are flexible, but inefficient to execute
  - High concurrency
  - A realistic simulation has billions of agents
  - Code heterogeneity
    - “Think like a vertex” is homogeneous
  - Communication-intensive
  - Existing frameworks assume little or no communication
  - For data management
    - Simulations generate a large amount of data
    - Image long-running simulations with billions of agents
    - Simulations form part of complex analytics pipelines
      - “How does the average wealth of the top 30% change?”
      - Simulations can be viewed as model samples
        - “What is the average wealth of the population if we increase the initial wealth by 10%, 20% and 50%, respectively?”

Contributions

- Formal models that define agents and their interactions
  - Programming model
    - Agents are sequential processes that communicate through messaging
    - A simulation is, conceptually, concurrent execution of interacting agents
    - “Simulate” as an operator for integrating with data science pipeline
  - Computational model
    - Weighted hierarchical BSP model
  - Optimizations
    - Thread merging
    - Tame high-concurrency
    - Direct memory accesses
      - Bypass messaging overhead
    - Deforestation
      - Reduce the volume of generated data
  - Implementation
    - An eDSL in Scala for parallel agent programming
    - A system architecture based on the BSP-like model

Benchmark Description

- Population Dynamics
  - Simulate the game of life example in a 2D grid
  - Model each cell in the grid as an agent
- Economics
  - Simulate the bidding process in the stock market
  - Model traders and the stock market as agents
- Epidemics
  - Simulate individuals of states Susceptible, Infectious, Recovered, Hospitalized, or Deceased
  - Model the population and locations as agents
  - Use random graph models to simulate population connectivity
    - Erdos-Renyi Model (ERM)
    - Stochastic Block Model (SBM)

Performance

- Our system has on par or better performance than current BSP-like systems
- Effective optimizations

Data plane
- Agent
- RAM data
- Message
  - In-memory, Network
- Partition, begin/ends
- Deforestation
- Reduce the volume of generated data