Scaling up Social Simulation

Jing-Ya You[2], Alex Voss[1], Eric Yen[2], Simon Lin[2], Ji-Ping Lin[3], Andy Turner[4]

[1]School of Computer Science, University of St Andrews
[2]Academia Sinica Grid Computing, Academia Sinica, Taiwan
[3]Center for Survey Research, Academia Sinica, Taiwan
[4]School of Geography, University of Leeds

ISGC, Taipei, Taiwan
23th March 2011

Overview

- Social simulation and agent-based modelling
- Social simulation at scale and tractability
- SimTaiwan: a model of migration in Taiwan
- Making SimTaiwan scale
- Observations on compute infrastructures

sicsa* The Scottish Informatics & Computer Science Alliance

Agent-based Social Simulation

- A new approach to modeling social phenomena
- Based on empirical data
- Based on existing theories
- A new way to explore them, complementing other forms of modelling and prediction
- Used to understand and predict
- Not just one form of simulation: systems dynamics, microsimulation, queueing models, etc.

sicsa* The Scottish Informatics & Computer Science Alliance

Agent-Based Social Simulation

- Simulating interactions between dynamic populations in changing environment
- Heterogeneous populations – each individual has specific attributes such as age, gender, socio-economic status, health, etc.
- Stochastic process – each run can differ from previous
- Model and analyse phenomena too complex for closed form, can be used in absence of knowledge about causality
- Supported by Agent-Based Modelling Frameworks: Repast Simphony, Mason, Swarm, NetLogo, FLAME, etc.

Migration in Taiwan

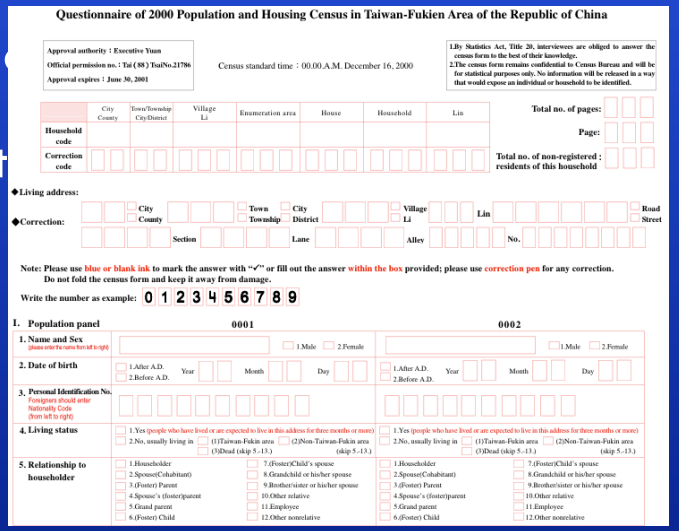
- Migration has been an important factor in Taiwanese social
  - Migration has been an important factor in Taiwanese social development and influenced by outside factors since the 1600s
  - Aim is to test existing theories of migration constructively

outward migration to China.

at Academia Sinica Center for Survey Research.

SimTaiwan: Migration in Taiwan

- Based on Taiwan 2000 Population and Housing Census
- Dataset is individual-level but with restricted variables, held at Academia Sinica
- Based on Taiwan 2000 Population an
  - Dataset is individual-level but wit
    Academia Sinica
- Individual-level data needed to estimate further socio-economic variables (e.g., income)
- Additional datasets providing data about geographic regions:
  - Employment and unemployment



Questionnaire of 2000 Population and Housing Census in Taiwan-Fukien Area of the Republic of China

Approval authority : Executive Yuan
Official permission no. : Tai (88) TouiNo.21786
Approval expires : June 30, 2001

Census standard time : 00.00 A.M. December 16, 2000

1.By Statistics Act, Title 20, interviewees are obliged to answer the census form to the best of their knowledge.
2.The census form remains confidential to Census Bureau and will be for statistical purposes only. No information will be released in a way that would expose an individual or household to be identified.

	City County	Town/Township City/District	Village Li	Enumeration area	House	Household	Lin
Household code							
Correction code							

Total no. of pages:
Page:
Total no. of non-registered residents of this household

◆Living address:
◆Correction: City/County, Town/Township, City/District, Village/Li, Lin, Road/Street, Section, Lane, Alley, No.

Note: Please use blue or blank ink to mark the answer with "✓" or fill out the answer within the box provided; please use correction pen for any correction. Do not fold the census form and keep it away from damage.

Write the number as example: 0 1 2 3 4 5 6 7 8 9

1. Population panel — 0001 — 0002

	0001	0002
1. Name and Sex (please enter the name from left to right)	1.Male 2.Female	1.Male 2.Female
2. Date of birth	1.After A.D. 2.Before A.D. Year Month Day	1.After A.D. 2.Before A.D. Year Month Day
3. Personal Identification No. (Foreigners should enter Nationality Code from left to right)		
4. Living status	1.Yes (people who have lived or are expected to live in this address for three months or more) 2.No, usually living in (1)Taiwan-Fukin area (2)Non-Taiwan-Fukin area (3)Dead (skip 5.-13.) (skip 5.-13.)	1.Yes (people who have lived or are expected to live in this address for three months or more) 2.No, usually living in (1)Taiwan-Fukin area (2)Non-Taiwan-Fukin area (3)Dead (skip 5.-13.) (skip 5.-13.)
5. Relationship to householder	1.Householder 2.Spouse(Cohabitant) 3.(Foster) Parent 4.Spouse's (foster)parent 5.Grand parent 6.(Foster) Child 7.(Foster)Child's spouse 8.Grandchild or his/her spouse 9.Brother/sister or his/her spouse 10.Other relative 11.Employee 12.Other nonrelative	1.Householder 2.Spouse(Cohabitant) 3.(Foster) Parent 4.Spouse's (foster)parent 5.Grand parent 6.(Foster) Child 7.(Foster)Child's spouse 8.Grandchild or his/her spouse 9.Brother/sister or his/her spouse 10.Other relative 11.Employee 12.Other nonrelative

sicsa* The Scottish Informatics & Computer Science Alliance

SimTaiwan Migration Model

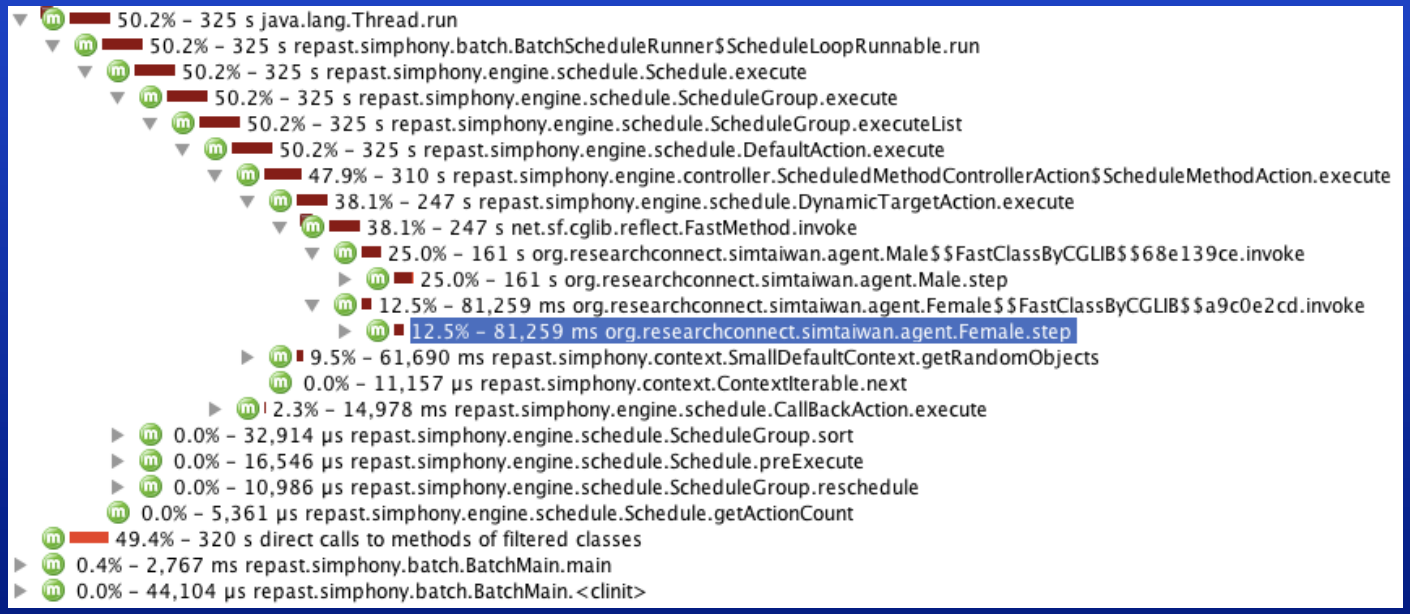
- Divided into two parts: departure and destination
- Probability of departure is a weighted sum of different factors such as: Sex, Age, Education, Marital status,
  - Divided into two parts: departure and destination
- Probability of departure is a weighted sum of different migration patterns measured by census.
- Note: we do not currently have linked socio-economic data such as income

Test Code and Parameters

- Simplified model with only fertility and mortality, same for all measured models
- 250k male and 250k female random initial population, running for 365 ticks (=days)
- Measurements taken using JProfiler 6.0.4
  - CPU sampling (5 sec intervals)
  - Memory allocations recording
- JVM Parameters: -Xmx8192M –Xss128M
- Simplified model with only fertility and mortality, same for

Naïve Serial Version

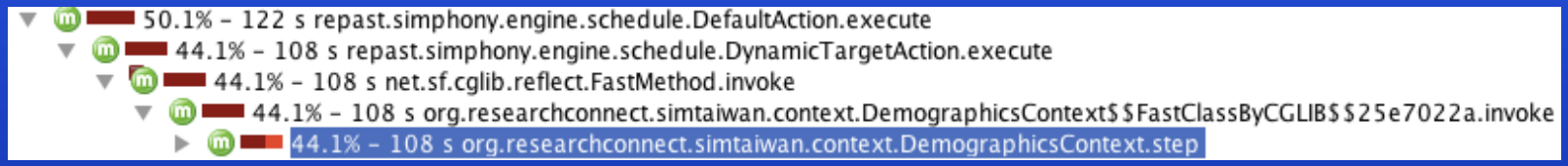
- More time spent in RePast scheduling code than in model code because events are scheduled for each individual agent every step.



sicsa* The Scottish Informatics & Computer Science Alliance

Improved Serial Version

- Event scheduled on DemographicsContext, code iterating through individual agents

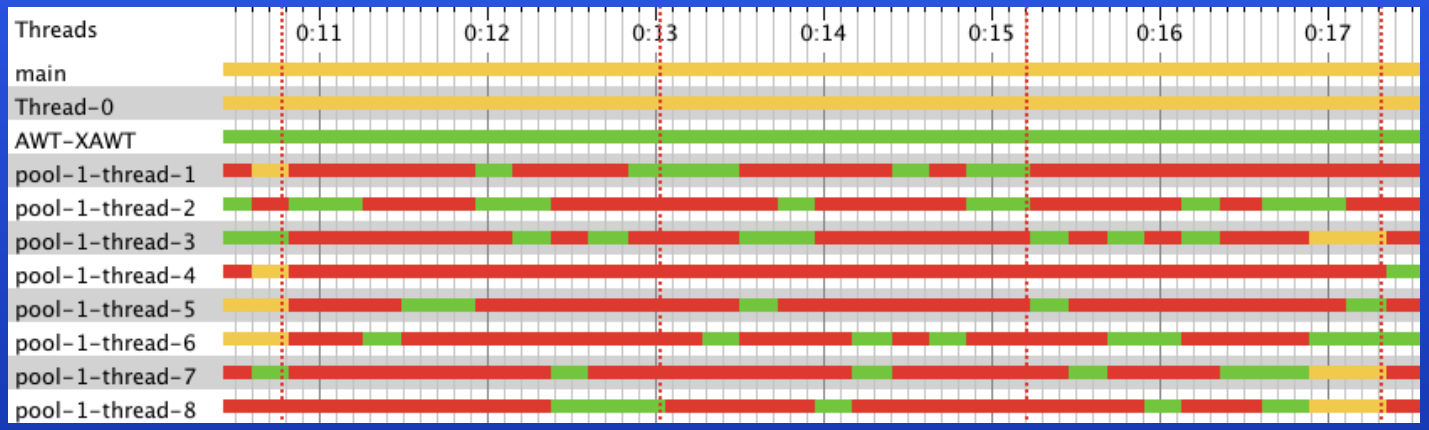


- Wallclock time down from 5:32 to 2:23
- Opens up opportunities for parallelising code as well…

Initial Parallel Version

- Need to partition data to allow multiple worker threads to exploit multiple CPUs & cores
- PartitionedContext keeping agents in separate HashSets that can return independent Iterators for use by multiple threads.
- ThreadPoolExecutor with configurable number of worker threads (here 8)
- Initial version brings only modest / no improvement, wallclock time in some runs > improved serial code
- Max. CPU utilisation ~ 200% (top)

Initial Parallel Version (II)



- Worker threads blocking a lot on monitors placed around RePast constructs.
- Worker threads blocking a lot on monitors placed around RePast constructs.
- Simulation schedule relatively minor issue

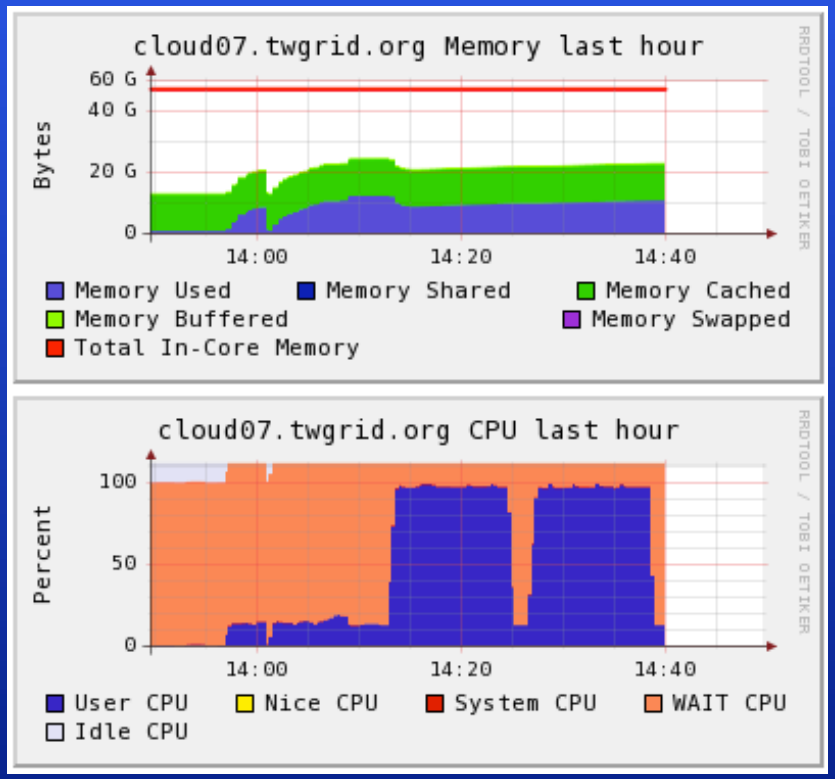
Improved Parallel Version

- Overloading some of RePast's code to make it thread safe.
- Reducing scope of monitor objects used and pulling code
  - Overloading some of RePast's code to make it thread safe.
- Reducing scope of monitor objects used and pulling code parts that are safe out of synchronized sections
  - Introducing thread-local variable containing a per thread

```
protected static ThreadLocal<Uniform> uniform = new ThreadLocal<Uniform>() {
  @Override
  protected Uniform initialValue() {
    RandomEngine generator = new MersenneTwister((int)System.currentTimeMillis());
    return new Uniform(generator);
```

Improved Parallel Version (II)

- Monitor contention is eased signficantly
- Wallclock running time down to 1:03 and max. CPU utilisation up to ~ 600%
- Time spent in serial code for analysis and production of charts is now significant.
- Issue: Iteration through contexts needs to be randomised



sicsa* The Scottish Informatics & Computer Science Alliance

Compute Infrastructures

- Current production grids do not support multi-threaded code with high memory requirements well.
  - Suitable for runs with smaller population sizes for testing and stability analysis
- Amazon EC2 provides a suitable instance type m2.4xlarge that provides about 64GB of RAM
  - Suitable for occasional larger runs or production runs that exceed local capacity
- High-memory local servers integrated into local cloud environments.

Conclusions

- SimTaiwan shows that population-scale models can be built using commodity computing platforms
- and relatively simple programming models to utilise multi core CPUs.
- We argue that SimTaiwan is a member of a class of social science investigations that thanks to increasing compute capacity of commodity hardware are becoming tractable.
- Further work is now focusing on:
  - the development of the model logic,
- SimTaiwan shows that population-scale models can be built using commodity computing platforms
- and relatively simple programming models to utilise multi