The University Of Sheffield.

Improving Transport Simulation Performance using Graphics Processing Units

Peter Heywood, Paul Richmond

The University of Sheffield

Transport Network Simulations

- Global transport demand is increasing
- Many constraints on transport networks
- Simulations can improve the use of limited resources
  - Planning
  - Management

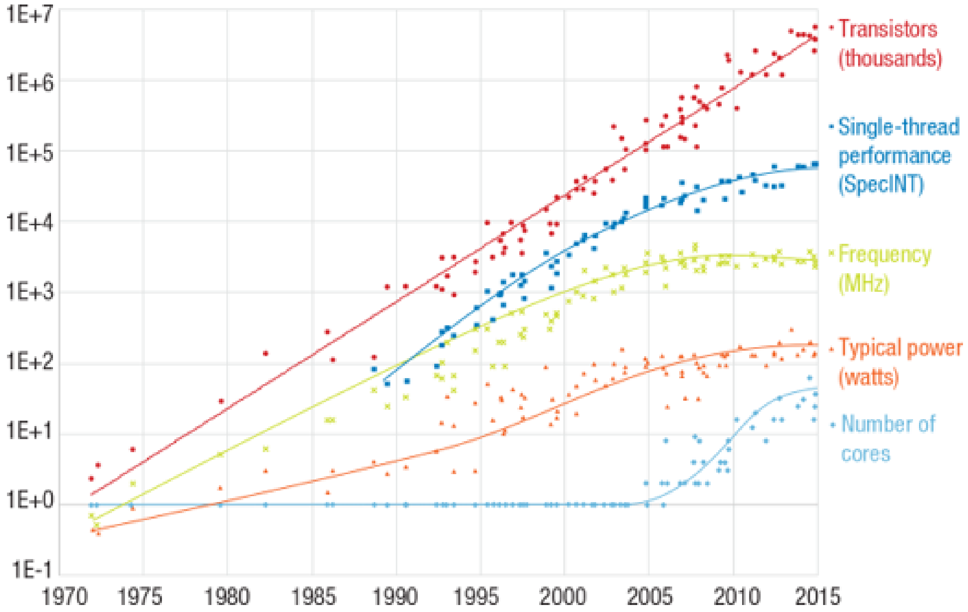
Transport Network Simulations

- Simulations are becoming more computationally expensive
  - **Larger**
    - City-scale, National-scale
  - **More Complex**
    - CAVs, Smarter Infrastructure
    - Multi-Mode
  - **More Permutations**
    - Weather, Demand, etc.

- Performance is limiting the use of simulation
- **Faster simulators are required**
  - Better-than-real-time

Central Processing Units (CPUs)

- Existing simulators use Central Processing Units (CPUs)
  - Aimsun, PTV simulators, SATURN, SUMO, ...
- General Purpose Processors
- Generational improvements have slowed
  - Individual cores not getting much faster
  - CPUs are becoming more parallel
- Very complex processing cores
- Multi-Core Processors
  - 10s of physical cores



Graphics Processing Units (GPUs)

- Originally developed for 2D and 3D computer graphics
  - Suitable for general purpose computing

- Many-Core Co-Processor
  - Massively Parallel
  - Thousands of processing cores
  - Processor cores are relatively simple
  - Connected over PCI-E bus
  - Power Efficient

2x Nvidia Titan Xp and 2x Titan V GPUs

Theoretical Peak Performance (and Power)

Serial Computing

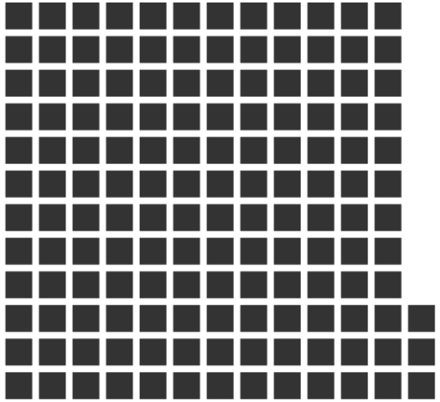
~53 GigaFLOPS

1 Core

Parallel Computing

~1.5 TeraFLOPS

28 Cores



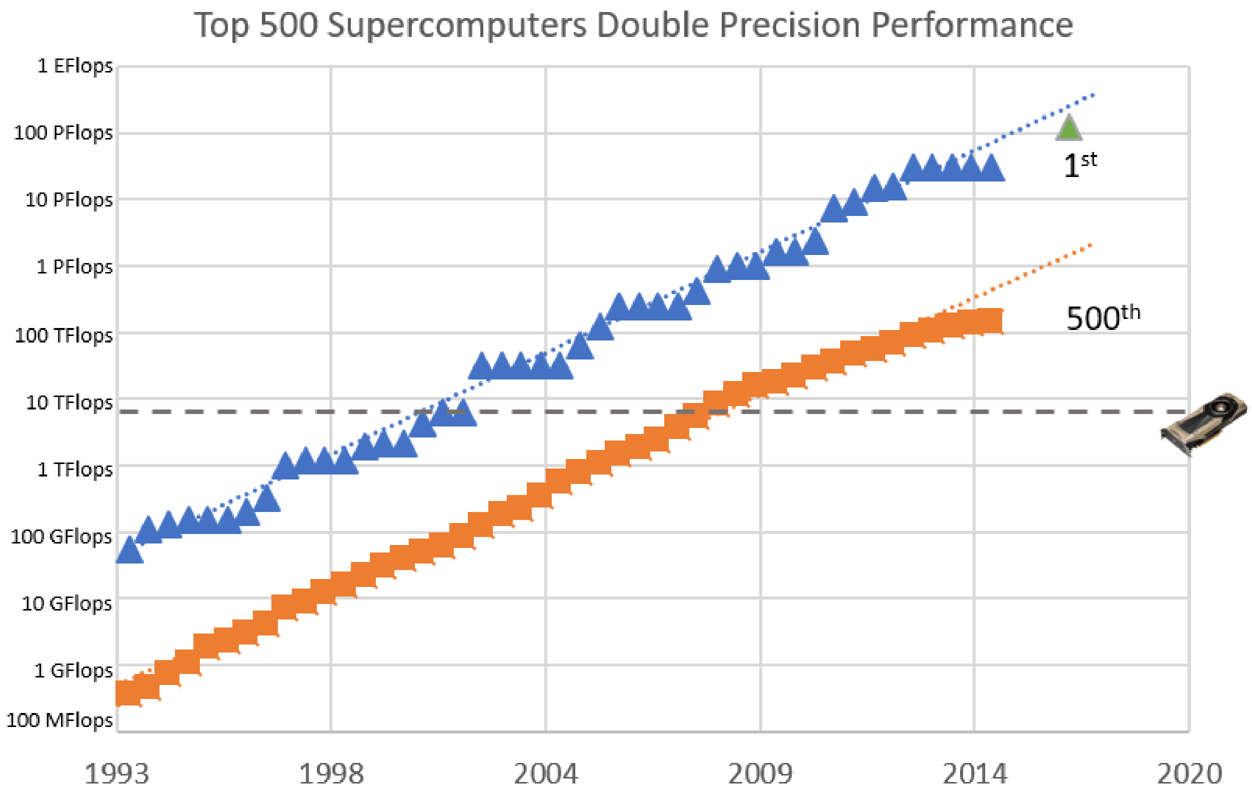
Accelerated Computing

7.8 TeraFLOPS

5120 Cores

- Theoretical Peak Performance
- Double-Precision
- Skylake 28 Core CPU
  - 1.5 TFLOPS
  - 165 W
  - 9 GFLOPS/Watt
- GPU
  - 7.8 FLOPS
  - 250 W
  - 31 GFLOPS/Watt
- GPUs even better at lower precision

Top 500 Supercomputers



- Fastest computers in the world
- 1 Titan V
  - 250W, 1kg, 26cm long
- Faster than No.1 in 2001
  - ASCI White
  - 6MW, 106 Tons, 200 Cabinates
  - 200 Cabinets
- 8 Titan Vs
  - Would have been 8th most powerful computer in 2007

Challanges of GPUs

- Switching from CPU to GPU is not a straight-forward
- Considerable changes to software
    - **New Algorithms**
    - **New Data Structures**
    - Data Locality and Data Transfer
- High level of parallelism required
    - If a problem is not parallel enough it **will not be faster**
- Specialist knowledge required to achieve high performance

GPU Accelerated Transport Simulations

1. Macroscopic Road Network Simulation and Assignment
2. Microscopic Road Network Simulation
3. Pedestrian Crowd Simulation
4. Multi-Modal Rail Network Simulation

GPU Accelerated Macroscopic Simulation

- SATURN - **S**imulation and **A**ssignment of **T**raffic to **U**rban **R**oad **N**etworks
- Macroscopic Assignment and Simulation
  - High level of abstraction
  - Relatively low computational requirements
- Large networks can take many hours per run
  - even using 32 CPU Cores

ATKINS

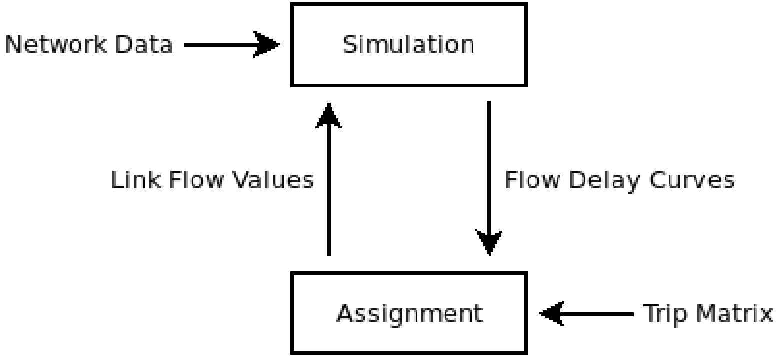
Member of the SNC-Lavalin Group

SATURN

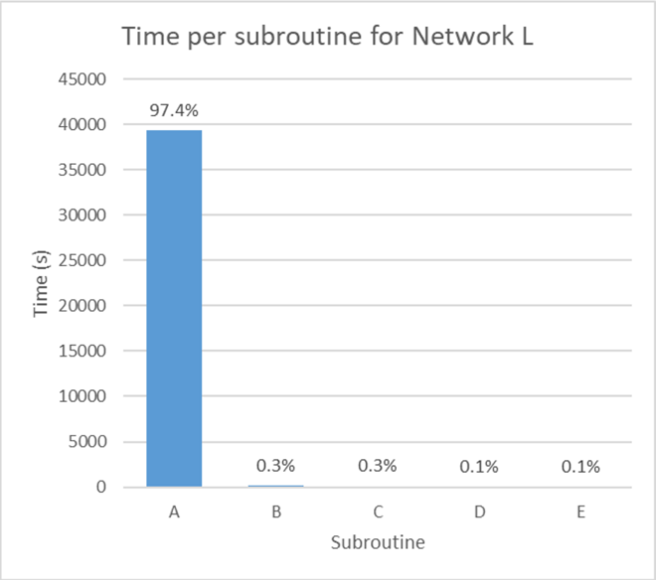
SATURN

- Iterative Equilibrium-based algorithm of Assignment and Simulation
- Used for Highways England Regional Transport Models (RTMs)
- Individual simulations can take many hours
- Profiling shows majority of work in Assignment Phase
  - Mostly calculating Shortest Paths

Network	Size	User Classes	Zones
E	Town	2	12
D	Small City	13	547
C	Large City	5	2548
L	Metropolitan	5	5194

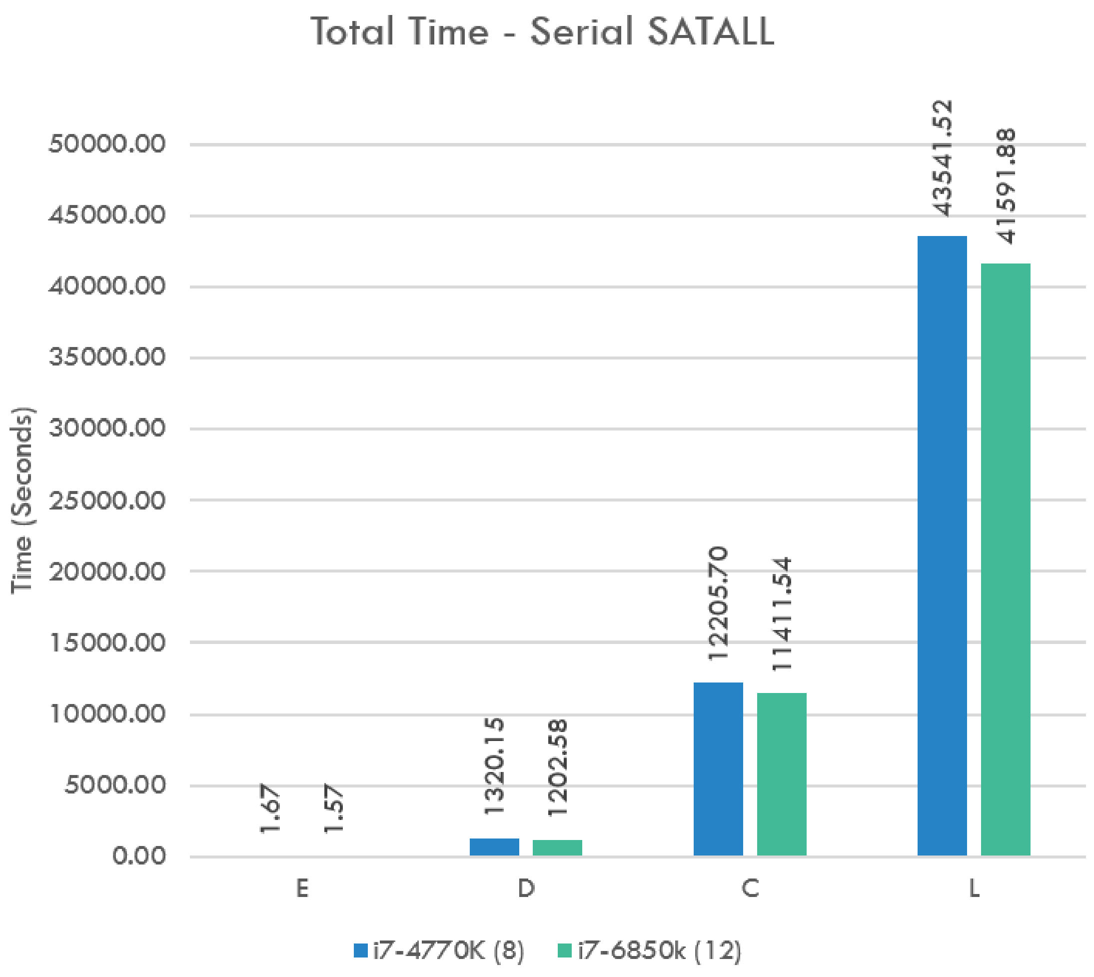


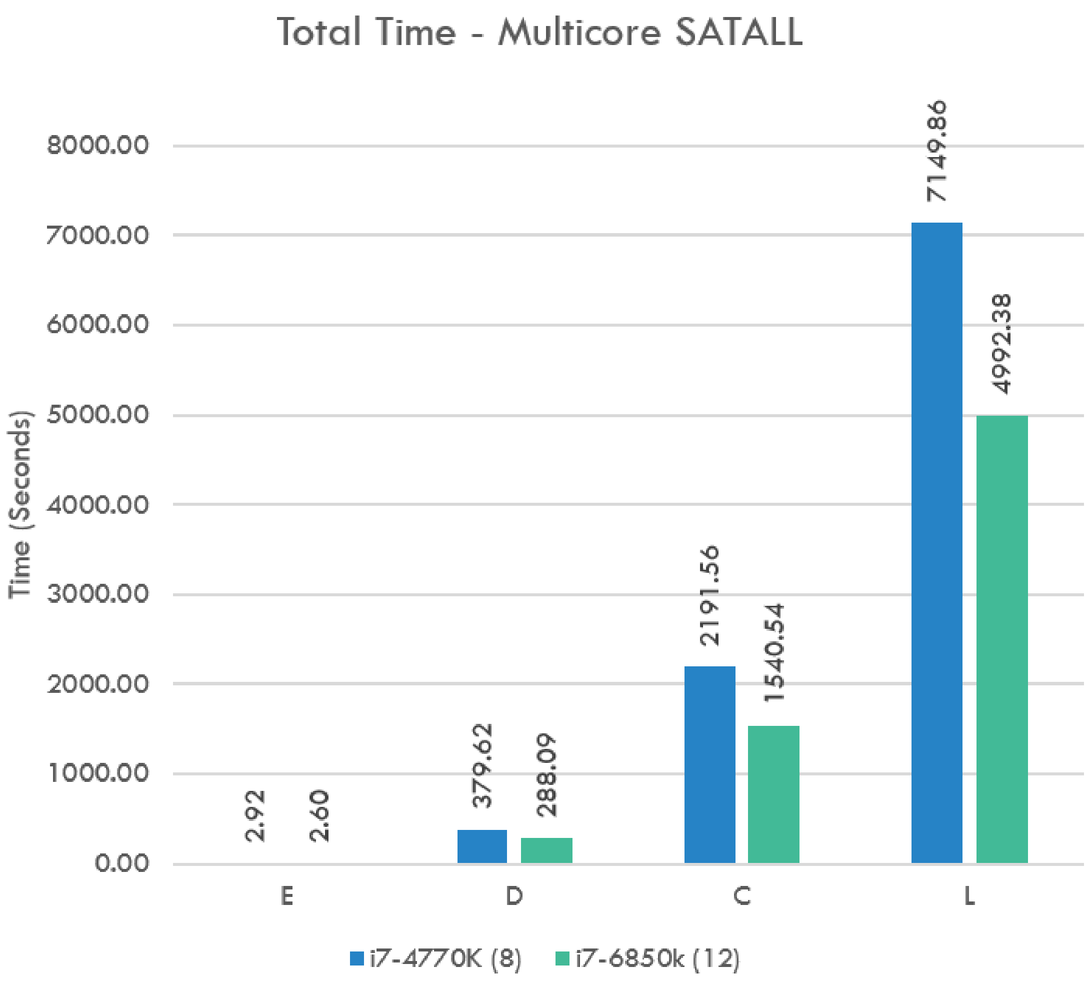
SATURN Assignment-Simulation Loop



SATURN Profiling

CPU Performance





Significant Algorithmic Changes - Shortest Paths

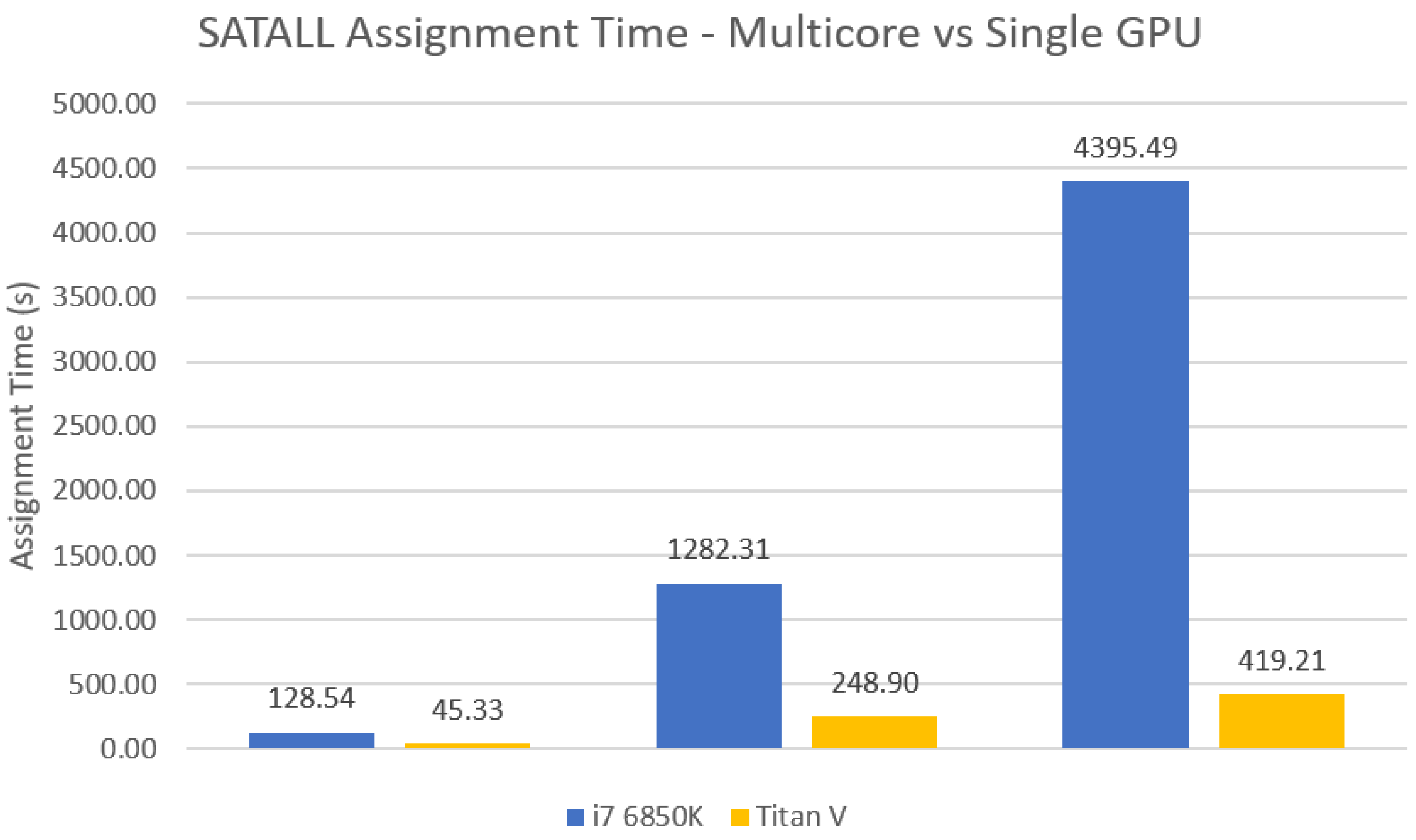
CPU

- Single Source Shortest Path (SSSP)
- Use the D'Esopo-Pape algorithm
  - Or Dijkstra's algorithm
- **Very Efficient** algorithms
- But **Highly Sequential**
  - Not Suitable for GPU

GPU

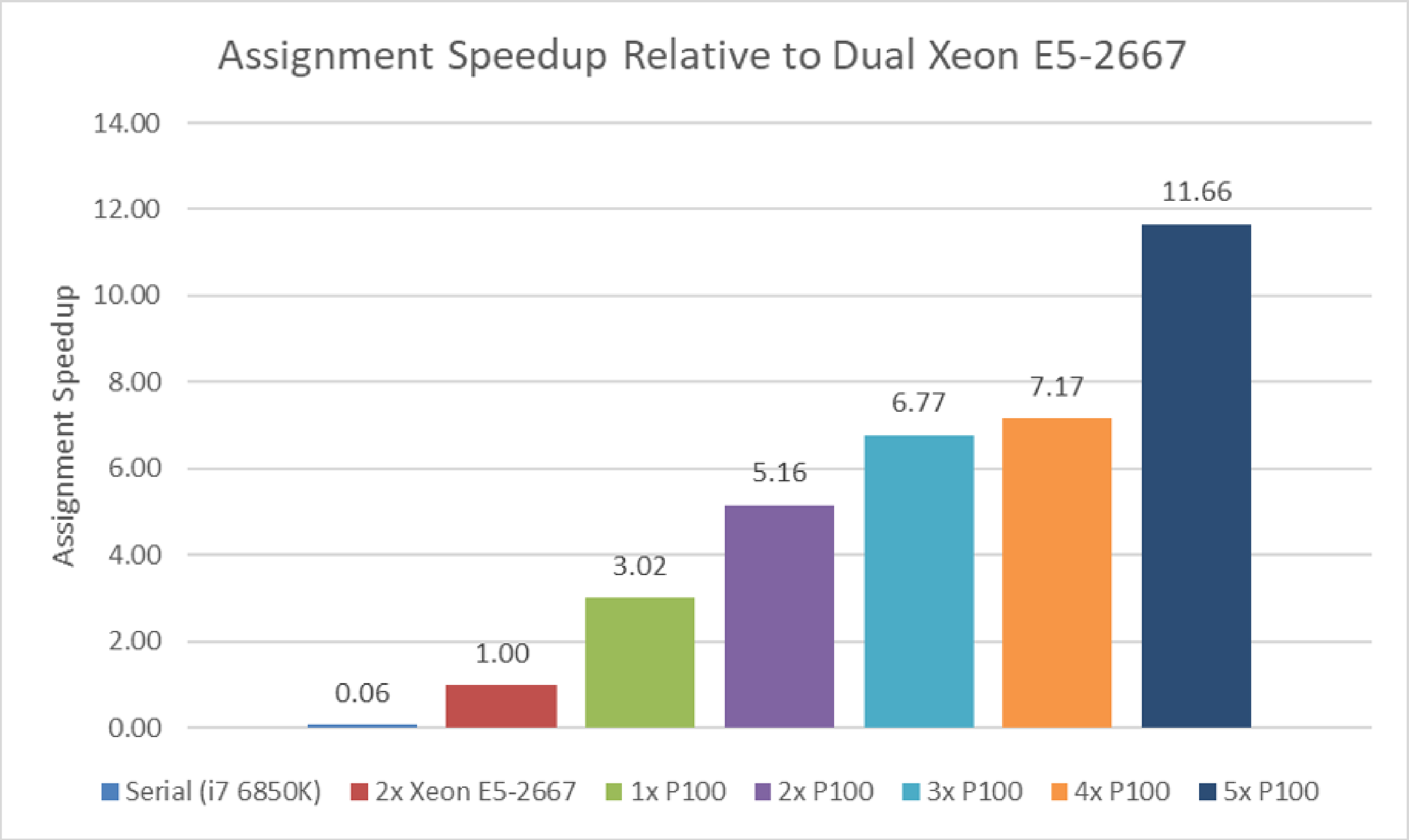
- Use the Bellman-Ford algorithm
- **Highly Parallel**
- But **Very inefficient**
  - Naive Implementation 360x slower
- *Many* optimisations for efficient GPU Algorithm
  - Vertex Frontier
  - Multiple Concurrent Sources
  - Multiple Concurrent User Classes
  - Cooperative Groups
  - etc.

Single GPU Performance vs Single CPU



- i7-6850k
  - 6 core CPU
  - 12 Threads
  - **4395.49** seconds (Assignment)
- Nvidia Titan V GPU
  - 5120 CUDA Cores
  - 12 GB HBM2
  - CUDA 9.0
  - **419.21** seconds (Assignment)
  - Up to **10.5** speed-up

Multi-GPU Performance vs Multi CPU

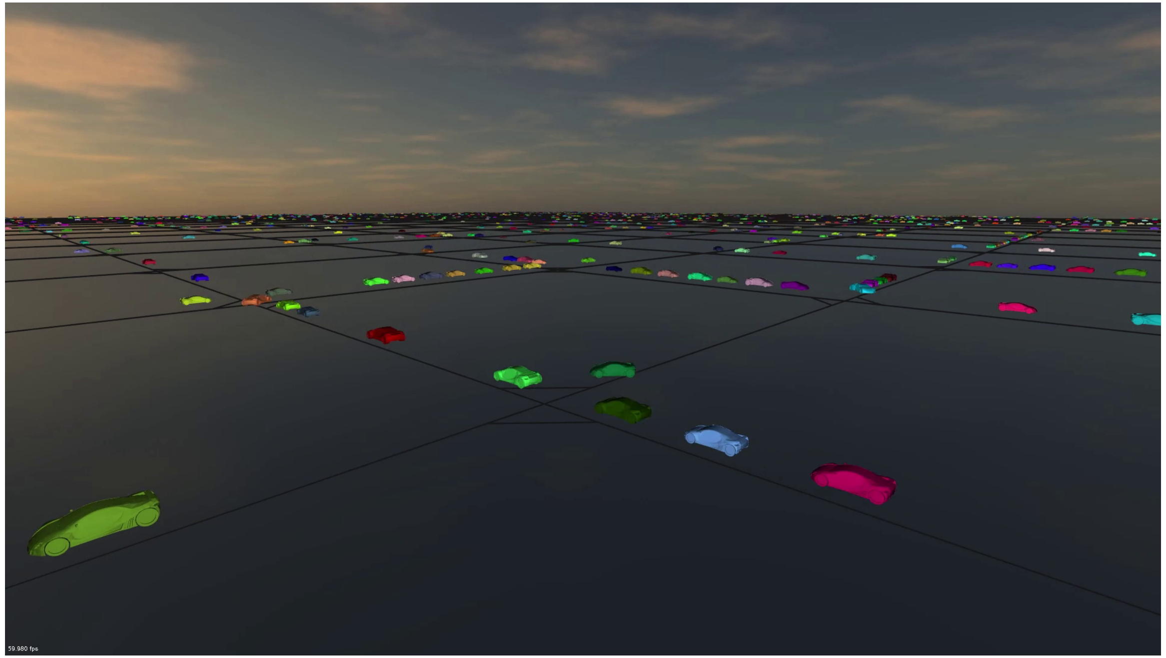


- LoHAM model
- Dual Socket Xeon E5-2667
  - 12 Cores, 24 Threads
  - **2633.25** seconds (Assignment)
- Nvidia Tesla P100
  - 5 User-classes of vehicle
  - **225.83** seconds (Assignment)
  - Up to **11.7x** speed-up

GPU Accelerated Microscopic Simulation

- Proof of concept GPU accelerated Microsimulation
  - Funded by Department for Transport T-TRIG grant
  - Worked in collaboration with Aimsun

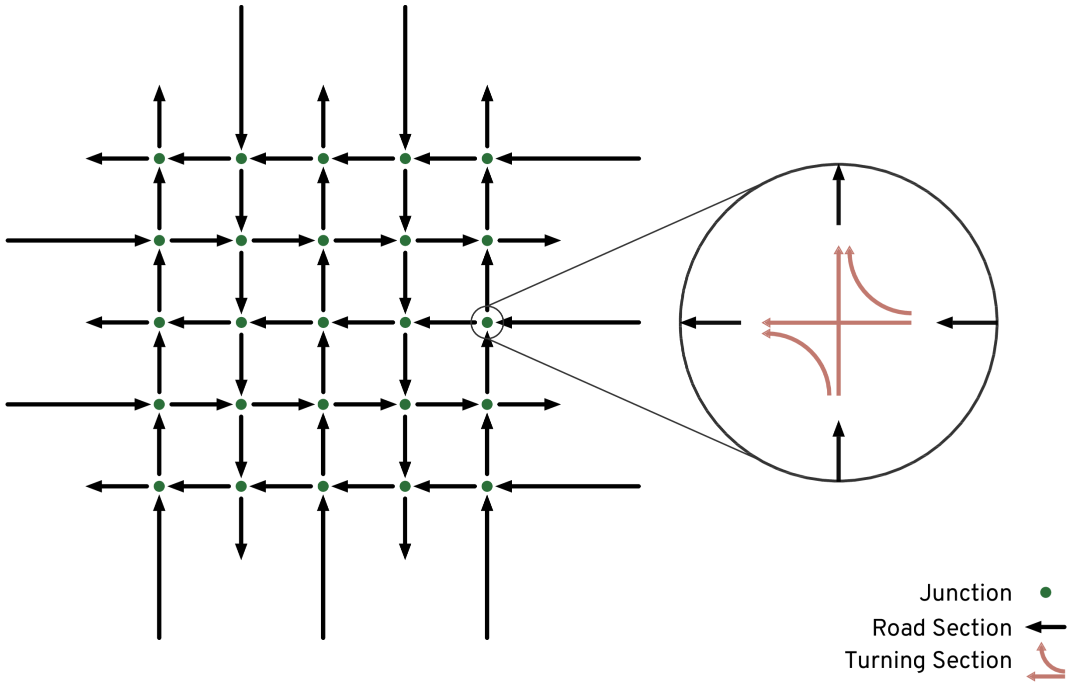
1. Implement subset of models for GPUs from existing simulator
2. Cross-validate each model and overall behaviour
3. Benchmark Performance using a scalable network



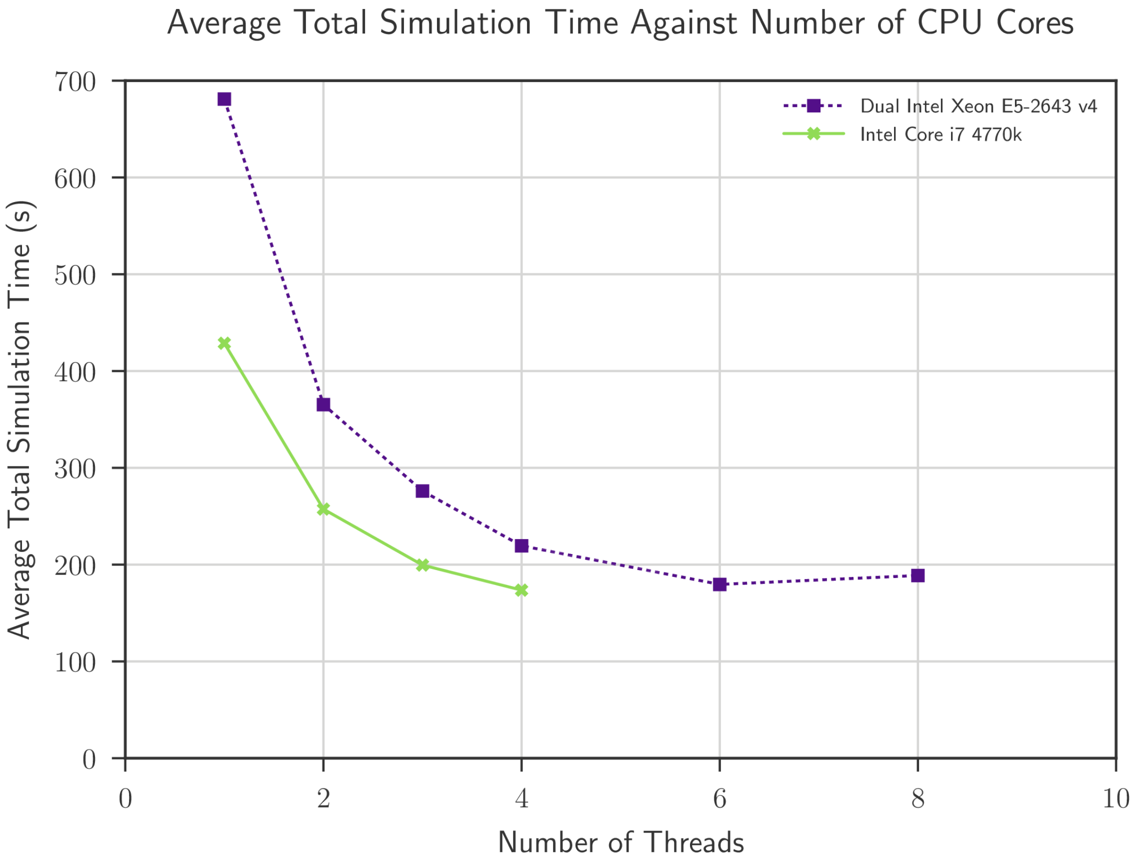
aimsun.

Artificial Network and CPU Performance

Benchmark Model



Aimsun 8.1 CPU Scaling



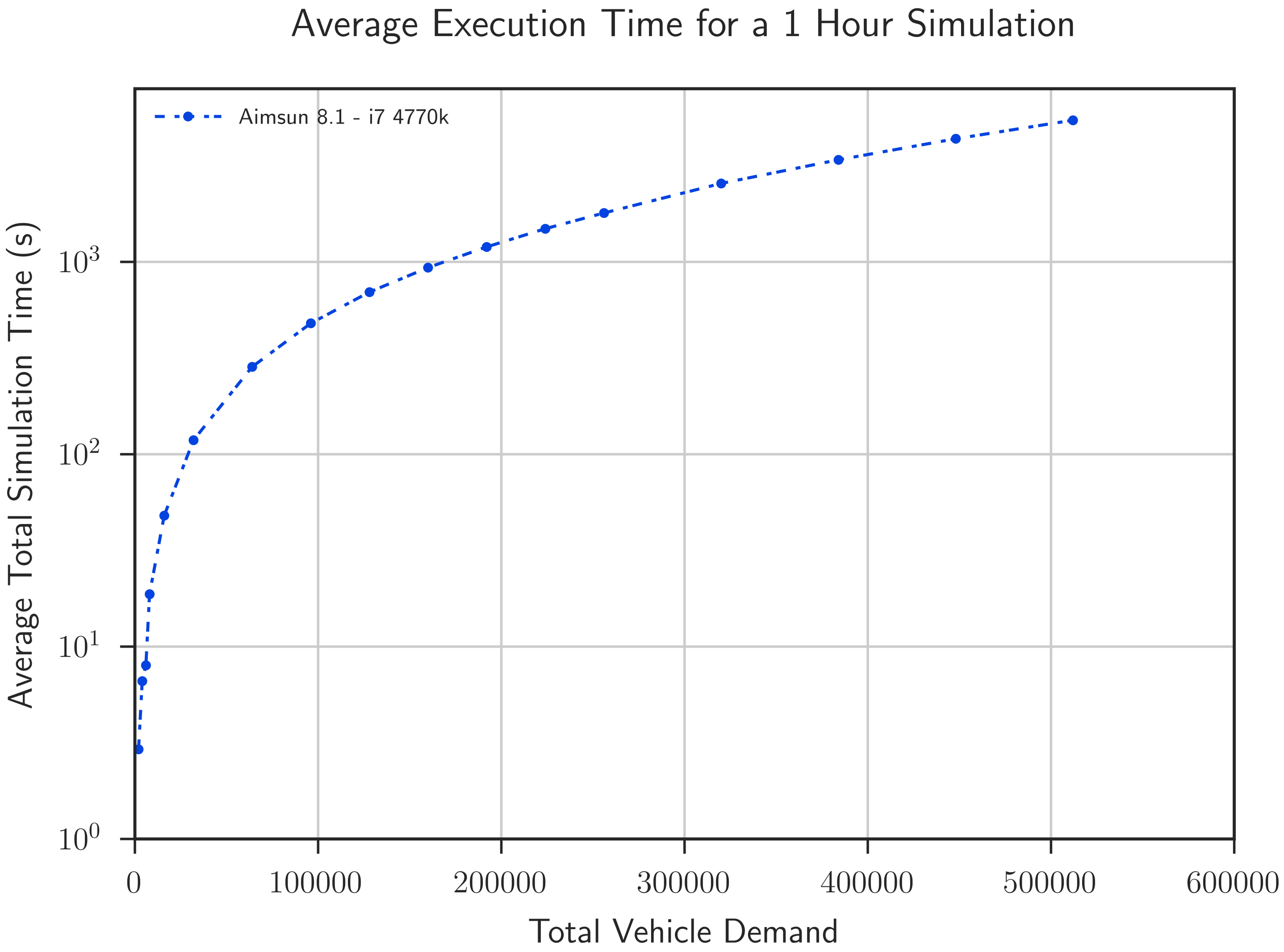
Implementation Details

- Implemented a subset of Aimsun 8.1 models
  - Gipps's Car Following Model
  - Gap Acceptance Model
  - Constant Vehicle Arrival
  - etc.
- Cross-validated
  - Individual Features
  - Overall network behaviour

- Implemented using **FLAME GPU**
- Template-based simulation environment for high performance simulation
- Agent Based Modelling
- No GPU knowledge required

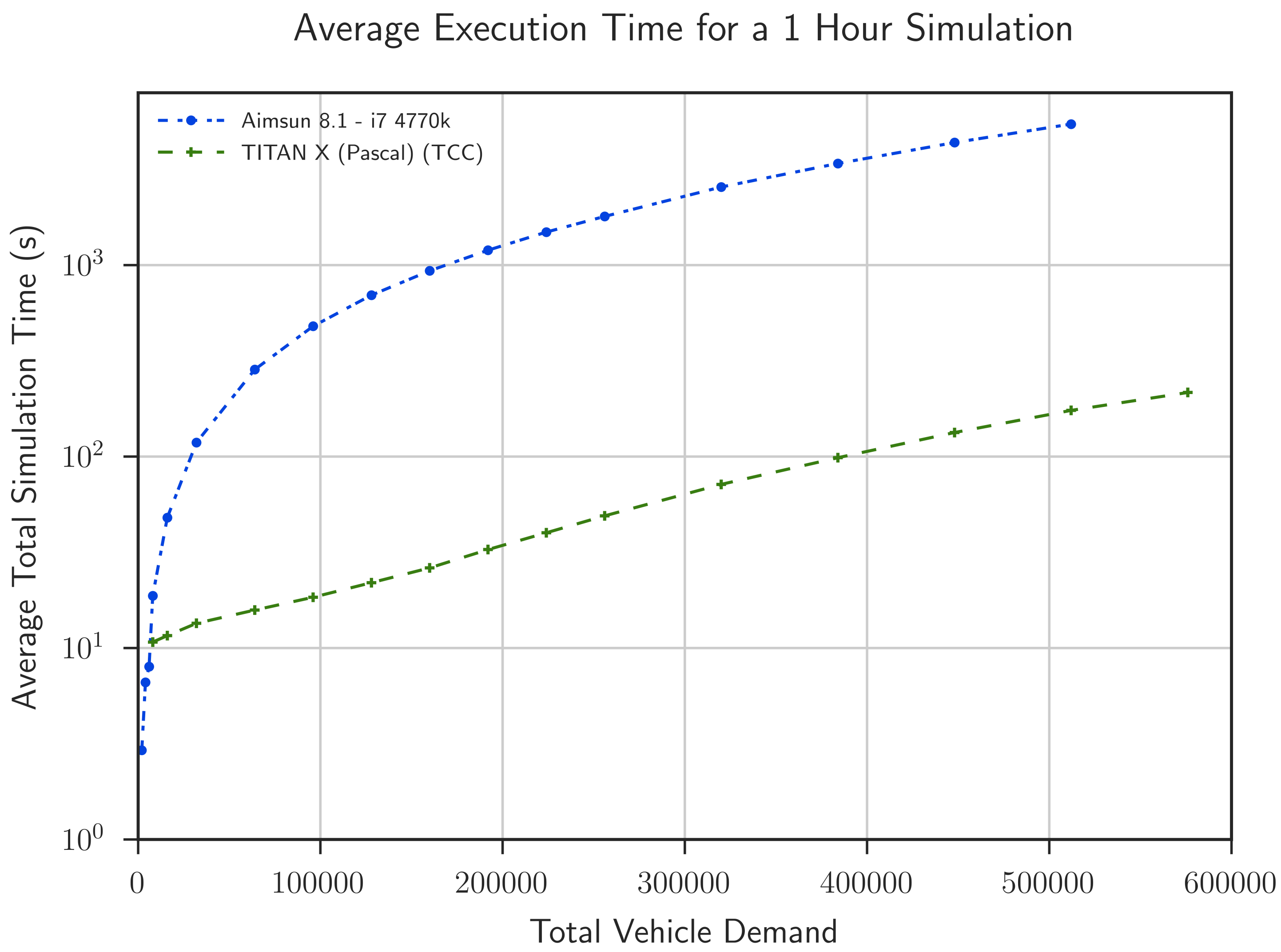
flamegpu.com

Scaling Population and Environment



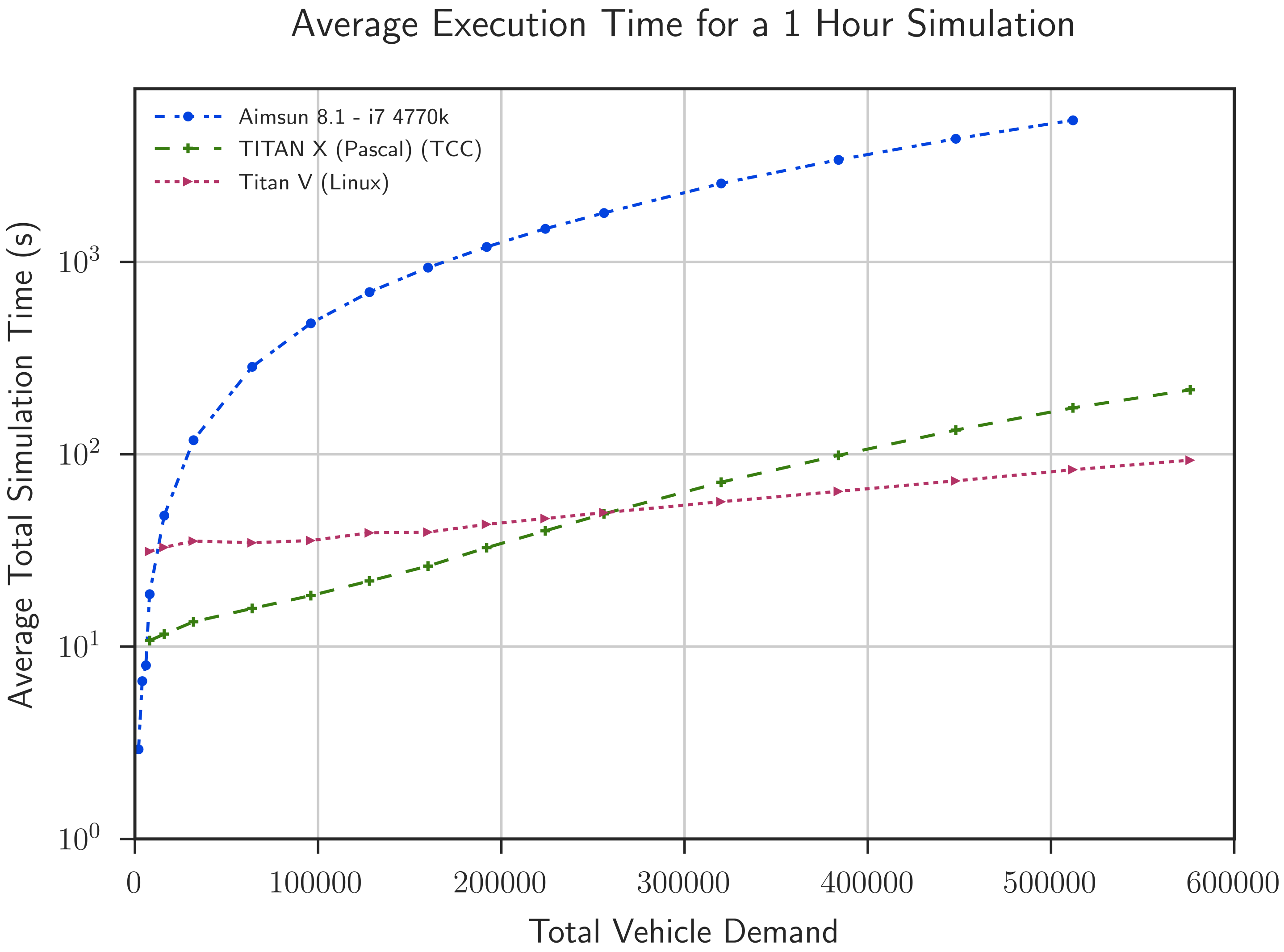
- 500,000 vehicles
- 60 minutes
- CPU - i7-4770k
  - Windows
  - **5447s**

Scaling Population and Environment



- 500,000 vehicles
- 60 minutes
- CPU - i7-4770k
  - Windows
  - 5447s
- GPU - Titan X (Pascal)
  - Windows TCC
  - **174.2s**
  - **31x** Speed Up

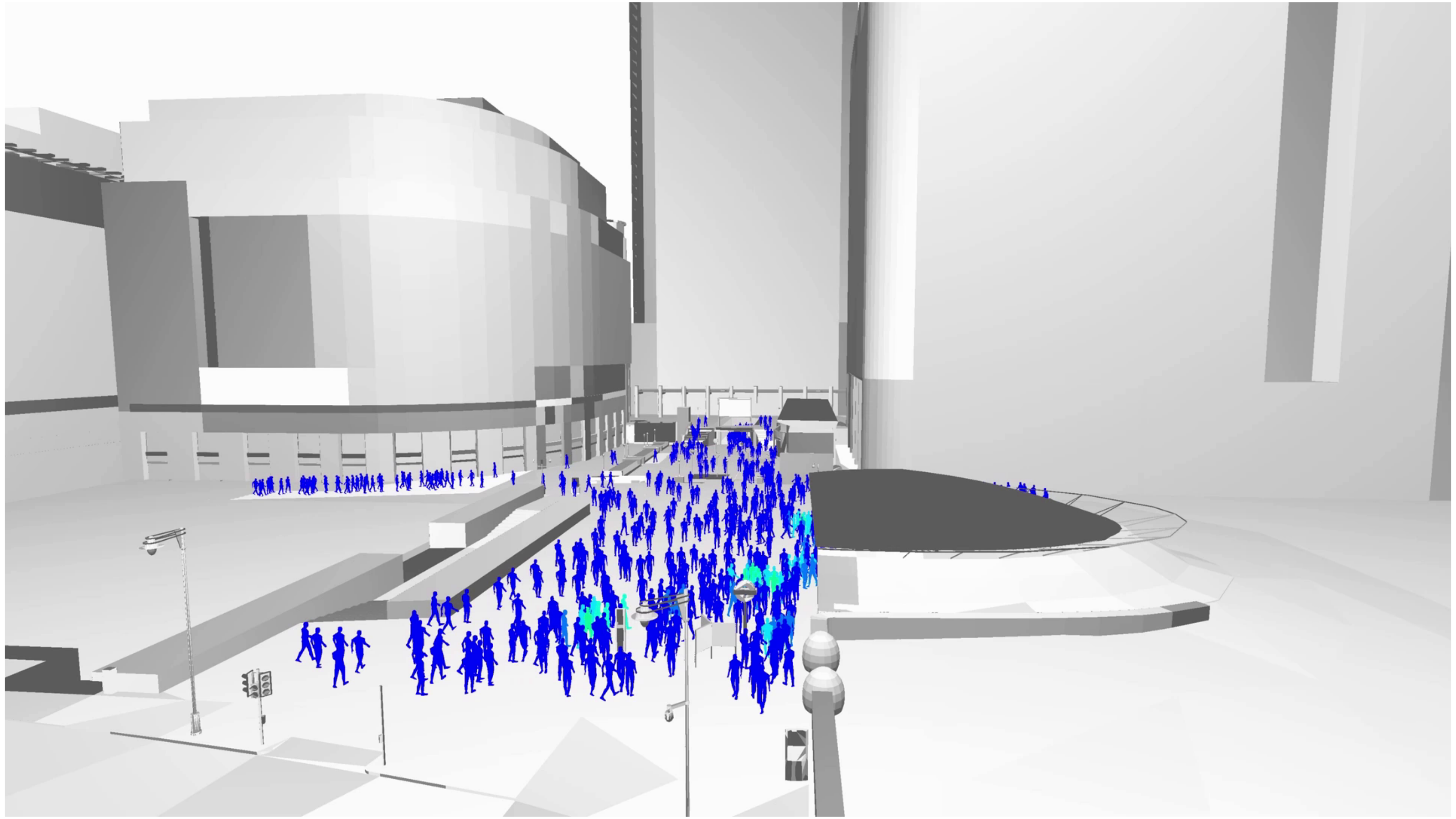
Scaling Population and Environment



- 500,000 vehicles
- 60 minutes
- CPU - i7-4770k
  - Windows
  - 5447s
- GPU - Titan X (Pascal)
  - Windows TCC
  - 174.2s
  - 31x Speed Up
- GPU - Titan V
  - Linux
  - **82.04s**
  - **66x** Speed up

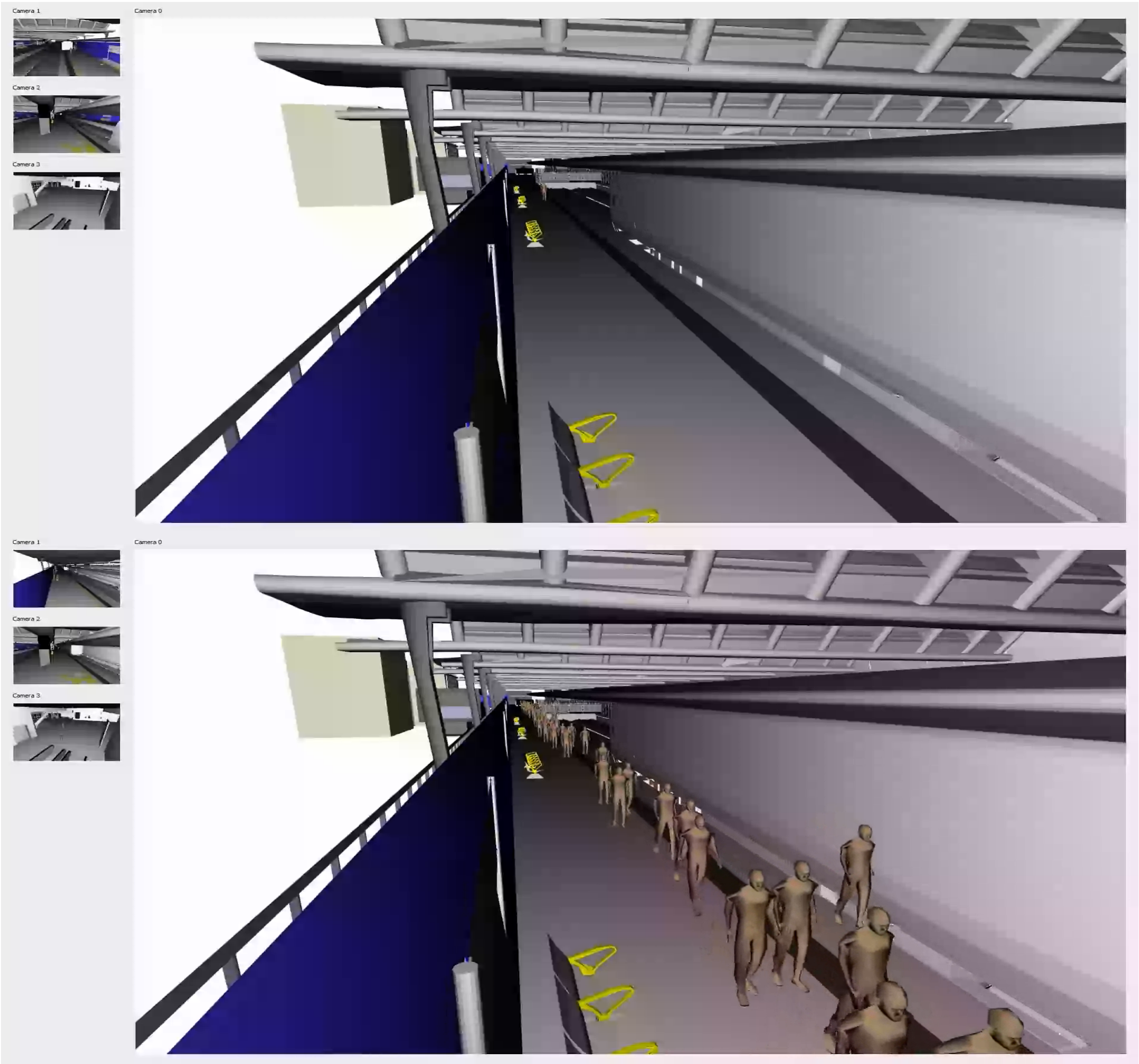
GPU Pedestrian Simulation

- GPUs suitable for many modes of transport
- I.e. Pedestrian Simulation
- Using FLAME GPU
- Can simulate 100,000s of pedestrians



Multi-Modal Simulation

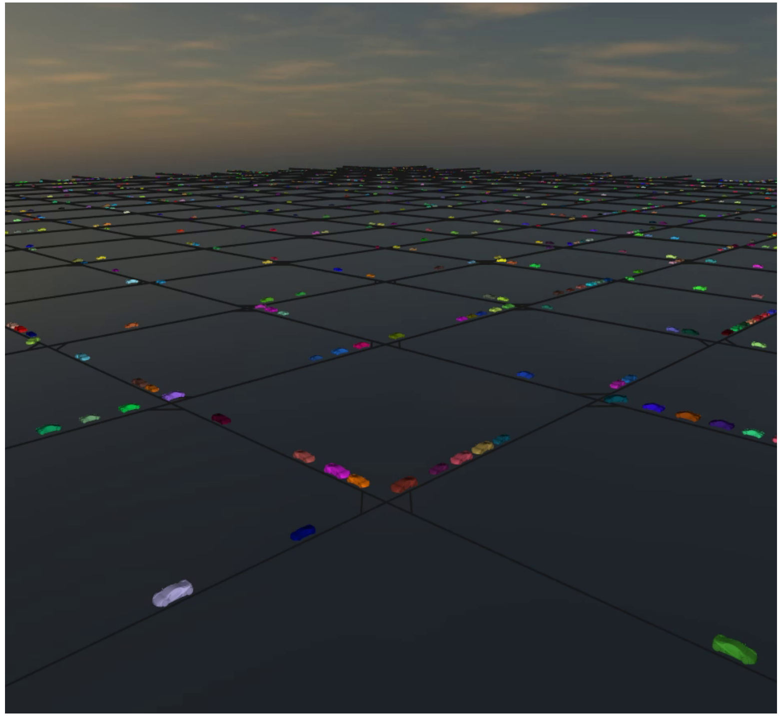
- Collaboration with Siemens
- Multi-modal Smart-City Simulation
  - CPU based rail simulation
  - GPU accelerated pedestrian simulation
  - CPU based road network simulation (SUMO)
- Evaluate rail network performance including pedestrian behaviours in station
- More information: youtube.com/watch?v=Rz_XzbZIMes



Conclusion

- Up to **11.7x** macroscopic simulation speedup
- Up to **66x** microscopic simulation speedup
  - **39x** faster than real-time for over 550,000 vehicles
- High performance pedestrian simulations

- Enables:
  - More simulations in less time
  - Larger-scale simulations
  - **Better than real time** microsimulation
  - **City-scale/national-scale multi-modal** simulations



Thank You

Contact

- Peter Heywood
  - p.heywood@sheffield.ac.uk
  - ptheywood.uk
- Paul Richmond
  - p.richmond@sheffield.ac.uk
  - paulrichmond.shef.ac.uk
- RSE Sheffield
  - http://rse.shef.ac.uk/

Research Software Engineering Sheffield.

Supported by

- EPSRC fellowship “Accelerating Scientific Discovery with Accelerated Computing” (EP/N018869/1)
- Atkins, STFC, TSC & Aimsun
- DfT Transport Technology Research Innovation Grant (T-TRIG July 2016)
- Siemens

More Information

- "Data-parallel agent-based microscopic road network simulation using graphics processing units"

  Heywood et al. 2017

  doi.org/10.1016/j.simpat.2017.11.002

Additional Slides

Models implemented and validated

Behaviours

- Gipps' Car Following Model
- Gap Acceptance Model
- Constant Vehicle Arrival
- Turning Probabilities
- Vehicle Detectors
- Stop Signs

FLAME GPU

- Template-based simulation environment for high performance simulation
- Agent Based Modelling
  - Define individual behaviours with local interactions
  - Complex behaviours emerge
- No GPU knowledge required
- Extended with transport network specific algorithms
- **flamegpu.com**

Validation

- Overall network behaviour over multiple runs
- Individual behaviours
  - I.e. Gipps' Car Following Model