Improved GPU Uniform Spatial Partitioning Through Modular Bin Strips

R. Chisholm, Dr P. Richmond, Dr S. Maddock. {R.Chisholm, P.Richmond, S.Maddock}@sheffield.ac.uk
Department of Computer Science, University of Sheffield, UK

Motivation

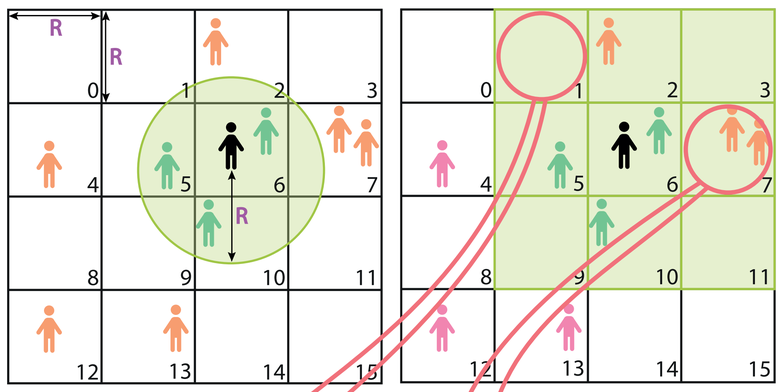
- Many complex systems contain mobile spatial actors which are influenced by their neighbours: particles, people, vehicles, etc.
- Uniform spatial partitioning is used to manage these actors in simulations (the Default technique) [1].
- Accessing this data structure is often more costly than model logic due to the high level of scattered memory accesses which create cache contention.
- This research focuses on techniques for optimising GPU uniform spatial partitioning to facilitate larger complex systems simulations.

Uniform Spatial Partitioning

Decide the interaction radius (R, green circle), and partition the environment into uniform cells with dimensions equal to R. Actors outside the environmental bounds have their locations clamped.

Implementation: Search

1. For each cell in the 3x3 Moore neighbourhood (green square area)
2. Locate the first index of the cell's data
3. Locate the first index of the next cell's data
4. For each agent in this range: If Euclidean distance < R:
   The neighbour is valid and can be further processed.

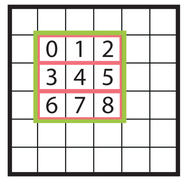


Implementation: Construction

1. Sort the agent data according to their cell indexes into an array
2. Build a boundary index of where each cell's actor storage begins

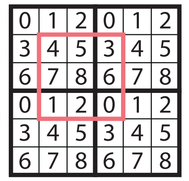
Optimisations

Strips



- The Moore neighbourhood of bins about a target location is searched. This consists of a 3x3 grid (green square).
- Each row of this grid contains 3 contiguous bins (as stored in memory), therefore we can treat each 3-bin strip as a single bin (pink rectangles).
- This modification decreases the number of bin changes from 9 to 3, reducing redundant memory reads and the branch divergence that occurs when threads within the same warp are operating on bins of differing sizes.

Modular



- In the Default technique each thread starts searching from the same point in the Moore neighbourhood, traversing the remainder of bins in the same order.
- Potentially every bin may be accessed simultaneously during a single iteration of the neighbourhood search.
- Instead the environment is subdivided into bins in 3x3 groups.
  - Each bin within these groups is labelled 0-8, such that any 3x3 Moore neighbourhood in the environment will encapsulate 9 bins labelled 0-8.
  - Bins of each Moore neighbourhood are now iterated in label order.
- Initial memory accesses are reduced to only access 1/9th of the bins.

Hybrid

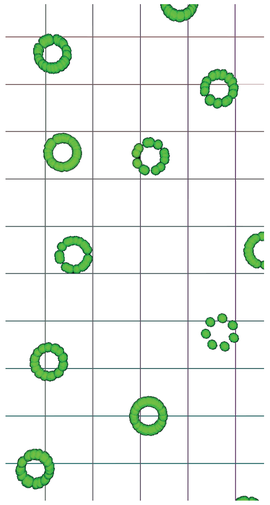
- The Strips technique only affects the first dimension.
- The Modular technique can be extended to higher dimensions. This reduces initial memory scatter to 1/27 in three dimensions.
- The Modular technique is applied to the remaining dimensions after the Strips technique has been applied to the first dimension.

Experiments

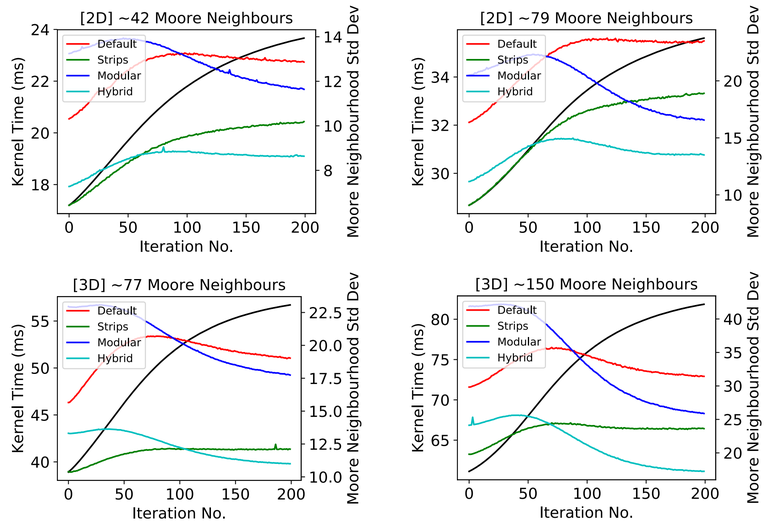
The Circles benchmark model [2] was used to evaluate the impact of these models. This benchmark model provides an analogue of general N-Body models.

Actors are initialised in a uniform random distribution. As the simulation progresses the actors move towards a steady state where they are arranged in circles (right).
A force modifier can be adjusted to change the time taken to reach the steady state.

Performance is captured alongside the standard distribution of neighbourhood sizes at each timestep of the simulation. This enables visualisation of how the uneven distribution of actors affects performance.



Results



- Strips consistently outperforms the Default technique.
- Strips is the fastest when actors are uniformly distributed.
- Modular and Hybrid improve, relative to Default, as the actor distribution becomes non-uniform. This improvement is emphasised more in higher density actor populations.
- In each case, Hybrid is the fastest by the end of the benchmark, when the population has formed clusters.
- The peak improvements are 1.19x and 1.28x in 2D and 3D respectively.
- There is a greater speedup in 3D due to the increased number of cells in each Moore neighbourhood.

Conclusions

Strips provides a near constant time speed up.

- Default and Strips are less capable of handling imbalanced workloads created by a non-uniformly distributed actor population.
- Modular and Hybrid performances improve as standard deviation increases.
- Future research will explore the automatic selection of optimisations at runtime based on actor distributions and hardware properties.

Acknowledgements

This work forms part of an ESPRC funded PhD studentship.

References

[1] P. Goswami, P. Schlegel, B. Solenthaler, and R. Pajarola, "Interactive sph simulation and rendering on the gpu," in Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. Eurographics Association, 2010, pp. 55–64.
[2] R. Chisholm, P. Richmond, and S. Maddock, "A standardised benchmark for assessing the performance of fixed radius near neighbours," in European Conference on Parallel Processing. Springer, 2016, pp. 311–321