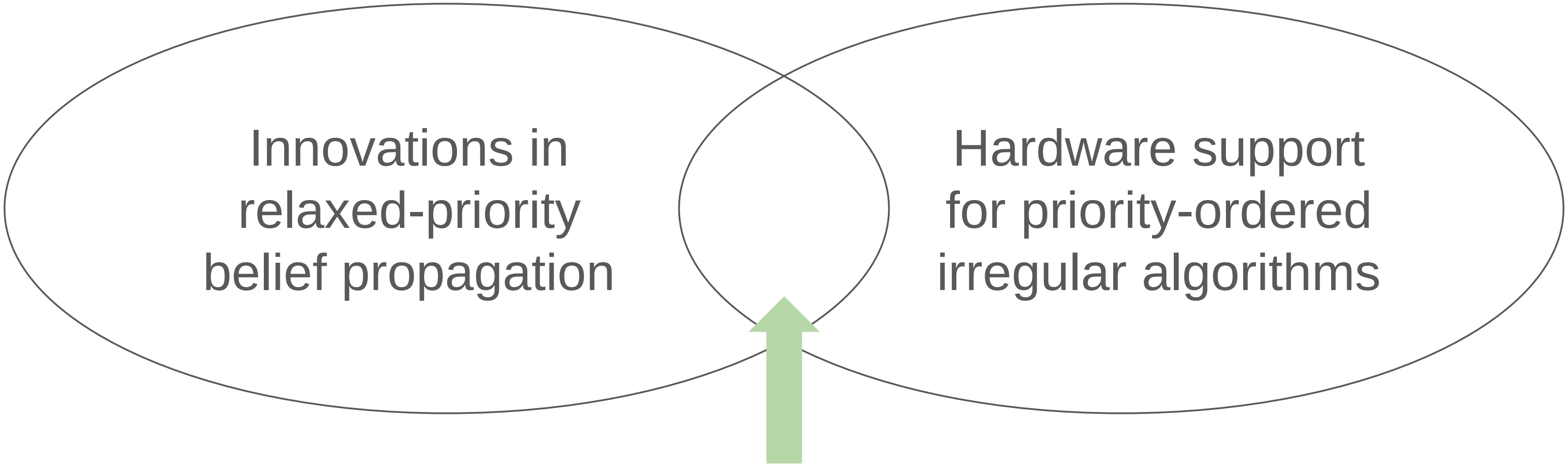
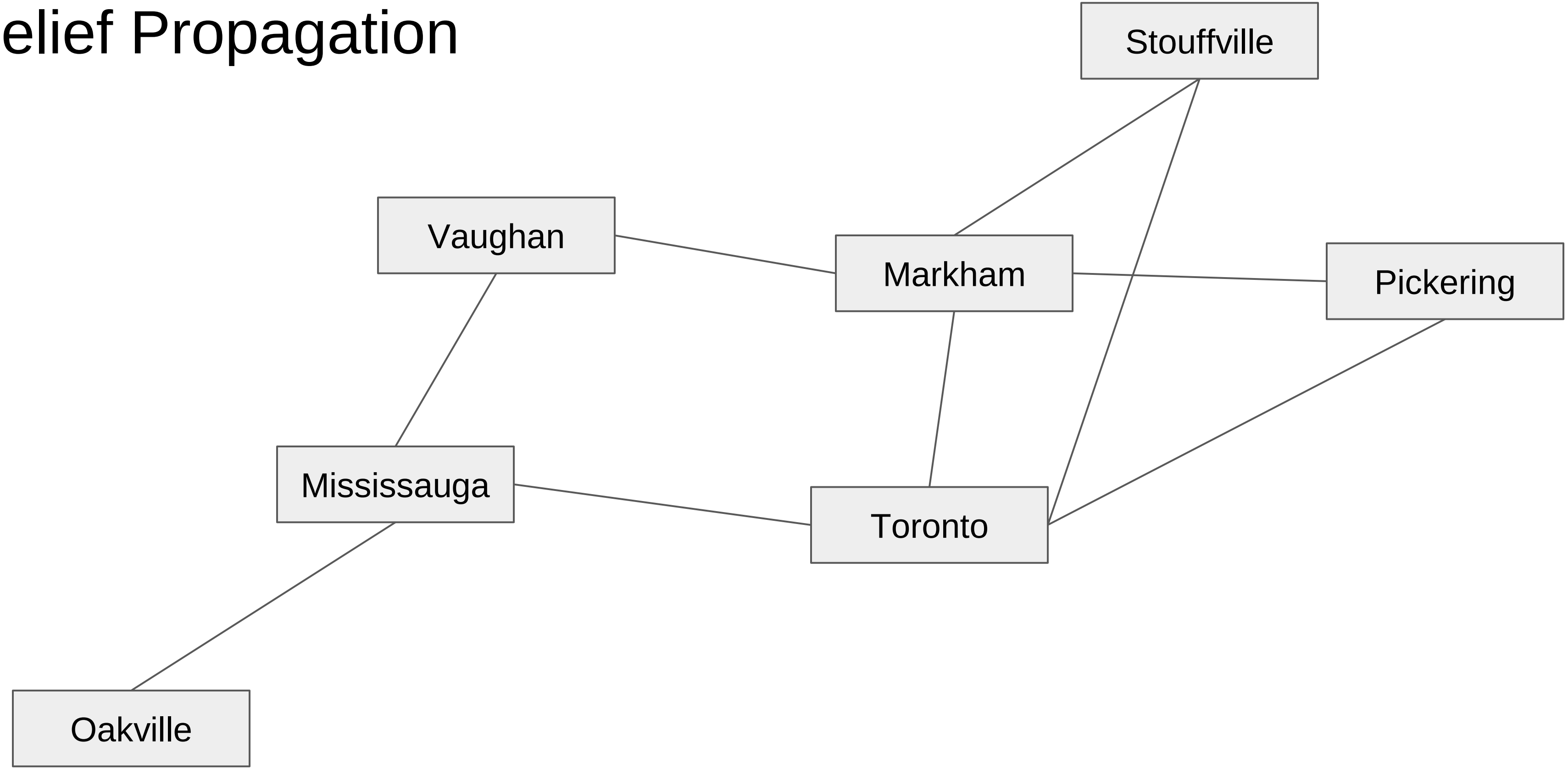
Accelerating Belief Propagation with Task-Based Hardware Parallelism

Balaji Venkatesh, Leo Han, Mark C. Jeffrey

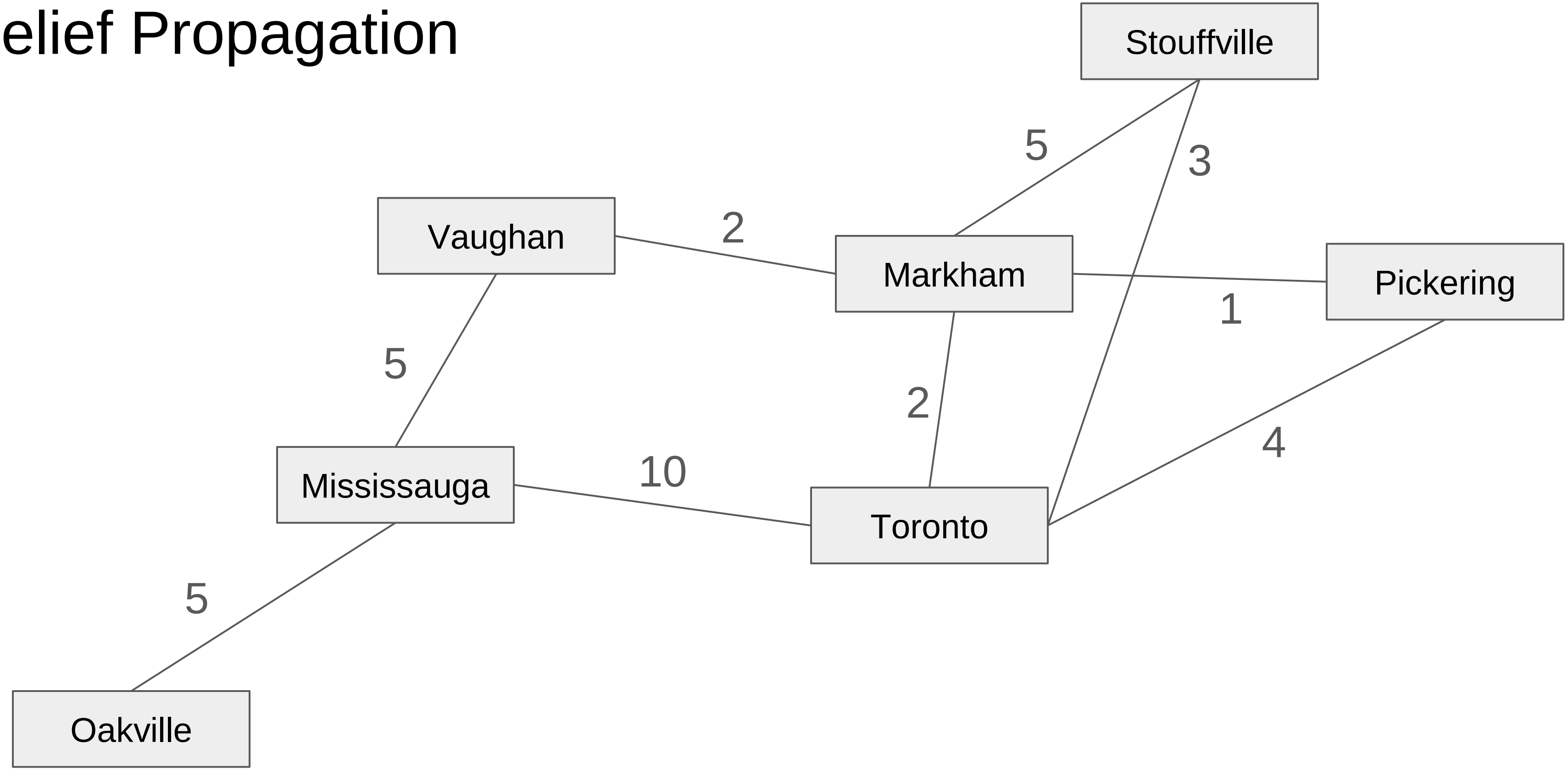
May 27th, 2025


Innovations in relaxed-priority belief propagation
Hardware support for priority-ordered irregular algorithms

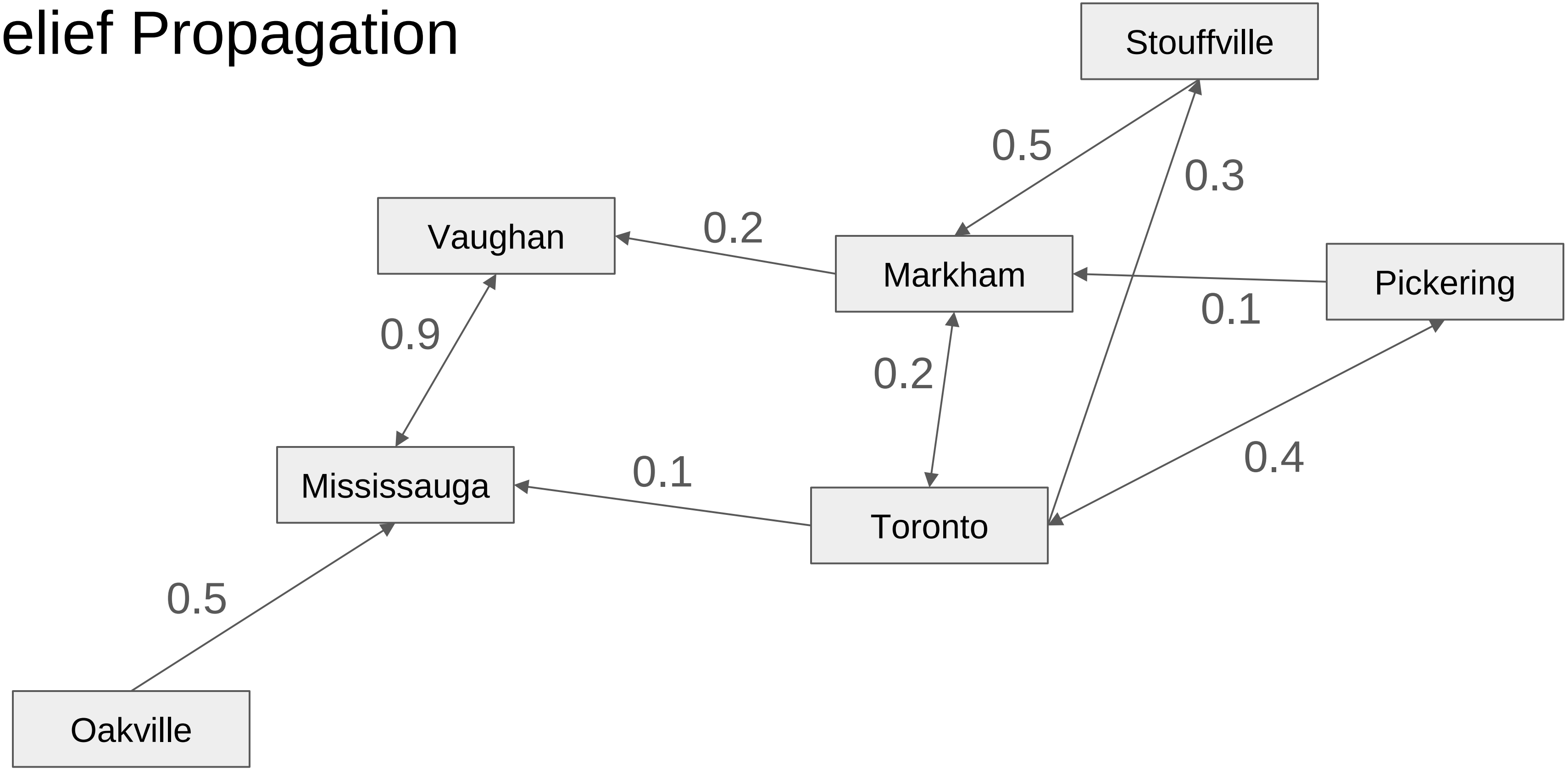
Belief Propagation



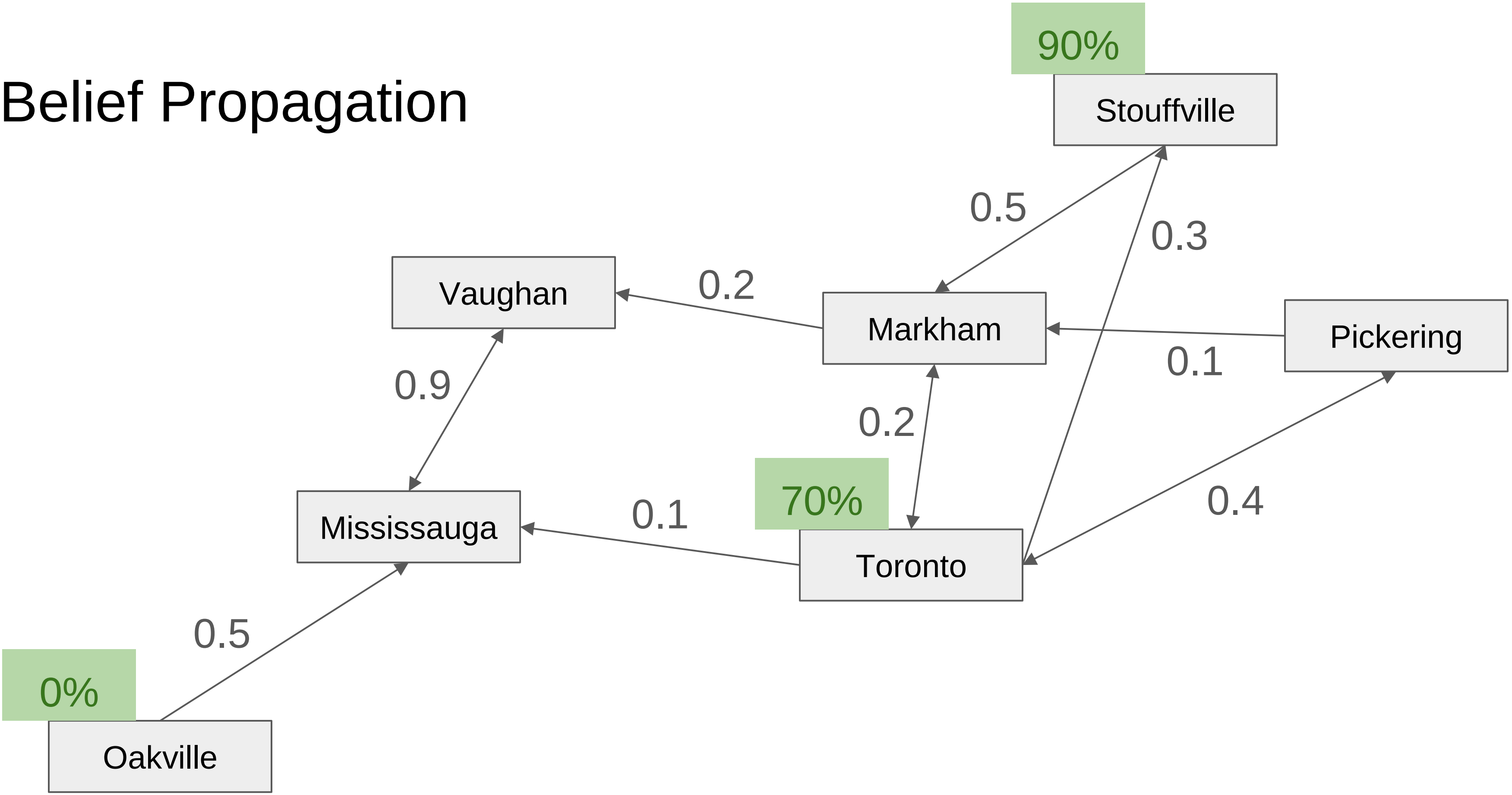
Belief Propagation



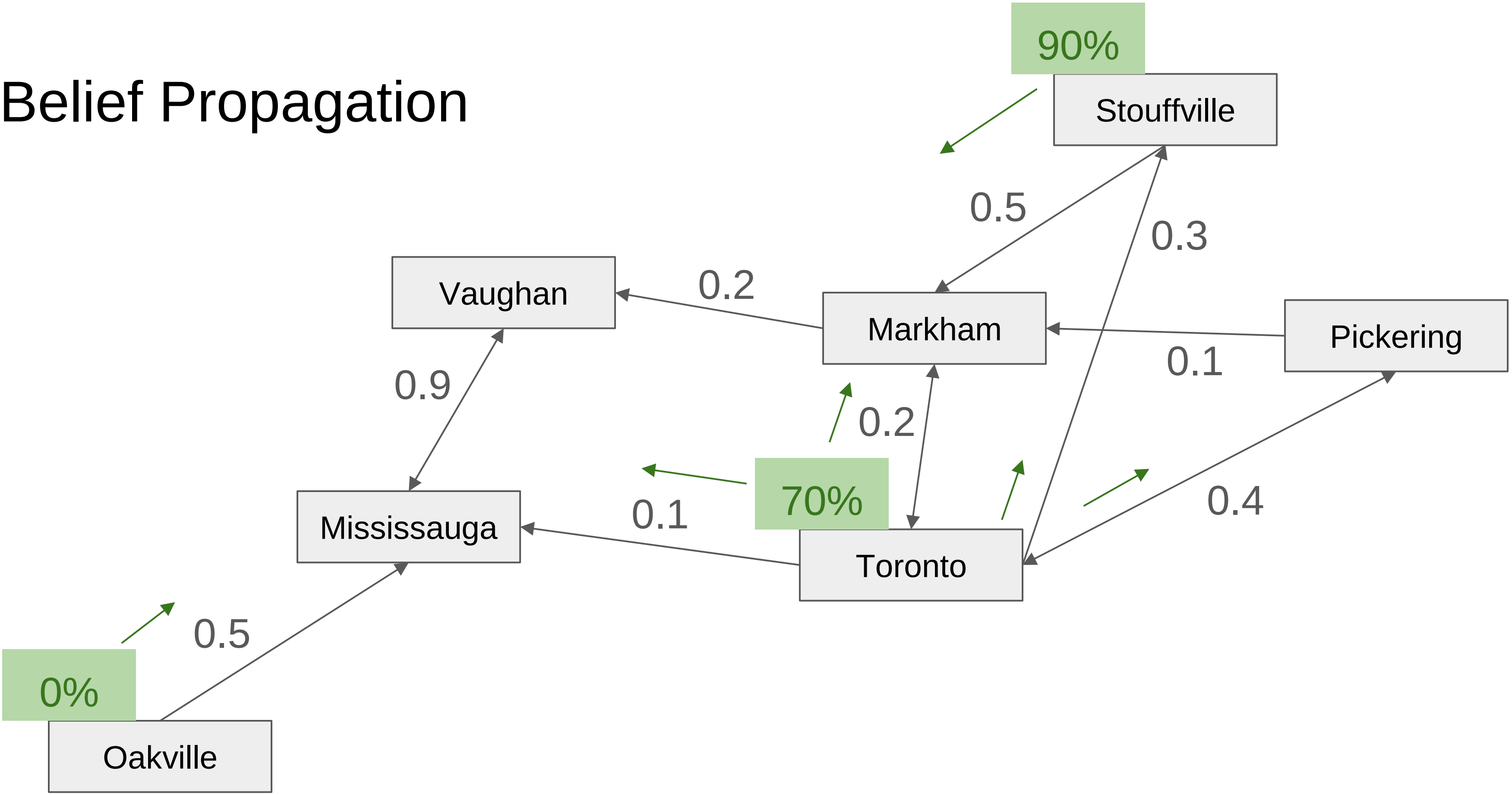
Belief Propagation



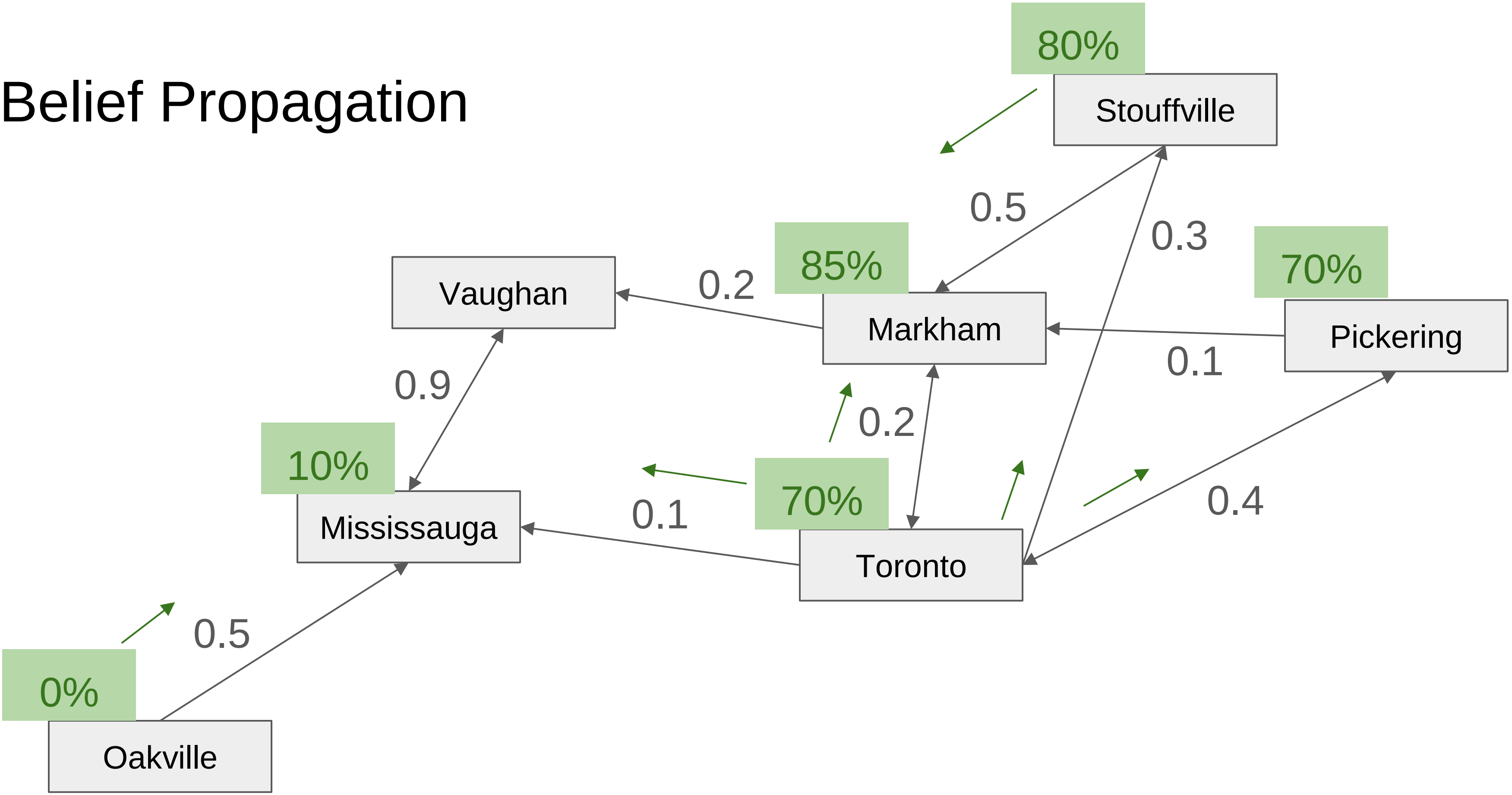
Belief Propagation



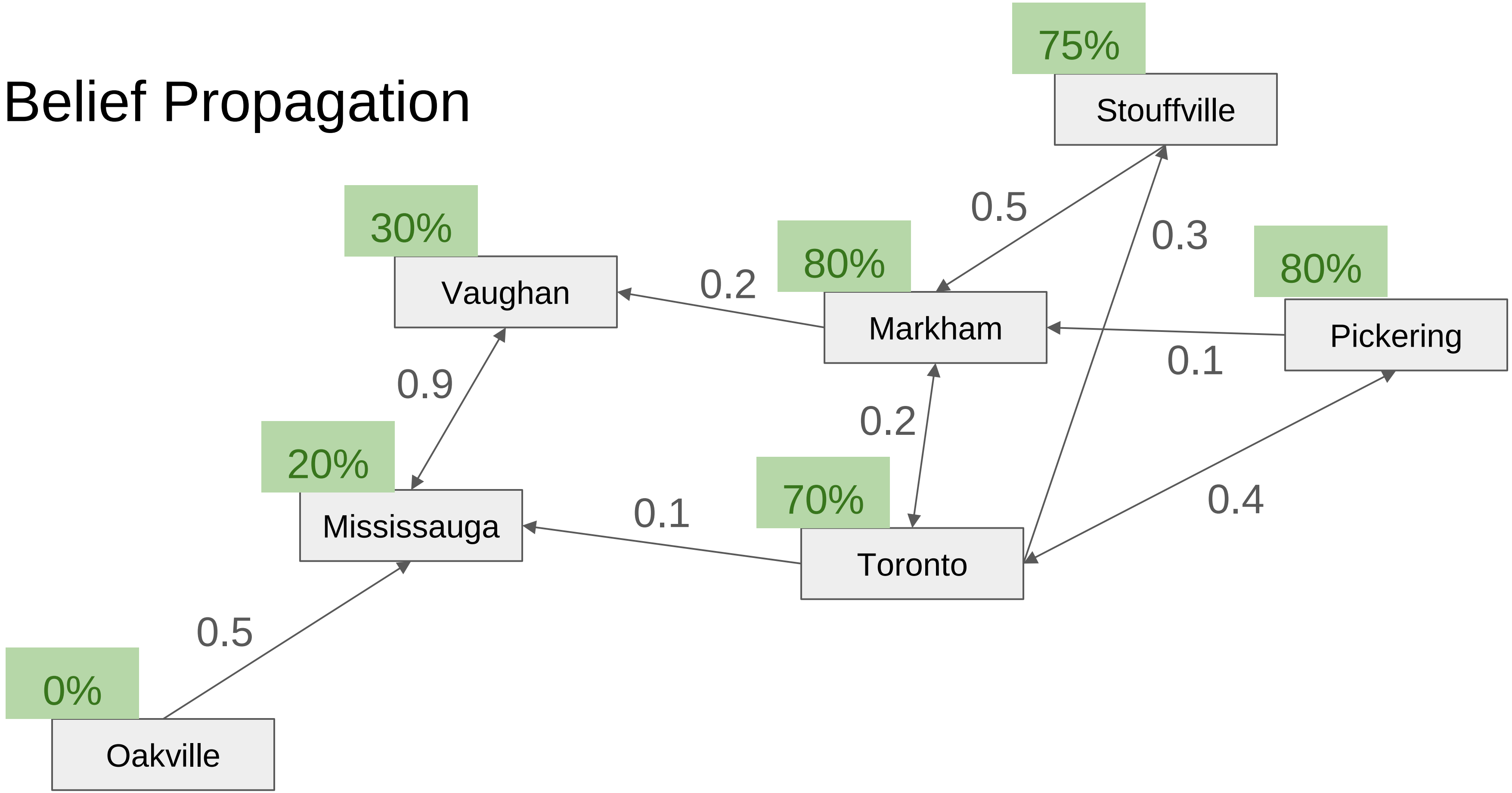
Belief Propagation



Belief Propagation



Belief Propagation



Applications and Significance

Stereo image processing [1]
Workplace safety predictions [2]
Hospital patient experience [3]
Insurance risk analysis [4]
Error correcting codes [5]

These applications can benefit from being able to make faster predictions on larger graphs.

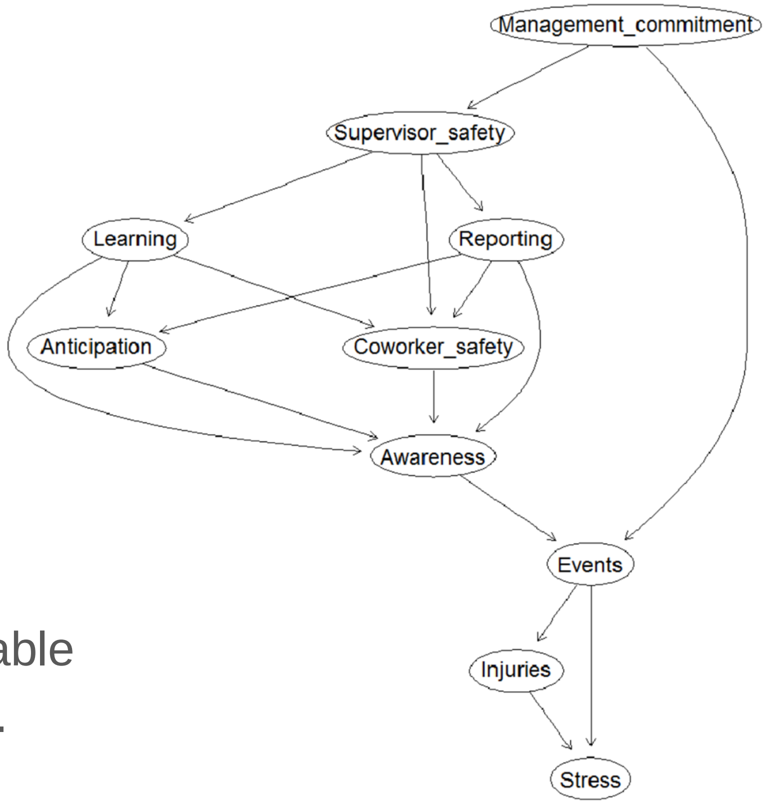


Figure 1: An example of a Markov Random Field being used for workplace safety [2]

Metrics

Convergence coverage
How big are the graphs that converge?

Convergence rate
How fast can we converge?

Scalability
How well does rate improve with more resources?

Efficiency
How well do we deal with priority queue overhead?

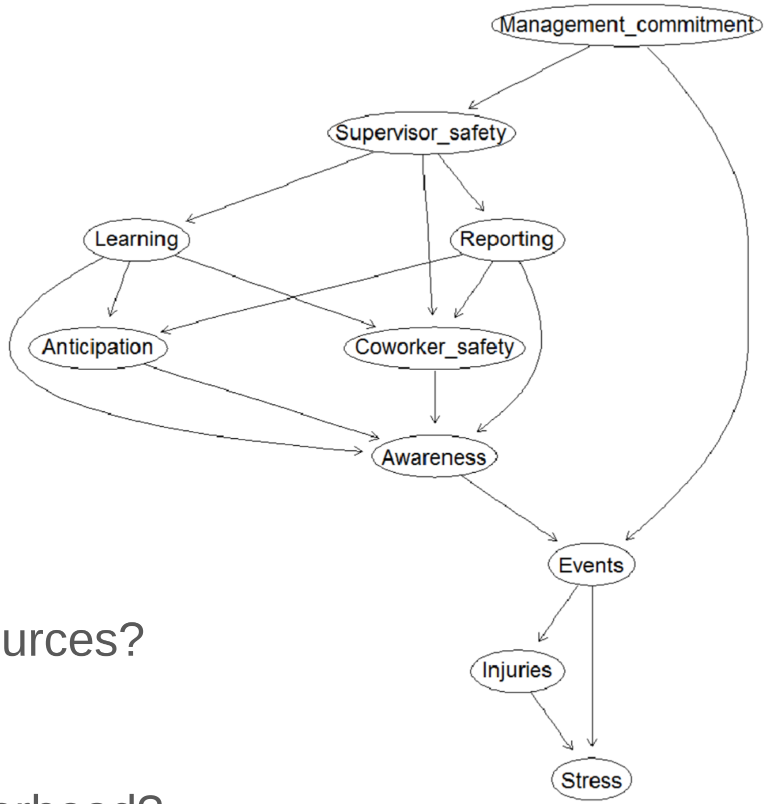
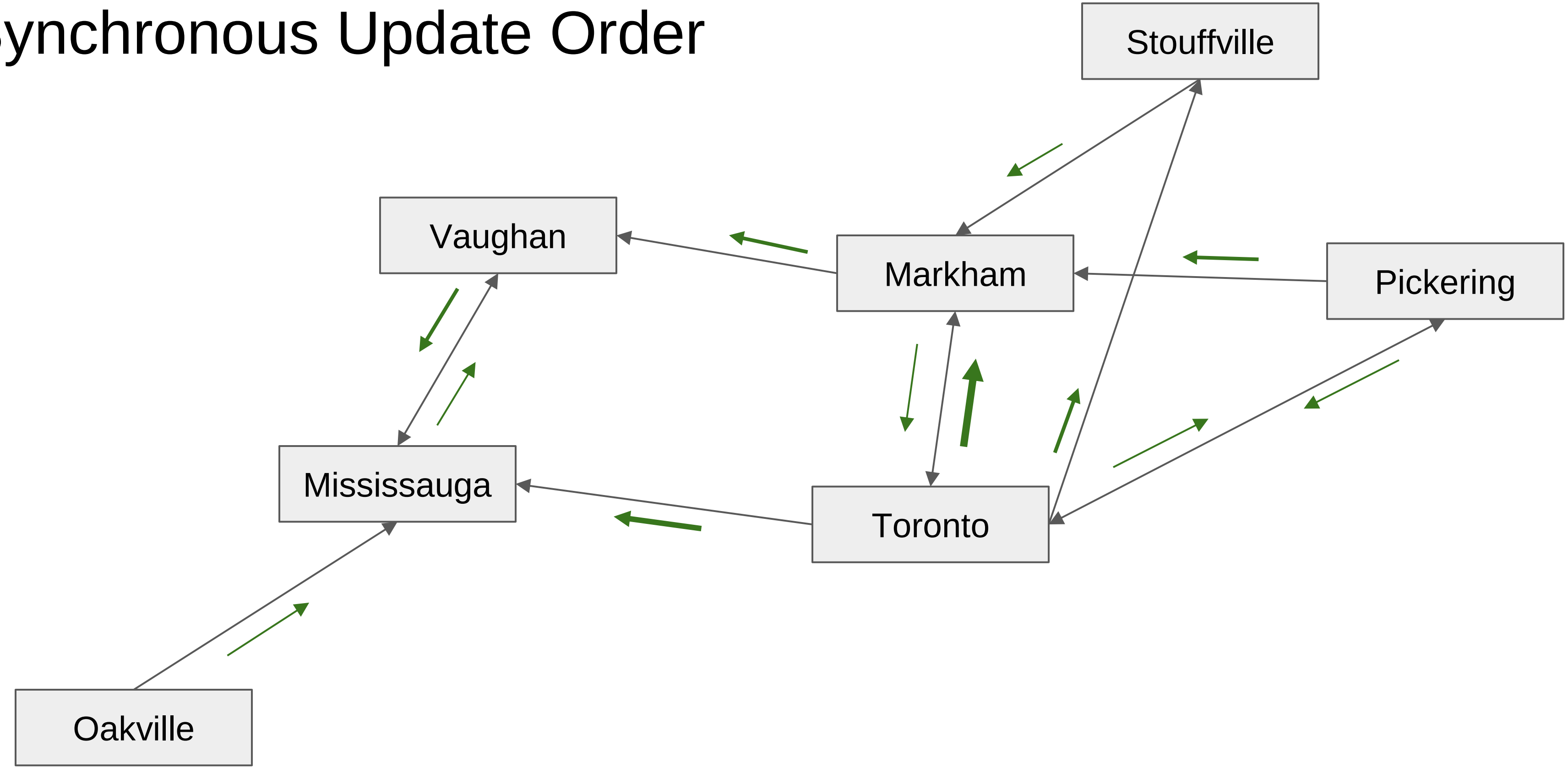


Figure 1: An example of a Markov Random Field being used for workplace safety [2]

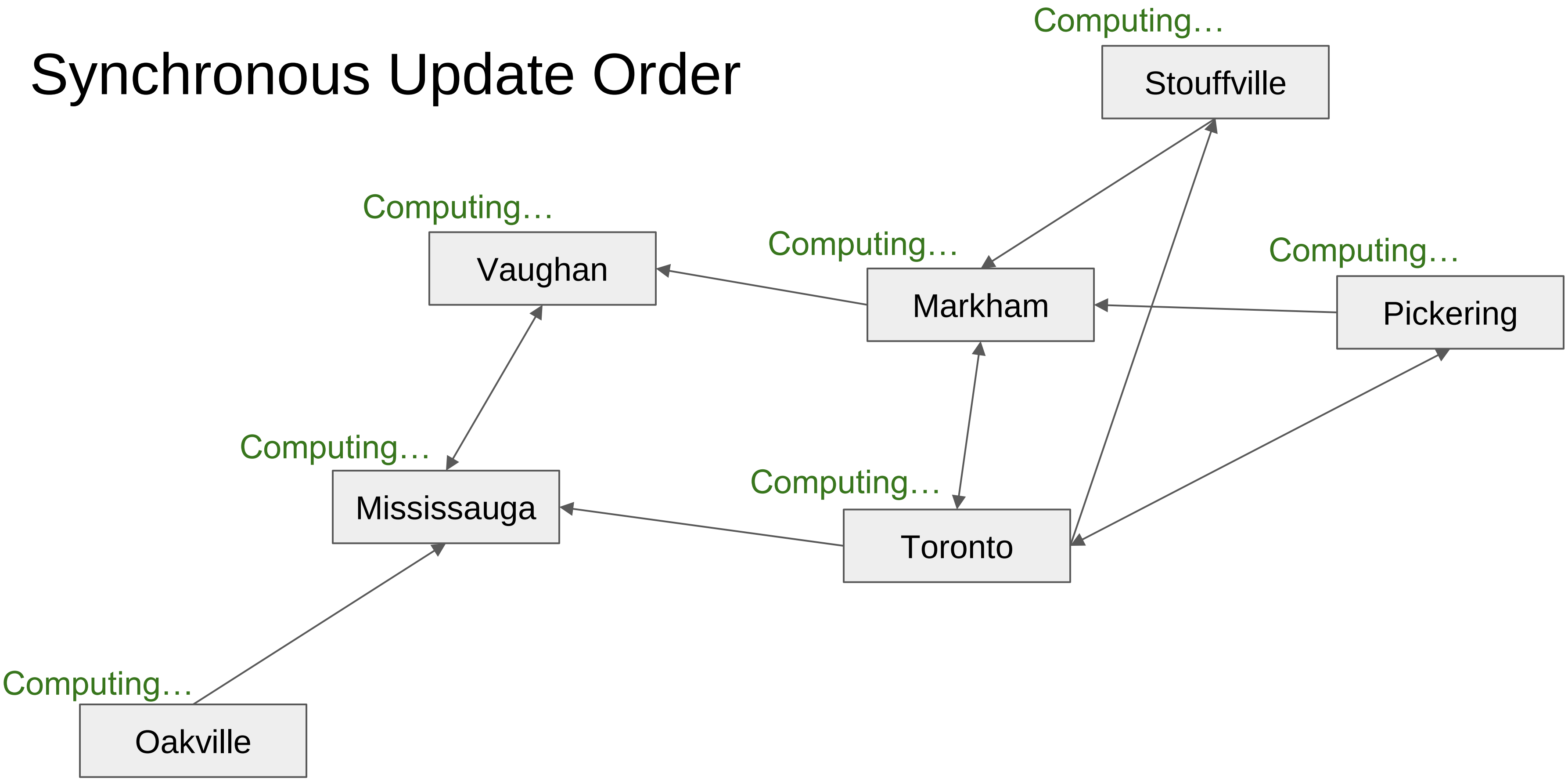
Program Flow

```
while (updates > convergence_criteria) {
      pick_updates();
      compute_beliefs();
      send_updates();
}
```

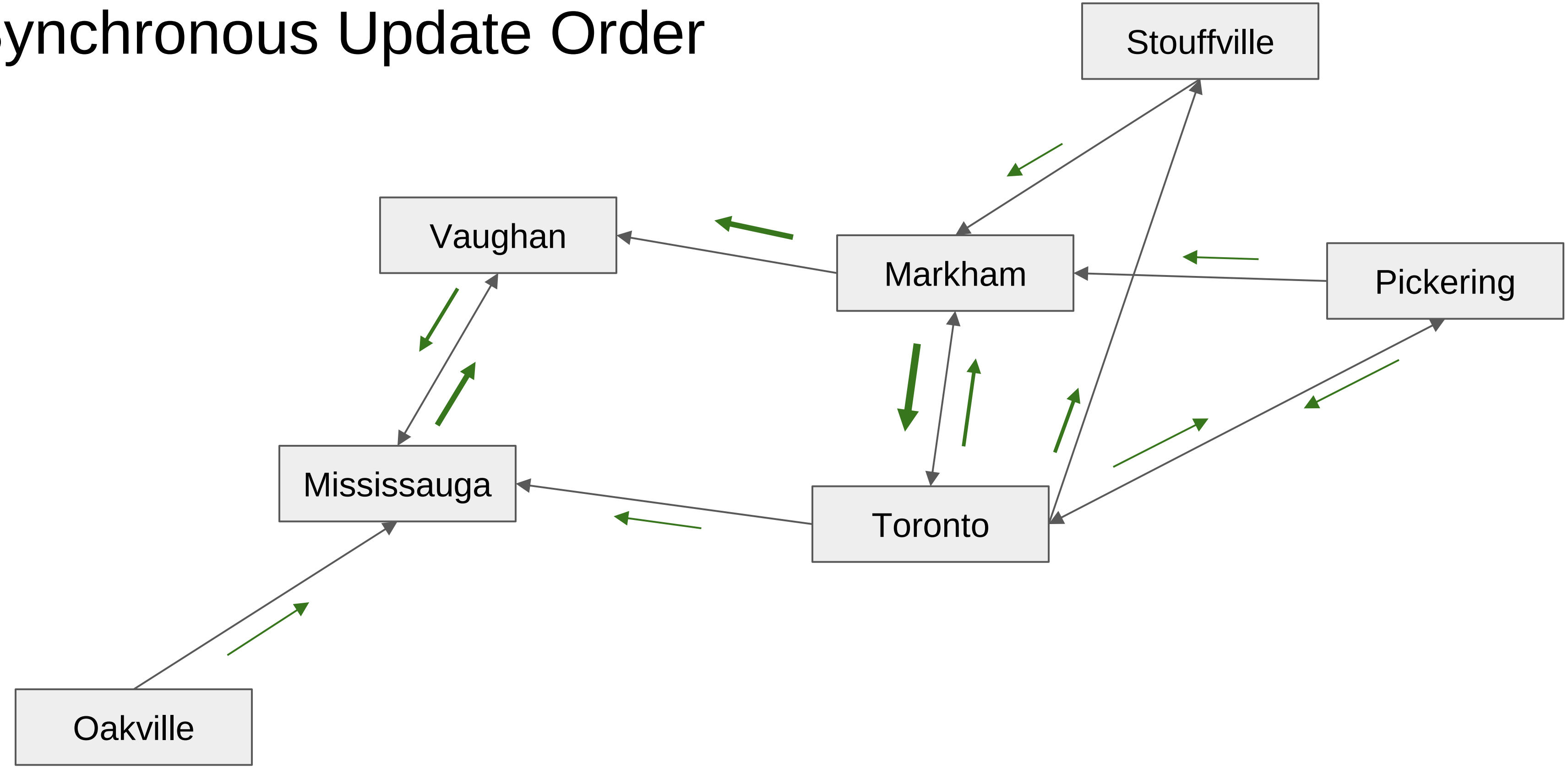
Synchronous Update Order



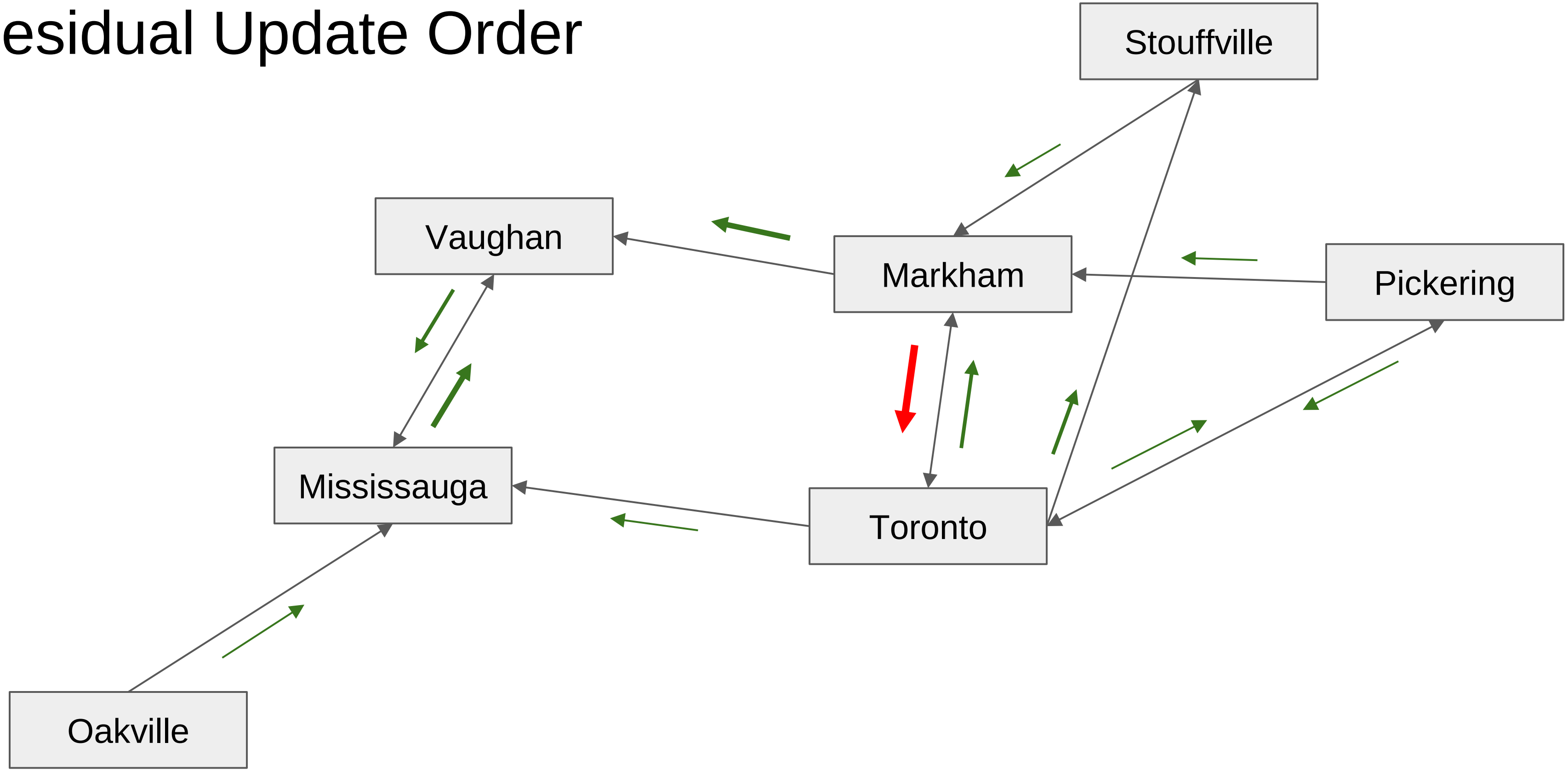
Synchronous Update Order



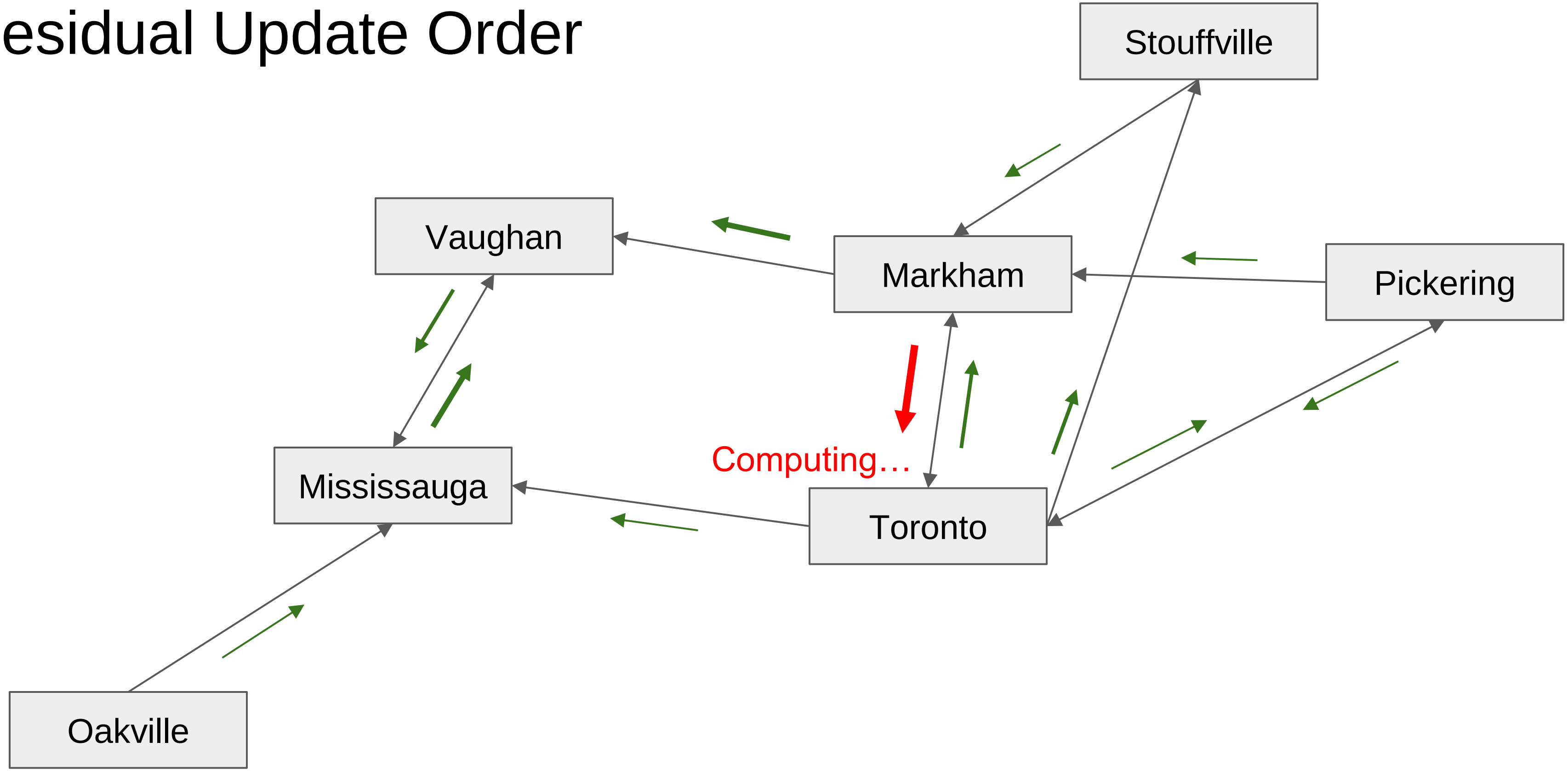
Synchronous Update Order



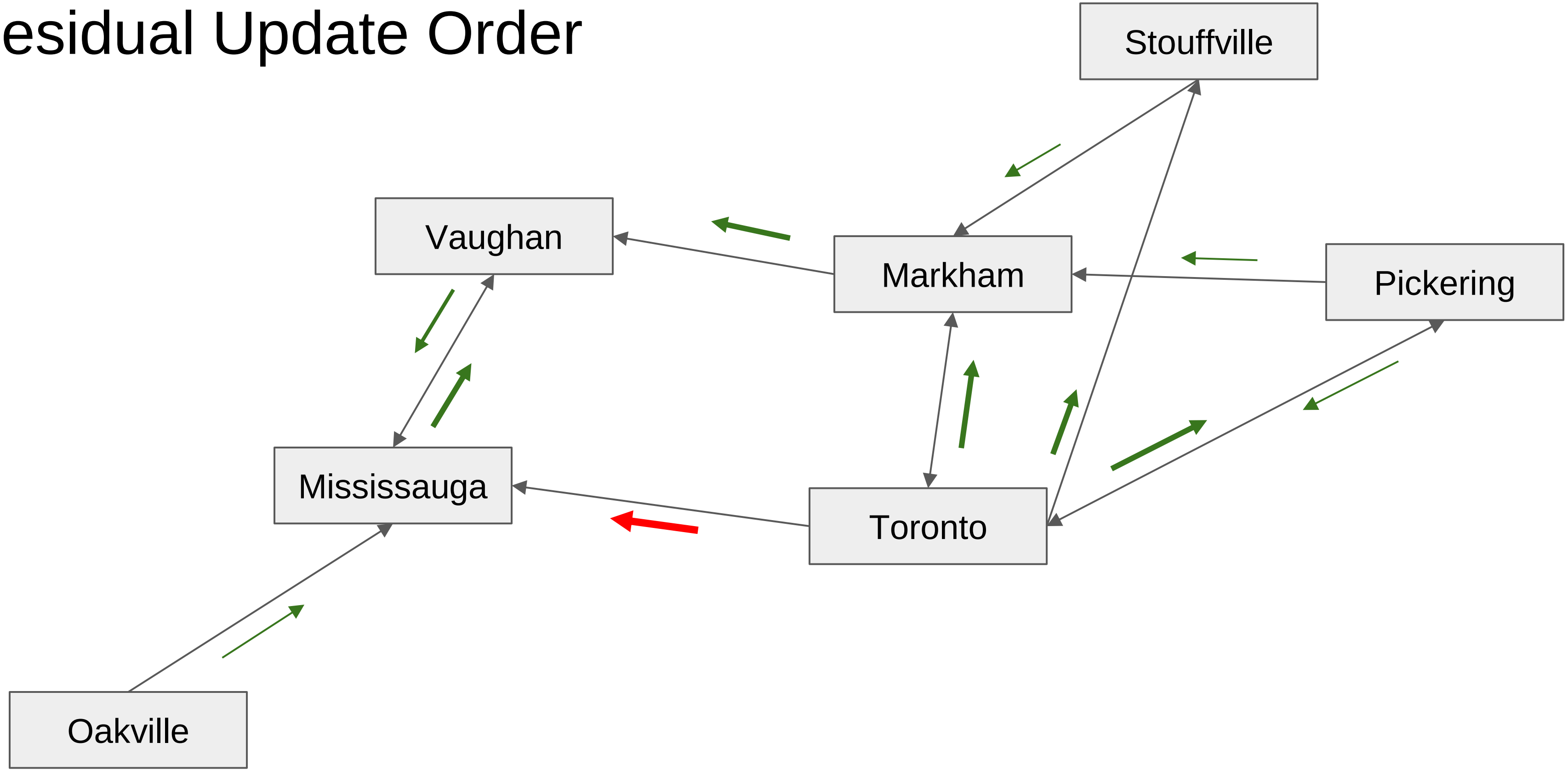
Residual Update Order



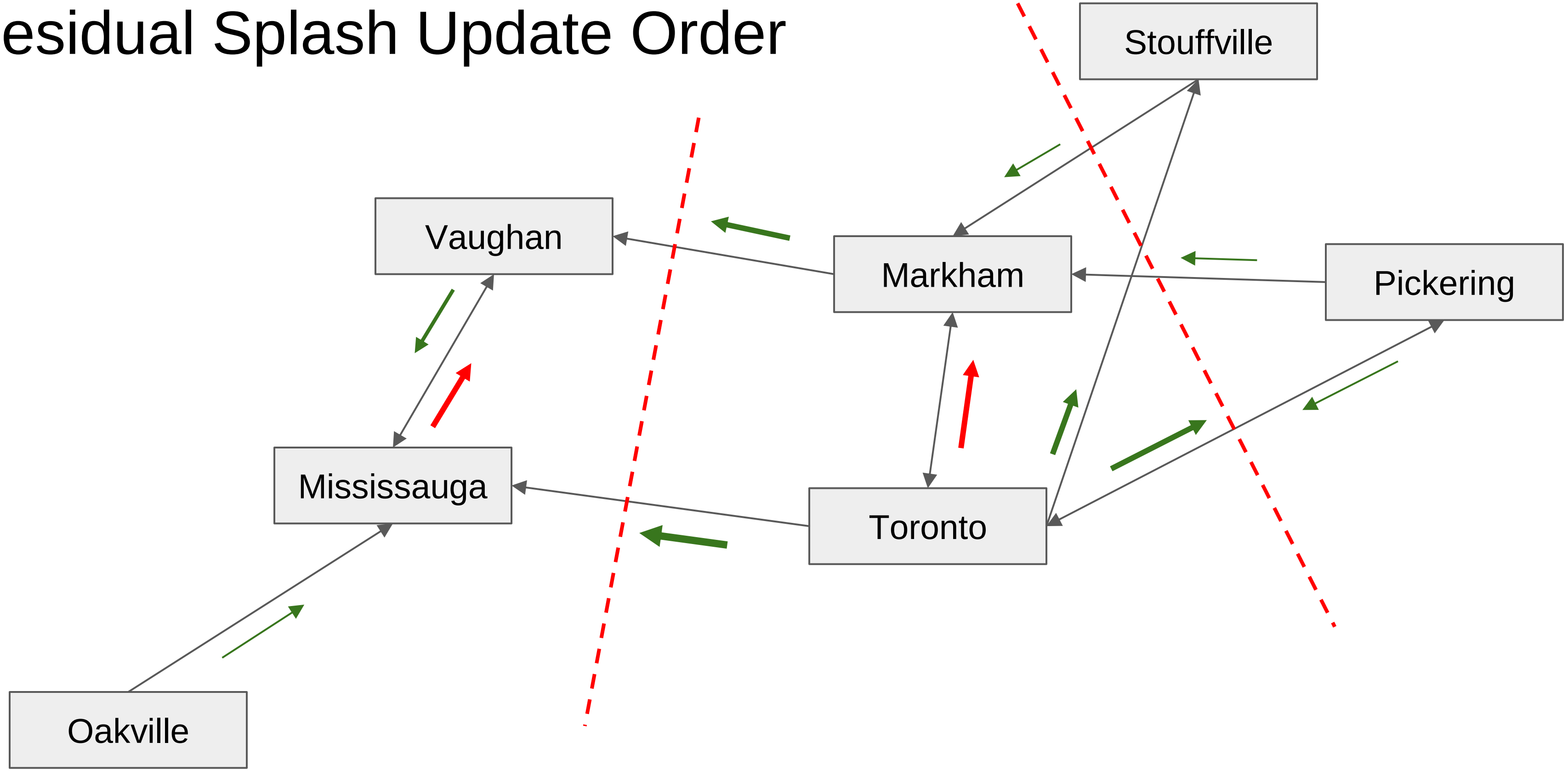
Residual Update Order



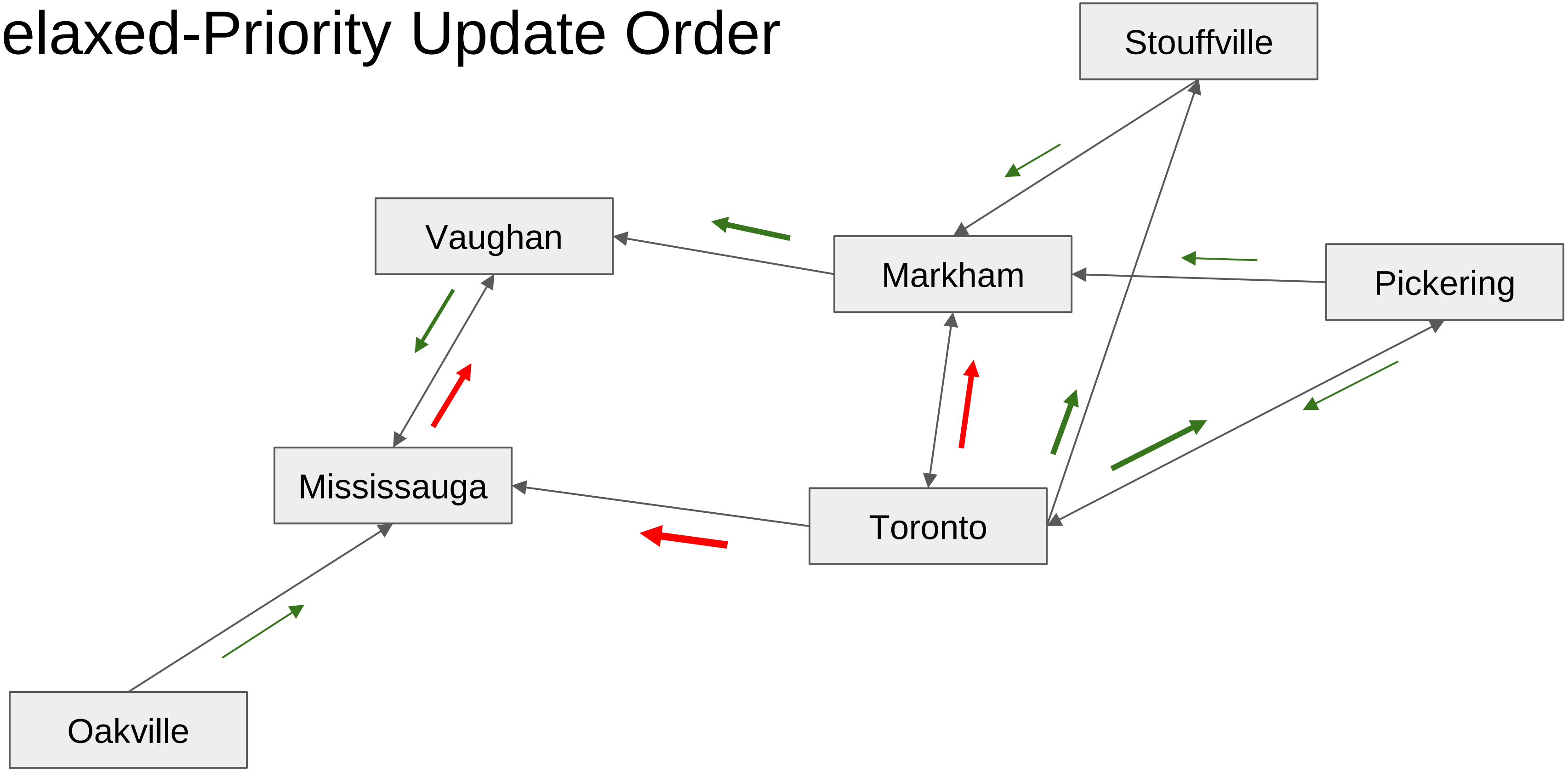
Residual Update Order



Residual Splash Update Order



Relaxed-Priority Update Order



Algorithmic Innovations

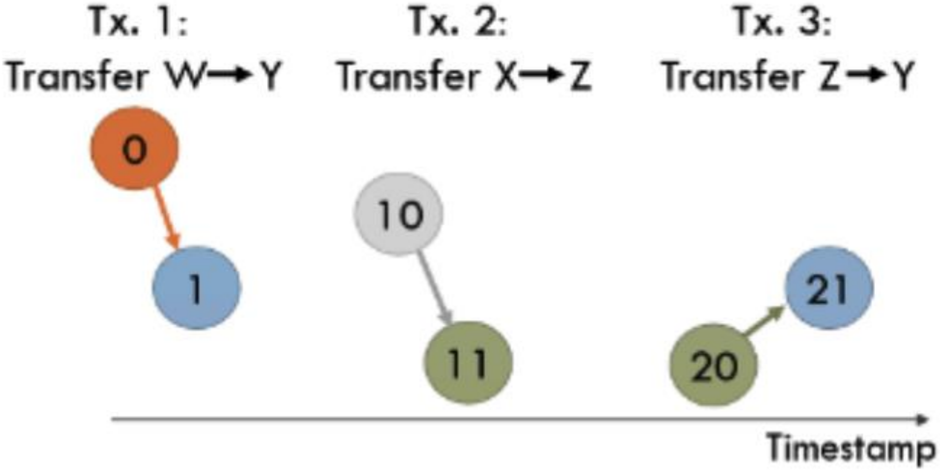
Algorithm	How do updates happen?	Coverage	Rate	Scalability	Efficiency
Bulk Synchronous [6]	Synchronous, single-threaded	Poor	Poor	None	N/A
Parallel [7]	Synchronous, multi-threaded	Poor	Poor	Linear	N/A
Residual [8]	Asynchronous, strictly-ordered	Good	Good	None	Poor
Residual Splash [9]	Asynchronous, strictly-ordered and partitioned	Good	Good	Sub-linear	Poor
Relaxed-priority [10]	Loosely-ordered asynchronous	Okay	Okay	Sub-linear	Okay
Speculative Parallel Residual [11]	Asynchronous, strict-order avoided using speculation	Good	Good	Linear	Good

Task-Based Hardware Parallelism

Spatially-Located Ordered Tasks

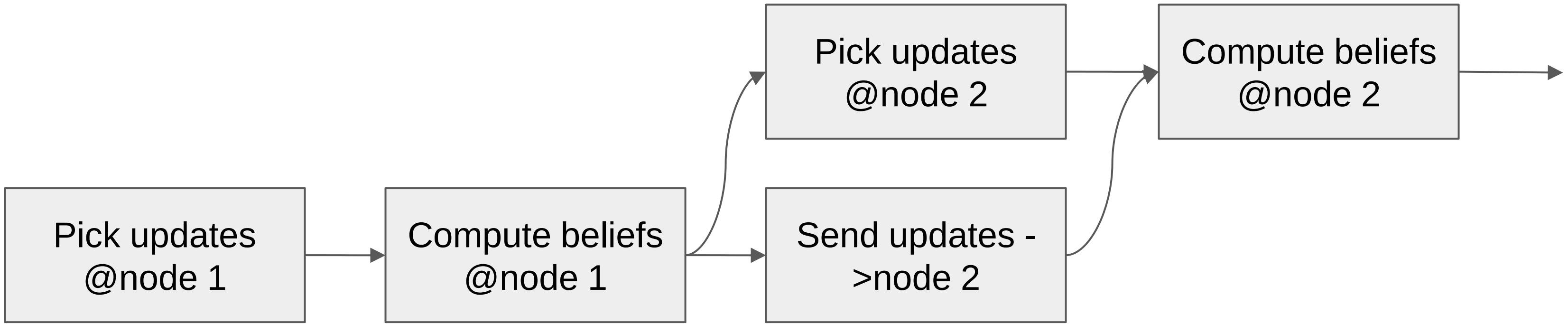
Chronos [12]

Account (object)	Balance
W	$100
X	$1500
Y	$200
Z	$400



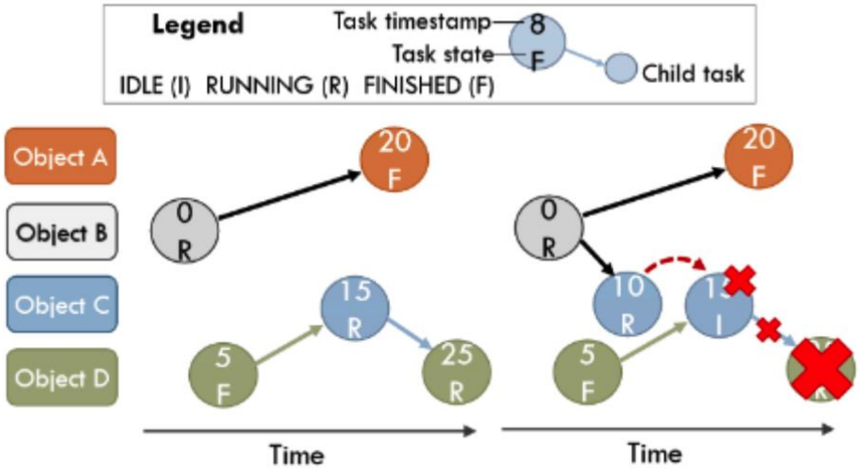
Task-Based Hardware Parallelism

```
while (updates > convergence_criteria) {
     pick_updates();
     compute_beliefs();
     send_updates();
}
```



Speculation Extracts Parallelism by Relaxing Order

Chronos [12]



(a) Tasks can run out of timestamp order

(b) Selective aborts: Task 10 aborts 15 and 25, but not 20

Research Gap

Existing accelerators are

- overly specific [5]
- too costly to implement [11]

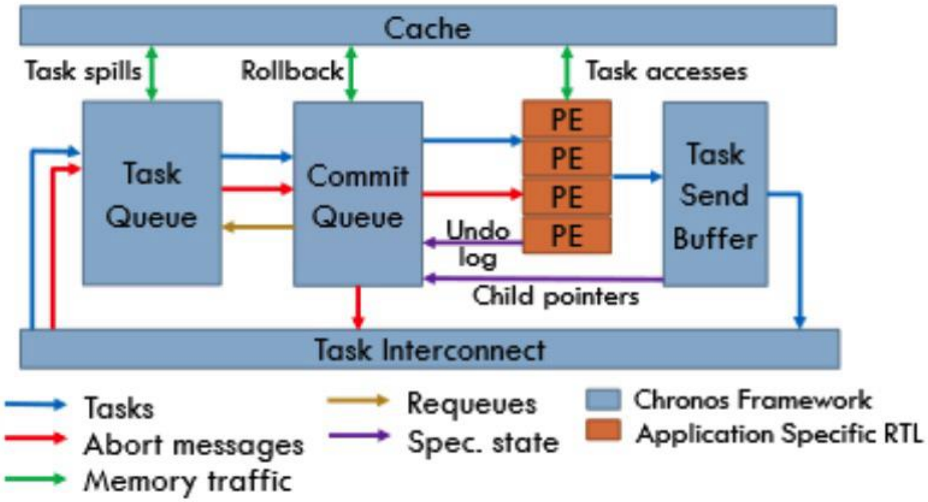
General Belief Propagation Accelerator on Chronos

Design Goal

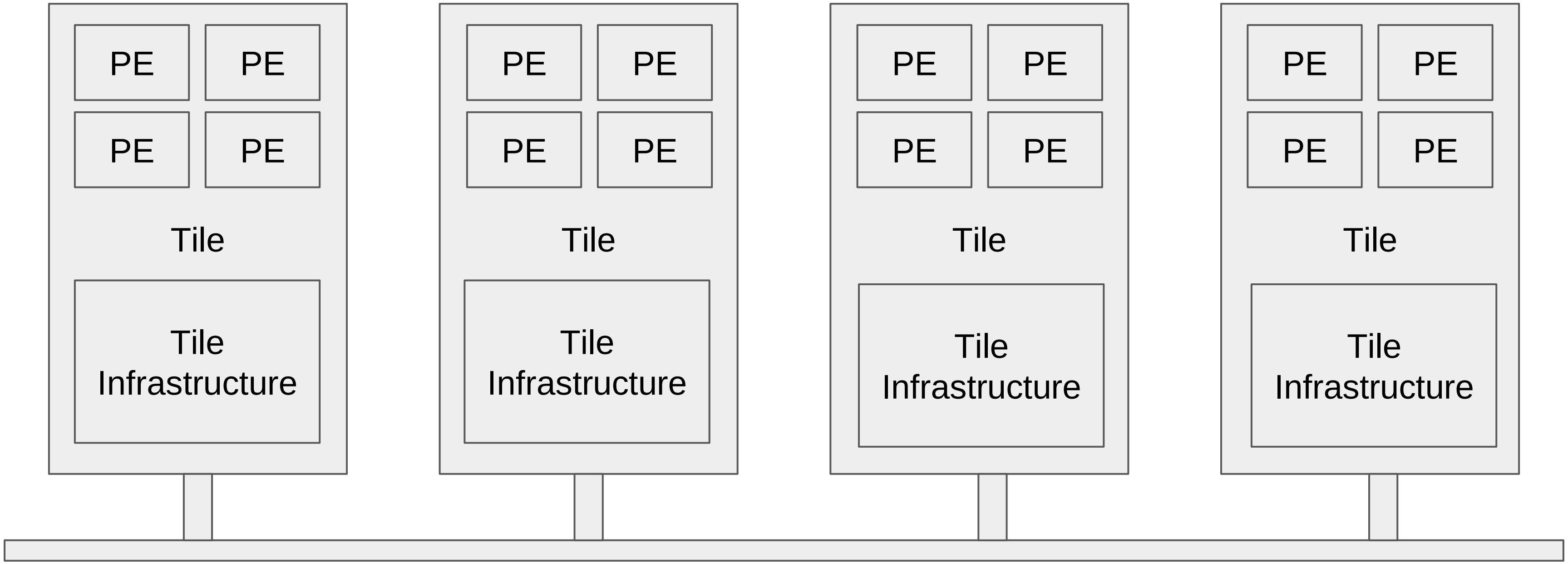
Eliminate deadlocks while retaining functional correctness

Scaling and optimizing to improve:

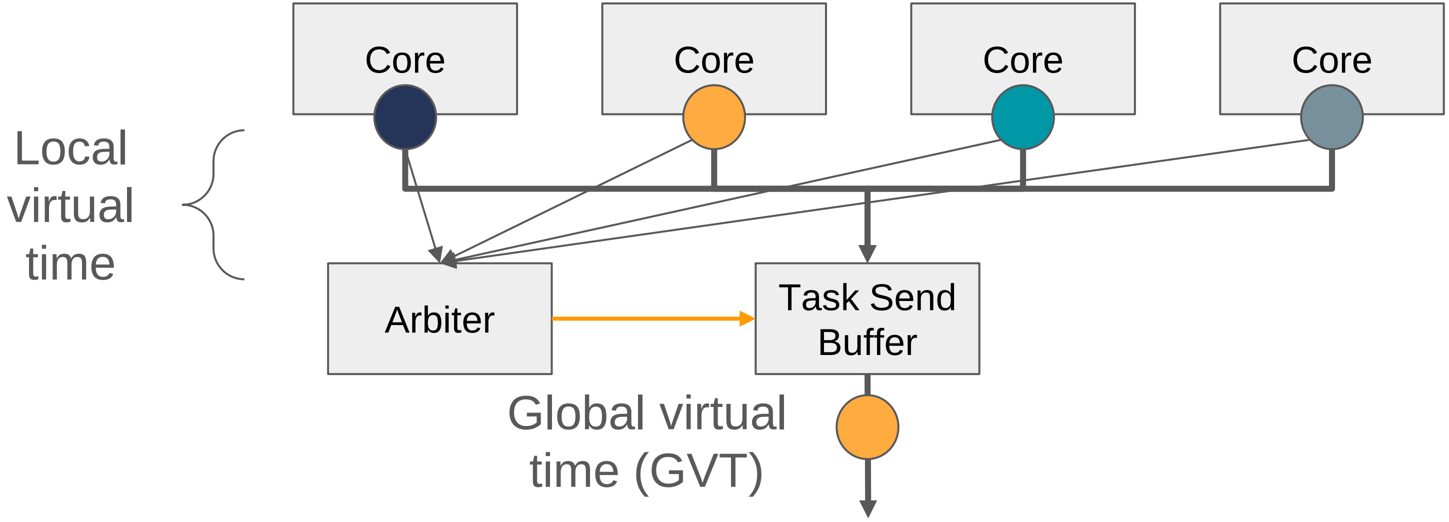
- Convergence coverage
- Convergence rate
- Scalability
- Efficiency



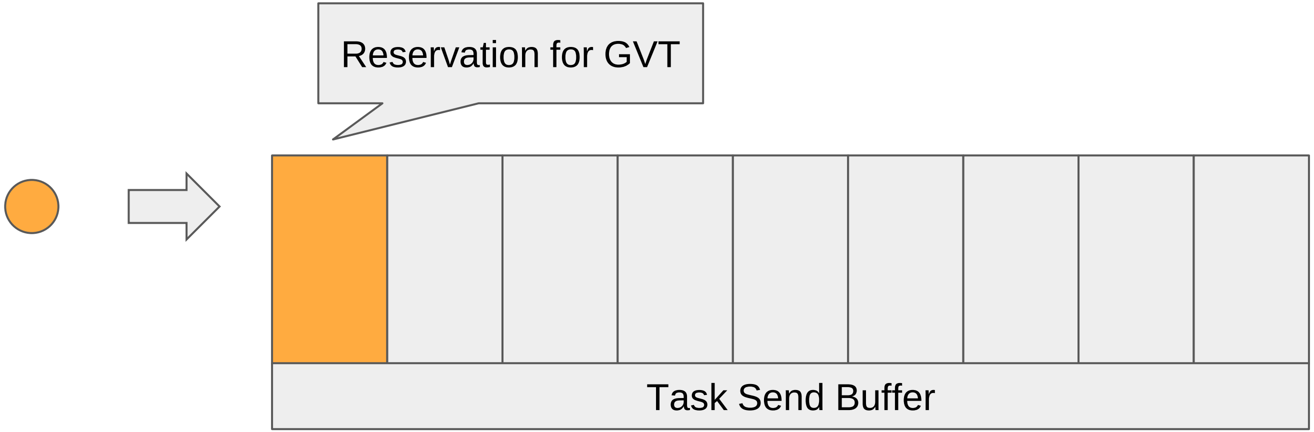
System Diagram



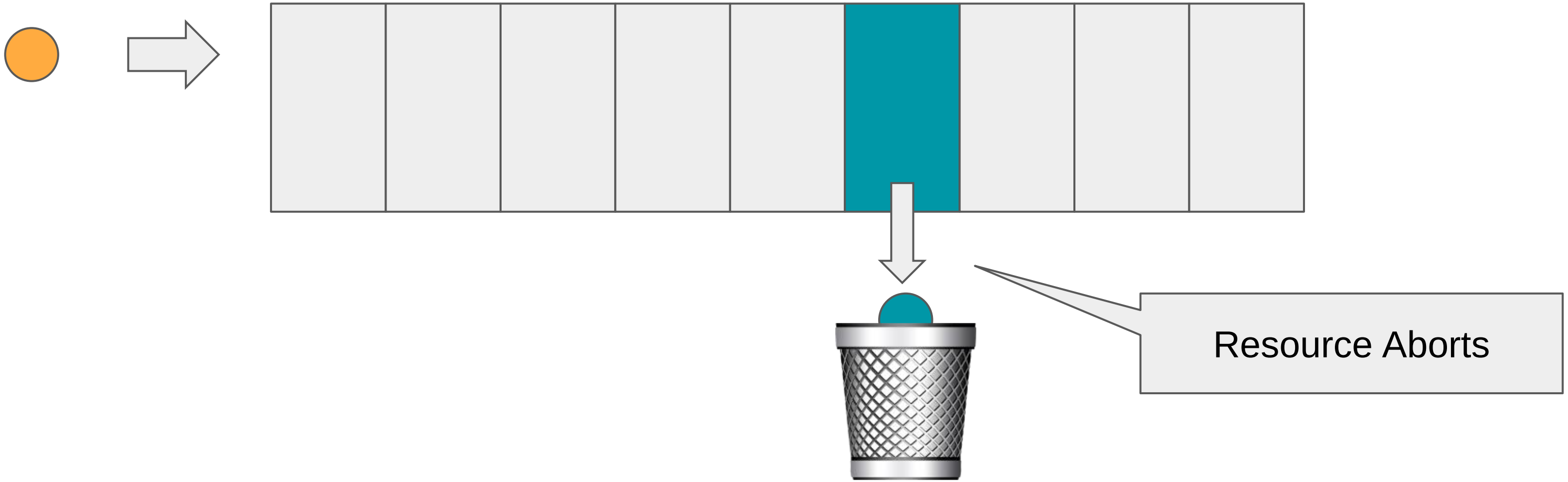
Deadlock Avoidance Prioritizes the GVT



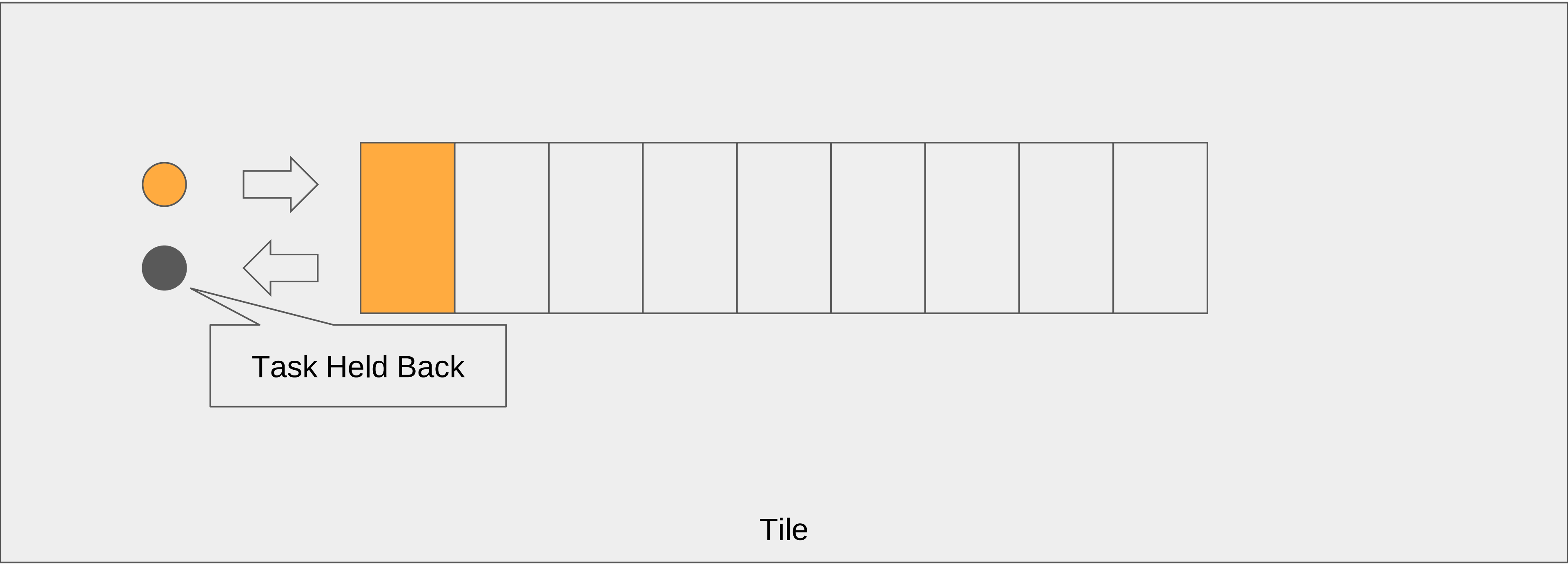
Prioritizing the GVT with Reservations



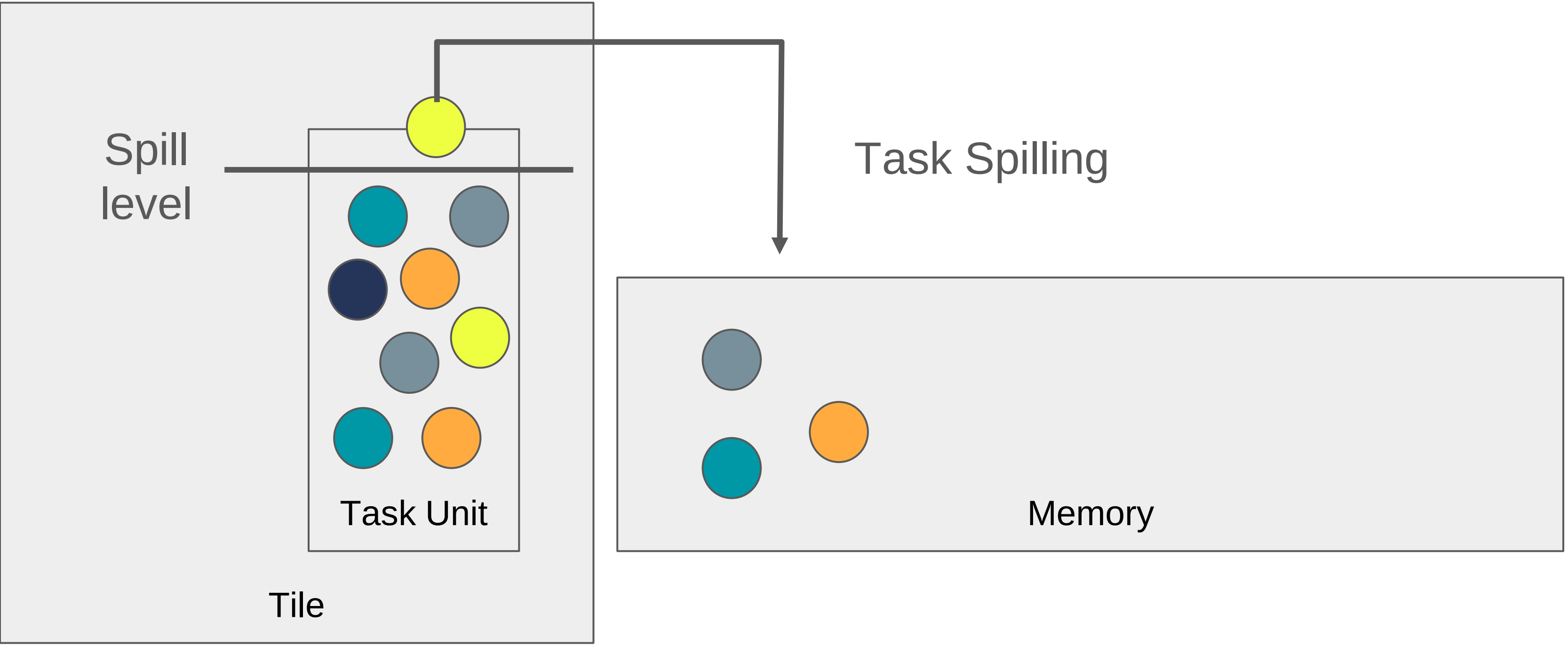
Prioritizing the GVT with Resource Aborts



Deadlock Avoidance



Deadlock Avoidance



Results

1. Coverage improved by removing deadlocks that occur with large graphs
2. Rate improved by optimizing size and configuration of accelerator
3. Scalability demonstrated with more PEs computing larger graphs
4. Efficiency used to extract parallelism by lowering priority queue overhead

Conclusions

Relaxed-priority BP and task-based parallelism can be combined to improve convergence coverage, convergence rate, and scalability of belief propagation through increased efficiency.

Implementing the accelerator on an FPGA makes it accessible for use in broader applications.

References

[0] L. Han, “Accelerating Belief Propagation with Hardware Speculative Parallelism,” Undergrad Thesis, University of Toronto, Toronto, Canada, 2023.

[1] T. Yan, X. Yang, G. Yang, and Q. Zhao, “Hierarchical Belief Propagation on Image Segmentation Pyramid,” IEEE Transactions on Image Processing, 2023, doi: 10.1109/TIP.2023.3299192.

[2] M. C. E. Simsekler and A. Qazi, “Adoption of a Data-Driven Bayesian Belief Network Investigating Organizational Factors that Influence Patient Safety,” Risk Analysis, vol. 42, no. 6, pp. 1277–1293, Jun. 2022, doi: 10.1111/risa.13610.

[3] A. Al Nuairi, M. C. E. Simsekler, A. Qazi, and A. Sleptchenko, “A data-driven Bayesian belief network model for exploring patient experience drivers in healthcare sector,” Ann Oper Res, Jun. 2023, doi: 10.1007/s10479-023-05437-9.

[4] L. Mkrtchyan, U. Straub, M. Giachino, T. Kocher, and G. Sansavini, “Insurability risk assessment of oil refineries using Bayesian Belief Networks,” Journal of Loss Prevention in the Process Industries, vol. 74, p. 104673, Jan. 2022, doi: 10.1016/j.jlp.2021.104673.

[5] Y. Sun, Y. Shen, W. Song, Z. Gong, X. You, and C. Zhang, “LSTM Network-Assisted Belief Propagation Flip Polar Decoder,” in 2020 54th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA: IEEE, Nov. 2020, pp. 979–983. doi: 10.1109/IEEECONF51394.2020.9443504.

[6] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Rev. 2. ed., Transferred to digital printing. in The Morgan Kaufmann series in representation and reasoning. San Francisco, Calif: Morgan Kaufmann, 2009.

[7] J.-F. Yan, J. Zeng, Y. Gao, and Z.-Q. Liu, “Communication-efficient algorithms for parallel latent Dirichlet allocation,” Soft Comput, vol. 19, no. 1, pp. 3–11, Jan. 2015, doi: 10.1007/s00500-014-1376-8.

[8] G. Elidan, I. McGraw, and D. Koller, “Residual Belief Propagation: Informed Scheduling for Asynchronous Message Passing,” 2006. [Online]. Available: http://www.robotics.stanford.edu/~galel/papers/ElidanRBP.pdf

[9] J. Gonzalez, Y. Low, and C. Guestrin, “Residual Splash for Optimally Parallelizing Belief Propagation,” in Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, PMLR, Apr. 2009, pp. 177–184. [Online]. Available: https://proceedings.mlr.press/v5/gonzalez09a.html

[10] V. Aksenov, D. Alistarh, and J. H. Korhonen, “Relaxed Scheduling for Scalable Belief Propagation,” Dec. 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/fdb2c3bab9d0701c4a050a4d8d782c7f-Paper.pdf

[11] G. Posluns, Y. Zhu, G. Zhang, and M. C. Jeffrey, “A scalable architecture for reprioritizing ordered parallelism,” in Proceedings of the 49th Annual International Symposium on Computer Architecture, New York New York: ACM, Jun. 2022, pp. 437–453. doi: 10.1145/3470496.3527387.

[12] M. Abeydeera and D. Sanchez, “Chronos: Efficient Speculative Parallelism for Accelerators,” Mar. 2020. doi: 10.1145/3373376.3378454.