

GPU Replica Exchange Monte Carlo

Lane College, Jackson TN
Department of Mathematics
Elijah MacCarthy, Ph.D
emaccarthy@lanecollege.edu

September 1, 2020

GPUs and the General Purpose GPU (*GPGPU*)

- GPUs have become very popular in scientific computing in the past few decades.
- Despite being originally developed for the gaming industry, it has spread its influence to numerous areas.
- Thus, GPGPUs are being used in astronomy, medicine, finance, mathematics and bioinformatics.
- GPUs are massively parallel and outperform even the best parallelized algorithms from CPUs.

Application Programming Interfaces (APIs) for GPU Programming

- CUDA is one API used mostly with GPUs.
- However, we use OpenACC here because:
 - OpenACC allows for preprocessor directives included in program.
 - Program does not have to be completely modified as in CUDA.
 - Allows for program to be implemented even without GPUs unlike CUDA programs which are solely for GPUs.
 - OpenACC supports all accelerators.

OpenACC Granularity

- There are three levels in OpenACC execution model.
- The gangs, workers and vector.
- This is supposed to map to any architecture.



Parallelization of Replica Exchange Monte Carlo (REMC)

- Monte Carlo simulations use random numbers to model populations
- REMC is a Monte Carlo Method that involves swapping of replicas at different temperatures
- In REMC, simulations of several replicas are implemented at different temperatures, T
- After some Monte Carlo time step, updates are performed

Replica Exchange Monte Carlo (REMC)

- Updates are accepted based on Metropolis criterion with probability
 - p given by:

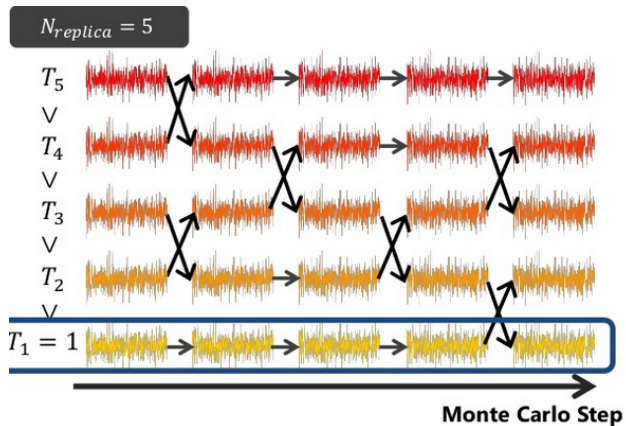
$$p = \min(1, \exp[-\beta(E_{new} - E_{old})]), \quad (1)$$

- Attempt to exchange neighboring replicas is initiated following a number of Monte Carlo updates
- Replica exchanges are accepted based on given probabilities

$$p = \min(1, \exp[(\beta_i - \beta_{i+1})(E_{new} - E_{old})]), \quad (2)$$

- With Parallel REMC, several replicas are simulated in Parallel to reduce computational time

Schema of REMC



Parallelization of Replica Exchanges

- REMC perform concurrent simulations of n different replicas of the Monte Carlo system, each running under different temperatures
- Systems at high temperatures are able to explore a larger volume of the phase space than at low temperatures
- During the swapping phase, replicas are exchanged between temperatures by a stochastic process that uses Eqn. (1).
- Replica Exchange simulation by itself requires a relatively small communication between replicas, thus, each replica can run on a single processor in multi-core settings.

Parallelization of Energy Computation of Conformations

- Though this works well with small to modestly sized Monte Carlo systems, it becomes a problem with many replicas and longer simulation times.
- This necessitates high-performance computing approach and GPUs.
- Energy computations are the most expensive of the REMC process.
- Replicas are moved from state i to j and the improvement of these movements are guided by the energy functions of the system.

Parallelization of Energy Computation of Conformations

- The energy function has two main categories
- those based on replicas ability to satisfy distance and contact restraints,
- and the other based on physical and statistical energy scores.
- The physical energy scores include van der Waals and electrostatic potentials.
- Whereas the statistical energy scores are derived from structural databases.

Parallelization of Energy Computation of Conformations

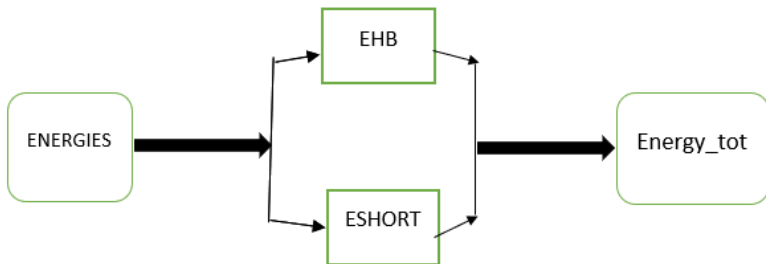
- For each replica in the system, we parallelize the operations by invoking several OpenACC gangs for the computations
- For each gang, several OpenACC threads are launched.
- A maximum of 1024 threads are launched per gang
- These launched threads partition the tasks in a parallel region among themselves
- Similarly for the energy scores, we assign each to its compute region

Parallelization of Energy Computation of Conformations

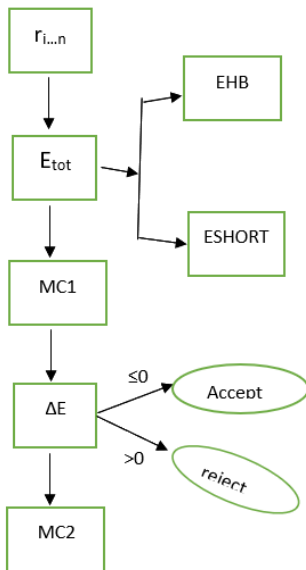
- The energy functions are summed for the potential energy of the system, thus, there is the need for communication of updated energy scores.
- These communication of updated energy scores and the data required for the computations introduces an enormous data transfer challenge with this optimization.

Parallelization of Energy Computation of Conformations

- EHB, ESHORT, Energycot calculate statistical energy component, energy based on contact and distance restraints and total energy.



Flow Chart of Parallelization of REMC



Parallelization of Monte Carlo Moves/Updates

- Several Monte Carlo moves/updates are attempted in a Monte Carlo process.
- Updates at high temperatures that change the energy of the system have a higher probability of being accepted based on Eqn (1).
- A move from state i to state j is denoted by the transition matrix:

$$M_{ij} = \beta_{ij} p_{ij} \quad (3)$$

- β_{ij} is the probability of attempting a move between the two states and p_{ij} is the probability of accepting the move, which is Eqn 1.

Parallelization of Monte Carlo Moves/Updates

- The moves change the torsional angle and any bond angles of the atom.
- We consider 2,3,4,5 and 6 bond moves for the system.
- Each move is assigned a compute region, meaning several thread blocks are used.
- Some of these threads are responsible for the calculation of the change in energy, used to determine whether a move/update should be accepted or rejected.
- There is therefore a communication of energy scores from the energy regions to the moves/updates.
- After several of Monte Carlo moves/updates, replica swaps are initiated which are also accepted based on energy changes.

Specifications of the Hardware used

-	reference CPU	GPU1	GPU2
name	Intel Xeon E5-2680v3	Pascal P100	Kepler K80
Node counts	1944	36	36
Cores per socket	12	14	12
RAM	128GB	128GB	128GB
clock speed	2.5GHz	2.4GHz	2.5GHz

- Speed-up is given by:

$$S_p = \frac{t_{CPU}}{t_{GPU}},$$

(4)

- The runtime on the GPU and CPU are t_{GPU} and t_{CPU} respectively

Case Study of REMC Method in I-TASSER

- We consider a performance comparison of just the REMC of top performing protein structure prediction method from our comprehensive review.
- I-TASSER is top performing protein structure predictor based on our review, hence we use it for our case study.
- Used sequence of length 146, obtained Average speedup of 3.6x.

Table: Performance Comparison of Energy Computations

Energy Computations	t_{CPU}/s	t_{GPU}/s	S_p
EHB	4458.28	1177.57	3.8
ESHORT	2427.56	709.71	3.4
All Energies	6885.84	1887.28	3.6

Case Study of REMC Method in I-TASSER

- We compare the moves in serial and GPU REMC

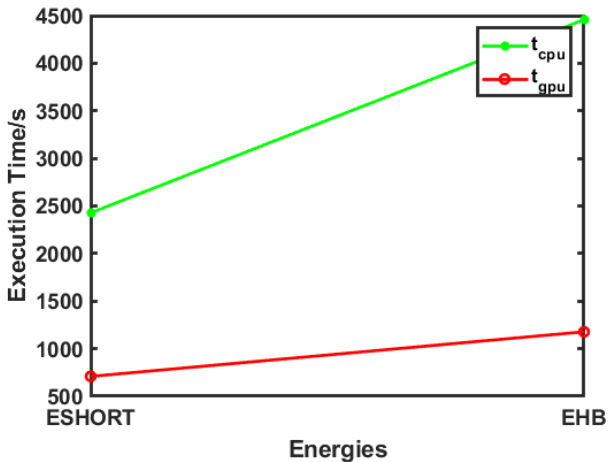
Table: Comparing Serial and GPU Moves

MC Moves	t_{CPU}/s	t_{GPU}/s	S_p
Move2	262.86	35.09	7.5
Move3d	107.56	11.83	9.1
Move3s	68.86	8.40	8.2
Move4d	60.10	7.09	8.5
Move4s	42.16	4.99	8.4
Move5d	52.97	6.98	7.6
Move5s	37.82	4.75	8.0
Move6	16.77	3.35	5.0
All Moves	649.1	79.75	8.1

- We observe a peak speedup of 9.1x with average 8.1x

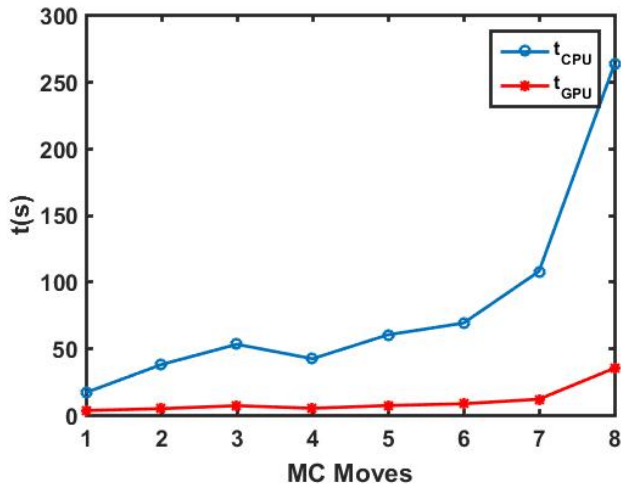
Serial REMC Vs. GPU REMC for Energies

- We compare the time from serial and GPU versions for the energy computations



Serial REMC Vs. GPU REMC for MC Moves





- We see execution time for the serial rising steadily while that for the GPU is contained.







Conclusion

- We have successfully parallelized REMC method on GPUs.
- We observed a peak speedup of 9.1x from the Monte Carlo moves.
- An average speedup of 8.1 across the moves.
- Peak speedup of 3.8 over the energy computations which is the most expensive.

References

-  Kyoung-Su and Jung, Keechul. "GPU implementation of neural networks." *Pattern Recognition* 37, no. 6(2004):1311-1314
-  Stantchev, George and Dorland, William and Gumerov, Nail. "Fast parallel particle-to-grid interpolation for plasma PIC simulations on the GPU." *Journal of Parallel and Distributed Computing* 68, no. 10(2008):1339–1349
-  Charalambous, Maria and Trancoso, Pedro and Stamatakis, Alexandros. "Initial experiences porting a bioinformatics application to a graphics processor." *Advances in Informatics* 2005:415–425
-  Xu, D., et al., "Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement." *Proteins: Structure, Function, and Bioinformatics*, 2011: 147-160.

References

-  Preis, Tobias and Virnau, Peter and Paul, Wolfgang and Schneider, Johannes J. "GPU accelerated Monte Carlo simulation of the 2D and 3D Ising model" *Journal of Computational Physics* 228, no. 12(2009):4468-4477.
-  Gross, Jonathan and Janke, Wolfhard and Bachmann, Michael."Massively parallelized replica-exchange simulations of polymers on GPUs." *Computer Physics Communications* 182, no. 8(2011): 1638-1644.
-  MacCarthy, Elijah and Perry, Derrick and KC, Dukka."Advances in Protein Super-Secondary Structure Prediction and Application to Protein Structure Prediction." *Methods in Molecular Biology* 2019:15-45
-  Chen, K. and L. Kurgan, "Computational prediction of secondary and supersecondary structures." *Protein Supersecondary Structures*. (2012):63-86

Thank You!
Questions?