Large-scale Molecular Dynamics Simulation with Forward Flux Sampling on Hadoop

Pengfei Xuan
Yueli Zheng
Sapna Sarupria
Amy Apon

Follow this and additional works at: http://tigerprints.clemson.edu/grads_symposium

Recommended Citation
Xuan, Pengfei; Zheng, Yueli; Sarupria, Sapna; and Apon, Amy, "Large-scale Molecular Dynamics Simulation with Forward Flux Sampling on Hadoop" (2013). Graduate Research and Discovery Symposium (GRADS). Paper 59.
http://tigerprints.clemson.edu/grads_symposium/59

This Poster is brought to you for free and open access by the Research and Innovation Month at TigerPrints. It has been accepted for inclusion in Graduate Research and Discovery Symposium (GRADS) by an authorized administrator of TigerPrints. For more information, please contact awesole@clemson.edu.
ABSTRACT
Simulating rare events is extremely difficult and requires massive computational resources and complex data processing workflow, which is determined by the nature of stochastic systems. To help computational scientists discover hard scientific problems in this area, we built a large-scale molecular dynamics simulation framework integrated with forward flux sampling (FFS) technique on Hadoop ecosystem. In this project, we port the customized FFS workflow to underlying MapReduce-based computing pipeline by using dataflow-driven design pattern and Gromacs application. The early works show that our framework is able to provide a scalable, fault-tolerance and efficient rare events simulation environment over varieties of computing infrastructures, while preserving the flexibility of the original scientific application.

BACKGROUND
Rare events are fluctuation-driven processes occurring infrequently such as the nucleation of crystals or protein folding events. It exists widely as kinds of different phenomena, such as earthquakes, financial crashes, telecommunication network failure, protein conformational change, activated chemical reactions and so on. It is notoriously difficult to simulate because the very few events are observed in the traditional simulation. The traditional simulation methods such as “brute-force” simulation is highly inefficient in the rare event simulation, because the waiting time between events is hundreds and thousands of times longer than the time scale of the event itself in these processes [1]. In this project, we use FFS (2) sampling method to improve the efficiency of rare events simulation.

STUDY CASE
We are trying to use FFS to simulate the nucleation and growth of gas hydrates to obtain detailed molecular-level picture of hydrate formation kinetics [3]. In the simulation, to perform FFS, Bash Script concurrently call GROMACS programs to generate a large volume of intermediate files (more than 100 thousands) with TB-level data size, and then calculate threshold to filter parts of results. The massive MD simulation gives a big overhead on the this approach.

RESULT
The data sets in Table 1 are obtained from the FFS implementation by using 64 computing nodes, which lists the cluster size at different interfaces and probabilities of the successful trajectories over all trajectories from current interfaces to next interfaces. With the cluster sizes at interfaces increasing from 70 to 380, the changing of the probability at every of forwarding interfaces can be observed. The original Bash Script-based implementation took more then 3 days to simulate only one branch for each move forward. Our new FFS-based implementation decreases simulation time for each move forward with using all branches from days-level to hours-level.

Table 1. FFS simulation result: list of cluster size at different interfaces and probabilities of the successful trajectories over all trajectories from previous interfaces to each different interface with simulation time.

<table>
<thead>
<tr>
<th>Interface</th>
<th>Cluster Size</th>
<th>Probability</th>
<th>Running Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>μ₀</td>
<td>μ₁</td>
<td>μ₂</td>
<td>μ₃</td>
</tr>
<tr>
<td>μ₀</td>
<td>μ₁</td>
<td>μ₂</td>
<td>μ₃</td>
</tr>
<tr>
<td>70</td>
<td>201</td>
<td>231</td>
<td>259</td>
</tr>
<tr>
<td>0.00106338</td>
<td>0.00655409</td>
<td>0.00583333</td>
<td>0.01638044</td>
</tr>
<tr>
<td>31 min</td>
<td>57 min</td>
<td>72 min</td>
<td>79 min</td>
</tr>
</tbody>
</table>

REFERENCES

ACKNOWLEDGEMENT
- Supported by NSF awards CNS-1228312 and CI-1212680CCIT
- OrangeFS Testbed and CITI Palmetto HPC resources