



**NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS
INTERDISCIPLINARY GRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"**

MSc THESIS

**StochMPI: An MPI-based parallel scalable stochastic
simulator for biomolecular networks**

Konstantinos P. Spyrou

Supervisor: Elias Manolakos, Professor

ATHENS

OCTOBER 2022

MSc THESIS

StochMPI: An MPI-based parallel scalable stochastic simulator for biomolecular networks

Konstantinos P. Spyrou

S.N.: piv176

SUPERVISOR: **Elias Manolakos**, Professor

**EXAMINING
COMMITTEE:** **Ioannis Emiris**, Professor
Martin Reczko, Research Professor

OCTOBER 2022



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**StochMPI: Παράλληλος στοχαστικός προσομοιωτής υψηλών
επιδόσεων για βιολογικά δίκτυα βασισμένος στο MPI**

Κωνσταντίνος Π. Σπύρου

Επιβλέπων: Ηλίας Μανωλάκος, Καθηγητής

ΑΘΗΝΑ

ΟΚΤΩΒΡΙΟΣ 2022

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

StochMPI: Παράλληλος στοχαστικός προσομοιωτής υψηλών επιδόσεων για βιολογικά δίκτυα βασισμένος στο MPI

Κωνσταντίνος Π. Σπύρου

A.M.: ΠΙΒ176

ΕΠΙΒΛΕΠΩΝ: **Ηλίας Μανωλάκος**, Καθηγητής

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Ιωάννης Εμίρης**, Καθηγητής
Martin Reczko, Ερευνητής

ΟΚΤΩΒΡΙΟΣ 2022

ABSTRACT

Stochastic simulation is a fundamental process for investigating the dynamics of noisy natural systems. In biology, we use it extensively to model and analyze kinetic reaction networks where multiple molecular species interact in non-linear ways through reaction channels in the presence of intrinsic or extrinsic stochasticity introducing uncertainty. We often represent such networks as bio-models in standardized formats (e.g., SBML) to aid their comprehension and facilitate their sharing among scientists and software tools. Several exact and approximate stochastic simulation algorithms have been developed in the literature. Still, as biomedical researchers aim to study *in silico* the stochastic dynamics of whole complex organisms, there is a growing need for high-performance parallel stochastic simulator software tools that can handle networks with thousands of species and reactions in a scalable manner.

In this graduate thesis, we have developed a scalable MPI-based stochastic simulator with an architecture that enables the exploitation of parallelism at two levels of granularity: Multiple repetitions of a simulation run with slightly different parameters and multiple reactions of the same simulation repetition performed in parallel. *StochMPI* implements two of Gillespie's well-known Stochastic Simulation Algorithms (SSAs), the Direct Method (DM) and the First Reaction Method (FRM), along with optimizations that we introduced to boost their performance. Its novel design makes it a high-performance stochastic simulation software tool suitable for multi-core CPUs and multi-computer cloud environments. Moreover, *StochMPI* favors extensibility, as it provides a framework aiding in the effortless integration of new stochastic simulation algorithms.

We have performed a multitude of experiments with both real-world and synthetic benchmark bio-models, scaling to thousands of species and reactions, to assess the performance scalability of *StochMPI*. In addition, we have validated the correctness of all incorporated simulation methods using the Discrete Stochastic Model Tests Suite (DSMTS) and well-established bio-models from the literature.

Our results suggest that the DM algorithm using repetition-level parallelization is the best performance option for small-scale bio-models with up to one hundred reactions. Furthermore, the optimized DMO algorithm is the best candidate for the stochastic simulation of larger bio-models. The DMO can reach a 2.2 GReactions/sec (billions of reactions per second) performance when deploying 8 processes on the 8-core Intel i7-5960X CPU, and 1 GR/sec when deploying 4 processes on the 4-core Intel 6700K CPU, for biomodels with 8K reactions. Moreover, we have confirmed the scalability of *StochMPI* using the HYPATIA EliXir Greece cloud infrastructure, especially on repetition-based simulations. Our simulation results using 90 processes on a Virtual Machine with 48 physical cores, running the FRMOv2 algorithm achieves performance of 1.3 GR/sec for all models ranging from 1K to 32K reactions. We hope our work will aid other scientists in their investigations and motivate them to further extend the capabilities of *StochMPI*, which will be released as an open-source project to the systems biology community.

SUBJECT AREA: Stochastic Simulation

KEYWORDS: parallelization, scalability, high-performance computing, Stochastic Simulation Algorithms, reaction networks, Systems Biology, MPI

ΠΕΡΙΛΗΨΗ

Η στοχαστική προσομοίωση αποτελεί θεμελιώδη διαδικασία για τη βαθύτερη κατανόηση της δυναμικής ενός συστήματος που μιμείται μια φυσική διεργασία σε ρεαλιστικές συνθήκες παρουσία θορύβου. Στη βιολογία τη χρησιμοποιούμε για να αναλύσουμε δίκτυα κινητικών αντιδράσεων με πολλαπλά μοριακά είδη που αλληλοεπιδρούν με μη γραμμικό τρόπο. Στη φύση αυτή η αλληλεπίδραση διακρίνεται από αβεβαιότητα και μπορεί να οδηγήσει σε διαφορετικά αποτελέσματα. Τα δίκτυα αντιδράσεων αναπαρίστανται ως βιολογικά μοντέλα (βιο-μοντέλα) και ακολουθούν συγκεκριμένες προδιαγραφές αναλόγως τον τύπο του αρχείου που θα αποθηκευτούν (π.χ. SBML). Αυτό αποσκοπεί στο να είναι κατανοητά μεταξύ επιστημόνων και συμβατά με προγράμματα ανάλυσης. Αρκετοί αλγόριθμοι στοχαστικής προσομοίωσης έχουν δημιουργηθεί που χωρίζονται σε δύο βασικές κατηγορίες, τους ακριβείς (exact) και τους προσεγγιστικούς (approximate). Είναι όμως ανάγκη να δημιουργηθούν γρηγορότερα λογισμικά με δυνατότητα αποδοτικής κλιμάκωσης, ώστε να επιτρέψουν την ανάλυση της στοχαστικής συμπεριφοράς δικτύων με χιλιάδες αντιδράσεις προκειμένου να γίνει εφικτή η ρεαλιστική προσομοίωση μέχρι και ολόκληρων οργανισμών.

Σε αυτή τη διπλωματική εργασία αναπτύξαμε τον παράλληλο στοχαστικό προσομοιωτή, *StochMPI* βασισμένο στο message passing interface (MPI) ο οποίος μπορεί να εκμεταλλευτεί αποδοτικά τους πόρους οποιουδήποτε μηχανήματος τον χρησιμοποιεί, ακόμα και μεγάλων υπολογιστικών υποδομών πλέγματος. Με την καινοτόμο αρχιτεκτονική του, παραλληλοποιεί τις προσομοιώσεις σε δύο επίπεδα: Παράλληλες στοχαστικές επαναλήψεις προσομοίωσης ενός μοντέλου, αλλά και παράλληλη εκτέλεση των αντιδράσεων της ίδιας επανάληψης του μοντέλου. Στην υλοποίηση έχουμε συμπεριλάβει τους δύο στοχαστικούς αλγόριθμους του Gillespie, τον Direct Method (DM) και τον First Reaction Method (FRM), καθώς και βελτιστοποιημένες εκδόσεις τους που δημιουργήσαμε για να αυξήσουμε την απόδοσή τους σε πολύ μεγάλα βιο-μοντέλα. Ένα ακόμη προτέρημα του *StochMPI* είναι πως ευνοεί την προσθήκη και νέων στοχαστικών αλγορίθμων χάριν στον τρόπο με τον οποίο έχει δομηθεί το λογισμικό.

Πραγματοποιήσαμε πληθώρα πειραμάτων με πραγματικά και συνθετικά βιο-μοντέλα ώστε να υπολογίσουμε την απόδοση και να εκτιμήσουμε την κλιμάκωσή της καθώς μεγαλώνει η πολυπλοκότητά τους σε μοντέλα με χιλιάδες αντιδράσεις. Τέλος, επιβεβαιώσαμε την ορθή λειτουργία όλων των αλγορίθμων του προσομοιωτή με τη βοήθεια της γνωστής σουίτας στοχαστικών μοντέλων Discrete Stochastic Model Tests Suite (DSMTS) και άλλων βιο-μοντέλων από τη βιβλιογραφία.

Τα αποτελέσματά μας φανερώνουν πως η αρχιτεκτονική παράλληλων επαναλήψεων είναι η καλύτερη επιλογή για το μέγεθος των μοντέλων που δοκιμάσαμε. Πιο συγκεκριμένα, ο DM αλγόριθμος είναι η βέλτιστη επιλογή για μικρά βιο-μοντέλα που περιέχουν μέχρι 100 αντιδράσεις, ενώ ο βελτιστοποιημένος αλγόριθμος DMO είναι η κατάλληλη επιλογή για μεγαλύτερα μοντέλα. Οι μέγιστες αποδόσεις που καταφέραμε να επιτύχουμε για βιο-μοντέλα με 8K αντιδράσεις ήταν 2.2 GReactions/sec χρησιμοποιώντας 8 διεργασίες στον 8-πυρρηνο επεξεργαστή Intel i7-5960X, και 1 GR/sec με 4 διεργασίες στον 4-πυρρηνο Intel i7-6700K. Επιπλέον, επιβεβαιώσαμε την κλιμάκωση του προσομοιωτή χρησιμοποιώντας την υπολογιστική υποδομή νέφους HYPATIA του ελληνικού κόμβου του EliXir. Καταφέραμε να φτάσουμε απόδοση 1.3 GR/sec στα μοντέλα από 1K έως 32K αντιδράσεις χρησιμοποιώντας τον αλγόριθμο FRMOv2 με 90 διεργασίες σε ένα 48-πυρρηνο μηχάνημα. Πιστεύουμε ότι το *StochMPI* θα αποδειχθεί χρήσιμο εργαλείο για την επιστημονική κοινότητα και σκοπεύουμε να το παρέχουμε ως λογισμικό ανοιχτού κώδικα με την ελπίδα ότι θα εμπνεύσει κι άλλους ερευνητές για τη περεταίρω διεύρυνση των δυνατοτήτων του.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Στοχαστική Προσομοίωση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: παραλληλοποίηση, κλιμάκωση, Υπολογιστικά Συστήματα Υψηλών Επιδόσεων, Αλγόριθμοι Στοχαστικής Προσομοίωσης, δίκτυα αντιδράσεων, Συστημική Βιολογία, MPI