ИСПОЛЬЗОВАНИЕ ИМИТАЦИОННОГО МОДЕЛИРОВАНИЯ ДЛЯ НАСТРОЙКИ ПАРАМЕТРОВ МАСШТАБИРУЕМЫХ АЛГОРИТМОВ ПРИ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛЕНИЯХ

Б.М. Глинский, А.С. Родионов, М.А. Марченко, Д.А. Караваев, Д.И. Подкорытов, Д.В. Винс

Введение

Исследование масштабируемости параллельных алгоритмов является важной задачей при оценке эффективности их реализации на будущих экзафлопсных суперкомпьютерах. Данная проблема выходит из круга задач программирования и требует научно-исследовательского подхода к ее решению, поскольку вычислительные алгоритмы, как правило, являются более консервативными по сравнению с развитием средств вычислительной техники. Имитационное моделирование является перспективным подходом к решению такого рода задач.

Реальные экзафлопсные компьютеры по прогнозам экспертов появятся в 2018-2020 гг., однако оценить поведение алгоритмов, путем реализации их на имитационной модели, содержащей тысячи и миллионы вычислительных ядер, можно уже сейчас. Имитационная модель позволит выявить узкие места в алгоритмах, понять, как нужно модифицировать алгоритм, какие параметры необходимо настраивать при его масштабировании на большое количество ядер. В работе [1] показана возможность применения агентно-ориентированной системы имитационного моделирования для решения некоторых проблем, возникающих при создании суперЭВМ экзафлопсной производительности.

В данной работе рассматриваются особенности масштабирования двух типов алгоритмов: распределенного статистического моделирования и численного моделирования 3D сейсмических полей.

Моделирование исполнения параллельных алгоритмов для анализа масштабируемости проводилось на гибридном кластере ССКЦ, который состоит из 40 вычислительных узлов HP SL390s G7. Каждый узел содержит: два 6-ядерных CPU Xeon X5670 (2.93GHz); 96 ГБ оперативной памяти; три карты NVIDIA Tesla M2090. Каждая карта содержит GPU с 512 ядрами и 6 ГБ оперативной памяти. Суммарно гибридный кластер содержит 80 процессоров (480 ядер) CPU и 120 процессоров (61440 ядер) GPU. Пиковая производительность – 85 TFlops, на тесте Linpak – 38 TFlops. Моделирование проводилось с использованием распределённой агентно-ориентированной системы имитационного моделирования AGNES, разработанной в ИВМиМГ СО РАН [2].

Агентно-ориентированная система имитационного моделирования AGNES

Пакет AGNES базируется на Java Agent Development Framework (JADE) [3]. JADE – это мощный инструмент для создания мульти-агентных систем на JAVA, и он состоит из 3-х частей: среда исполнения агентов; библиотека базовых классов, необходимых для разработки агентной системы; набор утилит, позволяющих наблюдать и администрировать МАС (мульти-агентная система). Для моделирования больших вычислений важно, что JADE – это FIPA-совместимая, распределенная агентная платформа, которая может использовать один или несколько компьютеров (узлов сети), на каждом из которых должна работать только одна виртуальная JAVA машина.

AGNES использует преимущества, предоставляемые JADE, и расширяет мульти-агентную систему до системы моделирования. AGNES состоит из двух типов агентов:

- Управляющие агенты (УА), которые создают среду моделирования.
- Функциональные агенты (ФА), которые образуют модель, работающую в среде моделирования.

Приложение AGNES – это распределенная МАС, называемая платформой. Платформа AGNES состоит из системы контейнеров, распределенных в сети. Обычно на каждом хосте находится по одному контейнеру (но при необходимости их может быть несколько). Агенты существуют внутри контейнеров.

Мульти-агентный подход органично подходит для задачи имитации вычислений. В качестве атомарной, независимой частицы в модели вычислений выбран вычислительный узел и исполняемый на нем код алгоритма. Каждый функциональный агент эмулирует поведение вычислительного узла кластера, и программу вычислений, работающую на этом узле. Вычисления представляются в виде набора примитивных операций (вычисление на ядре; запись/чтение данных в память; парный обмен данными; синхронизация данных между вычислителями) и временных характеристик каждой операции.

Далее рассмотрим применение системы AGNES для исследования масштабируемости распределенного статистического моделирования и численного моделирования сейсмических полей в 3D неоднородных упругих средах.

Исследование масштабируемости распределённого статистического моделирования

С целью изучения возможности масштабирования распределенного статистического моделирования на большое число вычислительных ядер производилось имитационное моделирование работы экзафлопсного

суперкомпьютера, загруженного такого рода задачами. Мы имеем в виду задачи статистического моделирования, требующие всей вычислительной мощи многопроцессорного суперкомпьютера, т.е. требующие моделирования экстремально большого количества независимых реализаций [4]. К числу таких проблем относятся задачи моделирования течений разреженного газа с учетом химических реакций, задачи переноса излучения и теории дисперсных систем.

Имитационное моделирование проводилось с использованием мульти-агентной системы AGNES [2]. Для имитации вычислений методов Монте-Карло созданы два класса функциональных агентов:

- DataAgregator: ядро-«сборщик», собирает информацию об вычислениях, обрабатывает и агрегирует её. Возможно иерархическое построение «сборщиков», которые на нижнем уровне обрабатывают данные непосредсвенно вычислителей, а затем передают их вышестоящему агенту DataAgregator. На вершине этой пирамиды всегда стоит одно главное ядро-«сборщик», подготавливающее итоговые данные обо всех вычислениях и сохраняющее их на жесткий диск.
- MonteCarlo: агент, имитирующий расчет методов Монте-Карло, ядро-«вычислитель». Каждый агент проводит независимые вычисления согласно схеме вычислений и взаимодействует только с соответствующим DataAgregator. Основными характеристиками агента являются временные и статистические свойства, оценки которых получены на основе реальных вычислений.

В результате работы модели собираются следующие отчеты:

- Набор времен, потраченных на каждую итерацию вычислений каждым агентом. Эти времена позволяют получить статистические характеристики протекающих в модели вычислений, для оценки правдоподобия модели.
- Информация о количестве итераций вычислений, совершенных каждым агентом MonteCarlo. При помощи данной статистики можно, например, отследить, как влияет количество вычислителей на скорость расчетов.
- Информация об интенсивности получения данных агентами DataAgregator от вычислителей, либо нижестоящих DataAgregator, в данном случае регистрируется количество полученных за равные промежутки времени пакетов.

Исходные данные для имитационного моделирования получены с использованием библиотеки PARMONC, предназначенной для использования на современных суперкомпьютерах тера- и петафлопсного уровня [5]. Область применения библиотеки: «большие» задачи статистического моделирования в естественных и гуманитарных науках (физика, химия, биология, медицина, экономика и финансы, социология и др.) Библиотека PARMONC установлена на кластерах Сибирского суперкомпьютерного центра (ССКЦ КП СО РАН) и может использоваться на вычислительных системах с аналогичной архитектурой. При этом использование библиотеки не привязано к каким-то определенным компиляторам языков C и FORTRAN или MPI. Инструкции по использованию библиотеки с примерами можно найти по ссылкам [6, 7].

Как известно, теоретическое ускорение при распараллеливании для методов статистического моделирования практически "идеальное", что подтверждается численными расчетами при числе вычислительных ядер порядка нескольких тысяч [8]. Тем не менее, при числе ядер порядка сотен тысяч или нескольких миллионов вопросы организации счета требуют серьезного исследования, поскольку при этом возникают проблемы с большой загрузкой ядер-"сборщиков", которые периодически собирают статистику с ядер-"вычислителей". А именно, проведенное имитационное моделирование показало, что при большом числе используемых вычислительных ядер (больше 10000) реальное ускорение от распараллеливания существенно отличается от теоретического, что связано с большой загрузкой выделенных ядер-«сборщиков», которые обрабатывают поступающие пакеты данных с ядер-«вычислителей». При этом до 1000 ядер ускорение в модели совпадает с ускорением в реальных расчетах. С целью повышения эффективности распараллеливания исследовались различные варианты организации обмена данными между ядрами.

Целесообразно осуществлять периодическую пересылку результатов промежуточного осреднения реализаций, независимо полученных на загруженных ядрах (ядрах-"вычислителях"), на выделенные ядра (ядра-"сборщики"), объединенные в многоуровневую структуру. Ядра-"сборщики" будут периодически получать переданные им данные и осреднять их, передавая затем результаты на ядро (с номером 0), соответствующее вершине многоуровневой структуры. Будем называть такое ядро главным ядром-"сборщиком"; в числе его задач - сохранение осредненных данных на диск. Рассчитанные на главном ядре-"сборщике" осредненные значения будут соответствовать выборке, полученной совокупно на всех ядрах-"вычислителях". Распределенное статистическое моделирование на разных вычислительных ядрах-"вычислителях" производится в асинхронном режиме. Отправка и получение результатов статистического моделирования также осуществляется в асинхронном режиме [1, 8].

Далее приводятся некоторые результаты по оценке масштабируемости, полученные путем решения конкретной задачи динамики разреженного газа по методу прямого статистического моделирования, связанному с моделированием реализаций ансамбля тестовых частиц. На кластере НКС-30Т Сибирского суперкомпьютерного центра с использованием библиотеки PARMONC, был произведен ряд расчетов для общего числа ядер от 48 до 968. Реальные затраты машинного времени на независимое моделирование

реализаций на ядрах-«вычислителях» и обмен данными (выборочными средними) с главным ядром-«сборщиком» были использованы для калибровки имитационной модели в AGNES. По результатам расчетов был сделан вывод, что требуемый уровень относительной статистической погрешности в 0.1% достигается при объеме выборки L, равном 240 000. Среднее время моделирования одной реализации составило 12 сек. Для ядер-«вычислителей» обмен данными с главным ядром-«сборщиком» происходил после каждой смоделированной на них реализации.

При имитационном моделировании с использованием AGNES предполагалось, что архитектура экзафлопсного суперкомпьютера не отличается от архитектуры кластера НКС-30Т[6]. Рассматривались два варианта организации обмена данными с главным ядром-«сборщиком»: одноуровневый и двухуровневый. В двухуровневом варианте ядра-«вычислители» были поделены на N равных частей (N = 10, 20, 100), для каждой из которых данные с ядер-«вычислителей» сначала отправлялись на свое выделенное промежуточное ядро-«сборщик». В свою очередь, N промежуточных ядер-«сборщиков» отправляли данные на главное ядро-«сборщик». В одноуровневом варианте (будем считать, что число промежуточных ядер-«сборщиков» равно нулю: N = 0) данные с ядер-«вычислителей» непосредственно отправлялись на главное ядро-«сборщик».

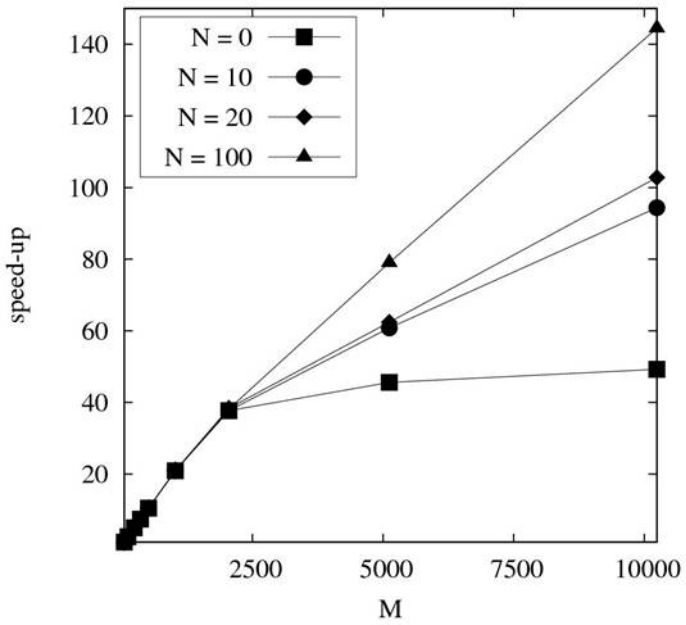

Рис. 1. Сравнение ускорения распределенного статистического моделирования для разных вариантов организации обмена данными для числа ядер M до 10 000.

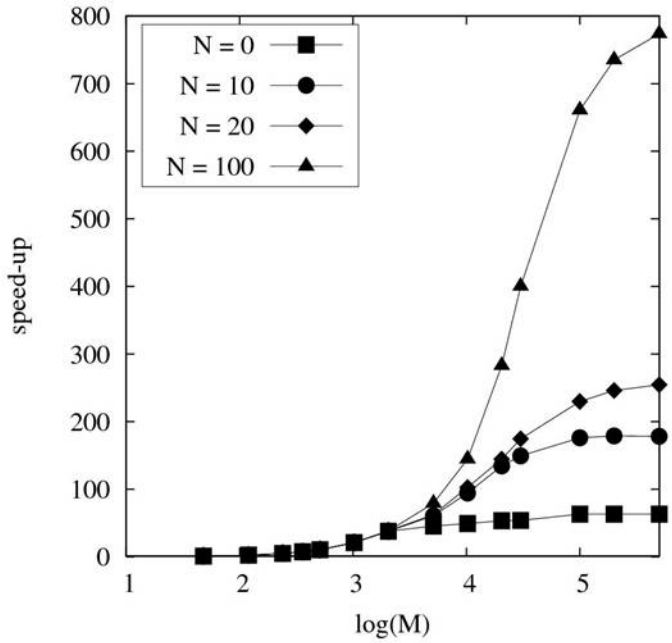

Рис. 2. Сравнение ускорения распределенного статистического моделирования для разных вариантов организации обмена данными для числа ядер M до 500 000 (горизонтальная ось – в логарифмическом масштабе).

Исследование масштабируемости алгоритма численного моделирования 3D сейсмических полей

Аналогичные исследования были проведены и для другого класса алгоритмов, основанного на применении разностного метода. В работе рассмотрен алгоритм численного моделирования 3D сейсмических полей в изотропной неоднородной упругой среде [9]. Такого вида задачи характеризуются большим объемом вычислений, поскольку область моделирования представляется достаточно подробно для проведения 3D моделирования. Предполагается, что вычислительные модули кластера состоят из нескольких CPU и GPU. Поэтому разработана программа на основе масштабируемого параллельного алгоритма при использовании комбинации технологий программирования CUDA и MPI. Для проведения расчетов различных 3D моделей была рассмотрена следующая организация параллельного алгоритма и программы: 3D область моделирования разделяется на слои, каждый слой рассчитывается независимо на выделенном GPU, а обмены данными между соседними GPU проводятся посредством MPI. При этом вычисления для слоя производятся посредством CUDA в 2D.

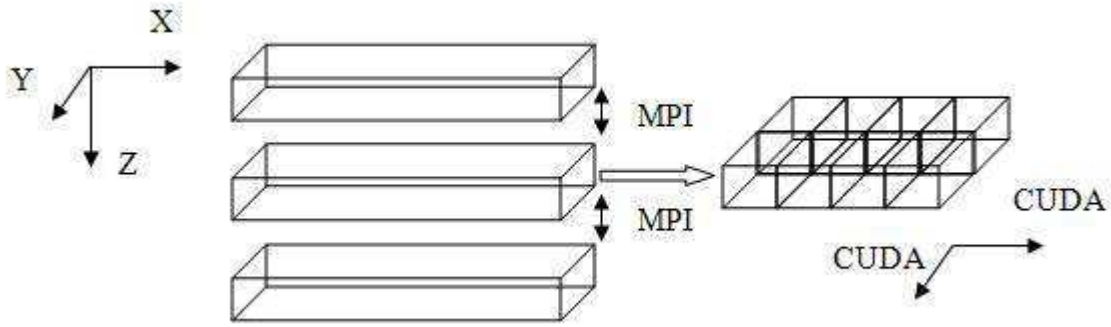

Рис.3 Схема декомпозиции расчетной области.

Способ декомпозиции расчетной области для организации параллельных вычислений представлен на рисунке 3.

Для исследования реализации алгоритма численного моделирования на предполагаемую модель экзафлопсного компьютера выбран следующий критерий масштабируемости – время счета алгоритма меняется незначительно при следующих допущениях: размер 3D модели увеличивается пропорционально количеству вычислительных узлов; каждый вычислительный узел совершает одно и то же количество итераций для своей подобласти. Исследование проводилось с использованием имитационного моделирования работы экзафлопсного суперкомпьютера в вышеупомянутой системе AGNES. Для проверки адекватности результатов имитационного моделирования проводилось их сравнение с результатами работы данного алгоритма на гибридном кластере НКС-30Т ССКЦ.

Для имитации сеточных методов реализован класс функциональных агентов Grid — узел-вычислитель, имитирующий расчет сеточных методов на одном вычислителе. Моделируются вычисления, когда область исследование режется вдоль одной оси, и полученные области загружаются на вычислители. Таким образом, получается, что у каждого вычислителя есть пересечение по данным максимум с 2-ми вычислителями («крайние» вычислители обмениваются только с одним соседом). Каждый вычислитель, на первом шаге рассчитывает свои граничные области, затем асинхронно передает насчитанные результаты соседям. Расчет внутренних областей идет на втором шаге, получив данные от соседей и просчитав изменение свой области, агент переходит к шагу один.

Общие результаты изменения времени счета в зависимости от количества доступных ядер GPU (при пропорциональном увеличении размера 3D модели) в логарифмическом масштабе приведены на рисунке 4. Показано хорошее соответствие экспериментальных и модельных результатов на начальном участке кривой (до 30720 ядер). При значительном увеличении количества вычислительных узлов с пропорциональным увеличением размера 3D модели время счета увеличивается, но несущественно (при росте числа узлов от 7680 до 1024000 время увеличилось на 17,5%).

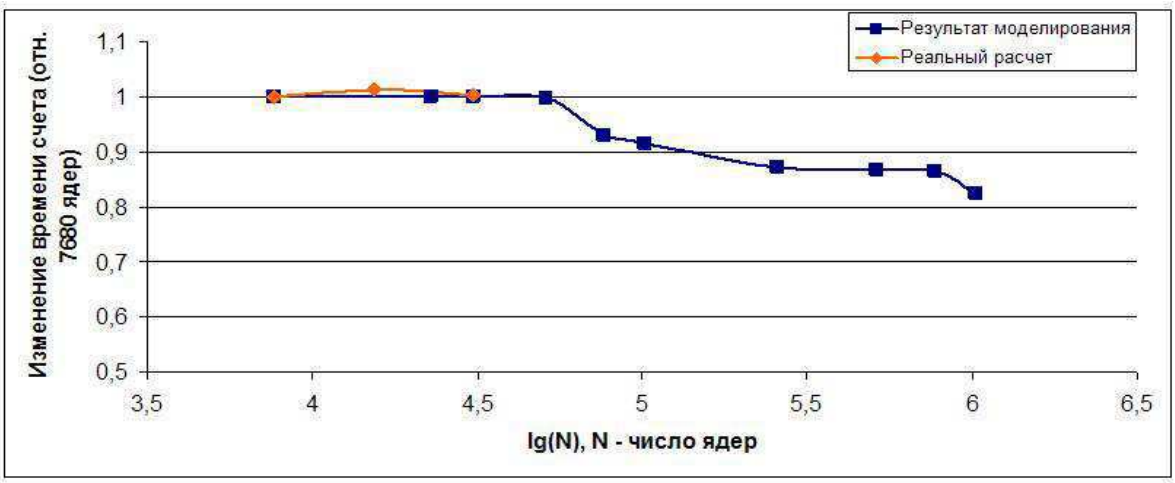

Рис. 4 Изменение времени расчета алгоритма численного моделирования в зависимости от числа вычислительных ядер (горизонтальная ось – в логарифмическом масштабе).

В заключение данного раздела приведем тестовый пример имитационного моделирования выполнения алгоритма решения этой задачи и реальный расчет на высокопроизводительном суперкомпьютере. На рис.5 приведено соотношение вычислительных ядер используемых для реальных расчетов и для имитационного моделирования. В реальном расчете понадобилось 30720 ядер, а для имитационной модели только 12 вычислительных ядер! А при моделировании работы данного алгоритма с использованием до 1,5 млн. ядер в системе AGNES нам понадобилось только 144 ядра!

Проведенное имитационное моделирование показало возможность масштабирования алгоритмов на большое число (сотни тысяч и даже миллионы) вычислительных ядер предполагаемого экзафлопсного суперкомпьютера, а также возможность исследования поведения алгоритмов при таком большом масштабировании.

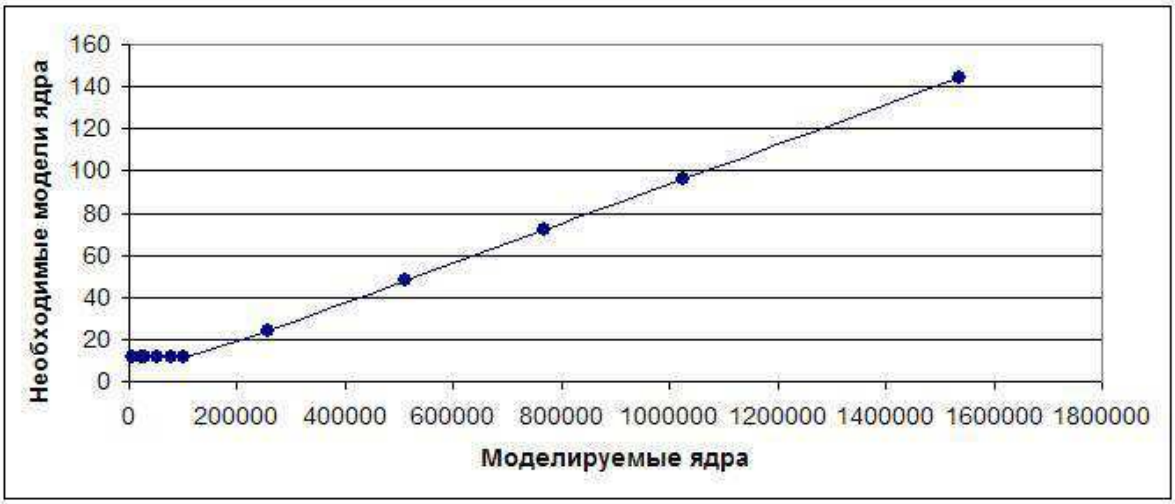

Рис. 5 Соотношение между количеством моделируемых ядер (горизонтальная ось) и количеством физических ядер (вертикальная ось), используемых для реализации имитационной модели.

Таким образом, в настоящее время для оценки масштабируемости вычислительного алгоритма на гибридном кластере можно рекомендовать следующее:

- составить граф выполнения программы;
- прогнать параллельную программу на небольшом количестве ядер от сотен до нескольких тысяч ядер;
- ввести в имитационную модель задержки, отображающие время счета на вычислительных ядрах, полученные из реальных расчетов;
- исследовать поведение алгоритма при большом количестве ядер (от сотен тысяч до миллионов вычислительных ядер);
- протестировать правильность расчета на имитационной модели путем сравнения на начальном участке реального и модельного расчетов;
- провести, при необходимости, коррекцию вычислительной схемы, реализующей данный алгоритм.

Система AGNES установлена в ССКЦ ИВМиМГ СО РАН и доступна по ссылке http://www2.sscc.ru/PPP/Mat-Libr/agnes.htm.

Заключение

В работе исследуется возможность масштабирования распределенного статистического моделирования и решения задачи распространения сейсмических волн в неоднородной изотропной среде сеточным методом на большое число (сотни тысяч и даже миллионы) вычислительных ядер предполагаемого экзафлопсного суперкомпьютера. Актуальность предмета исследования обосновывается необходимостью выяснения вычислительной эффективности алгоритмов в свете ожидаемого появления к концу десятилетия суперкомпьютеров экзафлопсного уровня производительности.

Исследование масштабируемости алгоритма на большое количество ядер проводилось с использованием агентно-ориентированной системы имитационного моделирования (AGNES). Исследование показало, что даже при явном распараллеливании алгоритма прямого статистического моделирования на большое количество ядер не происходит ожидаемого ускорения, близкого к линейному закону. Это связано с тем, что при числе ядер порядка сотен тысяч или нескольких миллионов возникают проблемы с большой загрузкой ядер-"сборщиков", которые периодически собирают статистику с ядер-"вычислителей". Следовательно, при масштабировании необходима модификация параллельной вычислительной программы, например, увеличение количества ядер-"сборщиков".

Аналогичные эксперименты проведены с численным моделированием сейсмических полей в 3D неоднородных упругих средах. В качестве метода решения используется сеточный разностный метод, а область моделирования представляется изотропной 3D неоднородной сложно построенной упругой средой. Моделирование показывает, что при решении этой задачи можно использовать 1млн. и более вычислительных ядер, следовательно, можно значительно ускорить время счета прямых задач, необходимых для интерпретации данных вибросейсмического зондирования [9].

Таким образом, проведенные исследования показывают эффективность имитационного моделирования при настройке параметров масштабируемых алгоритмов и исследовании их поведения при реализации на большом количестве вычислительных ядер.

Настоящая работа проводилась в рамках реализации федеральной целевой программы "Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы", государственный контракт 07.514.11.4016, а также при финансовой поддержке грантов РФФИ №№ 10-07-00454, 12-01-00034, 12-01-00727; МИП № 39 СО РАН, МИП № 47 СО РАН, МИП № 126 СО РАН, МИП № 130 СО РАН.

ЛИТЕРАТУРА:

1. Б.М. Глинский, А.С. Родионов, М.А. Марченко, Д.И. Подкорытов, Д.В. Винс Агентно-ориентированный подход к имитационному моделированию суперЭВМ экзафлопсной производительности в приложении к распределенному статистическому моделированию // Вестник ЮУрГУ, 2012. № 18 (277), Вып. 12., с. 93-106.
2. D. Podkorytov, A. Rodionov, H. Choo, Agent-based simulation system AGNES (AGent NEtwork Simulator) for networks modeling // Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication, ICUIMC'12, 2012.
3. F. L. Bellifemine, G. Caire, D. Greenwood, Developing Multi-Agent Systems with JADE // Wiley, 2007.
4. М.А. Марченко, Г.А. Михайлов Распределенные вычисления по методу Монте-Карло // Автоматика и телемеханика. 2007. Вып. 5. С. 157–170.
5. M.A. Marchenko PARMONC - A Software Library for Massively Parallel Stochastic Simulation // LNCS. 2011. V. 6873. P. 302-315.
6. Страница библиотеки PARMONC на сайте ССКЦ КП СО РАН [Электронный ресурс]. – Режим доступа: http://www2.sscc.ru/SORAN-INTEL/paper/2011/parmonc.htm
7. Документация к библиотеке PARMONC на сайте ССКЦ КП СО РАН [Электронный ресурс]. – Режим доступа: http://www2.sscc.ru/SORAN-INTEL/paper/2011/parmonc.pdf
8. Boris Glinsky, Alexei Rodionov, Mikhail Marchenko, Dmitry Podkorytov, Dmitry Weins. Scaling the Distributed Stochastic Simulation to Exaflop Supercomputers // Proceedings of 2012 IEEE 14th International Conference on High Performance Computing and Communications , p. 1131-1136.
9. Б.М. Глинский, Д.А. Караваев, В.В. Ковалевский, В.Н. Мартынов Численное моделирование и экспериментальные исследования грязевого вулкана «Гора Карабетова» вибросейсмическими методами. //Вычислительные методы и программирование. М.: Изд-во Моск. Гос. ун-та, 2010, Том 11, №1, С. 99-108