КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2017 Т. 9 № 6 С. 955–963
DOI: 10.20537/2076-7633-2017-9-6-955-963

Ки&М

МОДЕЛИ В ФИЗИКЕ И ТЕХНОЛОГИИ

УДК: 004.414.23, 519.876.5

Моделирование межпроцессорного взаимодействия при выполнении MPI-приложений в облаке

Н. А. Кутовский, А. В. Нечаевский, Г. А. Ососков[a], Д. И. Пряхина, В. В. Трофимов

Объединенный институт ядерных исследований,
Россия, 141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6

E-mail: [a] ososkov@jinr.ru

Получено 28.03.2017, после доработки — 10.10.2017.
Принято к публикации 11.10.2017.

В Лаборатории информационных технологий (ЛИТ) Объединенного института ядерных исследований (ОИЯИ) планируется создание облачного центра параллельных вычислений, что позволит существенно повысить эффективность выполнения численных расчетов и ускорить получение новых физически значимых результатов за счет более рационального использования вычислительных ресурсов. Для оптимизации схемы параллельных вычислений в облачной среде эту схему необходимо протестировать при различных сочетаниях параметров оборудования (количества и частоты процессоров, уровней распараллеливания, пропускной способности коммуникационной сети и ее латентности). В качестве тестовой была выбрана весьма актуальная задача параллельных вычислений длинных джозефсоновских переходов (ДДП) с использованием технологии MPI. Проблемы оценки влияния вышеуказанных факторов вычислительной среды на скорость параллельных вычислений тестовой задачи было предложено решать методом имитационного моделирования, с использованием разработанной в ЛИТ моделирующей программы SyMSim.

Работы, выполненные по имитационному моделированию расчетов ДДП в облачной среде с учетом межпроцессорных соединений, позволяют пользователям без проведения серии тестовых запусков в реальной компьютерной обстановке подобрать оптимальное количество процессоров при известном типе сети, характеризуемой пропускной способностью и латентностью. Это может существенно сэкономить вычислительное время на счетных ресурсах, высвободив его для решения реальных задач. Основные параметры модели были получены по результатам вычислительного эксперимента, проведенного на специальном облачном полигоне для MPI-задач из 10 виртуальных машин, взаимодействующих между собой через *Ethernet*-сеть с пропускной способностью 10 Гбит/с. Вычислительные эксперименты показали, что чистое время вычислений спадает обратно пропорционально числу процессоров, но существенно зависит от пропускной способности сети. Сравнение результатов, полученных эмпирическим путем, с результатами имитационного моделирования показало, что имитационная модель корректно моделирует параллельные расчеты, выполненные с использованием технологии MPI, и подтвердило нашу рекомендацию, что для быстрого счета задач такого класса надо одновременно с увеличением числа процессоров увеличивать пропускную способность сети. По результатам моделирования удалось вывести эмпирическую аналитическую формулу, выражающую зависимость времени расчета от числа процессоров при фиксированной конфигурации системы. Полученная формула может применяться и для других подобных исследований, но требует дополнительных тестов по определению значений переменных.

Ключевые слова: MPI, параллельные вычисления, облачные вычисления, имитационное моделирование

Работа выполнена при финансовой поддержке гранта РФФИ № 15-29-01217.

COMPUTER RESEARCH AND MODELING
2017 VOL. 9 NO. 6 P. 955–963
DOI: 10.20537/2076-7633-2017-9-6-955-963

Ки&М

MODELS IN PHYSICS AND TECHNOLOGY

UDC: 004.414.23, 519.876.5

Simulation of interprocessor interactions for MPI-applications in the cloud infrastructure

N. A. Kutovskiy, A. V. Nechaevskiy, G. A. Ososkov[a], D. I. Pryahina, V. V. Trofimov

Joint Institute for Nuclear Research,
Joliot-Curie st. 6, Dubna, Moscow region, 141980, Russia

E-mail: [a] ososkov@jinr.ru

Received 28.03.2017, after completion — 10.10.2017.
Accepted for publication 11.10.2017.

A new cloud center of parallel computing is to be created in the Laboratory of Information Technologies (LIT) of the Joint Institute for Nuclear Research JINR) what is expected to improve significantly the efficiency of numerical calculations and expedite the receipt of new physically meaningful results due to the more rational use of computing resources. To optimize a scheme of parallel computations at a cloud environment it is necessary to test this scheme for various combinations of equipment parameters (processor speed and numbers, throughput of a communication network etc). As a test problem, the parallel MPI algorithm for calculations of the long Josephson junctions (LDJ) is chosen. Problems of evaluating the impact of abovementioned factors of computing mean on the computing speed of the test problem are solved by simulation with the simulation program SyMSim developed in LIT.

The simulation of the LDJ calculations in the cloud environment enable users without a series of test to find the optimal number of CPUs with a certain type of network run the calculations in a real computer environment. This can save significant computational time in countable resources. The main parameters of the model were obtained from the results of the computational experiment conducted on a special cloud-based testbed. Computational experiments showed that the pure computation time decreases in inverse proportion to the number of processors, but depends significantly on network bandwidth. Comparison of results obtained empirically with the results of simulation showed that the simulation model correctly simulates the parallel calculations performed using the MPI-technology. Besides it confirms our recommendation: for fast calculations of this type it is needed to increase both, — the number of CPUs and the network throughput at the same time. The simulation results allow also to invent an empirical analytical formula expressing the dependence of calculation time by the number of processors for a fixed system configuration. The obtained formula can be applied to other similar studies, but requires additional tests to determine the values of variables.

Keywords: MPI, parallel computing, cloud computing, simulation

Citation: *Computer Research and Modeling*, 2017, vol. 9, no. 6, pp. 955–963 (Russian).

The work was supported by RFBR grant No. 15-29-01217.

1. Введение

Тенденция развития компьютерных средств обработки больших массивов данных в области высокотехнологичных приложений заключается в увеличении разнообразия типов вычислительных ресурсов. В настоящий момент в качестве таковых выступают фермы процессоров, суперкомпьютеры, которые используются с помощью облачных сред.

Одним из важных приложений теоретической физики к нанотехнологиям является численное моделирование фазовой динамики системы длинных джозефсоновских переходов (ДДП) с расчетом их вольтамперных характеристик. Такое моделирование позволило предсказать ряд важных свойств ДДП, в частности их поведение в гистерезисной области, что необходимо для разработки новых сверхпроводящих наноприборов высокой точности на основе процессов ДДП в высокотемпературных сверхпроводниках [Shukrinov, 2015]. Для проведения этих расчетов, требующих значительных вычислительных ресурсов, в 2015 году в Лаборатории теоретической физики (ЛТФ) Объединенного института ядерных исследований (ОИЯИ) разработаны параллельный алгоритм и соответствующий комплекс программ для параллельных вычислений ДДП с использованием технологии MPI [Shukrinov, 2016]. Как правило, для расчетов по технологии MPI используются процессоры и фермы параллельной архитектуры. Работа с ними связана со значительными усилиями по изменению структуры программ. Для уменьшения этих усилий представляется интересным попытаться использовать для запуска MPI-программ облачные структуры. Наличие в Лаборатории информационных технологий (ЛИТ) облачной инфраструктуры [Baranov, 2016] позволило протестировать перенос параллельных вычислений ДДП в облачную среду.

Перед переносом параллельных вычислений в облако необходимо оценить влияние параметров облака на время выполнения задачи для соответствующего алгоритма расчета ДДП. В частности, надо оценить баланс скорости вычислений и оборудования связи. Множество средств анализа и оптимизации вычислительных процессов, использующих интерфейсы MPI (см., например, [Hassani, 2014], не дает аппарата для такой оценки.

Таким образом, возникла задача моделирования параллельных MPI-вычислений в облачной среде, которая может быть решена с помощью имитационной или аналитической модели.

Авторы обладают опытом моделирования процессов в грид-структурах. С целью использования этот опыта, полученного в рамках предыдущих проектов, для моделирования вычислительных процессов, использующих интерфейс MPI, было предложено использовать программу имитационного моделирования SyMSim, разработанную в ЛИТ [Кореньков, 2015]. Для грид-структур моделируется процесс обработки потока независимых заданий. Исходными данными для программы являются: количество процессоров, их производительность, топология связей между ними, а также скорости передачи данных между вычислительными узлами. Эти значения задаются в базе денных, описывающих грид. При моделировании MPI в облаке каждый шаг вычислений на одном процессоре можно рассматривать как отдельное задание, но логически связанное с другими, так, как этого требует моделируемый алгоритм. В случае вычислений ДДП алгоритм разбит на шаги. Вычисления на шаге могут быть поделены между несколькими процессорами, на каждом шаге новое задание выполняется только после получения данных от предыдущего шага вычислений. В связи с этим потребовалась модификация базовой версии программы SyMSim, в которой задания рассматриваются как независимые. Построение модели в этом случае сводится к созданию таблиц базы данных, где набор виртуальных машин облака представляется как связанные между собой процессоры, а шаги алгоритма — как задания.

Простота схемы вычислений позволяет разработать аналитическую модель и проверить на ней результаты моделирования SyMSim. Вычислительный эксперимент, проведенный на облачном полигоне, предоставленном администрацией Многофункционального информационно-вычислительного комплекса (МИВК) ОИЯИ [Multifunctional…], показал хорошее согласование результатов аналитической, имитационной моделей и эксперимента.

2. Описание эксперимента на облачном полигоне

Облачный полигон для MPI-задач создавался на оборудовании *Dell PowerEdge FX2*, представляющем из себя корзину с 8 лезвиями, каждое из которых — это сервер *Dell PowerEdge FC430* с 48 ядрами (два процессора *Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz*) и 256 ГБ оперативной памяти, а также два *SSD*-диска по 256 ГБ). Серверы внутри этой корзины взаимодействуют между собой через *Ethernet*-сеть с пропускной способностью 10 Гбит/с.

Часть серверов из данной корзины была включена в облачную инфраструктуру ЛИТ ОИЯИ, на которых из заранее подготовленного образа для виртуальной машины (ВМ) типа *KVM* со всем необходимым программным обеспечением (ПО) и настройками, с использованием облачного инструментария, были развернуты 10 ВМ, составившие облачный виртуальный кластер (далее — облачный параллельный кластер, облачный полигон).

Операционная система внутри ВМ — *CentOS 6.8 x64*, версия *OpenMPI-2.0.1*.

Был осуществлен запуск 10 прогонов программы моделирования ДДП при фиксированных параметрах алгоритма [Башашин, 2016] и изменяющемся количестве узлов в облачном параллельном кластере от 1 до 10. Время работы программы зависит от нескольких факторов, среди которых загрузка сети в конкретный момент времени (это важно для скорости обмена данными между частями программы, работающими на разных узлах виртуального кластера) и скорость работы сетевого хранилища, на который писались результаты работы программы (на этот аспект скорость работы сети тоже влияет, а также влияет скорость работы самого сетевого хранилища, которое в нашем случае было размещено в другой подсети, и канал между тестовым облачным полигоном и сетевым хранилищем вполне мог быть занят другим трафиком). Десятикратное повторение позволило получить среднее время и его разброс для каждого количества рабочих узлов облачного параллельного кластера при расчетах ДДП с 10 переходами.

Результаты тестов доступны на сайте [Firmware complex]. Усредненные значения времени выполнения программы показаны на рис. 1.

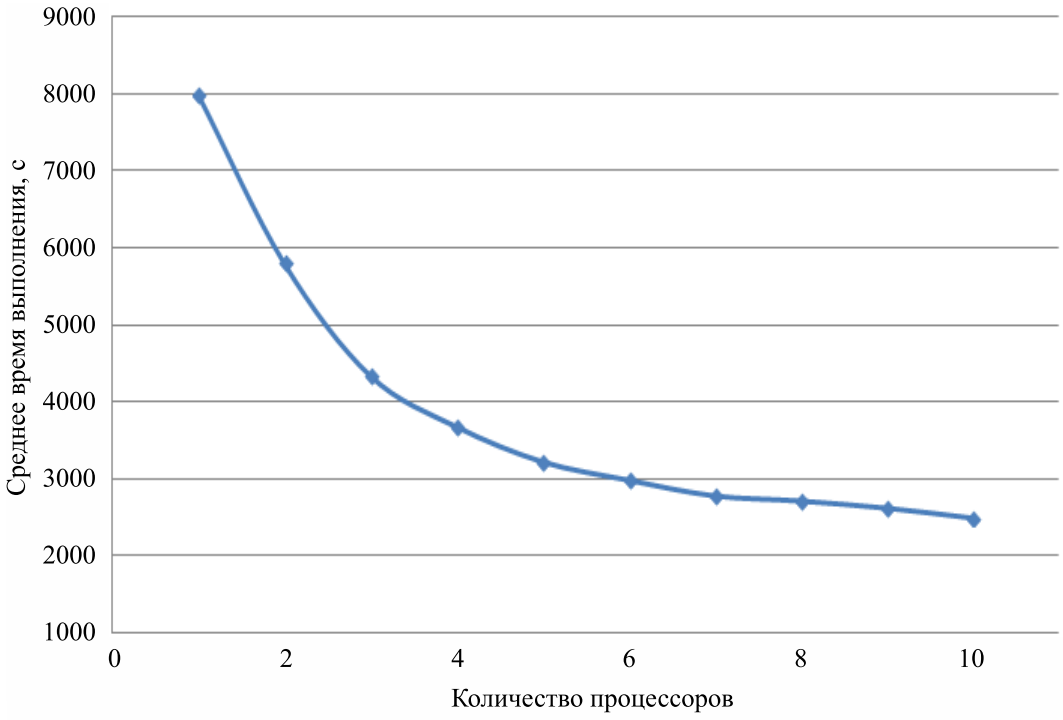

Рис. 1. Зависимость времени выполнения программы от количества однопроцессорных рабочих узлов полигона

3. Моделирование MPI

Для моделирования работы программы по расчету ДДП в гетерогенной среде, включающей многоядерные процессоры, объединенные между собой при помощи сети в единый кластер, а также в облачной инфраструктуре было применено дискретное моделирование событий, что позволяет в рамках единого подхода описать программный комплекс, использующий интерфейсы MPI, как взаимодействие процессов, запущенных на нескольких ядрах в рамках одного сервера, так и в виде виртуальных машин, взаимодействующих между собой в облачной архитектуре.

В настоящее время существует большое количество приложений, позволяющих моделировать MPI-программы. Например, MPI-PERF-SIM [Achour, 2011] или SMPI [Clauss, 2011]. Активно развиваются программы, основанные на распределенном параллельном выполнении процессов моделирования, что позволяет моделировать очень большие MPI-системы [Pelkey, 2011]. Разрабатываются системы для параллельного запуска программ моделирования в облачной среде [Cuomo, 2013]. Для выполнения моделирования MPI в облачной инфраструктуре были изучены результаты использования симулятора NetworkCloudSim [Garg, 2011].

В данной работе предметная область ограничена параллельными вычислениями ДДП. Выбор задачи ДДП для моделирования расчетов MPI обусловлен простотой верхнего уровня алгоритма. При построение модели SyMSim шаги алгоритма ДДП интерпретируются как несколько потоков заданий. Количество потоков совпадает с количеством процессоров, и каждый поток обрабатывается назначенным ему процессором. Задание (шаг алгоритма) может выполняться только после получения информации от предыдущих шагов.

Пусть параллельные процессы пронумерованы от 1 до N, число итераций — T. На первом шаге все процессы запускаются одновременно. Процесс m на текущей итерации t может быть запущен, если он получил данные от процесса $m-1$, выполненного на итерации $t-1$. Кроме того, существует процесс, который должен получить данные от всех процессов по окончании последней итерации. Время расчета одной итерации определяется случайным числом, распределенным по нормальному закону с известным средним значением. При таких упрощениях вычисления можно представить в терминах модели следующим образом. Процесс находится в состоянии ожидания до тех пор, пока не получает сигнал о готовности данных. После окончания работы процесса через случайный промежуток времени процесс посылает сигнал о наличии данных всем остальным процессам. После этого имитируется передача данных от одного процесса к другому, и алгоритм продолжается. Предварительные исследования [Kutovskiy, 2016] показали, что перенос MPI в облако имеет смысл для задачи ДДП только при достаточной скорости обмена процессоров, поэтому дальнейшие тесты и расчеты велись для скорости 10 Гбит/с. Для моделирования такой структуры потребовалась незначительная модификация базовой версии SyMSim, в которой задания рассматриваются как независимые. Для учета задержек, связанных с подготовкой буферов информации к передаче в программу, вводились дополнительные величины задержек, которые определялись по разнице времени выполнения на одном и двух процессорах. Логическая схема программы моделирования SyMSim приведена на рис. 2.

Имеет смысл сравнить результаты имитационного моделирования не только с экспериментом, но и аналитически. При таком простом построении алгоритма можно предложить аналитическую формулу времени выполнения расчетов и использовать ее для проверки работы программы вместе со значениями, полученными в результате тестовых прогонов. Сделаем следующие упрощающие предположения.

1. Сумма количества операций, выполняемых для полного расчета, постоянна и не зависит от количества процессоров.
2. Пропускная способность коммуникационной среды такова, что время обмена информацией не зависит от количества процессоров.
3. Размер буфера обмена постоянный и не зависит от количества процессоров.
4. Количество итераций постоянно и не зависит от количества процессоров.

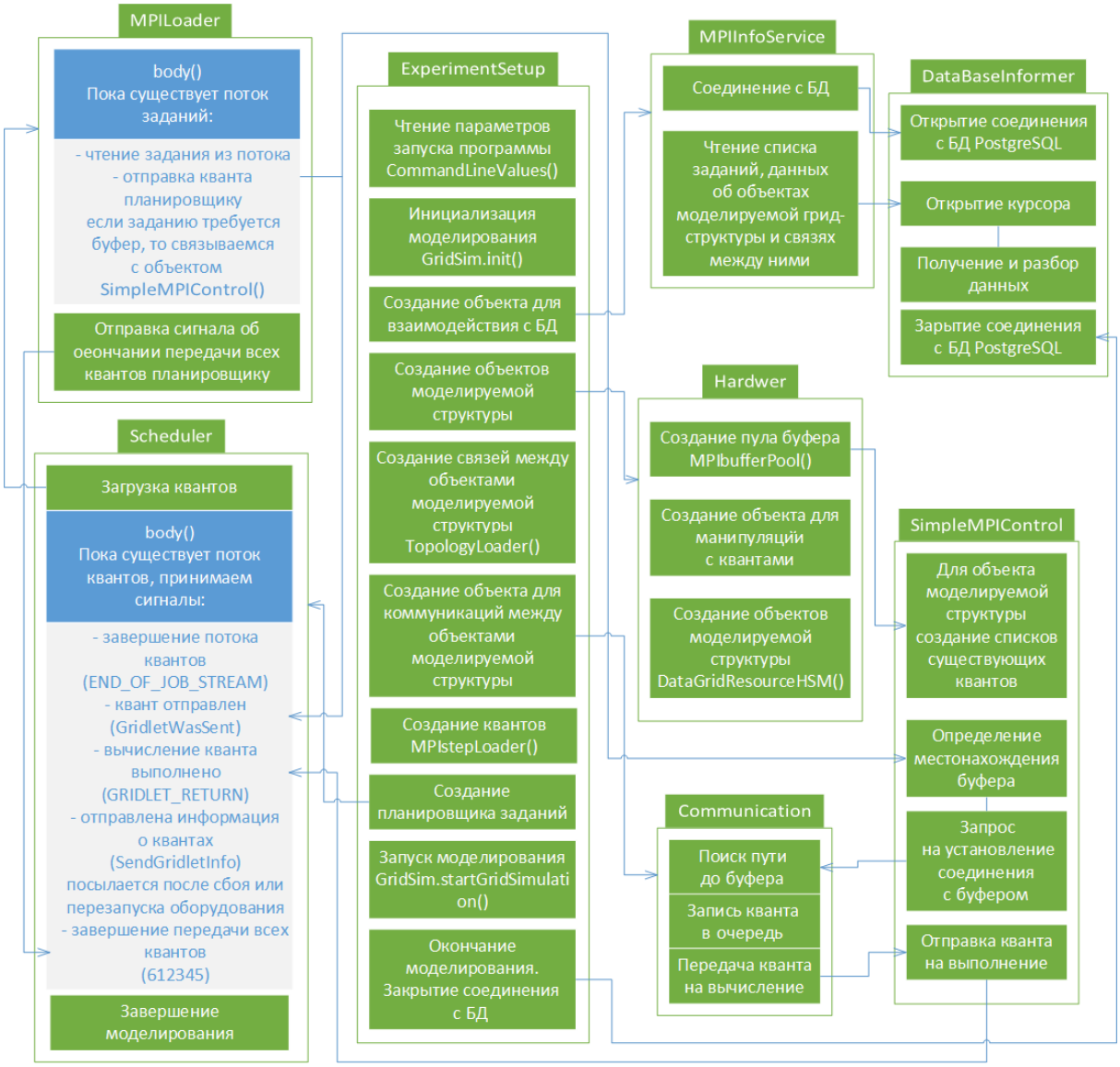

Рис. 2. Логическая схема программы SyMSim

5. Время, затраченное программой до начала итераций и после их завершения, мало, и им можно пренебречь.
6. Современные средства виртуализации при работе на современных процессорах с аппаратной поддержкой этой технологии имеют достаточно низкие накладные расходы (единицы процентов), которыми можно пренебречь.

Такой простейший случай можно описать аналитически. Время расчета прямо пропорционально временным затратам на выполнение всех итераций и обратно пропорционально числу процессоров, а с учетом затрат на буферизацию данных получаем простую формулу:

$$T = {}^{T_v \cdot I}/_{n} + I \cdot t \qquad (1)$$

для $n > 1$, где n — количество процессоров, T_v — время, которое затратит один процессор на одну итерацию без учета обмена, I — количество итераций, t — время передачи буферов между процессорами за итерацию.

Исходные данные для расчета определим на основе проведенных экспериментов [Башашин, 2016]. Вычислительная схема имеет два вложенных цикла: по времени и по току. За каждый шаг по времени выполняется 4 рекурсивные итерации Рунге–Кутты, т.е. четырежды происходит обмен между соседними вычислительными процессами. На каждом шаге по времени выполняется количество итераций, вычисленное по формуле $^{T_0}/_{T_P}$, где $T_0 = 100$ — время расчета, $T_P = 0.01$ — шаг по времени. Таким образом, на каждом шаге выполняется 10000 итераций. Количество шагов по току, для которых проводится расчет, равно $^{(I_{\max} - I_0) * 2}/_{I_0} = 218$, где $I_{\max} = 1.1$ — максимальное значение тока, $I_0 = 0.01$ — начальное значение тока. Исходя из вышеизложенного, общее число шагов по времени и току за один запуск $218 * 10000 = 2\,180\,000$, а количество итераций (I), т. е. событий, когда все процессоры обмениваются данными — $4 * 2\,180\,000 = 8\,720\,000$.

Вычислительная часть алгоритма построена так, что количество итераций, время расчета одной итерации, длина буфера не зависят от количества процессоров, но появляются дополнительные задержки, связанные с обменами между процессорами. «Накладные расходы» на передачу будут складываться из собственно времени передачи буфера и времени на его подготовку, т. е. сортировку и преобразование данных расчета. Анализ времени счета задачи показывает, что с количеством процессоров «накладные расходы» растут, но для фиксированного количества процессоров остаются постоянными. Время счета на одном процессоре равно 7937.241 с (согласно данным теста). При расчете на двух процессорах ожидается, что время счета без учета обменов будет 3968.5 с. По результатам теста время счета составило 5731.212 с, т. е. 1762.712 с ушло на обмены. Значит, передача одного буфера на итерацию занимает 202 мкс.

На рис. 3 приведены данные, полученные дискретным моделированием, аналитически и в результате тестовых прогонов задачи ДДП.

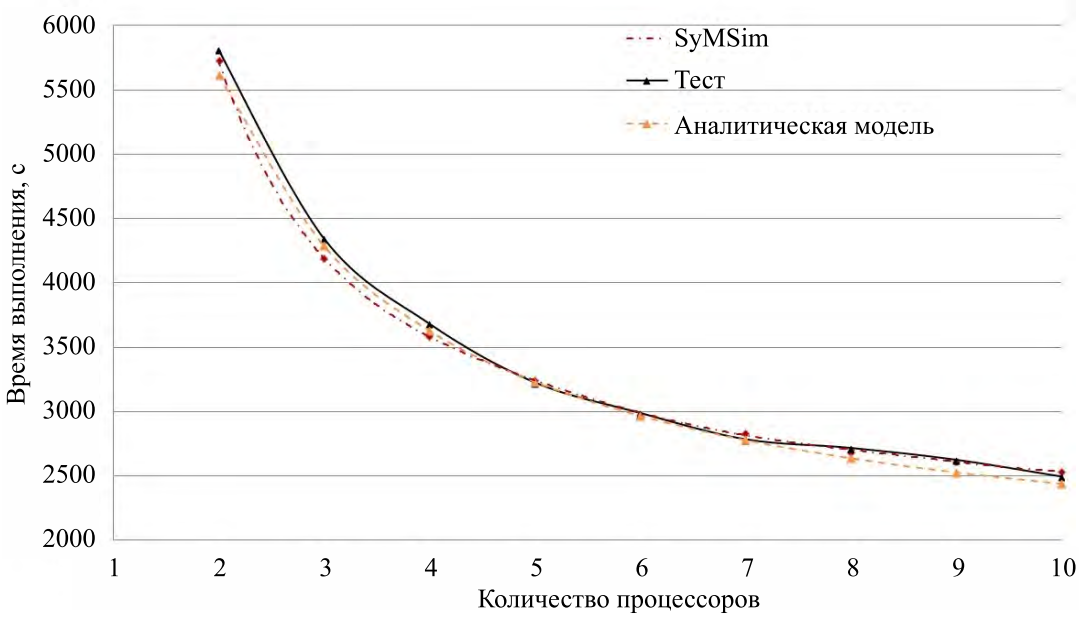

Рис. 3. Сравнение имитационной и аналитической моделей с результатами теста

Сравнение результатов, полученных эмпирическим путем, с результатами имитационного моделирования (см. рис. 3) показало, что имитационная модель корректно моделирует параллельные расчеты, выполненные с использованием технологии MPI.

4. Заключение

Основные параметры модели были получены по результатам вычислительного эксперимента, проведенного на специальном облачном полигоне для MPI-задач из 10 виртуальных машин, взаимодействующих между собой через *Ethernet*-сеть с пропускной способностью 10 Гбит/с. Вычислительные эксперименты показали, что чистое время вычислений спадает обратно пропорционально числу процессоров, но существенно зависит от пропускной способности сети. Сравнение результатов, полученных эмпирическим путем, с результатами имитационного моделирования показало, что имитационная модель корректно моделирует параллельные расчеты, выполненные с использованием технологии MPI, и подтвердило нашу рекомендацию, что для быстрого счета задач такого класса надо одновременно с увеличением числа процессоров увеличивать пропускную способность сети. В случае алгоритма вычисления ДДП удалось получить формулу, выражающую зависимость времени расчета от числа процессоров при фиксированной конфигурации системы. Приведенные результаты демонстрируют, что программу имитационного моделирования SyMSim можно успешно использовать для оценки времени выполнения MPI-алгоритмов в облачной среде с учетом межпроцессорных соединений. Это позволит без проведения серии тестовых запусков в реальной компьютерной обстановке определить целесообразность использования наличной облачной структуры, оптимальное количество процессоров при известном типе сети, характеризуемой пропускной способностью и латентностью. В случае планируемой модернизации облачного полигона этот подход дает возможность заранее предсказать, какое ускорение будет достигнуто. Совпадение результатов эксперимента, формулы и аналитического моделирования демонстрирует перспективность предлагаемого подхода.

Список литературы (References)

Башашин М. В., Земляная Е. В., Рахмонов И. Р., Шукринов Ю. М., Атанасова П. Х., Волохова А. В. Вычислительная схема и параллельная реализация для моделирования системы длинных джозефсоновских переходов // Компьютерные исследования и моделирование. — 2016. — Т. 8, № 4. — С. 593–604.

Bashashin M. V., Zemlyanaya E. V., Rahmonov I. R., Shukrinov Yu. M., Atanasova P. Kh., Volokhova A. V. Numerical approach and parallel implementation for computer simulation of stacked long Josephson Junctions // Computer research and modeling. — 2016. — Vol. 8, No. 4. — P. 593–604 (in Russian).

Кореньков В. В., Нечаевский А. В., Ососков Г. А., Пряхина Д. И., Трофимов В. В., Ужинский А. В. Синтез процессов моделирования и мониторинга для развития систем хранения и обработки больших массивов данных в физических экспериментах // Компьютерные исследования и моделирование. — 2015. — Т. 7, № 3. — С. 691–698.

Korenkov V. V., Nechaevskiy A. V., Ososkov G. A., Pryahina D. I., Trofimov V. V., Uzhinskiy A. V. Sintez protsessov modelirovaniya i monitoring dlya razvitiya system hraneniya i obrabotki bolshih massivov dannih v fizicheskih eksperimentah [Synthesis of the simulation and monitoring processes for the development of big data storage and processing facilities in physical experiments] // Computer Research and Modeling. — 2015. — Vol. 7, No. 3. — P. 691–698 (in Russian).

Achour S., Ammar M., Khmili B., Nasri W. MPI-PERF-SIM: Towards an automatic performance prediction tool of MPI programs on hierarchical clusters // Parallel, Distributed and Network-Based Processing (PDP). 19th Euromicro International Conference on, IEEE. — March 2011. — P. 207–211.

Baranov A. V., Balashov N. A., Kutovskiy N. A., Semenov R. N. JINR cloud infrastructure evolution // Physics of Particles and Nuclei Letters. — 2016. — Vol. 13, No. 5 (203). — P. 1046–1050.

Clauss P., Stillwell M., Genaud S., Suter F., Casanova H., Quinson M. Single node on-line simulation of MPI applications with SMPI // Parallel & Distributed Processing Symposium (IPDPS), 2011 IEEE International, IEEE. — 2011. — P. 664–675.

Cuomo A., Rak M., Villano U. Planting parallel program simulation on the cloud // Concurrency and computation: practice and experience. — 2013. — http://deal.ing.unisannio.it/perflab/assets/papers/ccpe2012.pdf

Firmware complex for numerical researches of Josephson nanostructures [Electronic resource]: http://jj.jinr.ru/results/

Garg S. K., Buyya R. NetworkCloudSim: Modelling Parallel Applications in Cloud Simulations // Fourth IEEE International Conference on Utility and Cloud Computing. — 2011. — P. 105–113.

Hassani R., Chavan G., Luksch P. Optimization of Communication in MPI-Based Clusters // IEEE Xplore, 2014. — http://ieeexplore.ieee.org/document/6984296/

Kutovskiy N., Nechaevskiy A., Ososkov G., Trofimov V. Simulation of Cloud Computation MPI Applications // Proc. of 7th International Conference Distributed Computing and Grid-technologies in Science and Education, Dubna, Russia, July 4–9, 2016, CEUR-WS.org, online URL: http://CEUR-WS.org/Vol-1787/343-348-paper-59.pdf

Multifunctional Information and Computing Complex [Electronic resource]: https://miccom.jinr.ru/ru/

Pelkey J., Riley G. Distributed Simulation with MPI in ns-3. — 2011. —

https://pdfs.semanticscholar.org/b555/7f4bb0b77ad7bb8f5e62ccdd941a8498fd4b.pdf

Shukrinov Yu. M., Rahmonov I. R., Nashaat M. Staircase structure of Shapiro steps // JETP Letters. — Dec. 2015. — Vol. 102, No. 12. — P. 803–806.

Shukrinov Yu. M., Rahmonov I. R., Plecenik A., Streltsova O. I., Zuev M. I., Ososkov G. A. Modeling of Intrinsic Josephson Junctions in High Temperature Superconductors under External Radiation in the Breakpoint Region // EPJ Web of Conferences. — Feb. 2016. — Vol. 108, 02042. DOI: 10.1051/epconf/201610802042