Моделирование вычислений MPI-приложений на облачной инфраструктуре

Н.А. Кутовский[1,2], А.В. Нечаевский[1], Г.А. Ососков[1,a], В.В. Трофимов[1]

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

[2] Российский экономический университет им. Г.В. Плеханова,
Россия, 117997, Москва, Стремянный пер., д. 36

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

В Лаборатории информационных технологий Объединённого института ядерных исследований проводятся работы по созданию программного комплекса для выполнения параллельных ресурсоемких счётных задач на облачных ресурсах и на вычислительном кластере. Ожидается, что созданный в результате реализации проекта облачный центр параллельных вычислений позволит существенно повысить эффективность выполнения численных расчетов и ускорить получение новых физически значимых результатов.

На данный момент на базе вычислительного центра в ОИЯИ разработан параллельный алгоритм и соответствующий комплекс программ для параллельных вычислений с использованием технологии MPI. Для оптимизации схемы параллельных вычислений необходимо протестировать работу алгоритма при различных сочетаниях параметров оборудования, количества процессоров, и уровней распараллеливания. Таким образом, возникает задача оценки влияния различных факторов (частоты процессора, пропускной способности коммуникационной сети, её латентности) на скорость вычислений конкретной задачи. Эту задачу предлагается решать методом имитационного моделирования. В работе представлены результаты моделирования вычислений с использованием технологии MPI на примере расчётов длинных Джозефсоновских переходов.

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

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

Введение

Тенденция развития сред для обработки больших массивов данных в области физики высоких энергий заключается в увеличении разнообразия типов вычислительных ресурсов. В настоящий момент в качестве таковых могут выступать фермы процессоров, облачные среды или суперкомпьютеры. Следуя этой тенденции, были опробованы облачные среды на предмет пригодности выполнения в них параллельных вычислений для разработки новых сверхпроводящих наноприборов высокой точности на основе процессов Джозефсоновских Переходов (ДП) в высокотемпературных сверхпроводниках. Такие исследования выполняются в Лаборатории теоретической физики ОИЯИ с использованием облачных вычислительных ресурсов Лаборатории информационных технологий (ЛИТ) ОИЯИ [Baranov, 2016]. Численное моделирование фазовой динамики системы длинных ДП с расчетом их вольт-амперных характеристик позволило предсказать ряд важных свойств ДП, в частности, их поведение в гистерезисной области. Для проведения этих расчетов, требующих значительных вычислительных ресурсов, в 2015 году на базе вычислительного центра HybriLIT [Alexandrov, 2015] в ОИЯИ разработан параллельный алгоритм и соответствующий комплекс программ для параллельных вычислений с использованием технологии MPI [Bashashin, 2016]. В настоящее время ведутся тестовые расчеты.

Для оптимизации схемы параллельных вычислений необходимо протестировать работу алгоритма при различных сочетаниях параметров оборудования, количества процессоров, и уровней распараллеливания. Таким образом, возникает задача оценки влияния различных факторов (частоты процессора, пропускной способности коммуникационной сети, её латентности) на скорость выполнения конкретной вычислительной задачи. Эту задачу предлагается решать методом имитационного моделирования. С целью использования опыта моделирования грид и облачных структур, выполненных в рамках предыдущих проектов, для моделирования вычислительных процессов, использующих интерфейс MPI, предложено использовать программу имитационного моделирования SyMSim, разработанную в ЛИТ [Korenkov, 2016].

Описание подхода к моделированию MPI

В данной работе предметная область ограничена вычислениями, алгоритм которых определён вышеуказанной схемой параллельных ДП расчетов.

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 расчёта одной итерации определяется случайным числом, распределённым по нормальному закону со средним значением G. 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.

С точки зрения реализации программы моделирования есть два пути. Один – создавать заранее поток шагов процессов с указанием, какие данные требуются, и какие будут на выходе. Второй – поток шагов формируется непосредственно в теле программы моделирования. Первый подход обладает большей универсальностью, второй проще в реализации. Предлагается использовать для создания модели компромиссный вариант. Работа программы

на основе MPI представляется, как расчёт потока шагов алгоритма. Сам алгоритм представляется как последовательность шагов, каждый из которых имеет общее для нескольких последовательных шагов имя и порядковый номер. При этом оборудование описано так, что каждый процессор может выполнять шаги только с одним именем. Под процессором в данном случае может подразумеваться физическое устройство, либо виртуальная машина.
Существует множество средств анализа и оптимизации вычислительных процессов, использующих интерфейсы MPI. В ЛИТ используется гетерогенная среда, включающая многоядерные процессоры, объединенные между собой при помощи сети в единый кластер, и облачная инфраструктура. Для моделирования такой среды требуется оригинальный подход. Оригинальность заключается в том, что используется дискретное моделирование событий, что позволяет в рамках единого подхода описать программный комплекс, использующий интерфейсы MPI как на нескольких ядрах, в рамках одного сервера, так и виртуальные машины, взаимодействующие между собой в облачной архитектуре.

Моделирование перечисленных выше схем потребовало внесения ряда изменений в программу SyMSim . Так в список полей таблицы, описывающей оборудование, введён параметр, который может принимать три значения: 0, 1 и 2. Каждая строчка таблицы описывает один вычислительный элемент, которому назначен соответствующий тип. При значении параметра 0 описываемая часть сайта либо весь сайт представляет собой классический кластер, состоящий из *n* счётных узлов (ядер). Каждый счётный узел может быть занят задачей, находящейся в состоянии выполнения или ожидания загрузки данных из внешнего источника. Время выполнения задания фиксировано и определяется из поля таблицы, описывающей входной поток. При значении параметра равном 1 сайт представляется виртуальным кластером, развёрнутым в облачной среде, с заранее сконфигурированными узлами в количестве *n*. Если заняты все узлы, задание помещается в очередь так же, как в случае классического кластера. При запуске задания ко времени расчёта добавляется фиксированная величина в секундах, которая определена статически в программе моделирования. При значении параметра равном 2 сайт представлен виртуальным кластером с динамическим созданием и уничтожением в облаке виртуальных машин. При этом делается допущение, что виртуальная машина создается и уничтожается каждый раз при запуске и завершении задания, независимо от состояния входной очереди. Значение величины *n* в этом случае определяет максимальное количество виртуальных машин, которые могут быть одновременно созданы. Если поток заданий требует загрузки файлов, то величина параметра *n* оказывает существенное влияние на результаты моделирования. В этом случае ожидание загрузки не приводит к простою физических процессоров и замедлению обработки всего потока в целом.

Проверка модели

Сделаем следующие 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 инфомацией 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.
5. Время, затраченное программой до начала итераций и после их завершения мало и им можно пренебречь.

Такой простейший случай можно описать аналитически. Время расчета будет определяться формулой:

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

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

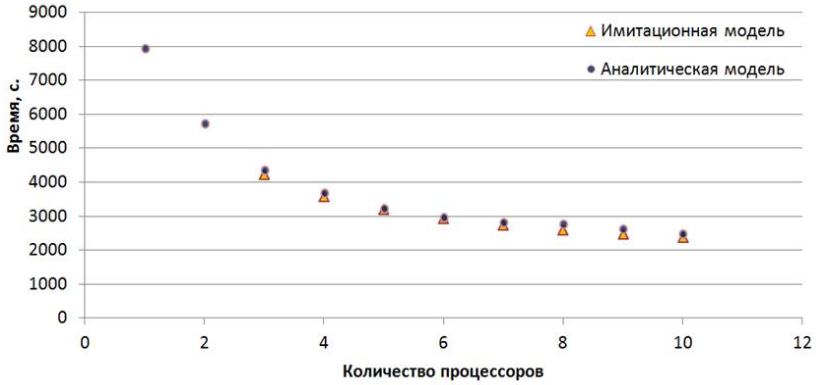

Рис. 1: Верификация модели

Сравнение результатов, полученных эмпирическим путем с результатами имитационного моделирования (рис.1), показало, что имитационная модель корректно моделирует параллельные расчёты, выполненные с использованием технологии MPI.

Результаты моделирования

В приведённых выше рассуждениях мы не учитывали эффект замедления скорости передачи данных с увеличением количества одновременно передаваемых буферов. Но с увеличением количества процессоров эта величина растёт линейно (рис.2).

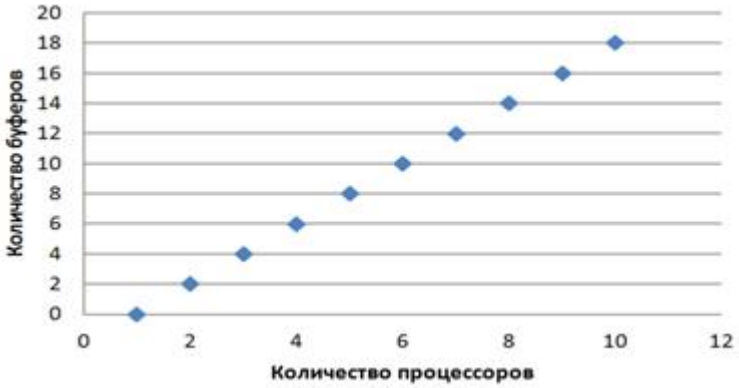

Рис. 2: Количество передаваемых за одну итерацию буферов

Поскольку в тесте невозможно разделить время, затраченное на передачу информации, от времени работы ЦПУ, то воспользуемся косвенным методом.

Логично предположить, что время собственно вычислений для процессора при фиксированном количестве машинных инструкций не меняется. Время передачи данных будет зависеть от величины нагрузки на сеть и будет величиной случайной. Тогда проведём несколько серий тестов, фиксируя в каждой количество процессоров. Интенсивность обмена по сети и количество операций ввода-вывода будет линейно увеличиваться, что следует из приведённого выше графика. Тогда, если время передачи буфера не зависит от нагрузки сети, то с увеличением количества передаваемых буферов стандартное отклонение времени

выполнения задачи, которое складывается из времени вычисления и времени передачи, будет монотонно уменьшаться. Ниже представлен график стандартного отклонения времени счёта от числа процессоров на серии из 10 тестов при количестве процессоров от 1 до 10

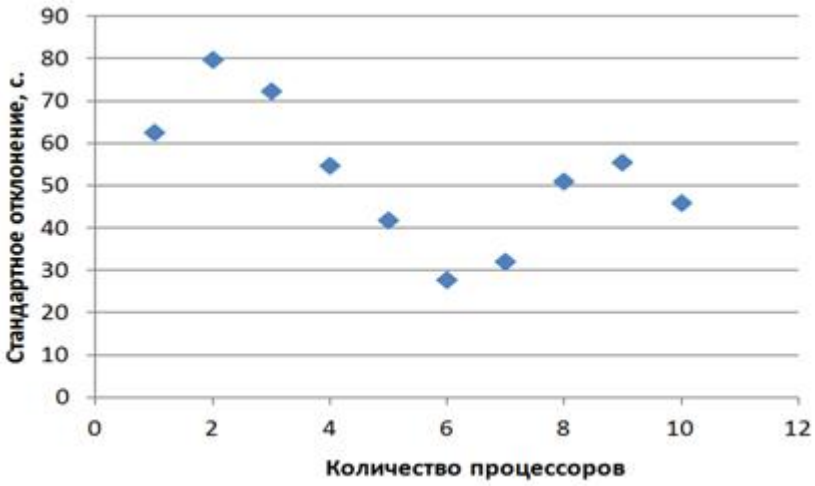

Рис. 3: Стандартное отклонения времени счёта от числа процессоров

Из графика на рис.3 видно, что до 6 процессоров стандартное отклонение времени вычисления монотонно уменьшается, если не учитывать вырожденный случай одного процессора, в котором стандартное отклонение имеет непонятный источник. Начиная с 7-ми процессоров оно растёт, что показывает на то, что время обмена становится менее стабильным. Наиболее вероятное объяснение – столкновения пакетов и повторные передачи.

Этот эффект учитывается в имитационной модели методом линейного увеличения величины случайного числа, добавляемого к среднему времени передачи буфера между процессорами. Коэффициент увеличения выбирается, исходя из результатов теста, и для количества процессоров меньше 6 равен 0.

Полученные при моделировании MPI-вычислений результаты представлены на рис. 4.

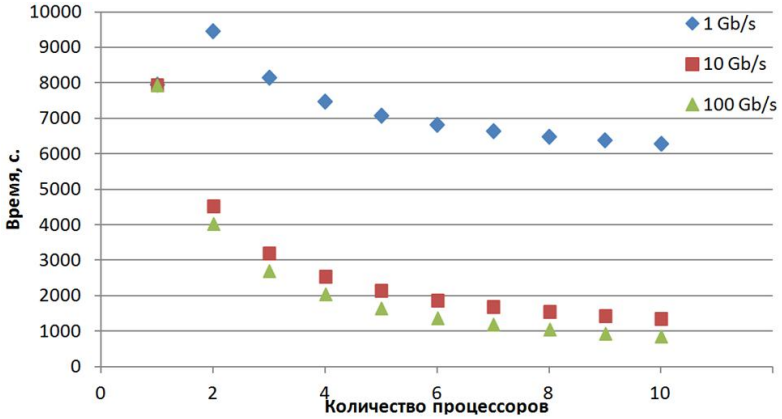

Рис. 4: Зависимость времени расчета от кол-ва CPU и скорости сети

При количестве процессоров более 6 существенного выигрыша во времени выполнения расчётов длинных ДП не наблюдается. Увеличение скорости сети с 1 Gb/s до 10 Gb/s дает существенный выигрыш, но увеличение скорости с 10 Gb/s до 100 Gb/s существенного выигрыша не даст.

Заключение

Работы, выполненные по имитационному моделированию расчётов длинных ДП с использованием MPI-технологий позволяют пользователям без проведения серии тестовых запусков подобрать оптимальное количество процессоров при известном типе сети, характеризуемой пропускной способностью и латентностью. Это может существенно сэкономить вычислительное время на счётных ресурсах, высвободив его для решения реальных задач.

В частности, моделирование вычислительных процессов длинных ДП, использующих интерфейс MPI, позволило определить оптимальные границы для количества ЦПУ и скорости сети в облачной реализации проводимых вычислений.

Список литературы

Alexandrov E.I., Belyakov D.V., Matveyev M.A., Podgainy D.V., Streltsova O.I., Torosyan Sh.G., Zemlyanaya E.V., Zrelov P.V., Zuev M.I. Research Of Acceleration Calculations In Solving Scientific Problems On The Heterogeneous Cluster Hybrilit // Bulletin of PFUR. Series Mathematics. Information Sciences. Physics. — 2015. — No 4. — P. 30–37.

Baranov A.V., Balashov N.A., Kutovskiy N.A., Semenov R.N. JINR cloud infrastructure evolution // Physics of Particles and Nuclei Letters, ISSN 1547-4771, eISSN: 1531-8567, 2016, vol. 13, No. 5, pp. 672 675. DOI: 10.1134/S1547477116050071.

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.

Korenkov V. V., Nechaevskiy A. V., Ososkov G. A., Pryahina D. I., Trofomov V. V., Uzhinskiy A. V. Simulation concept of NICA-MPD-SPD Tier0-Tier1 computing facilities // Particles and Nuclei Letters. — 2016. — Vol. 13, No 5. — P. 1074–1083.

Simulation of Cloud Computation MPI Applications

N.A. Kutovskiy[1,2], A.V. Nechaevskiy[1], G.A. Ososkov[1,a], V.V. Trofimov[1]

[1] Joint Institute for Nuclear Research, 6 Joliot-Curie street, Dubna, Moscow region, 141980, Russia

[2] Plekhanov Russian University of Economics,36 Stremyanny per., Moscow, 117997, Russia

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

At the Laboratory of Information Technologies of the Joint Institute for Nuclear Research new software for job parallel calculation on the cloud resources and computing cluster is in progress. It is expected that created cloud center of parallel computing will significantly improve the efficiency of the numerical calculations and expedite the receipt of new physically meaningful results.

At the moment, on the basis of the JINR computer center a parallel algorithm and corresponding software for parallel computing using MPI technology are developed. To optimize the scheme of parallel computations it is necessary to test the algorithm for various combinations of equipment parameters, number of processors and parallelization levels. Thus, the need to evaluate the impact of various factors (processor speed, throughput of communication network, its latency) to computing speed of particular job is arisen. Such evaluation can be achieved by simulating the MPI computing processes. Results of the MPI computing simulations is presented on the example of calculating long Josephson junctions.

Keywords: simulation, cloud computing, parallel computing, MPI

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