Анализ структуры задержек передачи информации в вычислительном кластере*

А. А. Горелов, А. И. Майсурадзе, А. Н. Сальников

Московский государственный университет имени М. В. Ломоносова

Предлагается схема сбора и анализа задержек пересылки данных от одного процесса программной модели MPI к другому. Предлагаемый подход абстрагируется от конкретной реализации сетевого взаимодействия между узлами вычислительного кластера, он опирается на специальные программные системы, собирающие данные о коммуникационной среде. Предложена адекватная информационная модель задержек и специальный способ её настройки. Показано, как использовать эту модель в задачах диагностики коммуникационной среды вычислительного кластера. Все этапы предлагаемой схемы проиллюстрированы на реальных данных о задержках, собранных на суперкомпьютерных системах МГУ имени М. В. Ломоносова.

1. Введение

При распараллеливании сложных задач на вычислительном кластере встаёт проблема планирования порядка и места выполнения подзадач исходной задачи. Современные распределённые программы реализуют такие методы, состав и число подзадач в которых становится известным только в ходе расчёта. Поэтому если раньше достачно было методов статического составления расписания, то есть планирование проводилось до запуска задачи на кластере, то в настоящий момент всё чаще используются динамические методы, выполняющие планирование в ходе выполнения задачи. При планировании необходимо учитывать не только время работы узлов кластера над подзадачами, но и время подготовки узлов для выполнения очередной подзадачи. При этом для оптимизации общего времени решения задачи на кластере необходимо учесть множество факторов, специфичных для каждого вычислительного кластера, в том числе и задержки пересылки данных от одного узла к другому. Соответственно, системы планирования должны иметь информацию о задержках (модели задержек), причём такая информация должна храниться компактно, а доступ к ней должен быть быстрым.

Отметим, что причины задержек передачи информации в сети многочисленны и специфичны для каждой системы, поэтому построение моделей задержек на основании изучения факторов их возникновения (функциональных моделей) предполагает подробное исследование программного и аппаратного обеспечения на каждом уровне сетевого протокола, что оказывается возможным только для чрезвычайно простых систем. В данной работе предлагается использовать информационные модели задержек, позволяющие абстрагироваться от конкретной реализации сетевого взаимодействия. Показано, что возможно построить адекватную информационную модель задержек, которую можно использовать, например, в задачах динамического планирования расписания или в задачах диагностики коммуникационной среды кластера, отталкиваясь только от «сырых данных», полученных в результате тестирования сети специальными программными системами.

В данной публикации работа предложенных методов демонстрируется на задержках передачи сообщений между процессами в программной модели MPI. Однако предлагаемые методы анализа легко обобщить на передачу сообщений между узлами разной природы, поэтому в теоретической части работы термины процесс и узел кластера не различаются и обозначают узлы с точки зрения среды передачи сообщений.

Типичным результатом работы существующих систем тестирования сети (network bench-

*Работа выполнена при финансовой поддержке РФФИ (гранты №13-01-00751а и №15-07-09214а).

marks) являются некоторые стандартные числовые показатели (статистики) полученной в ходе тестов выборки задержек. Например, на основании выборки задержек передачи сообщений фиксированной длины из одного узла в другой узел вычисляется эжмпирическое среднее, медиана, стандартное отклонение, максимум и минимум. При этом выбор вычисляемых статистик не основывается на свойствах и природе описываемых распределений, а значит, их набор, возможно, не является удачным описанием нужных распределений. Соответственно, в данной работе рассматривается проблема построения модели задержек при фиксированных контролируемых пользователем параметрах: длине сообщения, узле-отправителе и узле-получателе. Такая модель должна по компактности соответствовать набору числовых характеристик, а по информативности— хранению выборки задержек.

На основе модели задержек предложен метод диагностики вычислительного кластера на основании коммуникационных свойств задержек. Все предложенные методы будут проиллюстрированы реальными задержками сообщений в суперкомпьютерных системах МГУ им. М. В. Ломоносова. В целом, исследование проводится с целью обеспечить системы динамического планирования информацией о задержках. Метод агрегирования отдельных конкретных задержек для повышения эффективности их использования уже разработан и реализован, однако в данной работе из-за ограничений по размеру не рассматривается.

2. Трёхмерное аналитическое пространство задержек

Опишем общую модель вычислительной системы, используя терминологию по [8]. Класс систем со множественным потоком команд и множественным потоком данных (MIMD) предполагает, что в вычислительной системе есть несколько устройств обработки команд (процессоров), объединённых в единый комплекс и работающих каждое со своим потоком команд и данных. Вычислительные системы класса MIMD по организации памяти делятся на два подкласса: с общей памятью и с распределённой памятью. В системах второго подкласса память (в смысле адресного пространства) некоторым образом распределена между процессорами системы. Таким образом, все процессоры вычислительной системы с распределённой памятью можно разбить на группы процессоров, работающих в одном адресном пространстве. Такие группы называются вычислительными узлами. Для обмена информацией узлы объединяются друг с другом некоторой коммуникационной средой. Процессоры имеют доступ к памяти только своего узла, и получение информации с других узлов возможно только через коммуникационную среду данной системы. Если рассматривать узлы как вершины, а соединения в рамках коммуникационной среды между ними как рёбра, мы получим коммуникационную сеть вычислительной системы. Вычислительный кластер— объединённый коммуникационной сетью набор узлов, выполняющий вычисления и представляемый пользователю как единая система.

На каждом узле вычислительного кластера операционная система управляет работой процессов. Процесс— это такой контейнер для ресурсов (адресное пространство, стек и т. д.), который включает хотя бы один поток команд. С точки зрения адресного пространства процесс не может выйти за рамки своего узла. При распараллеливании задачи для систем с распределённой памятью она разбивается на несколько подзадач, выполняющихся в виде процессов на отдельных узлах. Необходима передача данных между процессами. Существует много программных интерфейсов для коммуникации процессов, самая распространённая из таких технологий — MPI [2].

MPI рассматривает все коммуникации между процессами на узлах как сообщения некоторой структуры. В нашей работе мы не будем исследовать влияние содержания MPI-сообщений на их передачу через коммуникационную сеть. Будем варьировать только размер сообщений, процесс-отправитель и процесс-получатель. Таким образом, для нас сообщение— это последовательность произвольных байтов определённой длины с определённым

узлом-отправителем и определённым узлом-получателем[1]. Получается, что каждое сообщение характеризуется тремя параметрами, которые в анализе данных принято называть измерениями. Таким образом, вводится трёхмерное аналитическое пространство. В каждой ячейке этого пространства хранится и анализируется информация о задержках передачи сообщений с фиксированным набором значений параметров.

В данном исследовании сначала будет предложена модель задержек для одного элемента аналитического пространства. Агрегирование элементов аналитического пространства (понижение его размерности, кластеризация) не описывается в данной публикации. когда оно было проведено, оказалось, что число таких кластеров для реальных систем составляет единицы, т. е. удаётся информацию о задержках представить компактно.

3. Модель задержек

Задержка передачи сообщения в коммуникационной сети — это интервал времени с момента принятия заявки на отправку сообщения узлом-отправителем до момента полного получения сообщения узлом-получателем. В конкретных вычислительных системах с помощью специальных программных продуктов можно измерять задержки при фиксированных контролируемых параметрах. В нашем исследовании сбор исходных данных о задержках производился с помощью модифицированной программной системы Network Test из пакета Parus [7]. Система Network Test позволяет анализировать коммуникационную сеть вычислительного кластера в нескольких режимах: one-to-one, режим рассылки, режимы get и put и так далее. Данная работа иллюстрируется на данных, полученных в режиме one-to-one. В этом режиме один процесс посылает сообщение одному другому процессу, пока остальные процессы бездействуют. Отметим, что в рамках работы была предложена и внедрена модульная архитектура для системы Network Test.

В результате работы системы тестирования в каждой ячейке трёхмерного аналитического пространства мы получаем некоторую выборку задержек. В данной разделе нашей задачей является построение удачного описания набора задержек независимо в каждой ячейке аналитического пространства, то есть при фиксированных контролируемых параметрах сообщения. Разумеется, задержки зависят не только от контролируемых нами параметров сообщения, но и от многих других характеристик состояния сети. Будем считать, что в каждой ячейке аналитического пространства задержка является случайной величиной. Тогда для стохастического моделирования задержки требуется выбрать семейство распределений и оценить параметры распределения по выборке.

В некоторых работах уже предпринимались попытки создания моделей задержек при передаче информации. В [4] проводился анализ задержек пакетов IP в сети Интернет и получилось, что большинство исследованных автором распределений задержек близки либо к трёхпараметрическому гамма-распределению, либо к трёхпараметрическому логнормальному распределению в зависимости от получателя и отправителя. Также в [1] и в [5] было показано, что задержки в сети Интернет хорошо описываются трёхпараметрическим логнормальным распределением. Отметим, что условия проведения исследований во всех рассматриваемых работах отличаются от таковых в нашем случае.

Изучение собранных нами данных позволяет заметить некоторые особенности изучаемых задержек. Во-первых, «мультимодальность» распределений задержек: на крупномасштабных гистограммах можно визуально выделить от одного до трёх унимодальных сгущений («горбов»). Во-вторых, «атомарность» распределений задержек: большинство задержек имеют значения из небольшого множества значений. Иными словами, задержки разбиваются на довольно большие группы равных между собой. Например, для суперкомпьютера Blue Gene/P в выборке получается 6% уникальных значений. Однако следует от-

[1] Напомним, что в теоретической части под узлом мы понимаем узел сети передачи сообщений. Для MPI это процесс.

Таблица 1: Показатели качества работы стандартного и предложенного методов на одном наборе данных. Предложенный метод по всем показателям превосходит стандартный. Суперкомпьютер Blue Gene/P.

	стандартный алгоритм	минимизация расстояния
Время работы, с, меньше — лучше	111.1	4.540
Правдоподобие (для станд.), больше — лучше	0.9773	1.519
Расстояние (для предл.), меньше — лучше	0.1235	0.0330

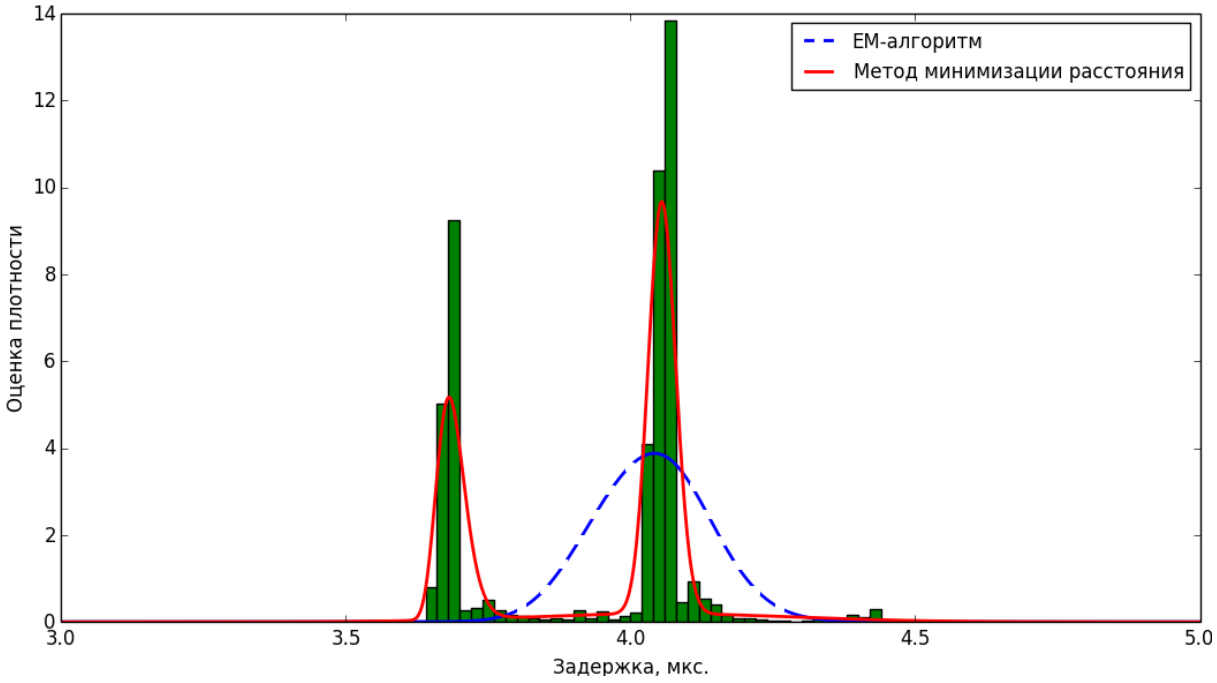

Рис. 1: Результаты работы стандартного и предложенного методов на одном наборе данных. Суперкомпьютер Blue Gene/P.

метить нерегулярность интервалов между уникальными значениями задержек, т. е. атомы распределения не получается объяснить дискретностью времени в системе. В-третьих, имеются выбросы, отстоящие далеко от сгущений в обе стороны.

Из анализа данных, подкреплённого существующими публикациями, получилось, что распределение задержек в ячейке аналитического пространства в крупном масштабе допустимо описывать смесью трёхпараметрических логнормальных распределений. Другие смеси, например смесь нормальных распределений, хуже согласуются с наблюдениями.

Серьёзным теоретическим вызовом оказалась разработка метода оценки параметров такой смеси (метода разделения). В [3] было показано, что оптимизирующие последовательности для оценок максимального правдоподобия для параметров трёхпараметрического логнормального распределения расходятся. Следовательно, мы лишены важного компонента для использования наиболее распространённых подходов к разделению смесей. Более того, тестирование показало, что применение методов разделения напрямую к данным, без какой-либо их предварительной обработки, не приводит к желаемому результату: компоненты смеси в этом случае сходятся не к крупномасштабным сгущениям, а к отдельным

Таблица 2: Результаты тестирования систем (число непройденных / всего тестов).

Вид теста	СК Ломоносов	Blue Gene/P	Regatta
Монотонность	53746 / 513408	9044 / 321328	297 / 7728
Нер-во тр-ка	0 / 10752	0 / 8512	0 / 1344
Неделимость	106862 / 255360	0 / 46564	1146 / 3696
Всего тестов	160608 / 779520	9044 / 376404	1443 / 12768

Таблица 3: Сравнение результатов тестирования по выборкам и по предложенному компактному представлению (число различий / всего тестов).

Вид теста	СК Ломоносов	Blue Gene/P	Regatta	Всего тестов
Монотонность	106 / 513408	253 / 321328	6 / 7728	365 / 842464
Нер-во тр-ка	0 / 10752	0 / 8512	0 / 1344	0 / 20608
Неделимость	464 / 255360	0 / 46564	0 / 3696	464 / 305620
Всего тестов	570 / 779520	253 / 376404	6 / 12768	829 / 1168692

самым большим атомам эмпирического распределения. С подобным эффектом приходилось, конечно, сталкиваться и другим исследователям — например, он описан в [6].

Нами был разработан метод разделения смеси на основе минимизации расстояний между распределениями. Этот метод превосходит попытки разделить смесь стандартными методами из семейства EM как по скорости работы, так и по качеству оценки, даже если брать оценку, которую оптимизирует стандартный алгоритм, см. табл. 1.

Таким образом, на данном этапе обработки каждая ячейка аналитического пространства содержит выборку задержек, оценки параметров смеси, числовые показатели эмпирических распределений. Ниже демонстрируется, что оценки параметров смеси являются одновременно компактным и достаточным дескриптором для анализа коммуникационной среды.

4. Анализ задержек на маршрутах

Пусть $t_{s,a\to b}$ — задержка передачи сообщения длины s от узла a узлу b. Основываясь на интерпретации задержек, логично потребовать для качественно настроенной коммуникационной среды выполнения следующих трёх условий.

1. $t_{s_1,a\to b} \geq t_{s_2,a\to b}$ при $s_1 \geq s_2$ (свойство монотонности) — длинное сообщение идёт не быстрее короткого;
2. $t_{s,a\to b} \leq t_{s,a\to c} + t_{s,c\to b}$ (неравенство треугольника) — лучше посылать сообщение сразу в конечный пункт, чем явно указывать промежуточные пункты;
3. $t_{s_1+s_2,a\to b} \leq t_{s_1,a\to b} + t_{s_2,a\to b}$ (свойство неделимости) — лучше посылать сообщение целиком, чем явно разбивать на части.

Эти свойства можно назвать обобщёнными метрическими требованиями. Напомним, что речь идёт о сравнении случайных величин, т. е. о стохастическом доминировании. По-

казано, что все три рассматриваемые свойства независимы.

Провести тестирование указанных требований только по числовым показателям эмпирических распределений не получается, посколько требования содержат арифметические операции, в статистике известно, для таких ситуаций мощных критериев нет. Сравним результаты тестирование по большой выборке и по предложенному компактному представлению. Результаты диагностики вычислительных систем Regatta, BlueGene/P и Ломоносов, установленных в МГУ имени М. В. Ломоносова, показаны в табл. 2. Видно, что системы работают не идеально. Результаты сравнения методов диагностики представлены в табл. 3. Видно, что компактное представление даёт результаты, практически неотличимые от результатов на богатом представлении.

5. Заключение

Предложена схема сбора и анализа задержек передачи сообщений в коммуникационной среде вычислительного кластера. Предложена вероятностная модель задержек и способ её настройки, что позволяет компактно описать взаимодействие между узлами. Предложен метод диагностики коммуникационной среды. Показано, что предложенное описание задержек позволяет решать указанные задачи предметной области с достаточным качеством. Реализованы система сбора информации и система анализа собранной информации. Рассматриваемые методы прошли апробацию на реальных данных с суперкомпьютерных систем МГУ имени М. В. Ломоносова. Оказалось, что для реальных систем не всегда выполняются естественные требования к задержкам.

Литература

1. Corlett A., Pullin D., Sargood S. Statistics of one-way internet packet delays // 53rd IETF. — 2002.

2. Gropp W., Lusk E., Skjellum A. Using MPI: portable parallel programming with the message-passing interface. — MIT press, 1999. — Vol. 1.

3. Hill B. M. The three-parameter lognormal distribution and Bayesian analysis of a point-source epidemic // Journal of the American Statistical Association. — 1963. — Vol. 58, no. 301. — P. 72–84.

4. Karakaş M. Determination of Network Delay Distribution over the Internet : Ph. D. thesis / Mehmet Karakaş ; MIDDLE EAST TECHNICAL UNIVERSITY. — 2003.

5. Mukherjee A. On the dynamics and significance of low frequency components of internet load // Technical Reports (CIS). — 1992. — P. 300.

6. On convergence problems of the em algorithm for finite gaussian mixtures. / Cédric Archambeau, John Aldo Lee, Michel Verleysen et al. // ESANN. — Vol. 3. — 2003. — P. 99–106.

7. Salnikov A. N. Parus: A parallel programming framework for heterogeneous multiprocessor systems // Recent Advances in Parallel Virtual Machine and Message Passing Interface. — Springer, 2006. — P. 408–409.

8. Воеводин В. В., Воеводин Вл. В. Параллельные вычисления. — СПб.: БХВ-Петербург, 2002.

Delay structure mining in computing cluster

Alexey Gorelov, Archil Maysuradze and Alexey Salnikov

Keywords: communication environment of computing cluster, MPI programming model, message passing delays, information model of delays, communication environment diagnostics, delay data aggregation

We propose a new scheme for the collection and analysis of message passing delays from one MPI process to another. The proposed approach is abstracted from a specific implementation of network interaction between computing cluster nodes. It is argued that it is possible to collect data with special software systems and build adequate information model of delays. It is shown how to use this model in the communication environment diagnostics and the dynamic scheduling problems. All stages of the proposed scheme are illustrated with real data collected at Lomonosov Moscow State University supercomputer systems.