УДК 004.416.6

АВТОМАТИЗИРОВАННАЯ НАСТРОЙКА ПАРАМЕТРОВ БИБЛИОТЕКИ S-MPI

Д. В. Донцов, С. И. Сапронов
(ООО "Центр компетенций и обучения", г. Саров)

Предлагается инструмент для настройки параметров библиотеки S-MPI, реализующей стандарт MPI-2. Данный инструмент позволяет автоматизировать подбор оптимальных значений параметров библиотеки S-MPI с учетом специфики как кластера, так и параллельного приложения. Рассматриваются подходы, применяемые при автоматизированной настройке, включая возможность анализа зависимости производительности параллельного приложения от значений настраиваемых параметров без анализа его исходного кода.

Ключевые слова: MPI, параллельные вычисления, методы оптимизации.

Введение

Производительность параллельных приложений, разработанных с использованием стандарта MPI (Message Passing Interface) [1] для высокопроизводительных кластерных систем, зависит от множества факторов, основные из которых можно разделить на три группы:

1) архитектура кластерной системы:
   - архитектура процессоров;
   - объем оперативной памяти;
   - коммуникационное оборудование, используемое для объединения вычислительных узлов кластерной системы;
   - программное окружение;
2) параметры запуска приложения:
   - число запускаемых MPI-процессов;
   - порядок раскладки процессов по узлам кластерной системы;
3) архитектура приложения:
   - способы организации взаимодействия процессов приложения во время его выполнения;
   - используемые MPI-операции;
   - размеры передаваемых сообщений.

Библиотека S-MPI [2] обеспечивает универсальность выполнения приложения, т. е. приложение, однажды собранное с библиотекой S-MPI, может быть перенесено и выполнено на любой кластерной системе с поддерживаемой архитектурой. Для достижения максимальной производительности в библиотеке S-MPI реализованы средства автоматического определения архитектуры кластерной системы, на которой происходит запуск приложения. В зависимости от архитектуры кластерной системы и параметров запуска приложения библиотека S-MPI пытается установить оптимальные для конкретных условий запуска значения параметров, определяющие, например, коммуникационные среды, которые требуется инициализировать для взаимодействия процессов приложения, алгоритмы коллективных и двухточечных операций, которые нужно использовать для различных размеров сообщений, размеры внутренних буферов для передачи и т. д.

Однако в силу разнообразия архитектур кластерных систем и, в частности, характеристик конкретных коммуникационных сред выбранные библиотекой значения параметров не всегда могут быть оптимальными. Кроме того, библиотека не обладает сведениями об архитектуре приложения. Поэтому подбор оптимальных значений параметров MPI-библиотеки с учетом специфики архитектуры конкретной кластерной системы, параметров запуска приложения и его архитектуры является важным вопросом, решение которого может дать заметный выигрыш в производительности.

Так как подбор оптимальных значений для параметров реализации MPI является долгим и трудоемким процессом, а его выполнение вручную сопряжено с возможными ошибками из-за присутствия человеческого фактора, в последнее время стали разрабатываться программные инструменты, автоматизирующие процесс настройки.

Инструменты автоматической настройки параметров реализаций MPI

К наиболее известным инструментам автоматической настройки параметров реализаций MPI относятся MPI Tuner библиотеки Intel MPI [3] и OTPO (Open Tool for Parameter Optimization) библиотеки OpenMPI [4].

MPI Tuner, как и вся библиотека Intel MPI, является коммерческим продуктом и служит только для настройки параметров самой библиотеки Intel MPI.

OTPO является свободно распространяемым продуктом и предназначен для настройки параметров библиотеки OpenMPI, которая используется как кодовая база для библиотеки S-MPI. Однако этот инструмент обладает рядом серьезных недостатков, основные из которых:

- отсутствие режима настройки параметров библиотеки на специфику архитектуры кластерной системы*;
- возможность настройки только тех параметров, значения которых выбираются из предопределенного списка или диапазона значений, и невозможность настройки композиционных параметров, которые могут содержать множественные значения и условия их выбора;
- изначальная поддержка работы только с некоторыми тестовыми приложениями, такими как NetPipe [5], Skampi [6] и NAS Parallel Benchmarks [7]. Для поддержки других приложений требуется модификация исходного кода OTPO.

С учетом указанных обстоятельств для библиотеки S-MPI был разработан собственный инструмент, получивший название MPIBoost, который не требует модификации исходного кода для настройки любого приложения и может функционировать в двух режимах: настройки параметров на специфику кластера и на специфику приложения.

Следует заметить, что оптимальные значения параметров библиотеки могут существенно зависеть от параметров запуска приложений, в частности, числа вычислительных узлов и порядка раскладки процессов по этим узлам (например, *round robin* или *group round robin*). Поэтому настройка параметров как на специфику кластера, так и на специфику приложения должна осуществляться отдельно для различных сочетаний числа узлов и числа процессов с учетом, если необходимо, порядка раскладки процессов по этим узлам.

Режим настройки параметров на специфику кластера

Настройка на специфику кластера предназначена в первую очередь для поиска оптимальных значений параметров библиотеки с учетом архитектуры кластера. В этом режиме используется набор тестов производительности IMB (Intel® MPI Benchmarks) [8], который позволяет MPIBoost получать данные о времени выполнения необходимых MPI-функций для различных размеров передаваемых сообщений. С помощью единичных тестов определяются оптимальные значения параметров библиотеки для основных двухточечных и коллективных MPI-операций на конкретной кластерной системе, что повышает производительность библиотеки S-MPI, а следовательно, приложений, которые выполняются на этом кластере с ее использованием.

Можно выделить три группы параметров настройки S-MPI, значения которых:

1) выбираются из определенного диапазона;
2) выбираются из предопределенного списка допустимых значений;
3) являются композиционными.

Нахождение значений для параметров первой и второй групп основано на хорошо известных методах перебора и бисекции. Для композиционных параметров оно осуществляется следующим способом.

Значения композиционных параметров библиотеки S-MPI представляют собой пары *значение — диапазон размеров сообщений* и используются, например, для задания различных алго-

*Такой режим позволяет получать значения параметров реализации MPI, оптимальные или субоптимальные для многих приложений, причем он заметно дешевле режима настройки параметров на специфику конкретных приложений и может использоваться для настройки значений по умолчанию параметров самой библиотеки MPI.

ритмов коллективных операций для разных размеров сообщений.

Для того, чтобы определить оптимальный алгоритм выполнения коллективной операции при передаче сообщений, имеющих размер в определенном диапазоне, применяется схема, состоящая из следующих шагов:

1. Получение данных о времени выполнения коллективной операции для каждого алгоритма в нескольких точках требуемого диапазона размеров сообщений. Данные получаются автоматическим запуском соответствующего теста IMB с нужными размерами сообщений.
2. Определение на основе полученных данных *точек переключения алгоритмов*, т. е. точек диапазона размеров сообщений, в которых время выполнения коллективной операции для двух (в общем случае нескольких) алгоритмов совпадает.
3. Формирование результата. На каждом участке, полученном в результате деления диапазона размеров сообщений точками переключения алгоритмов, определяется алгоритм с минимальным временем выполнения и формируется результирующая строка вида:

   номер алгоритма: *диапазон*; *номер алгоритма*: *диапазон*; ...; *номер алгоритма*: *диапазон*.

Использование этой схемы демонстрируется на рис. 1 (см. также цветную вкладку) на примере коллективной операции MPI_Allgather. Здесь приведены графики зависимости времени выполнения операции по различным алгоритмам от размеров сообщений в диапазоне от 1 байта до 4 Мбайт. Видно, что это время для различных алгоритмов сильно зависит от размера сообщений, и, следовательно, правильный выбор алгоритма для заданного размера сообщений очень важен. Так, точками переключения алгоритмов в приведенном примере будут размеры сообщений 228, 1 540, 2 555, 12 142, 22 259, 168 506, 4 194 304 байт, а оптимальным значением для параметра, который задает правила выбора алгоритмов для коллективной операции MPI_Allgather, будет следующая строка: 2:1-228; 4:228-1 540; 3:1 540-2 555; 4:2 555-12 142; 3:12 142-22 259; 4:22 259-168 506; 3:168506-4194304.

Заметим, что настройка параметров библиотеки S-MPI производится самими ее разработчиками для наиболее распространенных архитектур кластерных систем. Полученные настройки вносятся во встроенную базу данных библиотеки, содержащую специфику различных кластерных платформ. Содержимое этой базы данных используется библиотекой для обеспечения оптимального функционирования приложения в конкретной вычислительной среде.

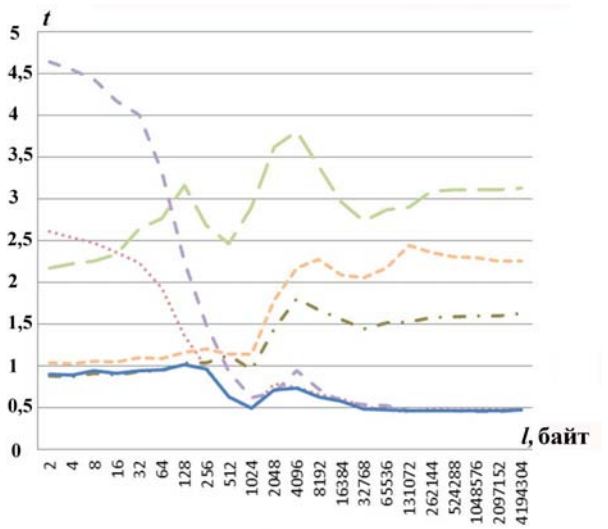

Рис. 1. Зависимость времени выполнения различных алгоритмов коллективной операции MPI_Allgather от размеров сообщений: — алгоритм 5; ····· — алгоритм 4; — — — алгоритм 3; — · — алгоритм 2; — — — алгоритм 1; — найденное значение

Режим настройки параметров на специфику приложения

Основная особенность настройки параметров библиотеки S-MPI на специфику приложения заключается в том, что изначально не известно, какие MPI-функции и с какими параметрами будет вызывать приложение и, в частности, сообщения какого размера будут передаваться при выполнении двухточечных и коллективных операций. Перебор всех возможных комбинаций даже только самых важных операций для различных размеров сообщений практически невозможен из-за больших затрат времени. Для решения этой проблемы в инструментах настройки параметров существующих MPI-реализаций применяются разные подходы. Так, например, OTPO требует от пользователя список параметров с указанием их допустимых (или реально используемых в приложении) значений.

MPIBoost использует для этой цели режим сбора статистической информации, обеспечивае-

мый библиотекой S-MPI. При включении такого режима собирается и по завершении приложения выдается на печать вся требуемая информация о выполнении приложением MPI-функций, включая размеры передаваемых сообщений и времена выполнения отдельных операций. В общем виде порядок работы MPIBoost в режиме настройки параметров на специфику приложения заключается в следующем: MPIBoost многократно запускает приложение в режиме сбора статистической информации, анализирует эту информацию и на основе полученных данных формирует оптимальные значения параметров библиотеки.

Следует заметить, что производительность приложений может зависеть не только от времени выполнения отдельных MPI-операций, но и от их сочетания при выполнении. Так, например, заметное влияние на производительность приложения могут оказывать моменты входа/выхода различных процессов приложения в/из функции, выполняющей коллективную операцию. А так как для разных алгоритмов эти моменты могут быть различными, то и время выполнения разных алгоритмов двух коллективных операций в сочетании может заметно различаться. При этом выполнение алгоритмов, выбранных в качестве оптимальных в режиме настройки параметров на специфику кластера, в сочетании может показать не лучшее время. Чтобы существенно повысить производительность приложения, MPIBoost пробует подобрать оптимальные алгоритмы не только для отдельных коллективных операций, но и для их сочетаний.

На рис. 2 приведен пример результата использования подобранных параметров библиотеки для теста IS класса B, входящего в состав NAS Parallel Benchmarks [7]. Необходимо отметить, что настройка на специфику приложения в приведенном примере осуществлялась с использованием библиотеки S-MPI, в которую в качестве значений по умолчанию были включены настройки параметров, полученные MPIBoost в режиме настройки на специфику кластера.

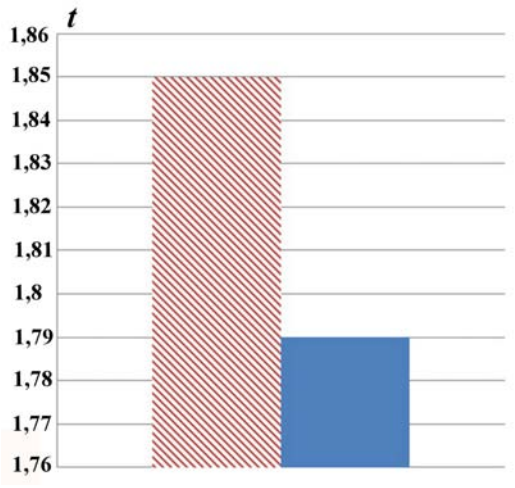

Рис. 2. Результат использования параметров, полученных в ходе настройки теста IS.B из набора NAS Parallel Benchmarks: левый столбец — настройки по умолчанию; правый столбец — оптимизированные настройки

Заключение

Представленный в статье инструмент MPIBoost позволяет производить настройку параметров библиотеки S-MPI в двух режимах: с учетом специфики кластера и специфики приложения. При этом в самой библиотеке S-MPI используются значения параметров по умолчанию, полученные с помощью MPIBoost для наиболее распространенных архитектур кластеров и параметров запуска приложений.

Таким образом, процесс настройки параметров библиотеки S-MPI является многоуровневым:

1. Настройка значений параметров библиотеки S-MPI по умолчанию для наиболее распространенных архитектур кластеров и параметров запуска приложений. Этот уровень настройки позволяет библиотеке эффективно работать сразу после ее установки на большинстве кластерных систем.
2. Настройка параметров библиотеки S-MPI на специфику конкретной кластерной системы. Предполагается, что эту настройку будет производить администратор кластера при установке пакета библиотеки S-MPI и при изменениях архитектуры кластера. Этот уровень настройки позволяет учесть специфические особенности конкретной кластерной системы и обеспечить высокую производительность большинства приложений, которые будут выполняться на данном кластере. Наибольший эффект достигается на кластерах со специфической архитектурой, настройки параметров для которой не были включены в библиотеку S-MPI как значения по умолчанию.

3. Настройка параметров библиотеки S-MPI на специфику конкретного приложения. Этот уровень настройки позволяет достичь максимальной производительности выполнения конкретного приложения на конкретной кластерной системе.

В итоге указанный процесс позволяет получить оптимальное соотношение между затраченным на настройку параметров временем и достигнутой производительностью выполнения приложений.

В дальнейшем предполагается дополнить MPIBoost новыми алгоритмами поиска оптимальных значений для параметров библиотеки S-MPI, расширить число настраиваемых параметров и, в частности, реализовать механизм поиска оптимальной привязки процессов приложения к ядрам вычислительных узлов.

Работа выполнена в рамках контракта (№ 07.524.12.4020) с Министерством образования и науки РФ.

Список литературы

1. *Message Passing Interface Forum*, MPI: A Message Passing Interface // Proc. of "Supercomputing'93". Los Alamitos: IEEE Computer Society Press, 1993. P. 878—883.
2. *Воронов Г. И., Трущин В. Д., Шумилин В. В., Ежов Д. В.* Программный комплекс S-MPI для обеспечения разработки, оптимизации и выполнения высокопараллельных приложений на суперкомпьютерных кластерных системах // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. 2013. Вып. 3. С. 55—60.
3. Центр разработчиков Intel. Библиотека Intel® MPI Library. http://software.intel.com/en-us/intel-mpi-library/.
4. *Chaarawi M., Squyres J. M., Gabriel E., Feki S.* Recent Advances in Parallel Virtual Machine and Message Passing Interface // Proc. of the 15th European PVM/MPI Users' Group Meeting. Dublin, Ireland. September 7—10, 2008. Springer, 2008. P. 210—217.
5. NetPIPE: A Network Protocol Independent Performance Evaluator. http://www.scl.ameslab.gov/netpipe/.
6. *Reussner R., Sanders P., Prechelt L., Muller M.* Recent Advances in Parallel Virtual Machine and Message Passing Interface // Proc. of 5th European PVM/MPI Users' Group Meeting. Liverpool, UK. September 7—9, 1998. Springer, 1998. P. 52—59.
7. *Bailey E., Barszcz E., Barton J. et al.* The NAS Parallel Benchmarks. Technical Report RNR-94-007. NASA Ames Research Center, March 1994.
8. Центр разработчиков Intel. Лицензионное соглашение на использование Intel® MPI Benchmarks. http://software.intel.com/en-us/articles/intel-mpi-benchmarks.

Статья поступила в редакцию 15.10.12.