

Предисловие

Стандарт MPI (Message Passing Interface) и библиотеки, реализующие его, остаются на протяжении нескольких десятков лет основным средством разработки параллельных программ для высокопроизводительных вычислительных систем с распределенной памятью. Значительное место в стандарте занимает описание коллективных, или групповых, глобальных коммуникационных операций, в которых участвуют все или часть процессов программы. Время их выполнения для многих приложений является критически важным.

Монография, которую вы читаете, — это попытка автора систематизировать и изложить наиболее значимые алгоритмы коллективных операций MPI. Большая часть книги посвящена алгоритмам, которые не учитывают архитектурных свойств вычислительных систем, а опираются лишь на пару двусторонних примитивов передачи `Send` и приема `Recv` сообщений. Такие алгоритмы формируют базис для построения более сложных архитектурно-ориентированных методов коллективных операций, построения высокопроизводительных коммуникационных сетей со специализированными топологиями и протоколами канального уровня. Актуальность этих алгоритмов не угасает, с каждым витком развития архитектуры вычислительных систем происходят возврат к базовым алгоритмам и их реализация с учетом текущего уровня развития коммуникационных и процессорных технологий.

Отражены основные подходы построения алгоритмов коллективных операций для систем определенных классов: для системы с общей памятью, иерархические алгоритмы для систем с многоуровневой иерархией памяти и коммуникационной сети. Изложена специфика реализации неблокирующих коллективных операций в библиотеках MPI — построение расписаний и продвижение их выполнения. В конце книги затрагивается вопрос экспериментальной оценки эффективности коллективных операций.

Часть монографии составили результаты исследований, выполненных автором в Сибирском государственном университе-

те телекоммуникаций и информатики (г. Новосибирск) и Институте физики полупроводников им. А.В. Ржанова Сибирского отделения РАН. Описанные практические аспекты реализации алгоритмов опираются на опыт участия автора и его аспирантов в международном проекте Open MPI (компоненты coll/base, coll/libnbc), а также на некоторые идеи, которые реализованы автором в серии библиотек структурно-ориентированных алгоритмов коллективных операций MPI по заказу компании Huawei.

Материал книги использовался в курсе лекций «Параллельные вычислительные технологии», который автор читал студентам бакалавриата и магистратуры в Сибирском государственном университете телекоммуникаций и информатики.

В своей профессиональной деятельности автору довелось общаться с большим количеством талантливых людей. Автор благодарен своему научному руководителю чл.-корр. РАН В.Г. Хорошевскому, коллегам по Кафедре вычислительных систем Сибирского государственного университета телекоммуникаций и информатики, а также сотрудникам Лаборатории вычислительных систем Института физики полупроводников им. А.В. Ржанова Сибирского отделения РАН.

Глубокую признательность и благодарность автор выражает всем, кто принял участие в обсуждении рукописи монографии: канд. тех. наук Полякову Артему Юрьевичу (стандарт PMIx, проекты OpenPMIx, Open MPI; NVIDIA), Аненкову Александру Дмитриевичу (YADRO), д-ру техн. наук Шидловскому Станиславу Викторовичу (Томский государственный университет), канд. тех. наук Кулагину Ивану Ивановичу (Институт системного программирования им. В.П. Иванникова РАН), д-ру техн. наук доценту Павскому Кириллу Валерьевичу (Институт физики полупроводников им. А.В. Ржанова СО РАН), канд. тех. наук доценту Пазникову Алексею Александровичу (National University of Singapore), канд. физ.-мат. наук Петрову Валентину Сергеевичу (проект OpenUCC, YADRO).

Автор выражает надежду, что возможные неточности и опечатки в тексте не станут непреодолимой преградой для восприятия основных идей. Замечания, предложения и комментарии можно направить по адресу mkurnosov@yandex.ru.

С уважением,
Михаил Георгиевич Курнос
профессор, доктор технических наук
Новосибирск, 2024

Введение

Как следует из названия, Message Passing Interface (www.mpiforum.org) определяет программный интерфейс библиотек высокопроизводительной передачи сообщений, включает операции двусторонних обменов (point-to-point), односторонних обменов (one-sided remote memory access) и коллективные операции, а также вводит понятия группы процессов, коммутаторов, виртуальных топологий и производных типов данных. Основное предназначение стандарта — обеспечить возможность создания параллельных программ, переносимых на уровне исходного кода между вычислительными системами различных архитектур и конфигураций: с общей и распределенной памятью, на базе заказного оборудования и свободно доступного на рынке, с процессорами общего назначения и/или специализированными ускорителями.

Изначальная мотивация создания стандарта — сложность переносимости параллельных программ между различными архитектурами вычислительных систем. К 1992 году каждый производитель высокопроизводительных вычислительных систем (Thinking Machines, NEC, Intel, Cray, Fujitsu, Hitachi), как правило, поставлял свою реализацию сред параллельного программирования, которые включали механизмы управления процессами и предоставляли операции двусторонних и коллективных обменов информацией. Можно выделить следующие проекты и программные системы, которые оказали значительное влияние на становление MPI: проект PVM (Parallel Virtual Machine) — открытая библиотека и среда времени выполнения для создания параллельных программ в модели передачи сообщений (двусторонние и ряд коллективных операций, уведомления об отказах), разработка Окриджской национальной лаборатории США (ORNL); проекты p4 (Аргонская национальная лаборатория США, ANL), Intel NX/2, IBM EUI, nCUBE Vertex, Zipcode, PICL, PARMACS, Express, Chimp, Chameleon.

Желание обобщить накопленный опыт и сформировать устойчивый интерфейс для создания параллельных программ

привел в 1992 году к возникновению инициативной группы специалистов — MPI Forum, которая включала более 40 организаций и 60 человек. В мае 1994 года была опубликована первая версия стандарта, которая включала блокирующие и неблокирующие функции двусторонних обменов, коллективные операции, управление группами процессов и коммутаторами, виртуальные топологии и производные типы данных. Эталонная реализация стандарта (reference implementation) вышла в 1996 году — библиотека MPICH. В июле 1997 года закончена работа и опубликован стандарт MPI 2.0, который включал функции односторонних обменов (one-sided communication), динамического порождения процессов, параллельный ввод-вывод (MPI I/O), интерфейсы для C++ и Fortran 90. Значительные изменения в стандарте произошли в версии MPI 3.0, которая опубликована в сентябре 2012 года. Основные изменения коснулись введения неблокирующих коллективных операций (nonblocking collective operation), разреженных коллективных операций (sparse collectives), расширения возможностей односторонних обменов, удалена поддержка интерфейса для C++. В июне 2021 года вышел MPI 4.0, который расширил интерфейс операциями с поддержкой large-count, решающий проблему переполнения знакового типа int для элементов count, коллективные операции с не меняющимися аргументами (persistent collectives), разделенные коммуникационные операции для многопоточных программ (partitioned communications) и модель сессий для инициализации MPI-программ (MPI Session).

Автор считает важным отметить, что в СССР и России велись и ведутся работы по созданию отечественных вычислительных систем с распределенной памятью и систем параллельного программирования для них: вычислительные системы с программируемой структурой (Минск-222, семейство МИКРОС) [1–5], семейство систем МВС [6], вычислительные системы со специализированными топологиями коммуникационных сетей [7, 8], реконфигурируемые вычислительные системы [9]. Автор монографии является членом научной школы по распределенным вычислительным системам с программируемой структурой, основателем которой является чл.-корр. РАН В.Г. Хорошевский. В начале 1960-х годов в Институте математики Сибирского отделения РАН были инициированы работы по созданию вычислительных систем с распределенной памятью на базе серийных ЭВМ. Для их сопряжения в систему были разработаны аппаратные системные устройства, поддерживающий основные типы коммуникацион-

ных операций: двусторонние обмены и коллективные операции: трансляционный обмен (one-to-all), коллекторный прием (all-to-one) и трансляционно-циклический обмен типа «каждый-всем» (all-to-all). Заинтересованный читатель может найти дополнительную информацию в [1–3, 5].

В центре внимания данной книги часть стандарта MPI — алгоритмы коллективных операций обмена информацией. В следующих главах дается краткое изложение основных понятий стандарта, приводится описание моделей для оценки эффективности алгоритмов и начинается систематическое описание алгоритмов.