0.1. *Третьяков Г.Н.* Исследование модели MapReduce в сравнении с библиотекой MPI

В последнее время среди технологий Big Data популярен Hadoop - свободно распространяемый набор утилит, библиотек и фреймворков для разработки и выполнения распределенных программ, работающих на кластерах из сотен и тысяч узлов. Среди основных характеристик Hadoop выделяют отказоустойчивость, масштабируемость и работу с распределенными вычислениями. Подобными характеристиками обладает и программный интерфейс MPI (Message Passing Interface, библиотека, добавляющая поддержку механизма передачи сообщений в стандартные языки программирования), разработанный для обмена данными между процессами в параллельном программировании.

Как правило, MPI наиболее распространен для обмена данными в параллельном программировании и применяется при разработке программ для кластеров и суперкомпьютеров, а также при решении задач, связанных с научным моделированием. MapReduce - это модель программирования, которая абстрагирует параллельные программы с помощью двух операторов - Map и Reduce. При больших наборах данных и невозможности обработки на одном компьютере применяются такие реализации MapReduce, как Hadoop.

Модель программирования MapReduce можно понимать, как подмножество функциональной части MPI, так как она представляет из себя стандартный функционал MPI с пользовательскими операциями. Таким образом, можно использовать MPI вместо MapReduce, но не наоборот, так как MPI описывает гораздо больший операций. Основным преимуществом же MapReduce и технологии Hadoop, в которой эта модель используется, является концентрация на единой параллельной концепции, что позволяет изучить ее в сроки, гораздо более короткие, чем MPI.

В данный момент функционал Hadoop и его применение в научном моделировании изучено не в полной мере, т.к. технология считается относительно новой. Отсюда вытекает и последующая проблема – свойства Hadoop, требующие дальнейшего исследования, и их особенность по сравнению с MPI.

В работе было проведено исследование отказоустойчивости и масштабируемости для параллельных задач с использованием инструментов и утилит Hadoop. Определены направления для дальнейшего его развития.

Работа выполнена в рамках государственного задания ИФП СО РАН (ГЗ 0242-2021-0011).

Научный руководитель — д.т.н., доц. Павский К. В.