Разработка унифицированного метода передачи параметров и запуска задач на вычислительных кластерах

М.В. Воробьёв[1], А.Н. Сальников[1]

МГУ имени М.В. Ломоносова[1]

Для обеспечения коллективного доступа к ресурсам кластерной вычислительной системы используются системы параллельной обработки заданий (системы ведения очередей). Основные функции системы ведения очередей: **1)** приём входного потока параллельных заданий от разных пользователей, **2)** ведение очереди параллельных заданий, **3)** выделение ресурсов массово-параллельной ВС для параллельного задания и освобождение занятых ресурсов после выполнения задания. [1]

Если имеется несколько вычислительных систем, для более полного использования ресурсов полезна единая система ведения очередей, которая имела бы возможность ставить задачи на разные вычислители в зависимости от их загрузки.

На вычислительных комплексах МГУ им. М. В. Ломоносова: Ломоносов, Bluegene и Regatta используются системы ведения очередей SLURM [2] и LoadLeveler [3] (на последних двух).

В разных системах ведения очередей имеется разный набор возможностей и одинаковые действия выполняются разными программами с разным набором параметров. Кроме различных систем управления очередями вычислительные комплексы могут иметь и другие особенности. Например, на Ломоносове используется программа Environment Modules [4], а на Regatta и Bluegene — нет. Вычислительные комплексы могут иметь различные политики хранения программ и пользовательских данных.

Modules позволяет иметь в системе несколько, возможно несовместимых программ, выполняющих схожие функции, и легко между ними переключаться при необходимости. Такими программами могут быть компиляторы различных производителей, несколько версий библиотек и т. п. Перед использованием программ, управляющихся системой модулей, нужно загрузить соответствующий этой программе модуль. Например, чтобы иметь возможность запустить программу, использующую MPI, на вычислительном комплексе Ломоносов, нужно загрузить модуль системы управления очередями SLURM и одной из реализаций MPI. Без этого необходимые команды не будут доступны.

На системе Regatta и Bluegene Modules не используется. Поэтому команды доступны сразу после входа в систему. Но здесь нужно знать, как называется необходимая команда с учётом префикса (суффикса) версии.

На Ломоносове используется специальная распределённая файловая система (Lustre). Только она доступна для вычислительных узлов. Поэтому перед выполнением программы все исполняемые файлы и данные необходимо скопировать на неё. Для доступа к ней пользователем, предусмотрена директория _scratch в домашней директории пользователя.

На Regatta вычислительному узлу доступна та же файловая система, что и интерфейсной машине. Поэтому перед запуском ничего копировать не надо.

Задачи пользователей могут принимать параметры как аргументы командной строки, стандартный поток ввода, файлы и переменные окружения. При чём одни агрументы могут зависеть от других. В простейшем случае аргумент командной строки задаёт путь к файлу. Аргуметны командной строки обычно имеют простой формат для разбора наборами функций getopt и argp_parse из стандарта POSIX [5] и библиотеки GNU C Library [6] соответственно. Но в некоторых случаях параметры могут иметь более сложную структуру, как, например, в наборе программ для молекулярной биологии EMBOSS [7].

Разрабатываемая система призвана облегчить постановку задач на выполнение и скрыть эти особенности от пользователя. Система состоит из сервера и адаптеров. Сервер устанавливается на выделенной машине. Он принимает имя программы, которую пользователь хочет поставить на выполнение, и параметры этой программы. Сервер проверяет их корректность по файлу описания программы, преобразует в универсальный формат и передаёт адаптеру через SSH-соединение. Адаптер устанавливается на вычислительном комплексе. Он преобразует описание задачи в универсалльном формате в команды для данного вычислительного комплекса и ставит задачу на выполнение.

Данная работа является дополнением работ [8] и [9] и преследует цель создания единой системы управления несколькими кластерами.

Литература

1. Сравнение систем пакетной обработки с точки зрения организации промышленного счёта, Баранов А. В., Киселёв А. В., Старичков В. В., Ионин Р. П., Ляховец Д. С., Научный сервис в сети Интернет: поиск новых решений: Труды Международной суперкомпьютерной конференции (17-22 сентября 2012 г., г. Новороссийск). Изд-во МГУ Москва. С. 506–508.

2. Сайт проекта SLURM — http://slurm.schedmd.com (дата обращения: 12.06.2016).

3. Документация LoadLeveler — https://www.ibm.com/support/knowledgecenter/SSFJTW/loadl_welcome.html (дата обращения: 12.06.2016).

4. Сайт проекта Environment Modules — http://modules.sourceforge.net/index.html (дата обращения: 12.06.2016).

5. Стандарт POSIX.2 (IEEE Std 1003.2-1992)

6. Документация GNU C Library, разбор опций программы с помощью argp — https://www.gnu.org/software/libc/manual/html_node/Argp.html (дата обращения 27.07.2016).

7. Сайт проекта EMBOSS, синтаксис командной — http://emboss.sourceforge.net/developers/acd/commandline.html (дата обращения: 12.06.2016).

8. Адаптация системы ведения задач пользователей SLURM с целью учёта лицензий ANSYS, А. Н. Сальников, В. Д. Никоноров, П. И. Каледа, Научный сервис в сети Интернет: многообразие суперкомпьютерных миров: Труды Международной суперкомпьютерной конференции (22-27 сентября 2014 г., г. Новороссийск). Изд-во МГУ Москва, 2014. С. 23–25.

9. Программная система для моделирования активности пользователей вычислительного кластера на основе системы ведения очередей SLURM, Сальников А. Н., Бойко А. Н., Параллельные вычислительные технологии (ПаВТ'2015): труды международной научной конференции (31 марта - 2 апреля 2015 г., г. Екатеринбург). Издательский центр ЮУрГУ Челябинск, 2015. С. 463–470.