Моделирование потока задач вычислительного кластера НИУ ВШЭ с использованием SLURM Simulator

Р.М. Мишенин, П.С. Костенецкий

Высшая школа экономики

Задача эффективного распределения ресурсов вычислительной системы широко известна. Она становится критически важной в многопользовательских многопроцессорных системах, таких как суперкомпьютеры. В НИУ ВШЭ функционирует высокопроизводительный вычислительный кластер «cHARISMa», состоящий из 48 вычислительных узлов шести типов. Основными характеристиками узлов являются наличие графических ускорителей и их модели, типы центральных процессоров и объем оперативной памяти. На суперкомпьютере используется планировщик задач SLURM версии 23.02.4.

При подборе параметров планировщика необходимо заранее определить, как они повлияют на поток задач вычислительного кластера. Одним из способов решения данной проблемы является моделирование вычислительного кластера с учетом его топологии и многообразия выполняемых задач. В настоящее время существует множество симуляторов высокопроизводительных кластеров, таких как AccaSim, GridSim, CloudSim, SimGrid. В результате сравнения было принято решение использовать SLURM Simulator [1].

Для проведения моделирования был подготовлен комплект конфигурационных файлов и трасс событий. В набор входят конфигурационные файлы SLURM, файл настроек менеджера учетных записей, файл, содержащий перечень аккаунтов и пользователей, зарегистрированных в системе управления заданиями. Каждое событие содержит временную метку постановки задания в очередь, количество запрошенных ресурсов, класс обслуживания и другие стандартные атрибуты задачи, необходимые для дальнейшего анализа и моделирования. В качестве источника данных для формирования файла с трассами событий был использован поток задач, выполненных за год на вычислительном кластере «cHARISMa».

Под модельным временем выполнения потока задач будем понимать время, необходимое для выполнения заданного набора задач на моделируемой конфигурации вычислительного комплекса. Задача сводится к минимизации модельного времени путем оптимизации конфигурации моделируемого суперкомпьютера. Ключевыми параметрами, существенно влияющими на работу алгоритма планирования, являются тип планировщика (builtin или backfill), а также весовые коэффициенты, варьирующиеся от 0 до 2 147 483 645 и определяющие приоритет задач. При помощи моделирования будет подобран оптимальный комплект вычислительного оборудования для апгрейда суперкомпьютера в 2026 г., который наиболее эффективно снизит время ожидания реальных вычислительных задач в очереди.

В качестве проверки модели было рассчитано влияние небольшого апгрейда, выполненного в 2024 г. В кластер были добавлены два вычислительные узла на базе GPU Nvidia H100. Среднее модельное время ожидания задач в очереди уменьшилось на 2.1 % относительно базовой конфигурации, что подтвердилось реальными статистическими значениями, полученных на суперкомпьютере – 1.95 %.

Дальнейшими задачами исследований является подстройка алгоритмов планировщика задач на моделируемой копии суперкомпьютера «cHARISMA». В результате моделирования будут определены изменения в параметрах конфигурации SLURM, внедрение которых в планировщик реального вычислительного кластера позволит снизить среднее время ожидания задач в очереди и дополнительно повысить эффективность использования суперкомпьютера.

Литература

1. Simakov N. et al. Slurm Simulator: Implementation and Parametric Analysis // High Performance Computing Systems. Performance Modeling, Benchmarking, and Simulation. PMBS 2017. Vol. 10724 / eds. by S. Jarvis, S. Wright, S. Hammond. Springer, Cham, 2017. P. 197–217. Lecture Notes in Computer Science. DOI: 10.1007/978-3-319-72971-8_10.

Моделирование потока задач вычислительного кластера НИУ ВШЭ с использованием SLURM Simulator

Р. М. Мишенин, П. С. Костенецкий
Национальный исследовательский университет «Высшая школа экономики»

Алгоритм работы с симулятором

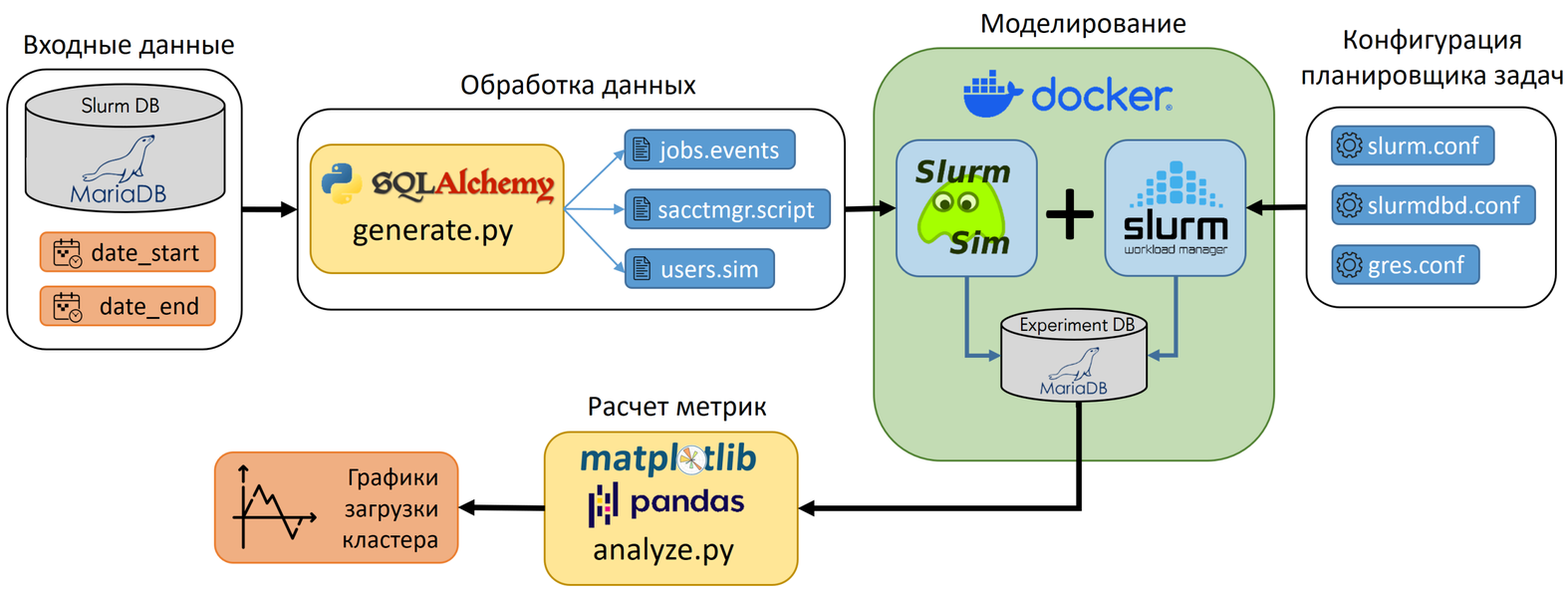

Рисунок 1. Схематическое представление процесса работы с симулятором

1. **Извлечение:** данные извлекаются из базы данных Slurm – планировщика задач, используемого на кластере cHARISMa НИУ ВШЭ.
2. **Обработка:** полученные данные обрабатываются с помощью скрипта generate.py. Создаётся 3 файла:
jobs.events: массив реальных вычислительных задач, которые предстоит моделировать в симуляторе ;
sacctmgr.script: список команд для утилиты saccmgr (создание аккаунтов и пользователей);
users.sim: список пользователей кластера
3. **Перенос:** в контейнер Docker, содержащий SLURM Simulator, загружаются файлы конфигурации SLURM и эксперимента.
4. **Запуск симуляции:** в результате создаётся база данных с результатами эксперимента.
5. **Анализ:** по результатам работы планировщика строятся графики среднего времени ожидания задач и загрузки CPU кластера.

Основные параметры планировщика задач SLURM

Параметр	Значение	Описание
SchedulerType	sched/backfill	Алгоритм планировщика – заполнение пустот
SchedulerParameters	pack_serial_at_end, bf_max_job_test=2000	Последовательные задачи ставятся в конец очереди
SelectType	select/cons_tres	Выбор с учетом Trackable Resources
SelectTypeParameters	CR_Core	Ресурсы выделяются на уровне ядер
PriorityDecay HalfLife	0	Приоритет задачи не снижается со временем
PriorityUsage ResetPeriod	None	Сброс истории приоритетов отключен
PriorityType	priority/multifactor	Многофакторная система приоритета заданий

Таблица 1. Ключевые параметры планировщика задач SLURM

Параметр	Значение	Описание
PriorityWeightFairshare	500	“Справедливость” – выше у тех, кто потреблял меньше ресурсов
PriorityWeightAge	500	Задания, дольше стоящие в очереди, получают больший приоритет
PriorityMaxAge	6-0	Возраст задания, до которого приоритет растёт
PriorityWeightPartition	1	Приоритет очереди – вес минимальный, почти не влияет
PriorityWeightJobSize	500	Масштабирует влияние размера задачи на приоритет
PriorityWeightQOS	500	Значимость класса обслуживания (QOS) в расчете приоритета
PriorityWeightTRES	CPU=500, GRES/gpu =2400	Приоритет в зависимости от выделяемых ресурсов: CPU и GPU

Таблица 2. Веса факторов, используемых для расчета приоритета задачи

Цель исследования — **выявить** такие параметры планировщика SLURM, которые позволяют **сократить среднее время ожидания** задач в очереди и **повысить загрузку суперкомпьютера**. Поскольку эксперименты на реальной системе сопряжены с рисками и ограничениями по времени, предлагается использовать **ускоренную симуляцию** потока задач. После **каждой итерации** симуляции оценивается, привело ли изменение **вышеописанных параметров** к улучшению ключевых метрик, что позволяет эффективно исследовать влияние **настроек планировщика задач SLURM** без риска для системы, находящейся в эксплуатации.

Представление вычислительной задачи в симуляторе

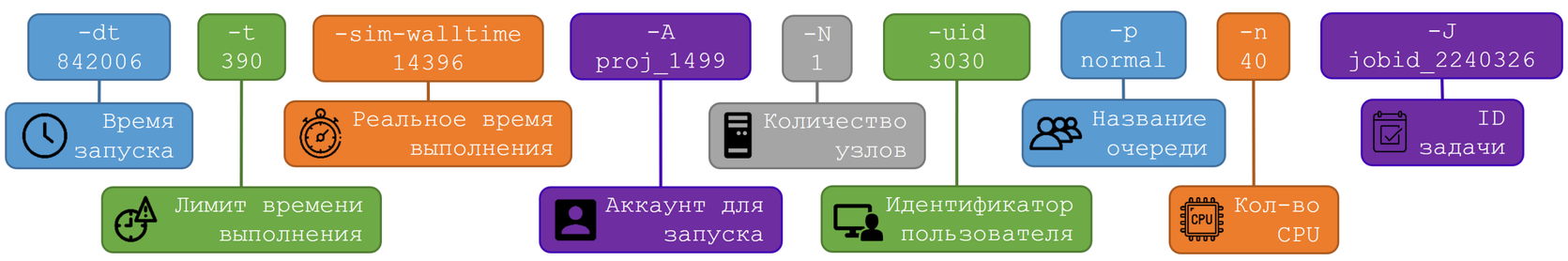

Рисунок 2. Пример представления вычислительной задачи в симуляторе

Основным объектом моделирования является **вычислительная задача**. Каждая строка соответствует **одной задаче** и содержит набор параметров, определяющих её характеристики. В отличие от стандартных параметров, используемых при запуске задач в SLURM, симулятор также оперирует такими атрибутами, как **фактическое время запуска** и **реальная длительность выполнения**.

Результаты

Рисунок 3. Загрузка кластера, полученная в результате моделирования

Рисунок 4. Загрузка кластера, полученная из исторических данных

Вычислительные эксперименты проводились на узле, включающем **48** процессоров **Intel Xeon Gold 5118 (2.30 GHz)** и **92 ГБ** оперативной памяти. Рисунки 3 и 4 иллюстрируют результаты моделирования на основе **7053** вычислительных задач, выполненных в **сентябре 2024** года на узлах **типа D** кластера cHARISMa (НИУ ВШЭ). Расчёты заняли **107** минут и **39** секунд. Эксперимент на этом наборе данных продемонстрировал наибольшее соответствие поведению реального планировщика задач SLURM на суперкомпьютере: **коэффициент корреляции Пирсона** составил **85,5%**, среднеквадратичная ошибка (**RMSE**) — **5,19%**, а средняя абсолютная ошибка (**MAE**) — **4,19%**.