СРАВНЕНИЕ КАЧЕСТВА ПЛАНИРОВАНИЯ ЗАДАНИЙ В СИСТЕМАХ ПАКЕТНОЙ ОБРАБОТКИ SLURM И СУППЗ

А.В. Баранов, Д.С. Ляховец

Системы пакетной обработки кластерных вычислительных систем

Любая современная кластерная вычислительная система (ВС), как правило, состоит из управляющей ЭВМ и решающего поля вычислителя. Вычислитель представляет собой совокупность вычислительных модулей (ВМ), объединённых некоторой коммуникационной средой.

Для обеспечения коллективного доступа к ресурсам кластерной вычислительной системы служат так называемые *системы пакетной обработки заданий* (СПО). Основными функциями СПО являются:

1. Приём входного потока параллельных заданий от разных пользователей.
2. Ведение очереди параллельных заданий.
3. Выделение на решающем поле вычислительных ресурсов, необходимых для выполнения параллельного задания, и их освобождение после окончания выполнения задания.

Компоненты СПО могут располагаться как на управляющей ЭВМ, так и на ВМ.

Во многих отечественных многопроцессорных вычислительных системах в качестве системы пакетной обработки более десятилетия используется Система управления прохождением параллельных заданий (СУППЗ) [1], созданная в Институте прикладной математики им. М.В. Келдыша РАН и Межведомственном суперкомпьютерном центре (МСЦ) РАН. Первая рабочая версия СУППЗ была развёрнута в 1999 году. За время эксплуатации и развития СУППЗ зарекомендовала себя надёжной и эффективной системой, способной круглосуточно обрабатывать поток пользовательских заданий в режиме промышленного счёта.

В последние годы на мировом рынке появился ряд СПО, создатели которых декларируют характеристики, сравнимые или превосходящие СУППЗ. Одна из таких СПО – Simple Linux Utility for Resource Management (SLURM, простая Linux-утилита для управления ресурсами). SLURM – отказоустойчивая и масштабируемая система управления кластером и диспетчеризации заданий для больших и малых Linux–кластеров [2]. Разработкой и поддержкой SLURM занимается международная группа разработчиков SchedMD LLC [3].

Несмотря на широкую распространённость SLURM, экспериментального сравнения этой системы с другими СПО до настоящего времени не проводилось. Работа по сравнению возможностей различных СПО была начата авторами в [4], настоящая работа является её логическим продолжением.

Особенности системы управления прохождением параллельных задач (СУППЗ)

Каждое параллельное задание, обрабатываемое в очереди СУППЗ, содержит следующую обязательную информацию:

- имя пользователя – владельца задания;
- число требуемых процессоров для выполнения задания;
- максимально допустимое время выполнения задания.

Пользователю СУППЗ предоставляет возможность контролировать своё задание на всех стадиях выполнения: постановки в очередь, наблюдения в очереди с прогнозом времени начала выполнения, удаления из очереди, просмотра стандартного вывода, снятия задания с выполнения.

Информация о поступивших в систему заданиях, времени их запуска и останова, пользователях и состоянии решающего поля вычислителя заносится СУППЗ в базу данных (БД) специально выделенной подсистемы «Статистика». Подсистема «Статистика» предоставляет возможность получения ряда типовых статистических отчётов о работе СУППЗ. Помимо стандартных отчётов, используя SQL-запросы, возможно напрямую получать информацию из БД.

СУППЗ ведёт очередь параллельных заданий с учётом требований полной загрузки вычислителя. В системе используется backfill-планировщик (т.н. сервер очередей), реализованный в СУППЗ на несколько лет раньше зарубежных аналогов (первые разработки относятся к 1994 году, существующий алгоритм планирования был реализован в 1999 году). Сервер очередей СУППЗ позволяет запускать задания, стоящие в очереди позже, если это не повлияет на время запуска заданий, стоящих в очереди ранее. Такое возможно, например, в случае, если для задания А, стоящего в очереди ранее, недостаточно ресурсов, и при этом задание Б, стоящее в очереди позже, успеет завершиться до момента, когда освободится достаточное количество ресурсов для запуска задания А.

В СУППЗ все задания пользователей делятся на три категории – *отладочные, пакетные* и *фоновые.*

Отладочные задания – это короткие по времени задания, которые запускаются исключительно в целях отладки на малом числе процессоров.

Пакетные задания – это средние по времени задания, которые производят реальные расчёты и выполняются, не прерываясь.

Фоновые задания – задания с большим временем счёта, которые могут прерываться системой. Для фонового задания пользователь должен явно указать квант – минимальное время счёта фонового задания, в течение которого задание прерывать нельзя.

Существование отладочных заданий и некоторого количества зарезервированных под них ВМ позволяет пользователям провести проверку работоспособности своего задания, не тратя длительное время в очереди. Фоновые задания эффективно используют время простоя ВМ, выполняются в течение длительного времени, не занимая при этом ресурсы ВС, если они нужны для расчёта других заданий.

В СУППЗ поддерживаются динамические приоритеты пользователей: существует понятие учётного периода, за который суммируется время выполненных пользователем заданий. От суммарного времени за учётный период напрямую зависит приоритет пользователя, что позволяет автоматически организовать саморегулирующуюся «справедливую» систему и делает невозможным захват решающего поля вычислителя одним пользователем на длительное время. Чем больше считал пользователь за учётный период, тем ниже его приоритет, при понижении приоритета задания пользователя начинают выполняться реже, суммарное время выполнения уменьшается, приоритет увеличивается – и так циклически повторяется процесс изменения приоритета.

Особенности системы пакетной обработки SLURM

Задания в СПО SLURM имеют аналогичные СУППЗ характеристики: имя пользователя-владельца, число процессоров и максимальное время выполнения задания. Так же, как и СУППЗ, СПО SLURM предоставляет пользователю все возможности для контроля прохождения задания в системе.

В SLURM возможно подключение разных планировщиков как внешних модулей. Для сравнения SLURM и СУППЗ в схожих условиях, здесь и далее будет рассматриваться backfill-планировщик с дополнением многофакторного приоритета (multifactor priority).

В SLURM задания не делятся на различные типы, но при этом существует деление вычислителя на секции. Возможно создание отладочной секции вычислителя (debug), на которой будут выполняться отладочные задания пользователей. Фоновые задания в СПО SLURM не поддерживаются.

Динамические приоритеты могут быть реализованы в SLURM при помощи подключаемого модуля Multi-factor Job Priority plugin. При его использовании приоритет задания зависит от пяти факторов, у каждого из которых есть вес, заданный администратором при настройке системы. Приоритет – взвешенная сумма факторов. Задания с высоким приоритетом запускаются раньше заданий с низким приоритетом. Одним из пяти факторов является динамический или справедливого распределения (fair-share) – ставит приоритет задания в зависимость от уже занятых пользователем ресурсов и тех ресурсов, что были ему выделены ранее; пользователь, активно использующий ресурсы, получает меньший приоритет, и его задания ставятся в конец очереди.

Показатели качества планирования заданий в СПО

Для сравнения СУППЗ и SLURM выделим следующие показатели качества планирования.

1. Утилизация ресурсов (загрузка вычислителя).

Под утилизацией ресурсов будем понимать отношение задействованного при выполнении заданий параллельного ресурса ко всему объёму параллельного ресурса. Если в течение 5 часов из 10 процессоров для выполнения заданий было использовано только 9, т.е. в течение каждого часа минимум один процессор простаивал в ожидании задания (рис. 1), то утилизация составит 80%.

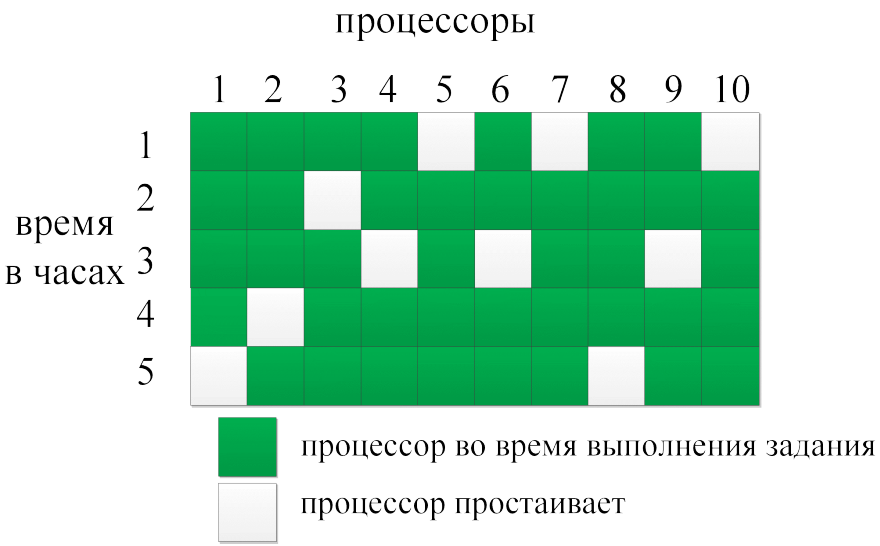

Рис. 1. Утилизация ресурсов вычислителя

2. Масштабируемость по количеству ВМ.

На время планирования влияет число ВМ решающего поля, распределяемое планировщиком между заданиями. У планировщика может существовать ограничение на число ВМ, при превышении которого неприемлемым становится либо время планирования, либо утилизация ресурсов.

3. Максимальные размеры планирования по количеству заданий.

Планировщик может иметь ограничения на число находящихся в очереди заданий. Последствия превышения этого числа заданий аналогичны последствиям превышения числа ВМ из пункта 3.

4. Среднее значение времени нахождение задания в очереди относительно заказанного времени её счёта.

Предположим, что Q_i – время, прошедшее с момента попадания *i*-го задания в очередь до начала его выполнения, R_i – заказанное время выполнения *i*-го задания. Величину *времени нахождения задания в очереди относительно заказанного времени его счёта*, обозначим как T_i:

$$T_i = \frac{Q_i}{R_i}$$

Тогда для *k* заданий можно определить величину T_{mid} – среднее значение времени нахождения задания в очереди относительно заказанного времени счёта:

$$T_{mid} = \frac{\sum_{i=1}^{k} T_i}{k} = \frac{1}{k} \cdot \sum_{i=1}^{k} \frac{Q_i}{R_i}$$

Приведение ко времени счёта осуществляется потому, что для разных заданий одно и то же время ожидания может означать разное качество планирования. Например, время ожидания 30 минут будет вполне удовлетворительным для задания с заказанным временем счёта 10 часов, но то же время ожидания (30 мин) будет неприемлемым для задания со счётом в 5 минут.

Очевидно, что чем меньше T_{mid}, тем выше качество планирования. Будем считать, что если справедливо условие $T_{mid} < 1$, то качество планирования является приемлемым, в противном случае – нет.

Определение показателей качества планирования СУППЗ

Утилизация ресурсов (загрузка вычислителя) за определённый период для СУППЗ вычисляется стандартными средствами (построение стандартного отчёта по утилизации ресурсов) подсистемы «Статистика». Определение среднего значения времени нахождения задания в очереди относительно заказанного времени счёта не входит в набор стандартных средств подсистемы «Статистика». Данный показатель качества может быть определен только путём выполнения SQL-запросов к БД.

Определение показателей качества планирования СУППЗ и SLURM производилось на статистических данных отечественной супер-ЭВМ МВС-100К, расположенной в МСЦ РАН. Кластер МВС-100К содержит в своём составе около 1 500 8-процессорных вычислительных модулей общей пиковой производительностью 100 TFlops.

Статистические данные по работе МВС-100К были взяты за период с 6 по 12 декабря 2012 года. В этот период под планирование заданий в системе было отведено 5 824 процессора (728 8-процессорных вычислительных модулей).

Стандартный отчёт подсистемы «Статистика» показал, что утилизация ресурсов за рассматриваемый период составила 97%.

Результаты SQL-запросов к БД «Статистика» показали, что за рассматриваемый период в системе было запущено 13 748 заданий, а среднее значение времени нахождения задания в очереди относительно заказанного времени счёта составило 0,23, т.е. задания ожидали в очереди в среднем 23% от заказанного времени счёта.

Максимальное число одновременно планируемых заданий в СУППЗ ограничено разработчиками системы и равняется 512. Задания, поступившие в систему сверх этого числа, принимаются СУППЗ, но на планирование не поступают, ожидая освобождения места в очереди.

Опыт эксплуатации СУППЗ свидетельствует о том, что система способна обслуживать решающее поле, состоящее из 1 500 ВМ с общим числом процессоров, равным 12 000. При этих значениях подсистема планирования достигает предела своих возможностей, и дальнейшее расширение решающего поля приводит к неприемлемому увеличению времени планирования.

Определение показателей качества планирования СПО SLURM

Под управлением СПО SLURM в мире работает значительное число кластерных ВС. Одна из них, с решающим полем в 2 500 ВМ, расположена в Суперкомпьютерном центре Барселоны (Barcelona SuperComputing Center, BSC, Испания). Множество вариантов настройки СПО SLURM (в т.ч. параметров планировщика) поставило перед инженерами BSC задачу определения оптимальных параметров, обеспечивающих наивысшее качество планирования для потоков задач, поступающих на вход ВС BSC.

Поскольку тестирование новых параметров планировщика на реально работающем кластере может негативно отразиться на качестве обслуживания текущих заданий, инженер BSC Алехандро Луцеро (Alejandro Lucero) разработал и реализовал следующий способ тестирования параметров планировщика [5]. Разработанное им программное средство (ПС) SLURM Simulator (симулятор SLURM) позволяет организовать работу СПО SLURM в режиме симуляции. В этом режиме реальное системное время заменяется на модельное, что позволяет ускорить проведение эксперимента.

Во время эксперимента реального выполнения заданий на вычислителе не производится. Для СПО SLURM достаточно, чтобы ВМ помечался как занятый на время выполнения пользовательского задания. Более того, в процессе эксперимента возможна эмуляция произвольного количества виртуальных ВМ в составе исследуемой кластерной ВС. При этом гарантируется совпадение результатов работы планировщика симулятора и реальной СПО SLURM.

Для определения показателей качества планирования СПО SLURM авторами был собран экспериментальный стенд SLURM Simulator с виртуальным вычислителем с параметрами, аналогичными кластеру МВС-100К (728 8-процессорных ВМ и 5 824 процессора).

Для корректного сравнения показателей качества планирования разных СПО необходимо обеспечить для них одинаковые входные потоки заданий. С этой целью из базы данных подсистемы «Статистика» СУППЗ МВС-100К была извлечена следующая статистическая информация за рассматриваемый период с 6 по 12 декабря 2012 года:

- имя пользователя, поставившего задание в очередь;
- время поступления задания в систему;
- число процессоров, требуемое для выполнения задания;
- время выполнения, заказанное пользователем;
- фактическое время выполнения задания (может быть меньше по сравнению с заказанным временем выполнения, т.е. задание может завершиться раньше запланированного времени).

На основе указанной информации был сформирован модельный поток заданий, поданный на вход SLURM Simulator на экспериментальном стенде.

При использовании SLURM Simulator А. Луцеро располагал журналами работы реальной кластерной ВС, по которым определялись параметры входного потока заданий. Авторами было модифицировано программное средство создания входного потока заданий для SLURM Simulator так, чтобы для задания параметров входного потока было возможно использовать информацию из БД «Статистика» СУППЗ.

В процессе вычислительного эксперимента SLURM Simulator сохраняет результаты симуляции (время поступления, запуска и останова заданий и др.) в собственную базу данных. После окончания симуляции показатели качества планирования могут быть рассчитаны с помощью SQL-запросов к БД SLURM Simulator.

При расчёте показателей качества планирования было необходимо учесть, что на момент начала симуляции виртуальный вычислитель SLURM, в отличие от реального вычислителя МВС-100К, не был загружен никакими заданиями, и первые выполненные задания будут запускаться на счёт без ожидания в очереди. Поскольку это напрямую влияет на показатели качества, анализ статистики был проведен не с момента запуска, а с момента заполнения виртуального вычислителя. Авторы анализировали статистику, начиная с третьего модельного дня работы стенда.

На вход SLURM Simulator был подан модельный поток из 17 392 фиктивных заданий. Первое поступившее задание по времени соответствовало заданию, поставленному в очередь СУППЗ МВС-100К 4 декабря 2012 года в 0 часов 0 минут и 1 секунду.

Показатели качества планирования для SLURM рассчитывались за тот же период времени, что и для СУППЗ, т.е. с 6 по 12 декабря 2012 года. За этот период СПО SLURM успела запустить 13 500 заданий против 13 748, выполненных СУППЗ.

Утилизация ресурсов в СПО SLURM составила 94% против 97%, достигнутых СУППЗ.

На официальном сайте SLURM [2] указано, что СПО SLURM может управлять кластерными ВС размером до 65 536 ВМ и обрабатывать до 120 000 заданий в час. Хотя заявленные показатели значительно превосходят возможности СУППЗ (1 500 ВМ и 512 заданий в очереди), проверить их достоверность нет возможности даже в режиме симуляции.

Среднее значение времени нахождение задания в очереди относительно заказанного времени счёта в СПО SLURM за рассматриваемый период составило 0,34 (задания ожидали в очереди в среднем 34% от заказанного времени счёта) против значения 0,23 у СУППЗ.

Выводы

Авторами статьи впервые была поставлена и решена задача экспериментального сравнения показателей качества планирования отечественной системы управления прохождением параллельных заданий (СУППЗ) и популярной системы пакетной обработки SLURM, созданной международным коллективом разработчиков. Результаты проведённого сравнения сведены в таблицу:

Таблица 1

Показатель эффективности	SLURM	СУППЗ
Утилизация ресурсов за исследуемый период	94%	97%
Среднее значение времени нахождение задания в очереди относительно заказанного времени счёта за исследуемый период	0,34	0,23
Число обслуживаемых вычислительных модулей	1 500	65 536
Максимальное число планируемых заданий	120 000	512
Число запущенных заданий за исследуемый период	13 500	13 748

Полученные результаты позволяют говорить о примерном паритете рассмотренных систем и одинаково высоком качестве планирования параллельных заданий в обеих системах.

ЛИТЕРАТУРА:

1. СУППЗ – Система управления прохождением параллельных заданий // http://suppz.jscc.ru
2. SLURM overview // https://computing.llnl.gov/linux/slurm/overview.html
3. SchedMD // http://www.schedmd.com
4. А.В. Баранов, А.В. Киселёв, В.В. Старичков, Р.П. Ионин, Д.С. Ляховец. Сравнение систем пакетной обработки с точки зрения организации промышленного счёта // Научный сервис в сети Интернет: поиск новых решений: Труды Международной суперкомпьютерной конференции (17-22 сентября 2012 г., г. Новороссийск). М.: Изд-во МГУ, 2012. С. 506.
5. SLURM Simulator // http://www.bsc.es/marenostrum-support-services/services/slurm-simulator