СРАВНЕНИЕ СИСТЕМ ПАКЕТНОЙ ОБРАБОТКИ С ТОЧКИ ЗРЕНИЯ ОРГАНИЗАЦИИ ПРОМЫШЛЕННОГО СЧЕТА

А.В. Баранов, А.В. Киселев, В.В. Старичков, Р.П. Ионин, Д.С. Ляховец

Функционирование любой современной массово-параллельной вычислительной системы (ВС) невозможно представить без системы пакетной обработки (СПО). Предназначение любой СПО – обеспечение коллективного доступа к ресурсам массово-параллельной ВС. СПО выполняет следующие основные функции:

- прием входного потока параллельных заданий от разных пользователей;
- ведение очереди параллельных заданий;
- выделение ресурсов массово-параллельной ВС для параллельного задания и освобождение занятых ресурсов после выполнения задания.

Очень часто массово-параллельная ВС функционирует в режиме, который можно назвать режимом **промышленного счета** со следующими отличительными особенностями.

1. **Круглосуточная доступность**. Вычислительная система обслуживает пользовательские задачи 24 часа в сутки и 7 дней в неделю, за исключением вынужденных простоев и плановой профилактики.
2. **Универсальность.** На вид и характер пользовательских прикладных программ не накладывается никаких ограничений. Зарегистрированный в системе пользователь с помощью доступных инструментальных средств имеет право и возможности разработки и выполнения любой программы.
3. **Автоматическое обслуживание**. Функционирование системы в штатном режиме не предусматривает ручного вмешательства оператора или администратора.

Более 10 лет в Межведомственном суперкомпьютерном центре РАН (супер-ЭВМ МВС-100К), ИПМ им. М.В.Келдыша РАН (супер-ЭВМ К-100) и ряде других организаций в качестве СПО используется Система управлением прохождением параллельных задач (СУППЗ) [1]. Изначально созданная как круглосуточно функционирующая система промышленного счета, СУППЗ затем неоднократно модифицировалась и совершенствовалась.

В последние годы активно развиваются такие системы пакетной обработки, как OpenPBS/Torque, LoadLeveler, Cleo, SLURM, MS HPC Server и другие. Целью работы авторов является сравнение современных СПО с СУППЗ с точки зрения организации промышленного счета. Настоящий доклад посвящен начальному этапу работы, на котором авторами были рассмотрены сравнительно недавно появившиеся и набирающие популярность СПО SLURM [2] и Microsoft HPC Server 2008 [3].

За долгие годы эксплуатации СУППЗ показала себя надежной и стабильной системой со сложившимся определенным порядком работы пользователей и администраторов, к которому и те, и другие привыкли. Принципы работы и характерные свойства СУППЗ послужили авторам основой для определения критериев сравнения. Накопленный авторами многолетний практичекий опыт организации промышленного счета позволяет выдвинуть к сравниваемым СПО как обязательные, так и желательные требования. Обязательные требования определяют саму возможность реализации режима промышленного счета, а желательные – его эффективность. Рассмотрим обязательные требования.

1. **Гарантированное обслуживание без потерь**. СПО должна гарантировать, что поступившее на ее вход задание не будет потеряно (уничтожено) и рано или поздно будет выполнено. Требование гарантированного обслуживания выполняется всеми рассматриваемыми СПО.
2. **Изоляция пользовательских заданий**, исключающая взаимное влияние процессов разных заданий друг на друга. Нарушение изоляции заданий неизбежно повлечет за собой деградацию производительности, приведет к блокировкам и конфликтам заданий за доступ к ресурсам. Изоляция пользовательских заданий подразумевает соблюдение следующих практических принципов:

- предоставления пользователю доступа только к выделенным для его задания вычислительным ресурсам и только на время выполнения задания; в остальное время или к другим ресурсам доступ пользователю должен быть запрещен;
- неделимости вычислительного модуля массово-параллельной ВС; модуль, выделенный для выполнения одного задания, не может быть одновременно предоставлен еще какому-либо другому заданию, даже если на модуле остаются свободные простаивающие процессоры или другие ресурсы;
- принудительной очистки от процессов пользователя каждого вычислительного модуля после завершения параллельного задания; если на модуле после очистки все еще остаются пользовательские процессы, модуль должен быть помечен как неисправный и выведен из состава решающего поля (заблокирован).

Принцип изоляции заданий был заложен в СУППЗ при ее создании в 1999 году и с тех пор неукоснительно соблюдается. Все три составляющих изоляции заданий – предоставление доступа только к выделенным задания ресурсам, неделимость вычислительного модуля и принудительная очистка ВМ после окончания задания – реализованы как в SLURM, так и в MS HPC Server.

SLURM сам по себе не ограничивает доступ к вычислительным модулям, однако он предоставляет механизм для такого рода ограничений. Для этого используется Pluggable Authentication Module (PAM), который запрещает доступ к ВМ всем пользователям, кроме тех, которым этот ВМ был выделен для счёта их заданий. SLURM может распределять вычислительные модули, процессоры, сокеты, ядра или память в зависимости от настроек, выставленных администратором. При этом по умолчанию вычислительный модуль, также как и в СУППЗ, является неделимым ресурсом. Однако возможен режим, при котором пользователь может позволить разделять выделенные ему ВМ с другими заданиями.

Аналогичный режим можно реализовать в MS HPC Server, указав в настройках задания свойство – эксклюзивное использование вычислительного модуля (exclusive) и выбрав в настройках задания распределяемый ресурс – процессоры (cores). Задание будет запущено на ВМ, на которых больше никакое другое задание считаться не будет, на количестве процессоров, которое будет указано пользователем при запуске.

Как SLURM, так и MS HPC Server производят качественную очистку (уничтожаются любые процессы пользователя, даже запущенные в режиме демона или службы) вычислительных модулей после окончания выполнения задания.

Следующие желательные требования отражают сложившийся порядок работы пользователей и адинистраторов СУППЗ, практика эксплуатации показала, что их соблюдение позволяет упростить организацию и повысить эффективность промышленного счета.

1. **Автоматическая организация режимов профилактики в строго определенное время.** Планировщик СУППЗ имеет возможность обеспечить отсутствие выполняющихся заданий на решающем поле к заданному времени. При этом продолжает вестись очередь заданий, в которую пользователи могут добавлять новые задания. Указанное свойство очень удобно для организации регулярных профилактических работ по заранее заданному расписанию.

В SLURM такая возможность не предусмотрена. Частичное решение задачи организации профилактики состоит в загрузке исполняемого от имени суперпользователя командного файла на нужный ВМ. Командный файл сработает после завершения пользовательского задания и заблокирует вычислительный модуль, сделав его доступным для проведения профилактики. Здесь серьёзным недостатком является зависимость от времени завершения пользовательского задания.

MS HPC Server предоставляет, как минимум, два пути реализации профилактики: выведение вычислительного модуля из счета путем установления его в состояние «выведен из счета» (offline) и возможность запуска тестовых заданий в установленное администратором время наравне с остальными заданиями. При выведении из счета ВМ переходит в состояние, когда он больше не принимает задания, но завершает текущие на нем. После этого он переходит в состояние offline и становится готовым к профилактике. Посредством планировщика заданий операционной системы и командных файлов возможно организовать автоматическое освобождение ВМ от заданий, но гарантировать, что к определенному времени все ВМ будут свободны без потери результатов заданий (т.е. их принудительного завершения) нельзя.

2. **Разделение заданий на отладочные, пакетные и фоновые,** поддерживаемое СУППЗ. Пакетные задания составляют основной поток промышленного счета, выполняются в течение длительного, хотя и ограниченного времени и могут использовать произвольное число процессоров. Характерной особенностью пакетного задания является его однократное выполнение в системе. Отладочные задания характеризуются как малым временем выполнения, так и малым числом используемых процессоров. Скорость прохождения через очередь отладочных заданий значительно выше остальных, что дает их пользователям возможность быстрой отладки своих программ непосредственно на решающем поле ВС. Фоновые задания выполняются произвольное время и на произвольном числе процессоров, но могут многократно прерываться системой с возвратом в очередь. Для пользователя фоновые задачи привлекательны возможностью организации длительных расчетов, администратору они полезны, поскольку позволяют поднять загрузку решающего поля ВС.

Как SLURM, так и MS HPC Server поддерживают различные типы пользовательских заданий, но ни та, ни другая СПО не имеет возможности планирования фоновых задач, что с точки зрения организации длительных промышленных расчетов является несомненным недостатком.

3. **Динамические приоритеты пользователей**. В СУППЗ есть понятие учетного периода, за который суммируется время выполненных пользователем заданий. От суммарного времени за учетный период напрямую зависит приоритет пользователя, что позволяет автоматически организовать саморегулирующуюся «справедливую» систему и делает невозможным захват решающего поля одним пользователем на длительное время. Чем больше считал пользователь за учетный 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.

В SLURM существует подключаемый модуль 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.

В MS HPC Server задания обладают приоритетами, устанавливаемыми в так называемых шаблонах заданий. Шаблоны заданий создаются администратором ВС. Пользователям и группам пользователей предоставляется доступ к использованию тех или иных шаблонов, но возможно предоставление прав на создание своих шаблонов или изменение существующих. Статистика времени счета конкретного пользователя не ведется, понятия учетного периода нет. Возможности организовать эквивалент динамического приоритета пользователя, который есть в СУППЗ и SLURM, нет.

4. **Простота работы с MPI-программами.** В СУППЗ можно выделить три главные части – ядро (планировщик), служебные процессы, выполняющиеся с правами администратора и надстройки, выполняющиеся от имени пользователя. Одной из надстроек, изначально входившей в состав СУППЗ, является поддержка MPI-программ – СУППЗ имеет собственную команду mpirun, осуществляющую автоматическое оформление для MPI-программы паспорта задания и постановку в очередь на выполнение. Хотя возможно включение в состав СУППЗ любых других надстроек, именно MPI-надстройка оказалась наиболее востребованной пользователями и стала характерной особенностью СУППЗ.

SLURM, как и СУППЗ, поддерживает различные реализации MPI, что предоставляет свободу выбора, но при этом усложняет настройку системы администратору. С пользовательской точки зрения, в зависимости от выбранного варианта реализации MPI, выделяют три режима работы: когда пользователь командой srun запускает свое задание, не заботясь об работе с MPI (её курирует SLURM), когда пользователь запрашивает (salloc) ресурсы, и на выделенных ресурсах командой mpirun запускает свои задания, и когда пользователь выделяет ресурсы, сам запускает задания, при этом задания используют сторонние способы обмена сообщениями, нежели предоставляемые SLURM. SLURM предоставляет обширную свободу действий, однако в СУППЗ работа с MPI выглядит проще и понятнее.

В MS HPC Server 2008 имеется собственная спецификация MS MPI, основанная на MPICH2 и являющаяся совместимой с ним, за исключением отличных от MPICH2 загрузчика, системы управления заданиями, отсутствия возможности динамического создания процессов. Эта реализация входит в набор программ MS HPC Server 2008 и поставляется в виде SDK. Компиляцию программ можно производить в MS Visual Studio при включении дополнительной библиотеки из SDK, запуск производится посредством специальной команды mpiexec, что является несколько непривычным для опытного пользователя, обычно использующего для запуска команду mpirun.

Подводя итог сравнительному анализу, можно сделать вывод, что как в SLURM, так и в MS HPC Server 2008, в полной мере реализованы главные принципы промышленного счета – гарантированное выполнение и надежная изоляция пользовательских заданий. Все рассмотренные системы предоставляют достаточно удобный и простой порядок работы пользователя с MPI-программами. Наиболее серьезным недостатком, снижающим эффективность использования SLURM и MS HPC Server 2008 в режиме промышленного счета, является невозможность организации профилактики в строго определенное время. При этом стоит заметить, что указанный недостаток существенен далеко не для всех суперкомпьютерных центров обработки данных.

К преимуществам SLURM и СУППЗ следует отнести развитую поддержку динамических приоритетов, отсутствующую в MS HPC Server 2008. SLURM и MS HPC Server 2008 проигрывают СУППЗ, не предоставляя возможности организации фоновых заданий.

ЛИТЕРАТУРА:

1. Баранов А.В., Смирнов С.В., Храмцов М.Ю., Шарф С.В. Модернизация СУПЗ МВС-1000 // Материалы Всероссийской научной конференции «Научный сервис в сети Интернет». Новороссийск, 2008. http://agora.guru.ru/abrau2008/pdf/099.pdf
2. Simple Linux Utility for Resource Management // https://computing.llnl.gov/linux/slurm
3. Microsoft Windows High Performance Computing (HPC) Server // http://www.microsoft.com/ru-ru/server-cloud/windows-server/high-performance-computing-hpc.aspx