Эффективный запуск гибридных параллельных задач в гриде*

А.П. Крюков[1,2,+], М.М. Степанова[3], Н.В.Приходько[4], Л.В. Шамардин[1], А.П. Демичев[1,2]

[1]Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына Московского государственного университета имени М.В. Ломоносова,
[2]Национальный исследовательский центр «Курчатовский институт»,
[3]Санк-Петербургский государственный университет,
[4]Новгородский государственный университет имени Ярослава Мудрого

ПАВТ-2013. Челябинск, 4 апреля, 2013г.

*) Работа поддержана грантом РФФИ (№ 11-07-00434-а) и грантом Президента РФ (НШ-3920.2012.2)

+) E-mail: kryukov@theory.sinp.msu.ru

Введение

- Современные прикладные и фундаментальные исследования требуют выполнения высокопроизводительных расчетов, использующих параллельные вычисления.
- Для получения максимальной производительности и скорости исполнения программ в режиме пакетной обработки необходимо запускать такие программы оптимальным образом с учетом особенностей конкретного кластера.
- При непосредственной работе на кластере пользователю хорошо известна его архитектура и конфигурация, поэтому он может точно указать все параметры ресурсов и запуска задачи.
- В гриде среда является гетерогенной, а ресурс, на котором будет выполняться задание, в общем случае, заранее неизвестен.
- Данное исследование направлено на разработку для инфраструктуры ГридННС универсального способа запуска сложных параллельных задач. Предлагаемое решение позволяет использовать грид-среду для эффективного запуска не только распространенных вариантов на базе MPI- и OpenMP-технологий, но и наиболее сложного комбинированного типа - гибридных (MPI+OpenMP)-задач.

Запуск заданий в гриде

- Процесс запуска в гриде - это многоуровневая процедура, в которой участвует большое количество сервисов и компонентов промежуточного ПО.
- Описание задания для запуска в гриде имеет абстрактный вид и не зависит от типа ресурса.
- Конечным итогом прохождения задания всех грид-сервисов является формирование скрипта задачи, который будет запущен конкретным локальным менеджером (ЛРМ) на конкретном ресурсе.
- Чтобы ЛРМ мог выполнить корректный запуск, ожидаемый пользователем, описание задачи должно быть исчерпывающим, а алгоритмы его обработки на всех уровнях очень тщательно проработаны.

Запуск заданий в гриде

- Специфика запуска заданий в гриде по сравнению с обычным кластером:
  - возможности формата описания заданий;
  - полнота публикуемой информации и гибкость алгоритма поиска ресурса;
  - различия в типах и конфигурации ЛРМ на сайте;
  - реализация интерфейса к ЛРМ;
  - требования к способу запуска и установке окружения для разных типов задач.

Запуск заданий в гриде

- Все существующие реализации гридов ограничены фиксированным набором типов заданий, где параметры запуска автоматически рассчитываются из требований к ресурсам.
- Такой подход привлекает простотой с точки зрения описания заданий, но годится лишь под ограниченный круг задач. В частности, это относится к разработкам, в основе которых лежит ПО GlobusToolkit, в том числе и к российскому проекту ГридННС [1].
- Другие известные проекты без дополнительной конфигурации также пока обеспечивают лишь базовую функциональность для простейших MPI- и OpenMP-задач.

Запуск заданий в гриде

- Самым законченным решением на сегодняшний день следует признать проект UNICORE. Характерная особенность – четкое разделение функциональности:
  - параметры описания из раздела Resources полностью определяют резервирование,
  - механизм программных окружений ExecutionEnv позволяет точно задать параметры и опции запуска задачи.
- Более подробно методы запуска задач в различных гридах можно посмотреть в работе [2].

Особенности запуска гибридных задач

- Классические параллельные приложения реализуются на основе технологий MPI или OpenMP. Вариант на базе MPI хорошо зарекомендовал себя для работы на традиционных кластерных системах. OpenMP предназначен для распараллеливания на многоядерных узлах с общей памятью.
- Самым сложным из параллельных типов являются гибридные задачи, которые исполняются на SMP-кластерах. Гибридный подход предполагает, что алгоритм разбивается на параллельные процессы, каждый из которых сам является многопоточным.

Особенности запуска гибридных задач

- За счет укрупнения MPI-процессов и уменьшения их числа гибридная модель может устранить ряд недостатков MPI, таких как большие накладные расходы на передачу сообщений и слабая масштабируемость при увеличении числа процессов [3-7].
- Однако, производительность гибридной задачи очень сильно зависит от режима ее запуска и выполнения, который определяет соотношение числа MPI-процессов и OpenMP-потоков на одном вычислительном узле, а также способа привязки MPI-процессов к физическим процессорам системы.
- В случае неправильного запуска или некорректного выделения ресурсов производительность таких программ может существенно снижаться.

Особенности запуска гибридных задач

- Требования к ресурсам и, соответственно, оптимальный способ запуска гибридной задачи в значительной степени определяется ее кодом. Большинство программ разрабатываются без привязки к архитектуре кластера, на котором она будет компилироваться и выполняться. Для их нормального выполнения достаточно задать полное количество MPI-процессов (M) и количество потоков на один процесс (K), а также обеспечить монопольное резервирование на время выполнения MxK процессоров/ядер.

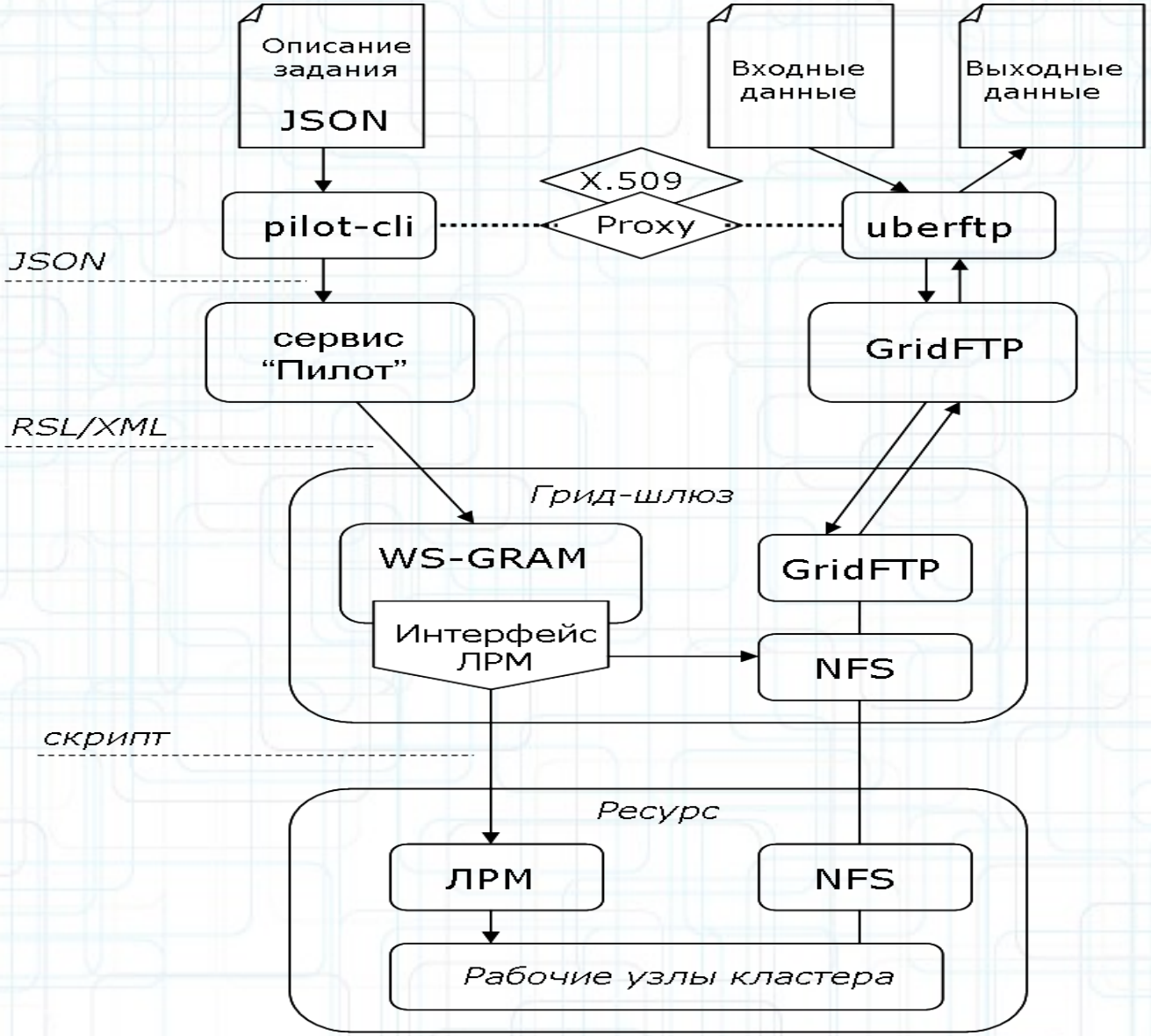
Особенности запуска гибридных задач

- Стандартный способ организации многопользовательского доступа к вычислительным кластерам – использование системы управления пакетной обработкой заданий (например, Torque, SLURM, PBSPro и др.)
- Для запуска параллельных и особенно гибридных задач необходима поддержка локальным менеджером корректного выделения и резервирования вычислительных ресурсов на время выполнения задачи.
- В частности, должно обеспечиваться управление динамическим распределением задач в больших системах и строгое ограничение процессов на подмножестве процессоров и памяти узла.
- Один из возможных механизмов такого типа – cpuset, реализованный на уровне ядра ОС Linux и доступный в Torque.

Методы запуска гибридных (MPI+OpenMP) задач на вычислительном ресурсе

	Метод запуска	Резервирование ресурсов (cpu) и параметры запуска
1	Запуск в режиме один MPI-процесс на узел при полном резервировании узла	#PBS –l nodes=N:ppn=K, где K=SMPSize; $MPIEXEC –pernode ./hyb_task
2	Запуск в режиме на один узел M MPI-процессов, каждый с числом потоков K	#PBS –l nodes=N:ppn=K, export OMP_NUM_THREADS=L, где M*L=K $MPIEXEC –npernode M ./hyb_task
3	Запуск в режиме отображения на ресурсы, задаваемого строкой опций	#PBS –l nodes=N:ppn=K, export OMP_NUM_THREADS=L, где M*L=K $MPIEXEC[special_option_string] ./hyb_task

Схема прохождения задания в ГридННС



Запуск параллельных задач в ГридННС

- Исходная реализация ГридННС обеспечивает запуск задач только двух типов:
  - single: простая задача с последовательным кодом; запускается на одном ядре (count=1);
  - mpi: параллельная MPI-задача; запускается в режиме один MPI-процесс на одно ядро (count>1);
  - для резервирования ресурса и запуска пользователю доступен один параметр описания – count, который определяет количество запрашиваемых ядер и задает число запускаемых процессов.
  - При формировании скрипта запуска для ЛРМ на грид-шлюзе выполнялась автоматическая подстановка способа запуска. Так в случае mpi - посредством $MPIEXEC –n count <executable>.

Типы задач

- Для адекватной работы большинства пользовательских приложений и специализированных пакетов можно выделить следующий набор базовых типов задач, которые должна обеспечивать современная грид-инфраструктура:
  - single - запуск задачи с последовательным кодом на одном ядре узла вычислительного кластера.
  - openmp - запуск многопоточной (например, использующей технологию OpenMP) задачи на одном узле с возможностью резервирования под задачу узла либо целиком, либо частично – резервирование запрошенного числа ядер на узле для монопольного использования задачей.
  - mpi - запуск MPI-задачи с одним процессом на ядро без каких-либо требований к распределению ядер по узлам кластера.
  - hybrid - запуск гибридной (MPI+OpenMP) задачи в режиме один многопоточный MPI- процесс на один узел.

Адаптация ГридННС к запуску гибридных параллельных задач

- Синтаксис языка описания задания в формате JSON - достаточно гибкий, и его расширение новыми атрибутами не вызывает затруднений. Однако, следует учитывать, что в дальнейшем это описание транслируется в RSL для обработки на грид-шлюзе сервисом WS-GRAM, а спецификация RSL допускает это только через механизм extensions.
- Изменение алгоритма подбора ресурсов сервисом распределения нагрузкой — Пилот.
- При выборе ресурса необходима исчерпывающая информация о текущем состоянии сайтов, которая предоставляется по средством информационного сервиса, использующий модифицированный вариант XML-реализации схемы GLUE 1.3.
- Модификация алгоритма обработчика заданий для интерфейсов ЛРМ.

Расширение языка описания задач

```
Секция definition:
  jobtype [single|openmp|mpi|hybrid]
  nodes [int]
  ppn [int]

Секция extensions:
  mpi_extra_arg [string]
```

Поддержка указанных расширений реализована:

- в сервисе Пилот;
- программах интерфейсов к ЛРМ на шлюзе;
- на клиенте.

Типы задач, поддерживаемые в ГридННС.

Тип задач	Обязательные параметры в описании задачи	Дополнительные (необязательные) параметры в описании задачи
single	Нет	Нет
openmp	`jobtype="openmp"` `ppn` или `count`	`environment:` `{"OMP_NUM_THREADS": <number>}`
mpi	`jobtype="mpi"` `count` или пара `{ppn,nodes}`	Нет

hybrid	`jobtype="hybrid"`
	любая пара из
	`{ppn,nodes,count}`
	*/ рекомендуется `{ppn,nodes}`
		`mpi_extra_args: <string>`
		`environment:`
		`{"OMP_NUM_THREADS": <number>}`

Адаптация ПО «Пилота» для запуска гибридных задач

- В схемы описания задач JSON Schema были добавлены определения новых полей ppn и nodes. Поскольку JSON Schema не позволяет задать сложные отношения между атрибутами, были также расширены правила верификации задач используемые загрузчиком описаний задач "Пилот".
- Были модифицированы модули генерации описания задач и выбора подходящих ресурсов на стороне сервера "Пилот". Поскольку RSL, используемый Globus Toolkit 4, не поддерживает атрибутов задач, логически соответствующих параметрам “ppn” и “nodes”, все атрибуты, связанные с запуском гибридных параллельных задач передаются через расширения в описании задачи.
- В модуле выбора подходящих ресурсов были внесены изменения, накладывающие ряд ограничений при выборе подходящих ресурсов. Например, количество процессов, запускаемых на одном узле (“ppn”) не должно превышать количества логических процессоров на узле (“logical slots” в информационной системе).

Взаимодействие грид-шлюза с менеджером локальных ресурсов

- Взаимодействия веб-сервиса запуска с локальным менеджером ресурсов организовано через грид шлюз, который использует GRAM ПО Globus Toolkit.
- При формировании задачи для ЛМР на стороне шлюза реализована поддержка двух вариантов обработки параметра ppn для целей резервирования:
  - резервирование части узла (в случае ЛРМ с поддержкой cpuset);
  - резервирование узлов полностью (в случае ЛРМ без поддержки cpuset).
- Способ обработки ppn задается в конфигурационном файле jobmanager-pbs.conf через параметр use_full_node (значения no или yes, по умолчанию yes).
- С целью поддержки дополнительных опций mpiexec также добавлена поддержка дополнительного параметра mpi_extra_args, в которой передается строка альтернативных аргументов для mpiexec.

Заключение

- В ходе выполнения работы были систематизированы и формализованы методы запуска гибридных задач на вычислительном ресурсе, определены необходимые расширения спецификации языка описания заданий набором атрибутов.
- Были разработаны спецификации и алгоритмы обработки новых атрибутов компонентами грид-сервисов на всех уровнях и выполнена программная реализация этих алгоритмов для грид-сервисов в среде ГридННС.
- Проведенные тестовые испытания на полигоне ГридННС показали, что предложенные методы обеспечивают корректное резервирование ресурсов на вычислительном кластере в соответствии с типом запускаемой задачи.
- Полученные результаты позволяют повысить эффективность использования суперкомпьютеров, подключенных к гриду, за счет учета специфики параллельных задач.

Литература

- 1. В.А. Ильин, В.В. Кореньков, А.П. Крюков. ГридННС: состояние и перспективы // Труды 5-й международной конференции "Распределенные вычисления и Грид-технологии в науке и образовании" (Дубна, 16-21 июля, 2012 г.).-Дубна: ОИЯИ, с. 332-336
- 2. M.M. Stepanova, O.L. Stesik. Running Parallel Jobs on the Grid // Distributed Computing and Grid-Technologies in Science and Education: Proceedings of the 5rd Intern.Conf. (Dubna, 16-21 July, 2012).– Dubna: JINR, 2012, pp.383-387, ISBN -5-9530-0345-2
- 3. Makris I. Mixed Mode Programming on Clustered SMP Systems // MSc in. High Performance Computing. The University Of Edinburgh, 2005.
- 4. R. Rabenseifner, G. Hager, and G. Jost. Hybrid MPI/OpenMP parallel programming on clusters of multi-core SMP nodes // In Proc. of 17th Euromicro Int'l Conference on Parallel, Distributed, and Network-Based Processing (PDP 2009), pages 427–236, 2009.
- 5. A. Rane and D. Stanzione. Experiences in tuning performance of hybrid MPI/OpenMP applications on quad-core systems // In Proc. of 10th LCI Int'l Conference on High-Performance Clustered Computing, 2009.
- 6. Глазкова Е.А., Попова Н.Н.Анализ эффективности гибридного параллельного программирования на примере системы BLUE GENE/P 2009 // URL: http://agora.guru.ru/abrau2009/pdf/36_NSSI_2009_Abrau-2009.pdf
- 7. MPI Forum Hybrid Programming Working Group (Lead: Pavan Balaji) // URL: http://meetings.mpi-forum.org/mpi3.0 hybrid.php (July 2010).

Спасибо за внимание!

Вопросы?