Вычислительная химия на российских грид-полигонах: текущее состояние, проблемы и перспективы *

В.М. Волохов[1], Д.А. Варламов[1,2], А.В. Волохов[1],
А.В. Пивушков[1], Г.А. Покатович[1], А.И. Прохоров[1]

Институт проблем химической физики РАН[1],
Институт экспериментальной минералогии РАН[2]

Проводится анализ текущего состояния вычислительных грид-сервисов в области квантовой химии и молекулярной динамики, реализованных в рамках существующих российских грид-полигонов. Кратко описаны принципы адаптации и функционирования основных прикладных программных пакетов (ППП) вычислительной химии. Сформулированы основные проблемы по работе с подобными сервисами в условиях российских грид-полигонов и возможные пути их решения и перспективы развития проблемно-ориентированных грид-сервисов.

1. Введение

Вычислительная химия и сопряженные с ней области знаний являются одними из наиболее заинтересованных в суперкомпьютерных и распределенных грид-вычислениях (в том числе и на входящих в состав грид-полигонов суперкомпьютерах) отраслями науки. Исследования, проводимые в области химии и смежных наук, в настоящее время, как правило, неэффективны без использования сверхмощных параллельных и распределенных вычислительных ресурсов для решения задач самых разных классов. Первыми научными задачами, использовавшими реальную петафлопсную производительность суперкомпьютеров, стали задачи вычислительной химии: расчеты электронный структуры высокотемпературных сверхпроводников и исследования эффекта магнитосопротивления в наночастицах методом Монте-Карло на суперкомпьютере “Jaguar” (Oak Ridge, USA – 1,64 Пф). В качестве ярких примеров можно привести ресурсные требования к достаточно тривиальным (на первый взгляд) химическим задачам: (а) исследования свойств воды в низкоразмерных системах – до $8 \cdot 10^6$ CPU-часов в Argonne National Laboratory; (б) исследования химических катализаторов в области нефтехимии – до $30 \cdot 10^6$ CPU-часов в год (“Jaguar”). Потенциально же задачи в области молекулярного моделирования и многоиерархического моделирования материальных объектов от квантово-механического уровня до уровня сплошных сред и конструкций могут выходить на уровень востребованности многих петафлопс. Например, некоторые задачи оптимизации крупных молекулярных структур требуют выполнения до 10^9 отдельных расчетов. Типичный докинг белковых лигандов с размерностью 200 атомов × 300 000 конфигураций × 1000 CPU-часов требует до 300 Пф вычислительных мощностей. Для построения многомерных потенциальных поверхностей, адекватно описывающих химические реакции, нужно провести $10^2 – 10^N$ независимых ресурсоемких *ab initio* расчетов. Для исследования динамики химической реакции с использованием классических траекторий зачастую нужно рассчитать до 10^7-10^9 независимых траекторий. Численное исследование многопараметрических функций $F(x_1,x_2…x_n)$ в области изменения параметров $x_1,x_2…x_n$ при разбиении диапазона изменения каждого параметра на 10 ячеек требует проведения 10^N независимых расчетов F.

Востребованность вычислительной химией все возрастающих вычислительных ресурсов подтверждается тем, что большинство суперкомпьютерных центров США (San Diego, Ohio, Illinois *etc.*) предоставляют до 40% вычислительных мощностей для нужд биохимии, молекулярного моделирования, квантовой химии, нанотехнологических расчетов. Создано достаточно много проблемно-ориентированных суперкомпьютерных центров, специализирующихся почти

* Исследовательские работы финансово поддерживаются Государственным контрактом (заявка №2012-1.1-12-000-2003-034, соглашение от 10.07.2012) в рамках ФЦП «Научные и научно-педагогические кадры инновационной России на 2009-2013 годы»

исключительно на квантово-химических и молекулярно-динамических расчетах: The UK National Service for Computational Chemistry Software (Великобритания, http://www.nsccs.ac.uk) – все виды вычислений; The National Resource for Biomedical and Chemistry Supercomputing (США, http://www.nrbsc.org) расчет молекулярных систем от 20000 до 120000 атомов, до 10^5 CPU-часов на структуру; Chemical Computing Group (http://www.chemcomp.com), Канада; Lawrence Berkeley National Laboratory (США, http://www.lbl.gov/csd), сегмент вычислительной химии до 450 Tflops; Lehrstuhl für Theoretische Chemie der Technische Universität München, Германия (http://www.lrz.de/services/software/chemie); Swiss National Supercomputing Centre (Швейцария, http://www.cscs.ch) – выделение до субпетафлопсных ресурсов для химических вычислений и т.п. К сожалению, при наличии в Российской Федерации достаточно мощных вычислительных суперкомпьютерных центров (МГУ, МСЦ, Саров и т.п.) подобные проблемно-ориентированные центры достаточной мощности в настоящее время практически отсутствуют.

Как правило, масштабные задачи химии требуют либо достаточно эксклюзивного использования суперкомпьютеров, что далеко не всегда приемлемо, или же вычислительных ресурсов, которые не может предоставить ни один из вычислительных центров. Это неизбежно приводит к необходимости использования для решения многих подобных задач мощностей крупных распределенных грид-полигонов, причем включающих суперкомпьютеры петафлопсной и выше производительности. Даже в условиях стремительного развития мощностей единичных установок (первые единицы и десятки петафлопс – в России это «Ломоносов», создаваемый суперкомпьютер МСЦ, Саровский центр и др.), значительная часть подобных задач может быть разрешена только в условиях распределенных вычислительных полигонов.

Для этого необходимы устойчиво работающие распределенные вычислительные системы, обеспечивающие как одновременный процессинг десятков и сотен тысяч разноплановых и разномасштабных задач, так и передачу данных объемом до сотен терабайт в сутки (и хранение многих петабайт данных).

Классический «грид» представляет собой географически распределенные инфраструктуры, объединяющие на основе единой технологии и установленных протоколов и правил множество ресурсов разных типов (расчетные узлы – включая суперкомпьютерные установки, хранилища и базы данных, высокоскоростные каналы связи), доступ к которым грид-пользователь может получить практически из любой точки независимо от места расположения. Грид-полигоны предлагают коллективный разделяемый доступ к ресурсам и к связанным с ними вычислительным грид-сервисам (в том числе проблемно-ориентированным) в рамках создаваемых распределенных виртуальных организаций, состоящих из ресурсных центров, сервисной грид-инфраструктуры и отдельных грид-пользователей, совместно использующих предоставляемые в общее пользование ресурсы. Каждая виртуальная организация устанавливает (в соответствии с общими правилами грид-полигона) собственную политику поведения участников, регламентирует используемое ПО (в том числе прикладное), обеспечивает контроль выполнения заданий и мониторинг и учет (аккаунтинг) использования ресурсов.

Современные грид-инфраструктуры (в том числе реализованные на российских вычислительных ресурсах) обеспечивают достаточно устойчивую интеграцию вычислительных ресурсов (и базирующихся на них сервисов), находящихся в разных организациях в единую вычислительную среду, позволяющую решать задачи по обработке большого количества задач и сверхбольших объемов данных.

2. Вычислительная химия на грид-полигонах в России

Авторы опираются на опыт использования грид-технологий в интересах вычислительной химии и смежных отраслей науки на суперкомпьютерных установках и в грид-средах, что является одним из основных направлений работы вычислительного центра Института Проблем Химической Физики в Черноголовке (ИПХФ РАН, http://www.icp.ac.ru). В настоящее время Институт располагает богатейшей в России библиотекой параллельных квантово-химических и молекулярно-динамических программ (авторских, «open source» и лицензионных), что позволяет проводить обширные эксперименты по адаптации подобных программ к грид-средам в интересах собственных пользователей. В течение года в институте проводится расчет от 3 до 4 тысяч вычислительных задач высокой сложности с публикацией более чем 400 печатных работ с

использованием результатов проведенных расчетов. Работы с использованием грид-вычислений в интересах квантовой химии и молекулярной динамики в ИПХФ ведутся с 2004 года, и в настоящее время осуществляются под эгидой нескольких государственных программ (включая Федеральные Целевые Программы, Программы Президиума РАН, гранты РФФИ). ИПХФ является инициатором по использованию средств вычислительной химии в ряде ранее созданных российских грид-полигонов разного масштаба : ГридННС (Национальная Нанотехнологическая Сеть, http://www.ngrid.ru), СКИФ-Полигон (http://skif-grid.botik.ru), EGEE-RDIG, сейчас EGI-[RU-NGI] - http://www.egee-rdig.ru), а также создаваемой в рамках ФЦП пилотной Российской грид-сети для высокопроизводительных вычислений.

2.1 Вычислительные грид-сервисы на базе проблемно-ориентированных прикладных пакетов

Одним из основных вариантов создания вычислительных грид-сервисов в рамках распределенных полигонов является адаптация прикладных программных пакетов (далее – «ППП») для их использования грид-пользователями. Адаптация прикладных пакетов проводилась авторами для различных распределенных сред, основанных на middleware gLite, Unicore, Globus Tools в условиях основных вышеперечисленных российских грид-полигонов

Как правило, реализация ППП в виде грид-сервисов помимо установки ППП на ресурсные сайты, адаптации их к распределенной среде и отладки подразумевает создание высокоуровневых дружелюбных конечному пользователю Web-интерфейсов, что значительно снижает трудоемкость работы пользователя с подобными пакетами в условиях распределенных вычислительных сред.

Для всех адаптированных прикладных пакетов были созданы подобные высокоуровневые web-интерфейсы к GRID средам с большим набором функций. Созданные интерфейсы позволяют грид-пользователю работать с распределенными средами через Интернет-браузеры и осуществлять следующие действия:

- авторизовать пользователя при входе в пространство грид-полигона на основе полученных им предварительно сертификатов полигона, далее – получать прокси-сертификаты грид-среды с учетом планируемой продолжительности заданий;
- подготовить задания (включая загрузку или создание начальных данных и конфигурационных файлов и их редакцию) в соответствие с требованиями пакета;
- запускать прикладные пакеты в инфраструктуре грид-полигона (при необходимости – на произвольном или выбранном пользователем грид-ресурсе);
- вести постоянный мониторинг выполнения задания (включая его останов и перезапуск при необходимости);
- по завершении – получить результаты счета или передавать их на указанный GridFTP сервис.

В последующей перспективе стоит построение интерактивных интерфейсов для формирования сложных конфигурационных файлов прикладных пакетов на основе систем иерархических меню и «деревьев» выбора, включая возможность работы с типовыми (для конкретных прикладных пакетов) шаблонами заданий, которые сводят определение задач к варьированию одного-двух наиболее значимых параметров, тем самым резко упрощая работу конечного грид-пользователя (например, выбор легирующих атомов металлов для фуллеренов или нанотрубок, указание фиксированных P-T условий расчета для стандартных молекул и т.п.) .

В той или иной степени для работы в указанных выше грид-средах были адаптированы следующие квантово-химические и молекулярно-динамические ППП: Gaussian 03, GAMESS US, Firefly, CPMD, Dalton 2012, NWChem, NAMD, AbInit, GROMACS, VASP, PWScf, Orca и др. Большинство ППП, их функциональность и особенности адаптации были описаны авторами ранее [1,2], в том числе в трудах данной конференции 2012 года.

Для всех указанных ППП в разной степени (в зависимости от лицензий) был организован доступ в рамках различных ВО (например, в ГридННС – «NanoChem» плюс ВО, относящихся к конкретным ППП – Gamess, AbInit, GROMACS и т.п.) через порталы ГридННС

https://ui.ngrid.ru и портал пилотной российской грид-сети для высокопроизводительных вычислений (поддерживается ФГУП «Восход»). Как уже говорилось, часть ППП была также ранее реализована авторами на пространстве грид-полигонов EGEE(EGI)-RDIG и СКИФ-Полигона, а в настоящее время большинство реализовано и в рамках созданной в 2011-2012 годах пилотной Российской грид-сети для высокопроизводительных вычислений. Как правило, адаптация пакетов для разных грид-сред различается особенностями реализации низкоуровневых интерфейсов между middleware, используемой грид-полигоном (gLite, Unicore, Globus Tools и т.п.) и собственно ППП. Большинство вышеперечисленных пакетов установлены на кластерах ресурсных центров, входящих в ГридННС (в том числе – практически все из вышеперечисленных установлены на ресурсном грид-центре ИПХФ) и доступны в качестве вычислительных грид-сервисов. Для части ППП сделаны тестовые инсталляции (в основном из-за проблем с лицензиями). Для пакетов после установки и тестирования на кластерах настроены шлюзы для приема входящих грид-заданий и работы с GridFTP ресурсами, установлены высокоуровневые интерфейсы (различной степени сложности) и проведено массовое тестирование (включая стресс-тесты) на различных задачах. ИПХФ выступал в качестве установщика ряда адаптированных пакетов (Gamess, Gaussian), предоставлял ресурсы своего грид-центра и производил тестирование созданных грид-сервисов через низкоуровневые и высокоуровневые web-интерфейсы грид-среды.

2.2 Разработка некоторых технологий для повышения эффективности работы с проблемно-ориентированными сервисами в грид-средах

Помимо адаптации ППП в качестве вычислительных грид сервисов, развитие вычислительной химии применительно к распределенным и суперкомпьютерным средам требует развития новых технологий по запуску и эксплуатации химического ПО. Для решения ряда проблем, возникших при использовании распределенных и суперкомпьютерных вычислений в интересах вычислительной химии авторами в рамках грид-сервисов разработаны (и продолжают развиваться) несколько технологий как оригинального плана, так и созданные на основе существующих разработок:

1. Проведено изучение способов применения современных методов проведения распределенных расчетов на грид-полигонах прикладных пакетов в области квантовой химии и молекулярной динамики с использованием GPU (Graphical Processor Units) устройств, что во многих случаях ведет к убыстрению расчетов от десятков процентов до первых порядков по времени, либо (как вариант) – возможности существенного повышения точности или детальности расчетов. В настоящее время резко интенсифицировался перевод программного кода прикладных пакетов на параллельные технологии с использованием технологий программирования CUDA™ и аналогичных. При этом программирование гибридного кода для квантово-химических и молекулярно-динамических пакетов занимает одно из ведущих позиций в этом направлении. Развитие GPU технологий достигло точки, когда множество существующих приложений реализуются с их использованием и работают значительно быстрее, чем на обычных многоядерных системах. Это позволяет проводить высокоинтенсивные вычисления с применением встраиваемых в расчетные узлы кластеров высокопроизводительных графических процессоров (Nvidia – Tesla и Kepler, AMD – ATI Radeon). В настоящее время авторами проводятся методические исследования на предмет широкого использования новых программных вариантов прикладных квантово-химических пакетов, тестирование и оптимизацию имеющихся вариантов с последующей адаптацией их к проведению расчетов в грид-средах. Одновременно разрабатываются методики запуска грид-заданий, ориентированных на использование ресурсов с поддержкой CUDA™ технологий, в том числе на доступных суперкомпьютерных установках. Более детально этот вопрос рассмотрен в статье авторов, представленной в данном сборнике трудов [3].
2. Проводится создание на новом технологическом уровне системы для работы с большими пулами параллельно-независимых грид-заданий с целью решения задач, требующих проведения расчетов на больших массивах равномерных данных или варьирующих входных параметров. Система предназначена как для решения большого класса многопараметрических задач квантовой химии (с многими варьирующими параметрами или на равномерных

независимых «сетках» данных), так и для проведения расчетов задач, предусматривающих использование массового запуска стандартных прикладных пакетов при значительном варьировании 1-2 входных параметров. Предусмотрено обеспечение следующих функций расчетов: разбиение первичной задачи на равномерных областях данных или входных параметров, автоматическое формирование пула независимых друг от друга грид-заданий, их «параллельный» запуск в грид-средах, мониторинг и контроль выполнения заданий в рамках пула, сборку результатов расчетов заданий с грид-ресурсов в конечный обобщающий результат. Система должна обеспечивать разные способы разбиения первичной задачи – по «сетке» области данных или с варьированием одного или нескольких входных параметров и обеспечивать работу с не менее чем 10^4 заданиями (в перспективе до 10^7). Данная технология в дальнейшем может быть также использована для проведения расчетов на суперкомпьютерных установках пета- и экзафлопсного масштаба. Идеология системы и пилотный вариант системы были ранее апробированы для промежуточного ПО gLite и Unicore [4] для числа заданий до 10^4, в настоящее время система адаптируется для работы в средах, основанных на использовании Globus Tools 4 и 5.

Все указанные технологии могут быть легко использованы для значительного масштабирования задач вычислительной химии на распределенных полигонах, и, в определенной степени, - в условиях суперкомпьютерных установок нового поколения.

3. Основные проблемы, стоящие перед работой проблемно-ориентированных грид-сервисов и возможные способы их решения

Несмотря на весьма успешное построение нескольких грид-полигонов как чисто российских, так и в качестве сегментов международных проектов, востребованность пользователями из областей науки, промышленности, бизнеса доступных грид-ресурсов, на наш взгляд, остается весьма низкой из-за ряда как объективных, так и субъективных проблем. Опыт работы авторов показал наличие достаточно масштабных проблем, возникших в процессе создания и эксплуатации грид-полигонов в интересах конечных пользователей из сфер науки, производства, бизнеса. Нижеперечисленные проблемы во многом ограничивают процессы роста и консолидации вычислительных ресурсов в рамках грид-полигонов, сокращают круг доступных вычислительных ресурсов и пользователей, желающих и могущих ими воспользоваться Охарактеризуем их кратко:

1. На уровне государства:

а) отсутствие заинтересованности со стороны науки и производства, включая общую низкую культуру проведения научных и технологических разработок в большинстве НИИ и промышленных организаций, а также неумение (и нежелание) исследователями и инженерами использовать результаты математического моделирования и высокопроизводительных расчетов. Неготовность большинства исследователей к постановке и решению масштабных задач, требующих ресурсоемких вычислений;

б) отсутствие в большинстве отраслей стратегических заказчиков в лице государства и бизнеса, особенно в рамках долгосрочных исследовательских и конструкторских программ;

в) острая нехватка квалифицированных специалистов как со стороны вычислителей (системщики, программисты), так и со стороны пользователей (химики, технологи);

г) практическое отсутствие отечественных прикладных пакетов и малая доступность зарубежных коммерческих высокоэффективных ППП --> высокая стоимость ПО, лицензионные ограничения, сокращенная область использования, трудности освоения;

д) высокие экономические затраты на создание и эксплуатацию вычислительных ресурсов и инфраструктуры, при этом – отсутствие устойчивой финансовой поддержки российских суперкомпьютеров и грид-полигонов после окончания проектов (обычно успешных) по их созданию и вводу в эксплуатацию;

е) неурегулированность правовых и денежных отношений между собственниками вычислительных сервисов (как грид-ресурсов, так и суперкомпьютеров) и потребителями.

2. На уровне собственно вычислительной химии:

а) высокие требования к аппаратно-программному оснащению расчетных узлов суперкомпьютеров и ресурсных грид-сайтов: RAM > 4 Gb на ядро, локальный дисковый массив от 1 Tb, наличие специализированных библиотек, организация доступа к внешним хранилищам данных и т.п., что зачастую не реализуется в составе суперкомпьютерных центров;

б) как правило, персонал непрофильных суперкомпьютерных центров и грид-ресурсов некомпетентен в проблемных областях и мало заинтересован в области организации квантово-химических и молекулярно-динамических расчетов;

в) отсутствуют стабильные, не зависящие от «капризов» разработчиков, площадки в рамках существующих грид-полигонов для отладки прикладного ПО и новых технологий вычислений;

г) только устойчиво и долговременно работающие виртуальные организации в области прикладных вычислений могут привлечь пользователей, «погрязших» в выполнении краткосрочных грантов.

Каковы же возможные пути решения возникших проблем и возможные политики государства в рамках развития суперкомпьютерных и грид-ориентированных вычислительных ресурсов:

1. создание государством (или долевое участие) вычислительной инфраструктуры (суперкомпьютеры, кластеры средних масштабов, высокоскоростные каналы связи и т.п.), последующая компенсация государством на долговременной основе части эксплуатационных расходов суперкомпьютерных центров и ресурсных грид-центров;
2. налоговые льготы на разработку и приобретение отечественного прикладного ПО, создание ориентированных на это инновационных парков;
3. создание в рамках проектов достаточного количества дружелюбных, с большим количеством базовых шаблонов, WWW-ориентированных интерфейсов ППП к вычислительным суперкомпьютерным и грид-сервисам;
4. стимулирование пользователя (промышленного в первую очередь) на использование суперкомпьютерных и грид-сервисов, например, отсутствие или минимальная оплата пользования государственными вычислительными ресурсами на первых стадиях инновационных разработок (при компенсации расходов на их эксплуатацию из бюджета);
5. введение государством в качестве обязательного условия сертификации применения средств математического моделирования и тестирования для разработки ряда товарных продуктов и наукоемкой продукции.
6. в области государственного образования: (а) расширение количества бюджетных мест в вузах для специальностей «системное администрирование/программирование», «прикладное программирование» и т.п.; (б) использование в процессе обучения широкого круга студентов и аспирантов вузов для разработки необходимого системного и прикладного ПО; (в) обучение пользователей – как очное в вузах и на курсах повышения квалификации, так и в режиме интерактивного on-line: создание информационных web-ресурсов по обучению работе с суперкомьютерами и грид-ресурсами;
7. введение в планы обучения прикладных специалистов (инженеров, химиков и т.п.) дополнительных курсов обучения работы с прикладными пакетами ПО и проведение студентами обязательных расчетов реальных научно-технических и инженерных задач с использованием суперкомпьютеров и грид-структур;
8. совершенствование нормативно-правового регулирования в области стратегических информационных технологий, в первую очередь – упорядочение денежно-правовых отношений между собственниками и потребителями вычислительных ресурсов..

Хотя бы частичное решение поставленных проблем позволит во многом решить проблемы существующих грид-полигонов и повысить эффективность их использования в проблемно-ориентированных ресурсоемких вычислениях, в том числе в области вычислительной химии.

3. Заключение

Всемерное развитие проблемно-ориентированных грид-сервисов (в том числе в области вычислительной химии, молекулярного моделирования и дизайна и большинства смежных отраслей науки), применение новых технологий применительно к грид-вычислениям позволяет существенно повысить эффективность научных исследований в данных отраслях и решать ранее недоступные из-за вычислительных проблем ресурсоемкие задачи. На это ориентировано как создание на базе грид-ресурсов проблемно-ориентированных ВО и специализированных грид-центров, так и вовлечение максимального количества исследователей в круг заинтересованных в супервычислениях. Однако, проблемы, стоящие перед продвижением грид-технологий в повседневную практику российских исследователей, во многом тормозят как собственно научно-исследовательские работы, так и развитие самих грид-полигонов в российской Федерации ввиду низкой заинтересованности научных исследователей и технических разработчиков. Авторами предложены некоторые пути возможных решений данных проблем.

Литература

1. Волохов В.М., Варламов Д.А., Волохов А.В., Пивушков А.В., Покатович Г.А., Сурков Н.Ф. Квантово-химические прикладные пакеты на Российских грид-полигонах // «Параллельные вычислительные технологии ПАВТ-2012»: труды международной научной конференции, (Новосибирск, 26-30 марта 2012 г.), Челябинск: Изд-во ЮУрГУ, 2012. с.394–399
2. Волохов В.М., Варламов Д.А., Пивушков А.В., Волохов А.В. Применение GRID технологий в области вычислительной химии // «Известия Академии наук. Серия химическая», 2011, № 7, с.1483-1490
3. Волохов А.В., Волохов В.М., Варламов Д.А., Пивушков А.В., Покатович Г.А., Прохоров А.И. Использование гибридных вычислительных узлов на базе GPU TESLA C2075 при проведении расчетов в области вычислительной химии и молекулярной динамики // ПАВТ-2013 (настоящий сборник трудов)
4. Волохов В.М., Варламов Д.А., Пивушков А.В., Покатович Г.А., Сурков Н.Ф. Технологии ГРИД в вычислительной химии // «Вычислительные методы и программирование», М.: МГУ, 2010, т.11, № 1, с.175-182