СЕРВИСЫ ВЫЧИСЛИТЕЛЬНОЙ ХИМИИ НА РОССИЙСКИХ ГРИД-ПОЛИГОНАХ: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

Волохов Вадим Маркович	Институт проблем химической физики РАН (ИПХФ), зав. отделом, д.ф.-м.н.
Варламов Дмитрий Анатольевич	ИПХФ, с.н.с.
Волохов Александр Вадимович	ИПХФ, ведущий инженер
Пивушков Александр Викторович	ИПХФ, с.н.с., к.ф.-м.н.
Покатович Геннадий Александрович	ИПХФ, зав. сектором
Прохоров Андрей Иванович	ИПХФ, н.с.

Доклад посвящен анализу текущего состояния вычислительных суперкомпьютерных и грид-сервисов, относящихся к областям квантовой химии и молекулярной динамики и реализованных в рамках существующих российских грид-полигонов. Сформулированы основные проблемы по работе с прикладными проблемно-ориентированными сервисами в условиях российских суперкомпьютеров и грид-полигонов, возможные способы их решения и потенциальные перспективы развития в будущем.

Ключевые слова: вычислительная химия, распределенные вычисления, грид-сервисы

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 х 300 000 конфигураций х 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 $F(x_1,x_2 \ldots x_n)$ в области изменения параметров $x_1,x_2 \ldots x_n$ 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 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.

Однако, как правило, столь масштабные задачи 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 и др.), значительная часть подобных задач может быть разрешена только в условиях распределенных вычислительных полигонов.

Авторы опираются на опыт использования прикладных программных пакетов (далее – «ППП») вычислительной химии 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). 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 квантово-химических пакетов в ряде ранее созданных российских грид-полигонов разного масштаба.

Адаптация прикладных пакетов проводилась авторами для различных распределенных сред, основанных на middleware gLite, Unicore, Globus Tools, в условиях основных российских грид-полигонов: 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 Российской грид-сети.

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.

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 и получать проккси сертификаты грид-среды;
- подготовить задания (включая создание и редакцию начальных данных и конфигурационных файлов) в соответствие с требованиями пакета;
- запускать прикладные пакеты в инфраструктуре грид-полигона (при необходимости – на произвольном или избранном грид- ресурсе);
- вести мониторинг выполнения задания (включая останов и перезапуск);
- по завершении – получить результаты счета.

В той или иной степени для работы в грид-средах были адаптированы следующие квантово-химические и молекулярно-динамические ППП:

Gaussian 03	GAMESS US	Firefly
CPMD	Dalton 2011	NWChem
NAMD	Abinit	GROMACS
VASP	PWScf	и др.

Все указанные ППП в разной степени (в зависимости от лицензий) доступны в рамках ВО «NanoChem», а также ВО, относящимся к конкретным ППП (Gamess, AbInit и т.п.), через портал ГридННС https://ui.ngrid.ru. Как уже говорилось, часть ППП была также ранее реализована авторами на пространстве грид-полигонов EGEE(EGI)-RDIG и СКИФ-Полигона. Как правило, адаптация пакетов для разных грид-сред различается особенностями реализации низкоуровневых интерфейсов между middleware и ППП. Большинство вышперечисленных пакетов установлены на кластерах ресурсных центров, входящих в ГридННС (в том числе ИПХФ) и доступны в качестве вычислительных грид-сервисов, для части сделаны тестовые инсталляции (в основном из-за проблем с лицензиями). Для пакетов после установки и тестирования на кластерах настроены шлюзы 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.

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 были разработаны или применены несколько технологий как оригинального плана, так и на основе существующих разработок:

1. Создание средств управления «пулами» («пучками») грид-заданий для работы с большими задачами на равномерных «сетках» данных или параметров (с использованием многопараметрических задач или прикладных пакетов типа GAMESS). Они могут быть представлены в виде объединения большого количества параллельно выполняемых независимых друг от друга задач, с масштабами «пулов» до 10^7 заданий;
2. «Виртуализация» грид-приложений и использование динамически формируемых «вычислителей» (контейнеров), обеспечивающих запуски проблемно-ориентированных ППП на неподготовленных заранее грид-ресурсах или суперкомпьютерах;
3. Технологии использования унифицированных виртуальных машин (**ВМ**) для размещения управляющих и сетевых сервисов ресурсных сайтов грид-полигонов;
4. Использование графических высокопроизводительных видеоадаптеров (**GPU**) на ресурсных грид-узлах для значительного повышения эффективности параллельных расчетов в областях квантовой химии и молекулярной динамики.

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.

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. 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 стратегических заказчиков в лице государства и бизнеса, особенно в рамках долгосрочных исследовательских и конструкторских программ;
   в) 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, непрофильные суперкомпьютерные центры и грид-ресурсы некомпетентны и мало заинтересованы в области организации квантово-химических и молекулярно-динамических расчетов;
   в) отсутствие стабильных, не зависящих от «капризов» разработчиков, площадок в рамках 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? 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.