Моделирование и анализ информационных систем. Т. 26, № 3 (2019), с. 360–364
Modeling and Analysis of Information Systems. Vol. 26, No 3 (2019), pp. 360–364

DOI: 10.18255/1818-1015-2019-3-360-364

УДК 004.4: 004.7

GRID-система на основе европейских стандартов EGI для крупномасштабных расчетов по оригинальному ускоренному методу квантовой химии

Аникин Н. А., Мускатин А.Ю., Кузьминский М.Б., Русаков А.И.

Поступила в редакцию 9 июля 2019
После доработки 2 сентября 2019
Принята к публикации 4 сентября 2019

Аннотация. На основе анализа современных средств создания ИС GRID-типа, входящих в ставший европейским EGI-"стандартом" репозиторий UMD (включая новые версии Globus Toolkit, ARC, dCache и др.), кратко рассмотрено применение GRID-систем для задач вычислительной химии. Созданная авторами GRID-система объединяет два кластера с Linux CentOS 7 и базируется на программном обеспечении из UMD-4. Актуальность и эффективность применения систем пакетной обработки (у нас используется Torque 4.2.10) в квантовохимических расчетах повышается для массовых расчетов докинг-комплексов (в т.ч. для задач моделирования лекарств), для чего был предложен усовершенствованный полуэмпирический метод с более эффективными аппроксимациями, реализованный в программном комплексе LSSDOCK на Fortran-95. Для таких расчетов разработаны новые методы аппроксимаций, в т.ч. для функционалов DFT, и осуществляется их программная реализация. Разработаны конверторы результатов расчетов по LSSDOCK в естественный для GRID, основанный на XML, формат CML версии 3. С использованием CML-формата на базе программных средств dCache реализовано единое дерево виртуальной файловой GRID-системы, распределённой между гетерогенными узлами, которое используется для хранения результатов расчетов по LSSDOCK.

Ключевые слова: GRID, UMD, Web-сервисы, распределенная файловая система, CML, квантовая химия, докинг-комплексы

Для цитирования: Аникин Н. А., Мускатин А.Ю., Кузьминский М.Б., Русаков А.И., "GRID-система на основе европейских стандартов EGI для крупномасштабных расчетов по оригинальному ускоренному методу квантовой химии", *Моделирование и анализ информационных систем*, **26**:3 (2019), 360–364.

Об авторах:
Аникин Николай Алексеевич, orcid.org/0000-0002-5724-8969, кандидат химических наук,
Институт органической химии им. Н. Д. Зелинского Российской академии наук,
Ленинский проспект, 47, г. Москва, 119991 Россия, e-mail: nikan@swf.chem.ac.ru

Мускатин Александр Юрьевич, orcid.org/0000-0002-3596-2782,
Институт органической химии им. Н. Д. Зелинского Российской академии наук,
Ленинский проспект, 47, г. Москва, 119991 Россия, e-mail: amus74@mail.ru

Кузьминский Михаил Борисович, orcid.org/0000-0002-3944-8203, кандидат химических наук,
Институт органической химии им. Н. Д. Зелинского Российской академии наук,
Ленинский проспект, 47, г. Москва, 119991 Россия, e-mail: kus@free.net

Русаков Александр Ильич, orcid.org/0000-0001-8893-4577, доктор химических наук, профессор,
Ярославский государственный университет им. П. Г. Демидова,
ул. Советская, 14, г. Ярославль, 150003 Россия, e-mail: alex@yars.free.net

Благодарности:
Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 18-07-00657.

Построение распределенных ИС с использованием GRID-технологии активно развивается в настоящее время для решения самых разных задач, например, для создания инфраструктуры хранения и обработки данных, и включает центры обработки данных вплоть до суперкомпьютерного уровня (см., например, [1]). Вычислительная химия стала традиционной областью высокопроизводительных вычислений (HPC), и особенно актуальными там представляются задачи моделирования лекарств, что планируется делать, например, на будущей экзафлопсной суперЭВМ мира Cray Frontier в знаменитой в суперкомпьютерном мире Национальной лаборатории Ок-Риджа Министерства энергетики США [2]. Для решения таких задач применяется молекулярный докинг, требующий расчетов энергий взаимодействия тысяч лигандов с протеином.

Традиционный для этого расчет методом молекулярной механики (см., например, [3]) не учитывает квантовых эффектов (в т.ч. межмолекулярный перенос электронной плотности), актуальных для повышения точности. Поэтому расчеты энергий взаимодействия квантовохимическими методами представляются более универсальными и перспективными [4–6]. Настоящая работа ориентирована на разработку и применение усовершенствованных методов и программных средств, относящихся к области квантовой химии.

При таком подходе встает задача обработки большого числа пакетных заданий, которые можно эффективно выполнять в GRID-среде. Одним из преимуществ такого подхода является возможность интеграции вычислительных мощностей различных кластеров, что особенно актуально в случае дефицита настроенных на задачи вычислительной химии суперкомпьютерных ресурсов, в т.ч. в РФ.

Для построения современных GRID-систем необходимо обеспечение высокого уровня интероперабельности и поддержание современных мировых стандартов используемых программных средств. По этой причине в нашей стране естественно ориентироваться на использование репозитория UMD, дающего максимальную интероперабельность на современном европейском уровне EGI [7] и высокое протестированное качество отобранных средств [8].

Современные программные средства построения GRID-систем быстро развиваются, причем одни из самых широко употребимых средств могут заменяться другими. Так, в 2018 году Globus Toolkit практически прекратили свое развитие и заменены рядом других программных продуктов, в т.ч. доступных в UMD [7]. А EGI теперь поддерживает еще и новые средства CMD для облачных вычислений. В 2019 году появились новые версии UMD 4.8 (последняя – 4.8.5). В их состав входят, например, новые версии классических средств управления заданиями на вычислительных GRID-узлах Globus Gram5 14.1.0, средств передачи данных по протоколу GridFTP 13.8.1 и промежуточного программного обеспечения (ППО) ARC 15.3.19 [9].

Следует отметить, что GRID-системы уже используются для задач вычислительной химии и не только для квантовой химии. В качестве примера можно указать на применение GRID-систем для молекулярной динамики [10]. То, что GRID естественно использовать для моделирования лекарств, указано, например, в [11]. Однако публикаций по применению GRID в задачах квантовой химии не так много. Известно, что обычные GRID-системы используются для выполнения расчетов по высокоэффективным комплексам квантовохимических программ, например, Gaussian [12]. Наиболее известной ориентированной именно на задачи вычислитель-

ной химии можно считать GRID-систему (научный портал) MoSGrid (Molecular Simulation Grid) [13], в которой Gaussian интегрирован со средствами молекулярной динамики GROMACS, а функции квантовохимических рабочих процессов реализуются через среду выполнения Parallel Grid и среду разработки Web-сервисов, и передаются в UNICORE (входившей ранее в состав UMD).

Необходимые для задач моделирования лекарств массовые квантовохимические расчеты докинг-комплексов, содержащих большие протеины и маленькие молекулы лигандов, требуют огромных вычислительных ресурсов. Так, наш одноточечный расчет наиболее быстрым полуэмпирическим квантовохимическим методом нулевого дифференциального перекрывания (НДП) только одного относительно небольшого протеина из 1663 атомов по программе Gaussian-09 с быстродействующим алгоритмом, заменяющим лимитирующую время расчета диагонализацию матрицы алгоритом с линейным масштабированием времени расчета от ее размера, требует примерно 6 часов на одном ядре старого процессора Xeon E3-1240/3,3 ГГц (для более точного метода DFT – больше месяца). Это делает массовые расчеты серий, состоящих из тысяч докинг-комплексов протеин-лиганд, требующими слишком больших времен, если не использовать специальную методику для кардинального ускорения таких расчетов.

В рамках методов НДП нами предложена новая методика для дополнительного ускорения, специализированная на массовые расчеты комплексов протеина с лигандами [4], дающая в нашем программном (на Fortran-95) комплексе LSSDOCK сверхвысокое быстродействие. Тогда указанный выше расчет требует порядка минуты на один докинг-комплекс. Эта методика основана на нашем общем аппроксимационном подходе с явным учетом в локальном гамильтониане лиганда с прилегающей активной частью протеина специфики данной задачи (расчеты тысяч комплексов одного и того же крупного протеина с большим числом разнообразных, но относительно небольших молекул-лигандов), при котором точность расчета практически не ухудшается.

Указанный выше подход может быть внедрен в более точные, но требующие значительно больше вычислительных ресурсов, неэмпирические квантовохимические методы, в первую очередь, в распространенный метод DFT и в моделирующие гамильтониан DFT более быстрые методы, в т.ч. в SCC-DFTB.

Такие массовые расчеты могут быть успешно разработаны в форме пакетной обработки заданий и их возможно эффективно реализовать в GRID-системах. Наш GRID-сервер был развернут на базе CentOS 7 с ППО ARC 15.3.19 (из UMD-4.8.2), которое остаётся на данный момент наиболее универсальным ППО из дистрибутива UMD, включая ARC Resource-coupled EXecution service (A-REX), обеспечивающий интерфейс с локальными пакетными системами; сервер GridFTP (GFS) и Enhanced Grid Information Indexing Service (EGIIS). ARC поддерживает все основные языки описания задач (с которыми обеспечивается интерфейс в системах пакетных заданий), используемые в GRID-системах – xRSL, JSDL и JDL. В качестве системы пакетной обработки у нас применяется Torque 4.2.10, которая используется и для интеграции в GRID-систему высокопроизводительного сервера ЯрГУ NVIDIA DGX-1 с 8 GPU архитектуры V100SXM (с пиковой производительностью 7.8 TFLOPS DP каждый), работающего с Ubuntu 16.04.

Были развёрнуты также программные средства dCache 3.2.21 из UMD-4.7.0, ко-

торые обеспечивают сохранность больших объёмов данных и оптимизируют скорость доступа к ним. Для хранения служебной информации используется СУБД PostgreSQL. dCache предоставляет возможность получить информацию о состоянии системы и данных в формате GRID-стандартов GLUE 1.3 и 2.0. Данные результатов квантовохимических расчётов размещаются в естественном для GRID, основанном на XML-формате CML версии 3 [14]. Разработана система поиска результатов квантовохимических расчетов по имени лиганда и его контекстной части, элементам периодической системы, индексам протеина и др.

Однако бывают ситуации, когда точности используемых в LSSDOCK методов может быть недостаточно, а требуется использование более точных неэмпирических методов квантовой химии [5]. Поэтому в настоящее время нами ведется разработка и программная реализация ускоренного метода с явной физико-математически корректной аппроксимацией самого гамильтониана DFT, а не его матричных элементов, что существенно точнее SCC-DFTB.

Предложенные методы и разрабатываемые программные средства для массовых квантовохимических расчетов докинг-комплексов можно эффективно применять в создаваемой GRID-среде, интегрирующей кластеры в ИОХ РАН и ЯрГУ.

Список литературы / References

[1] Биктимиров М. Р., и др., “Использование информационных технологий и инфраструктур для агрегации научной информации. Опыт Канады, Нидерландов, Германии”, *Моделирование и анализ информационных систем*, **22**:1 (2015), 114–126; [Biktimirov M. R., et al., “Information Technologies and Infrastructures for Research Data Aggregation. The Experience of Canada, Netherlands and Germany”, *Modeling and Analysis of Information Systems*, **22**:1 (2015), 114–126, (in Russian).]

[2] https://www.olcf.ornl.gov/frontier/.

[3] Фарков М. А., Легалов А. И., “Применение методов оптимизации для выполнения молекулярного докинга на графических процессорах”, *Моделирование и анализ информационных систем*, **21**:5 (2014), 93–101; [Farkov M. A., Legalov A. I., “Application of Numerical Optimization Methods to Perform Molecular Docking on Graphics Processing Units”, *Modeling and Analysis of Information Systems*, **21**:5 (2014), 93–101, (in Russian).]

[4] Аникин Н. А., и др., “Новый подход к ускорению массовых квантово-химических расчетов докинг-комплексов”, *Известия Академии наук. Серия химическая*, **6** (2018), 1100–1103; [Anikin N. A., et al., “A New Approach for the Acceleration of Large-scale Serial Quantum Chemical Calculations of Docking Complexes”, *Russian Chemical Bulletin*, 2018, № 6, 1100–1103, (in Russian).]

[5] Caldararu O., et al., “Binding Free Energies in the SAMPL5 Octa-acid Host–guest Challenge Calculated with DFT-D3 and CCSD(T)”, *Journal of Computer-Aided Molecular Design*, **31**:1 (2017), 87–106.

[6] Yilmazer N. D., Korth M., “Recent Progress in Treating Protein–Ligand Interactions with Quantum-Mechanical Methods”, *International Journal of Molecular Sciences*, **17**:5 (2016), Article ID 742.

[7] Мускатин А. Ю., Кузьминский М. Б., Русаков А. И., “На пути к унифицированным грид-системам”, *Открытые системы. СУБД*, **1** (2019), 10–14; [Muskatin A. Y., Kuz’minskii M. B., Rusakov A. I., “Towards Unified Grid Systems”, *Open Systems. DBMS*, **1** (2019), 10–14, (in Russian).]

[8] Fernandez P. O., et al., “umd-verification: Automation of Software Validation for the EGI Federated e-Infrastructure”, *Journal of Grid Computing*, **16**:4 (2018), 683–696.

[9] http://repository.egi.eu/category/umd releases/.

[10] Merelli I., "Infrastructure for High-Performance Computing: Grids and Grid Computing", *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 2018, 230–235.

[11] Grunzke R., et al., "Managing Complexity in Distributed Data Life Cycles Enhancing Scientific Discovery", *IEEE 11th International Conference on e-Science*, 2015, 371–380.

[12] Bhanwar S., Bawa S., "TUX-INTERO: A Portal for Secure Interoperation of Grids", *Int. J. Eng. Sci. Technol*, **2**:7 (2010), 3335–3343.

[13] Herres-Pawlis S., et al., "Quantum Chemical Meta-workflows in MoSGrid", *Concurrency and Computation: Practice and Experience*, **27**:2 (2015), 344–357.

[14] http://www.xml-cml.org/.

Anikin N. A., Muskatin A. Y., Kuzminsky M. B., Rusakov A. I., "GRID-system Based on European EGI Standards for Large-scale Calculations Using the Original Accelerated Method of Quantum Chemistry", *Modeling and Analysis of Information Systems*, **26**:3 (2019), 360–364.

DOI: 10.18255/1818-1015-2019-3-360-364

Abstract. Based on the analysis of modern tools for creating GRID-type information systems that are part of the European EGI "standard" – UMD repository (including new versions of Globus Toolkit, ARC, dCache, etc.), the applying of GRID systems for computational chemistry is briefly discussed. The GRID system created by the authors combines two clusters with Linux CentOS 7 and is based on software from UMD-4. The relevance and effectiveness of batch processing systems (we use Torque 4.2.10) in quantum chemical calculations is increased for mass calculations of docking complexes (including for drug modeling problems), for which an improved semiempirical method with more efficient approximations was proposed, implemented in the Fortran-95 LSSDOCK software package. For such calculations, new approximation methods have been developed, including for DFT functionals, and their software implementation is carried out. Converters of calculation results by LSSDOCK into a natural for GRID XML-based format CML version 3 are developed. Using the CML format based on dCache software, a single tree of a virtual GRID filesystem distributed between heterogeneous nodes is used to store the results of LSSDOCK calculations.

Keywords: GRID, UMD, Web services, distributed file system, CML, quantum chemistry, docking complexes

On the authors:
Nikolay A. Anikin, orcid.org/0000-0002-5724-8969, PhD,
N. D. Zelinsky Institute of Organic Chemistry RAS,
47 Leninsky Prospect, Moscow 119991, Russia, e-mail: nikan@swf.chem.ac.ru

Alexander Y. Muskatin, orcid.org/0000-0002-3596-2782, PhD,
N. D. Zelinsky Institute of Organic Chemistry RAS,
47 Leninsky Prospect, Moscow 119991, Russia, e-mail: amus74@mail.ru

Mikhail B. Kuzminsky, orcid.org/0000-0002-3944-8203, PhD,
N. D. Zelinsky Institute of Organic Chemistry RAS,
47 Leninsky Prospect, Moscow 119991, Russia, e-mail: kus@free.net

Alexandr I. Rusakov, orcid.org/0000-0001-8893-4577, PhD,
P. G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl 150003, Russia, e-mail: alex@yars.free.net

Acknowledgments:
This work was funded by the RFBR according to the research № 18-07-00657.